Haiping Huang

# Statistical Mechanics of Neural Networks

Higher Education Press

Springer

# Statistical Mechanics of Neural Networks

Haiping Huang

# Statistical Mechanics of Neural Networks

Higher Education Press

Springer

Haiping Huang
Sun Yat-sen University
Guangzhou, China

ISBN 978-981-16-7569-0 ISBN 978-981-16-7570-6 (eBook)
https://doi.org/10.1007/978-981-16-7570-6

Jointly published with Higher Education Press
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the print book from: Higher Education Press.



This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

*To my wife, Huihong Pan,*
*and to my children,*
*Qing Huang and Yun Huang*

# Preface

Neural networks have become a powerful tool in various domains of scientific research and industrial applications. However, the inner workings of this tool remain unknown, which prohibits us from a deep understanding and further principled design of more powerful network architectures and optimization algorithms. To crack the black box, different disciplines including physics, statistics, information theory, non-convex optimization and so on must be integrated, which may also bridge the gap between the artificial neural networks and the brain. However, in this highly inter-disciplinary field, there are few monographs providing a systematic introduction of theoretical physics basics for understanding neural networks, especially covering recent cutting-edge topics of neural networks.

In this book, we provide a physics perspective on the theory of neural networks, and even neural computation in models of the brain. The book covers the basics of statistical mechanics, statistical inference, neural networks, and especially classic and recent mean-field analysis of neural networks of different nature. These mathematically beautiful examples of statistical mechanics analysis of neural networks are expected to inspire further techniques to provide an analytic theory for more complex networks. Future important directions along the line of scientific machine learning and theoretical models of brain computation are also reviewed.

We remark that this book is not a complete review of both fields of artificial neural networks and mean-field theory of neural networks, instead, a biased-viewpoint of statistical physics methods toward understanding the black box of deep learning, especially for beginner-level students and researchers who get interested in the mean-field theory of learning in neural networks.

This book stemmed from a series of lectures about the interplay between statistical mechanics and neural networks. These lectures were given by the author in his PMI (physics, machine and intelligence) group during the years from 2018 to 2020. The book is organized into two parts—basics of statistical mechanics related to the theory of neural networks, and theoretical studies of neural networks including cortical models.

The first part is further divided into nine chapters. Chapter 1 gives a brief history of neural network studies. Chapter 2 introduces multi-spin interaction models and

the cavity method to compute the partition function of disordered systems. Chapter 3 introduces the variational mean-field methods including the Bethe approximation and belief propagation algorithms. Chapter 4 introduces the Monte Carlo simulation methods that are used to acquire low-energy configurations of a statistical mechanical system. Chapter 5 introduces high-temperature expansion techniques. Chapter 6 introduces the spin glass model where the Nishimori line was discovered. Chapter 7 introduces the random energy model which is an infinite-body interaction limit of multi-spin disordered systems. Chapter 8 introduces a statistical mechanical theory of the Hopfield model that was designed for associative memory of random patterns based on the Hebbian local learning rule. Chapter 9 introduces the concepts of replica symmetry and replica symmetry breaking in the spin glass theory of disordered systems.

The second part is divided into nine chapters. Chapter 10 introduces the Boltzmann machine learning (also called the inverse Ising problem in physics or maximum entropy method in statistics) and the statistical mechanics of the restricted Boltzmann machine learning. In this chapter, a variational mean-field theory for learning a generic RBM of discrete synapses is also introduced in depth. Chapter 11 introduces the simplest model of unsupervised learning. Chapter 12 introduces the nature of unsupervised learning with RBM (only two hidden neurons are considered), i.e., the unsupervised learning process can be understood in terms of a series of continuous phase transitions, including both weight-reversal symmetry breaking and hidden-neuron-permutation symmetry breaking. Chapter 13 introduces a single-layer discrete perceptron and its mean-field theory. Chapter 14 introduces the mean-field model of multi-layered perceptron and its analysis via the cavity method. In this chapter, a mean-field training algorithm of multi-layered perceptron with discrete synapses is introduced, together with mean-field training from an ensemble perspective. Chapter 15 introduces the mean-field theory of dimension reduction in deep random neural networks. Chapter 16 introduces the chaos theory of random recurrent neural networks. In this chapter, the excitatory-inhibitory balance theory of cortical circuits is also introduced, together with the backpropagation through time for training a generic RNN. Chapter 17 introduces how the statistical mechanics technique can be applied to compute the asymptotic behavior of the spectral density for the Hermitian and the non-Hermitian random matrices. Finally, perspectives on a statistical mechanical theory toward deep learning and even other interesting aspects of intelligence are provided, hopefully inspiring future developments of the interdisciplinary fields across physics, machine learning and theoretical neuroscience and other involved disciplines.

I am grateful for the students' efforts in drafting the lecture notes, including preparing figures. Here, I list their contributions to associated chapters. These students in my PMI group are Zhenye Huang (Chaps. 4 and 10), Zijian Jiang (Chaps. 2, 13 and 16), Chan Li (Chaps. 11, 15 and 16), Jianwen Zhou (Chaps. 5, 8 and 17), Wenxuan Zou (Chaps. 3, 6 and 14) and Tianqi Hou (Chap. 12). I also thank the other PMI members, Ziming Chen, Yiming Jiao, Junbin Qiu, Mingshan Xie, Xianbo Xu and Yang Zhao for their reading feedbacks on the draft. I also would like to thank Haijun Zhou, K. Y. Michael Wong, Yoshiyuki Kabashima and Taro Toyoizumi

for their encouragements and supports during my Ph.D. and Post-doctoral research career. I finally acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 11805284 Grant No. 12122515).

Guangzhou, China Haiping Huang
December 2021

# Contents

# About the Author

**Dr. Haiping Huang** received his Ph.D. degree in theoretical physics from the institute of Theoretical Physics, the Chinese Academy of Sciences. He works as an associate professor at the School of Physics, Sun Yat-sen University,China. His research interests include the origin of the computational hardness of the binary perceptron model, the theory of dimension reduction in deep neural networks, and inherent symmetry breaking in unsupervised learning.In 2021, he was awarded Excellent Young Scientists Fund by National Natural Science Foundation of China.

# Acronyms

| | |
|---|---|
| AMP | Approximate message passing |
| AT | De Almeida-Thouless |
| BA | Bethe approximation |
| BM | Boltzmann machine |
| BP | Belief propagation |
| BPTT | Backpropagation through time |
| CD | Contrastive-divergence |
| CLT | Central limit theorem |
| CNN | Convolution neural network |
| DG | Dichotomized Gaussian |
| EA | Edwards–Anderson |
| EM | Expectation-maximization |
| gBP | Generalized backpropagation |
| KL | Kullback–Leibler |
| LIF | Leaky-integrated firing |
| LSTM | Long short-term memory |
| MCMC | Markov chain Monte Carlo |
| MCS | Monte Carlo sweep |
| MFA | Mean-field approximation |
| MPM | Maximizer of the posterior marginal |
| PS | Permutation symmetry |
| PSB | Permutation symmetry breaking |
| RAP | Random active path |
| RBM | Restricted Boltzmann machine |
| RF | Receptive field |
| RG | Random guess |
| RNN | Recurrent neural network |
| RS | Replica symmetry |
| RSB | Replica symmetry breaking |
| SaS | Spike and slab |
| SDE | Saddle-point equation |

| | |
|---|---|
| SG | Spin glass |
| SK | Sherrington-Kirkpatrick |
| sMP | Simplified message passing |
| SSB | Spontaneous symmetry breaking |
| TAP | Thouless–Anderson–Palmer |

# Chapter 1
# Introduction


Neural network studies stemmed from the curiosity about how the brain works and even biological mechanisms of high-level intelligence [1]. This original curiosity has a very long history that is also a history of humans' endeavors to understand the brain. A modern artificial neural model was proposed by McCulloch and Pitts in 1943 [2], and the neuron of complex biological processes was abstractly modeled as a non-linear transfer function of simply weighted sum of inputs. A few years later, Donald Hebb proposed the Hebbian learning rule [3], i.e., "cells that fire together, wire together". This rule forms the basics of a later development, i.e., the abstract model of associative memory, the so-called Hopfield model [4, 5], where the Hebbian rule was used to construct the effective coupling between neurons in the model that can realize the retrieval of a memory item (e.g., a random pattern the Hebbian rule uses), under a less noisy neural dynamics from an initial state where the memory item is corrupted by a few bits. The Hebbian rule, despite its simplicity, plays a significant important role in the current status of both experimental and theoretical neuroscience studies [6].

Based on the McCulloch–Pitts model of artificial neurons, Frank Rosenblatt introduced the first perceptron model of supervised classification tasks [7]. At that time, this model can only be used to classify linearly separable patterns [8]. However, this abstract model plays an important role in neuroscience studies, as the perceptron model was popular in modeling the learning behavior of the cerebellar Purkinje cells [9, 10]. The perceptron model can also be easily generalized to multilayer feedforward neural networks, which are able to separate non-linearly separable patterns, due to the highly nested non-linear layer-wise computations. This nested non-linearity makes an analytic understanding of inner workings challenging in the academic community [11, 12]. However, the backpropagation of the error from top layers was shown to work in practical training of multi-layer neural networks [13], which establishes the algorithmic foundation of deep learning.

Fukushima introduced neocognitron in 1980, using the biological concept of simple and complex cells observed in visual pathways of a cat's visual cortex [14]. When



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_1

these neural computations are organized in a hierarchical way, the position-shift invariance can be achieved. This neocognitron model inspired the further development of multi-layer neural computation, for example, the powerful architecture—convolution neural network (CNN) proposed in 1990s [15], where the computation of simple cells corresponds to convolution while computation of complex cells corresponds to pooling, showing the power of multi-layer neural networks in practice, e.g., in solving computer vision tasks [16].

In 1985, Hinton and Sejnowski introduced the Boltzmann machine [17], where the model parameters, e.g., coupling and fields of an Ising model, can be learned directly from the data samples, matching only the first two moments of the data statistics [18]. This framework has a recently renewed interest in system neuroscience [19], being known as an inverse Ising problem in physics [20] with a wide application in different interdisciplinary fields ranging from neural activity modeling and protein structure prediction to financial data analysis [21]. Paul Smolensky later introduced the original two-layer neural network with stochastic activations [22], the so-called restricted-Boltzmann machine (RBM) [22, 23], where neurons in a traditional Boltzmann machine are divided into visible and hidden groups. In 2006, Hinton and Salakhutdinov proposed efficient methods to train a deep belief network composed of layer-wise stacking of RBMs [24], initializing the deep learning revolutionary in both academic and industrial neural network studies.

Another type of neural network architecture has salient features in its recurrent computation, incorporating temporal information. There appeared extensive research interests in algorithmic issues around 1990 [25–28]. However, training the recurrent neural network (RNN) is typically challenging, due to vanishing/exploding gradients of the objective function [29]. In the current deep learning era, many smart techniques are being proposed to tackle this challenge. In particular, Hochreiter and Schmidhuber introduced the long short-term memory (LSTM), using various kinds of information-gating mechanisms [30], to avoid the training difficulty of RNNs. The RNN structure is also considered as a canonical model of perception, learning, memory, action and other high-level cognition [31–33].

In the history of artificial neural networks, still a lot of important topics are not covered in the above retrospect. For complete reviews, we refer interested readers to several recent reviews of neural networks [34–36]. On the other side, the statistical mechanics plays a key role in understanding the emergent behavior of artificial neural networks, even real neuronal networks [37].

The first statistical mechanical theory of neural networks was published in 1985; providing a complete phase diagram of the Hopfield network [38, 39] and explaining low temperature and low memory load are necessary to guarantee a successful retrieval of one of the embedded random patterns (akin to memory items). The analytic techniques are rooted in studies of disordered systems, such as spin glass systems [40, 41]. One powerful technique is the replica trick, which introduces many copies of the original model, and the original complex spin-to-spin interactions are decoupled into overlaps between replicas, while the overlaps are exactly the order parameters of the statistical mechanical model. This technique was later generalized to the perceptron model, inventing the concept of the Gardner volume to determine

the capacity of a perceptron system [42, 43]. The Gardner method is still popular as a powerful physics tool in the theoretical neuroscience community [10].

In 1988, Sompolinsky et al. developed another powerful physics method for analyzing the recurrent dynamics of RNNs with random weights [44]. This method treats the behavior of a real RNN as an effective mean-field limit of a homogeneous system, whose first two moments of neural dynamics statistics are recursively established, resulting in a mean-field calculation of the Lyapunov exponent determining whether a transition-to-chaos is possible. This framework can be derived under the path-integral representation of the dynamics [45, 46] and is still popular in analyzing more complex RNNs with structured connectivity. The mean-field study of a random RNN was later generalized to neuronal networks of excitatory and inhibitory cells [47, 48], satisfying biological Dale's rule (a biological neuron cannot produce both excitatory and inhibitory synapses). In this study, the excitatory–inhibitory balance condition [49, 50], i.e., feedback inhibition cancels with strong excitatory recurrent inputs, can be identified in the mean-field limit, leading to a mechanistic explanation of the irregular asynchronous neural activity observed in cortical circuits. Brunel further studied the emergent behavior of spiking activity of a sparsely connected excitatory–inhibitory neural network [51]. These theoretical paradigms still benefit the computational and theoretical community even now. Therefore, the statistical physics methods, including equilibrium phase diagram analysis and non-equilibrium mean-field theory, are very promising in exploring the black box of deep neural networks, which may further connect to other branches, e.g., random matrix theory [52], etc.

In this book, we will provide our personal selections of statistical mechanical techniques applied to neural networks studies, and an in-depth introduction of these statistical physics methods, especially applications in simple toy models where learning mechanisms can be revealed in a mathematically concise way, even with theoretical predictions of new emergent behavior.

## References

1. R. Yuste, Nat. Rev. Neurosci. **16**(8), 487 (2015)
2. W.S. McCulloch, W. Pitts, Bull. Math. Biophys. **5**, 115 (1943)
3. D.O. Hebb, *The Organization of Behavior* (Wiley, New York, 1949)
4. J.J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982)
5. S.I. Amari, Biological Cybern. **26**, 175 (1977)
6. W. Gerstner, M. Lehmann, V. Liakoni, D. Corneil, J. Brea, Front. Neural Circ. **12**, 53 (2018)
7. F. Rosenblatt, Psychol. Rev. **65**, 386 (1958)
8. S. Papert, M.L. Minsky, *Perceptrons: An Introduction to Computational Geometry* (MIT Press, Cambridge, 1988)
9. C. Clopath, J.P. Nadal, N. Brunel, PLOS Comput. Biol. **8**(4), e1002448 (2012)
10. N. Brunel, Nat. Neurosci. **19**, 749 (2016)
11. A. Saxe, S. Nelli, C. Summerfield, Nat. Rev. Neurosci. **22**, 55 (2020)
12. D. Hassabis, D. Kumaran, C. Summerfield, M. Botvinick, Neuron **95**(2), 245 (2017)
13. D.E. Rumelhart, G.E. Hinton, R.J. Williams, Nature **323**, 533 (1986)
14. K. Fukushima, Biolog. Cybern. **36**(4), 193 (1980)
15. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Proc. IEEE **86**, 2278 (1998)

16. A. Krizhevsky, I. Sutskever, G.E. Hinton, in *Advances in Neural Information Processing Systems 25*, ed. by P. Bartlett, F. Pereira, C. Burges, L. Bottou, K. Weinberger (2012), pp. 1097–1105
17. D.H. Ackley, G.E. Hinton, T.J. Sejnowski, Cognit. Sci. **9**(1), 147 (1985)
18. E.T. Jaynes, Phys. Rev. **106**, 620 (1957)
19. E. Schneidman, M.J. Berry, R. Segev, W. Bialek, Nature **440**, 1007 (2006)
20. Y. Roudi, E. Aurell, J. Hertz, Front. Comput. Neurosci **3**, 1 (2009)
21. H.C. Nguyen, R. Zecchina, J. Berg, Adv. Phys. **66**(3), 197 (2017)
22. P. Smolensky, Information processing in dynamical systems: foundations of harmony theory (MIT Press, Cambridge, 1986), pp. 194–281
23. Y. Freund, D. Haussler, Unsupervised learning of distributions on binary vectors using two layer networks. Technical Report, Santa Cruz, CA, USA (1994)
24. G. Hinton, R. Salakhutdinov, Science (New York, N.Y.) **313**, 504 (2006)
25. J.L. Elman, Cognit. Sci. **14**(2), 179 (1990)
26. P.J. Werbos, Neural Netw. **1**(4), 339 (1988)
27. F.J. Pineda, Phys. Rev. Lett. **59**(19), 2229 (1987)
28. R.J. Williams, J. Peng, Neural Comput. **2**(4), 490 (1990)
29. R. Pascanu, T. Mikolov, Y. Bengio, in *Proceedings of the 30th International Conference on Machine Learning* (2013), pp. 1310–1318
30. S. Hochreiter, J. Schmidhuber, Neural Comput. **9**(8), 1735 (1997)
31. D. Sussillo, L. Abbott, Neuron **63**(4), 544 (2009)
32. D.V. Buonomano, W. Maass, Nat. Rev. Neurosci. **10**(2), 113 (2009)
33. S. Vyas, M.D. Golub, D. Sussillo, K.V. Shenoy, Ann. Rev. Neurosci. **43**(1), 249 (2020)
34. J. Schmidhuber, Neural Netw. **61**, 85 (2015)
35. Y. LeCun, Y. Bengio, G. Hinton, Nature **521**(7553), 436 (2015)
36. P. Mehta, M. Bukov, C.H. Wang, A.G. Day, C. Richardson, C.K. Fisher, D.J. Schwab, Phys. Rep. **810**, 1 (2019)
37. D.J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, Cambridge, 1989)
38. D.J. Amit, H. Gutfreund, H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985)
39. D.J. Amit, H. Gutfreund, H. Sompolinsky, Physical Review Letters **55**(14), 1530 (1985)
40. M. Mézard, G. Parisi, M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987)
41. P. Peretto, Biolog. Cybern. **50**(1), 51 (1984)
42. E. Gardner, Europhys. Lett. (epl) **4**, 481 (1987)
43. E. Gardner, J. Phys. A: Math. Gen. **21**, 257 (1988)
44. H. Sompolinsky, A. Crisanti, H.J. Sommers, Physical Review Letters **61**(3), 259 (1988)
45. A. Crisanti, H. Sompolinksy, Phys. Rev. E **98**(6), 62120 (2018)
46. M. Helias, D. Dahmen (2019). arXiv:1901.10416
47. C. van Vreeswijk, H. Sompolinsky, Science **274**(5293), 1724 (1996)
48. C. van Vreeswijk, H. Sompolinsky, Neural Comput. **10**(6), 1321 (1998)
49. M.N. Shadlen, W.T. Newsome, J. Neurosci. **18**(10), 3870 (1998)
50. M. Okun, I. Lampl, Nat. Neurosci. **11**(5), 535 (2008)
51. N. Brunel, J. Comput. Neurosci. **8**(3), 183 (2000)
52. M.L. Mehta, *Random Matrices* (Academic, San Diego, 2004)

# Chapter 2
# Spin Glass Models and Cavity Method


Spin glasses are magnets with two-spin interactions of random signs (Mézard et al., in Spin Glass Theory and Beyond. World Scientific, Singapore, 1987 [1]), e.g., an alloy with randomly localized magnetic moments. In spin glass models, the randomness emerges in spin interactions. For example, in the Sherrington–Kirkpatrick model (Sherrington and Kirkpatrick in Phys. Rev. Lett. 35(26):1792, 1975 [2]), all two-spin interactions follow independently a Gaussian distribution with variance $N^{-1/2}$ ($N$ is the system size); in the Edwards–Anderson model (Edwards and Anderson in J. Phys. F: Metal Phys. 5(5):965, 1975 [3]), the spins sit on a finite-dimensional lattice, and in the Bethe lattice model (Viana and Bray in J. Phys. C: Solid State Phys. 18(15):3037, 1985 [4]; Mézard and Parisi in Eur. Phys. J. B 20:217, 2001 [5]), the spins locate at a random lattice of finite connectivity for each spin. All these models belong to the category of multi-spin interaction models, originally studied in physics, later widely explored in the context of optimization problems in computer science, machine learning and computational neuroscience.

## 2.1 Multi-spin Interaction Models

Before going into details of the underlying physics, we would like to give a few seminal applications of the spin glass models. The first one is the random $K$-SAT problem. The random $K$-SAT problem is finding a solution, say an assignment of $N$ Boolean variables, to satisfy a random formula composed of logical AND of $M$ clauses. Each clause is expressed as a logical OR function of $K$ randomly selected distinct variables (either directed or negated with equal probability) from the variable set. For example, one short formula is given by

$$\mathcal{F} = (\overline{x}_3 \vee x_7 \vee \overline{x}_2) \wedge (\overline{x}_1 \vee x_5 \vee \overline{x}_6) \wedge (\overline{x}_4 \vee x_7 \vee \overline{x}_5). \tag{2.1}$$



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_2

From a physics viewpoint, the random $K$-SAT can be treated as a spin glass problem with a focus on the typical case analysis.[1] If $x_i$ is TRUE, then we transform it to an Ising spin with value 1 (spin up); otherwise, it is transformed to $-1$ (spin down). Given a configuration of spins, the number of violated clauses can be defined as an energy function in statistical physics [6],

$$E(\boldsymbol{\sigma}) = \sum_{m=1}^{M} \prod_{j=1}^{K} \frac{1 + J_j^m \sigma_{i_j^m}}{2}, \tag{2.2}$$

where $i_j^m$ is the $j$th variable appearing in the $m$th clause. The quenched disorder $J_j^m$ is 1 if the Boolean variable in the formula appears negated and $-1$ otherwise. Hence, the constraint satisfaction problem reduces to a physics problem of finding minima of the energy function.

Analogously, the random $K$-XOR SAT formula can be written as

$$\mathcal{F} = \bigwedge_{m=1}^{M} \left[ \left( \bigoplus_{j=1}^{K} x_{i_j^m} \right) \oplus y_m \right], \tag{2.3}$$

where the symbol $\oplus$ denotes the logical XOR operation, and $y_m$ is quenched random Boolean value. This formula corresponds to a linear system, with an efficient solver of the Gaussian elimination procedure. In physics, the diluted Ising $p$-spin model with coupling $\pm 1$ belongs to the class of random $K$-XOR problem. Similar to the random $K$-SAT Problem, one can easily write down the associated energy function [7]

$$E(\boldsymbol{\sigma}) = \sum_{m=1}^{M} \frac{1 - J_m \prod_{j=1}^{K} \sigma_{i_j^m}}{2}, \tag{2.4}$$

where $J_m$ is an Ising-mapping of the Boolean variable $y_m$.

The above two constraint satisfaction problems belong to multi-spin interaction models in physics. Physicists are interested in studying the mean-field limit $N \to \infty$ and $M \to \infty$ but keeping the ratio $M/N$ constant. One expects that rich phase transitions emerge due to complex interactions among spin variables. Next, we will illustrate how the cavity method can be used to compute the free energy function of this class of models. Cavity method was first proposed to reproduce the replica results of the Sherrington–Kirkpatrick model [8], and further reformulated in the study of neural networks [9], and was finally proposed at the concise physics level and systematic mathematical level on the Bethe lattice, a broad class of glass models of finite connectivities [5]. We will also explore the core physics assumption behind the cavity method in detail in a multi-spin interaction model. The multi-spin interaction models like the above two cases can be analogously treated.

[1] The computational complexity is defined in the worst-case setting.

The multi-spin interaction model can be also defined in the context of information transmission, for which we shall give a concrete example. Let us consider the case that we want to send a message to a receiver, and the message may be perturbed during transmission because of the noise in the channel. It is a highly non-trivial task for the receiver to retrieve the original message from the perturbed one. One solution is to introduce redundancy to the original message at the sender site. Then the receiver can correct some transmission errors according to the redundancy. In 1948, Claude Shannon proved that error-free transmission is possible when the code rate is below the channel capacity, which establishes a fundamental bound for designing engineering practical codes [10]. Numerous efforts have been devoted to design the codes approaching Shannon's bound (channel capacity). Among them, the Sourlas code is the first one in physics [11], which relates error-correcting codes to a spin glass model.

It is easy to figure out how to construct a Sourlas code. Supposed that we have an $N$-bit binary original message $\boldsymbol{\xi} \in \{\pm 1\}^N$, and then encode them into an $M$-bit transmitted message $\boldsymbol{J^0} = \{J_1^0, J_2^0, \ldots, J_M^0\}$. The $a$th bit of $J^0$ is the product of a subset $\partial a$ of randomly selected original message bits,

$$J_a^0 = \prod_{i \in \partial a} \xi_i. \tag{2.5}$$

We then denote $J_a$ as the $a$th bit of the received message, which may not be equal to the transmitted message due to the channel noise flipping message bits. We further assume that each bit of transmitted messages can be independently flipped with the same probability $p$. Hence, the conditional probability of a received message given a transmitted one reads

$$P(J_a|J_a^0) = p\delta(J_a + J_a^0) + (1-p)\delta(J_a - J_a^0). \tag{2.6}$$

To decode the sent message, we write the computational task as a statistical mechanics problem with the following Hamiltonian [12]:

$$H(\boldsymbol{\sigma}) = -\sum_{a=1}^{M} J_a \prod_{i \in \partial a} \sigma_i, \tag{2.7}$$

where $\sigma_i$ is the dynamical binary spin variable for decoding the original message $\{\xi_i\}$. What we need to do is computing the posterior probability $P(\boldsymbol{\sigma}|\boldsymbol{J})$ which is given by

$$P(\boldsymbol{\sigma}|\boldsymbol{J}) = \frac{\exp(-\beta H(\boldsymbol{\sigma}))}{Z}, \tag{2.8}$$

where $\beta$ is the inverse temperature, and $Z$ is the partition function. This decoding process amounts now to searching for the ground state of the statistical mechanics problem. The energy of the model takes a minimal value if $\prod_{i \in \partial a} \sigma_i = J_a$. According

to the canonical ensemble theory, all emergent properties of the decoding process are included in the partition function that is mathematically formulated as follows:

$$Z = \sum_{\{\sigma_i\}} \exp(-\beta H(\boldsymbol{\sigma})). \tag{2.9}$$

## 2.2 Cavity Method

In this section, we apply the cavity approximation to compute the partition function. Notice that a direct calculation of Eq. (2.9) involves in summing up $2^N$ terms, which is computationally impossible once $N > 30$. The cavity method can reduce the computational cost down to the order of $O(N)$ for a sparsely connected factor graph model. Let us explain this in detail as follows.

The model can be represented by a factor graph [13], illustrating how spins interact with each other (see Fig. 2.1). Because we aim at analyzing the Boltzmann distribution, i.e., the posterior [Eq. (2.8)], we use the probabilistic language, for the goal of computing the marginal probability as well. To achieve this, we should modify the original graph that allows strong correlations among variables. If we add one function node $a$ to the original system (see Fig. 2.2), the Hamiltonian of the new system can be written as the sum of the Hamiltonian of the original system and the change caused by the newly added function node. More precisely,

$$H^{\text{new}} = H^{\text{old}} - J_a \prod_{i \in \partial a} \sigma_i. \tag{2.10}$$

It then proceeds that

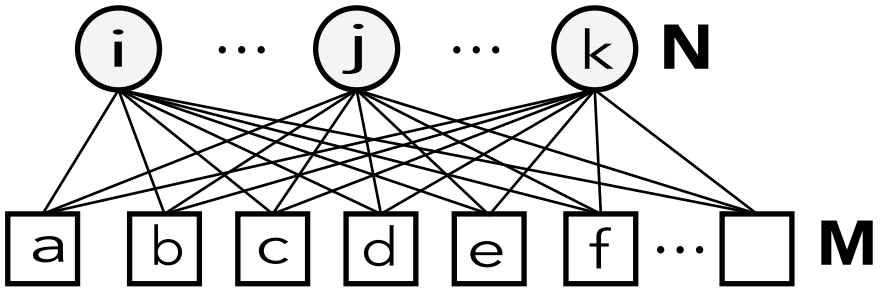



**Fig. 2.1** Factor graph representation of a random construction of the Sourlas codes. Circles are spin variables (variable nodes) $\{\sigma_i\}$, and squares are received message (function nodes) $\{J_a\}$. In the figure, we only show three message bits, and each square is connected to them. In fact, the square can be connected to other different message bits (not shown), thereby forming a sparse random graph, where the degree of variable nodes follows a Poisson distribution

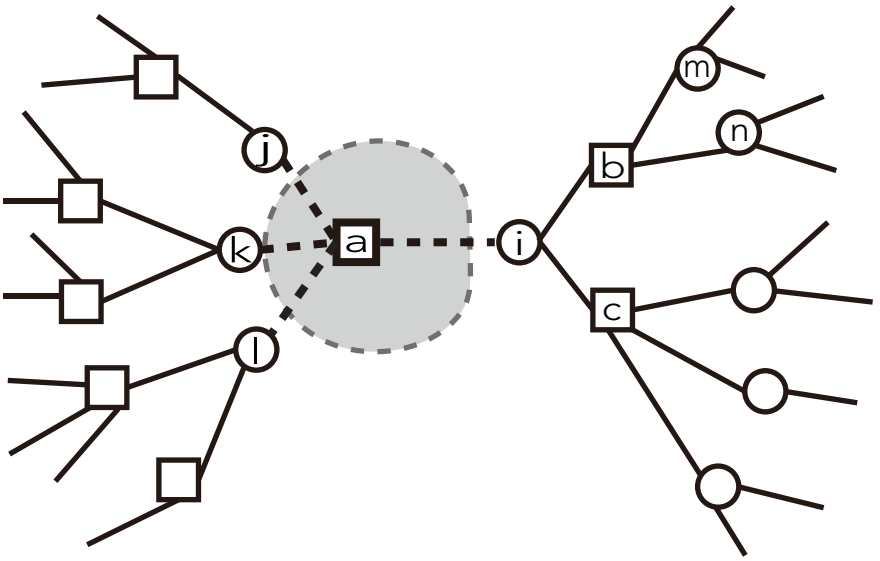



**Fig. 2.2** Addition of the function node $a$ to original system (outside the shadow part). We call the shadow part a cavity, and the nodes $\{i, j, k, l\}$ serve as the boundary of the cavity

$$\begin{aligned} Z^{\text{new}} &= \sum_{\{\sigma_i\}_{i=1}^N} \exp\left(-\beta H^{\text{old}} + \beta J_a \prod_{i\in\partial a} \sigma_i\right) \\ &= Z^{\text{old}} \sum_{\{\sigma_i\}_{i=1}^N} \frac{\exp(-\beta H^{\text{old}})}{Z^{\text{old}}} \exp\left(\beta J_a \prod_{i\in\partial a} \sigma_i\right). \end{aligned} \tag{2.11}$$

It is easy to see that $\frac{\exp(-\beta H^{\text{old}})}{Z^{\text{old}}}$ is exactly the joint probability distribution of $\{\sigma_i\}_{i=1}^N$ in the old system. One can find that $\exp(\beta J_a \prod_{i\in\partial a} \sigma_i)$ only relates to $\{\sigma_i | i \in \partial a\}$, and then we can divide the configuration sum into two parts: one involves in only variable nodes with direct connections to the newly added functional node $a$, and the other involves in the rest. We then have

$$\begin{aligned} Z^{\text{new}} &= Z^{\text{old}} \sum_{\{\sigma_i | i\in\partial a\}} \sum_{\{\sigma_i | i\notin\partial a\}} \frac{\exp(-\beta H^{\text{old}})}{Z^{\text{old}}} \exp(\beta J_a \prod_{i\in\partial a} \sigma_i) \\ &= Z^{\text{old}} \sum_{\{\sigma_i | i\in\partial a\}} \exp(\beta J_a \prod_{i\in\partial a} \sigma_i) \sum_{\{\sigma_i | i\notin\partial a\}} \frac{\exp(-\beta H^{\text{old}})}{Z^{\text{old}}}. \end{aligned} \tag{2.12}$$

The last summation in Eq. (2.12) is exactly the marginal probability distribution of $\{\sigma_i | i \in \partial a\}$ in the old system. Compared with the new system, the old system has a cavity in the position where the function node $a$ is added to the new system. Therefore, we denote the marginal probability as a cavity distribution $P_{\text{cavity}}(\{\sigma_i | i \in \partial a\})$ and

call $\{\sigma_i | i \in \partial a\}$ as the boundary of the cavity. It is reasonable to assume that variable nodes on the boundary of the cavity are weakly correlated, because of the weak couplings in a fully connected system or the sparsely connected topology of a sparse model. This assumption is exact if the underlying factor graph is a tree. Thus, the $P_{\text{cavity}}(\{\sigma_i | i \in \partial a\})$ can be factorized as

$$P_{\text{cavity}}(\{\sigma_i | i \in \partial a\}) \approx \prod_{i \in \partial a} q_{i \to a}(\sigma_i), \tag{2.13}$$

where $q_{i \to a}$ denotes the distribution of $\sigma_i$ without the presence of the function node $a$. Let us then define a cavity magnetization $m_{i \to a} \equiv q_{i \to a}(\sigma_i = +1) - q_{i \to a}(\sigma_i = -1)$, and thus, $q_{i \to a}(\sigma_i) = \frac{1+\sigma_i m_{i \to a}}{2}$. Then $Z^{\text{new}}$ can be rewritten as

$$\begin{aligned} Z^{\text{new}} &= Z^{\text{old}} \sum_{\{\sigma_i | i \in \partial a\}} \exp(\beta J_a \prod_{i \in \partial a} \sigma_i) \prod_{i \in \partial a} \frac{1 + \sigma_i m_{i \to a}}{2} \\ &= Z^{\text{old}} \cosh(\beta J_a) \left(1 + \tanh(\beta J_a) \prod_{i \in \partial a} m_{i \to a}\right). \end{aligned} \tag{2.14}$$

According to the free energy definition $F = -1/\beta \ln Z$, the free energy shift due to adding the function node $a$ is given by

$$-\beta \Delta F_a = \ln \frac{Z^{\text{new}}}{Z^{\text{old}}} = \ln \left[ \cosh(\beta J_a) \left(1 + \tanh(\beta J_a) \prod_{i \in \partial a} m_{i \to a}\right) \right]. \tag{2.15}$$

Similarly, if we add one variable node $i$ and its neighboring function nodes $\{b | b \in \partial i\}$ to the system (see Fig. 2.3),[2] the partition function of the new system reads

$$\begin{aligned} Z^{\text{new}} &= \sum_{\boldsymbol{\sigma}^{\text{old}}} \sum_{\sigma_i} \exp\left(-\beta H^{\text{old}} + \beta \sum_{b \in \partial i} J_b \prod_{j \in \partial b} \sigma_j\right) \\ &= \sum_{\boldsymbol{\sigma}^{\text{old}}} \sum_{\sigma_i} \exp\left(-\beta H^{\text{old}} + \beta \sum_{b \in \partial i} J_b \sigma_i \prod_{j \in \partial b \backslash i} \sigma_j\right) \\ &= Z^{\text{old}} \sum_{\boldsymbol{\sigma}^{\text{old}}} \sum_{\sigma_i} \frac{\exp(-\beta H^{\text{old}})}{Z^{\text{old}}} \exp\left(\beta \sum_{b \in \partial i} J_b \sigma_i \prod_{j \in \partial b \backslash i} \sigma_j\right), \end{aligned} \tag{2.16}$$

where $j \in \partial b \backslash i$ denotes the set of variable nodes with connections to the function node $b$, yet the node $i$ is excluded. The subset of nodes $\{\sigma_j | j \in \partial b \backslash i, b \in \partial i\}$ is the boundary of the cavity (see Fig. 2.3). We can first sum over the configuration of all

[2] This operation will make the definition of cavity probabilities reasonable.

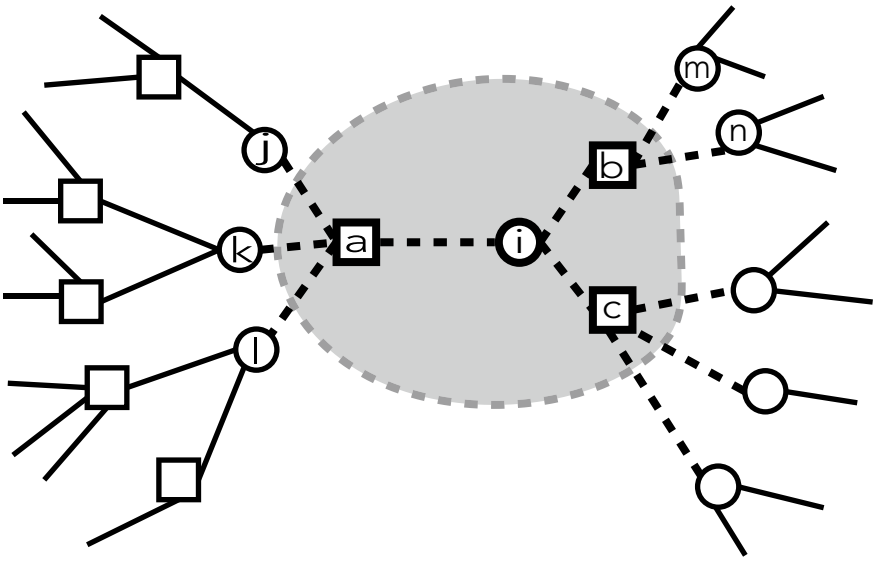



**Fig. 2.3** Adding a variable node $i$ together with its neighboring function nodes $\{a, b, c\}$ to the original system (outside the cavity). The subset $\{j, k, l, m, n \ldots\}$ denotes the boundary of the cavity

variable nodes that are not at the boundary of the cavity (except $i$), akin to what we have done in Eq. (2.12). Using Eq. (2.13), we then arrive at the following result:

$$
\begin{aligned}
Z^{\text{new}} &= Z^{\text{old}} \sum_{\{\sigma_j | j\in\partial b\backslash i; b\in\partial i\}} \sum_{\sigma_i} P_{\text{cavity}}(\{\sigma_j | j \in \partial b\backslash i; b \in \partial i\}) \exp\Big(\sum_{b\in\partial i} \beta J_b \sigma_i \prod_{j\in\partial b\backslash i} \sigma_j\Big) \\
&\approx Z^{\text{old}} \sum_{\{\sigma_j | j\in\partial b\backslash i; b\in\partial i\}} \sum_{\sigma_i} \prod_{b\in\partial i} \prod_{j\in\partial b\backslash i} q_{j\to b}(\sigma_j) \exp\Big(\sum_{b\in\partial i} \beta J_b \sigma_i \prod_{j\in\partial b\backslash i} \sigma_j\Big) \\
&= Z^{\text{old}} \sum_{\sigma_i} \sum_{\{\sigma_j | j\in\partial b\backslash i; b\in\partial i\}} \prod_{b\in\partial i} \prod_{j\in\partial b\backslash i} q_{j\to b}(\sigma_j) \exp\Big(\sum_{b\in\partial i} \beta J_b \sigma_i \prod_{j\in\partial b\backslash i} \sigma_j\Big) \\
&= Z^{\text{old}} \sum_{\sigma_i} \prod_{b\in\partial i} \sum_{\{\sigma_j | j\in\partial b\backslash i\}} \prod_{j\in\partial b\backslash i} q_{j\to b}(\sigma_j) \exp\Big(\beta J_b \sigma_i \prod_{j\in\partial b\backslash i} \sigma_j\Big) \\
&= Z^{\text{old}} \sum_{\sigma_i} \prod_{b\in\partial i} \sum_{\{\sigma_j | j\in\partial b\backslash i\}} \prod_{j\in\partial b\backslash i} \frac{1+\sigma_j m_{j\to b}}{2} \exp\Big(\beta J_b \sigma_i \prod_{j\in\partial b\backslash i} \sigma_j\Big) \\
&= Z^{\text{old}} \left( \prod_{b\in\partial i} \Lambda^{+}_{b\to i} + \prod_{b\in\partial i} \Lambda^{-}_{b\to i} \right),
\end{aligned}
\tag{2.17}
$$

where

$$\begin{aligned}\Lambda^{+}_{b\to i} &= \sum_{\{\sigma_j | j\in\partial b\backslash i\}} \prod_{j\in\partial b\backslash i} \frac{1+\sigma_j m_{j\to b}}{2} \exp\left(\beta J_b \times (+1) \times \prod_{j\in\partial b\backslash i} \sigma_j\right) \\ &= \cosh(\beta J_b)\left(1+\tanh(\beta J_b)\prod_{j\in\partial b\backslash i} m_{j\to b}\right),\end{aligned} \tag{2.18}$$

$$\begin{aligned}\Lambda^{-}_{b\to i} &= \sum_{\{\sigma_j | j\in\partial b\backslash i\}} \prod_{j\in\partial b\backslash i} \frac{1+\sigma_j m_{j\to b}}{2} \exp\left(\beta J_b \times (-1) \times \prod_{j\in\partial b\backslash i} \sigma_j\right) \\ &= \cosh(\beta J_b)\left(1-\tanh(\beta J_b)\prod_{j\in\partial b\backslash i} m_{j\to b}\right).\end{aligned} \tag{2.19}$$

Hence, the free energy shift due to adding the variable node $i$ together with its neighboring function nodes $\{b \in \partial i\}$ is given by

$$-\beta\Delta F_i = \ln\frac{Z^{\text{new}}}{Z^{\text{old}}} = \ln\left[\prod_{b\in\partial i}\Lambda^{+}_{b\to i} + \prod_{b\in\partial i}\Lambda^{-}_{b\to i}\right]. \tag{2.20}$$

Finally, the total free energy is given by

$$F = \sum_i \Delta F_i + \sum_a \Delta F_a - \sum_a |\partial a|\Delta F_a, \tag{2.21}$$

where $|\partial a|$ is the number of variable nodes connecting to the function node $a$. The last term of Eq. (2.21) is to ensure that each node's contribution to the total free energy has been counted only once. Once we have access to $\{m_{j\to b}\}$, we can calculate the free energy function. In the next section, we explain how to calculate $\{m_{i\to a}\}$.

## 2.3 From Cavity Method to Message Passing Algorithms

According to the cavity assumption, the cavity magnetization $\{m_{i\to a}\}$ can be iteratively constructed, because the local structure of a random factor graph is statistically homogeneous. Note that $m_{i\to a}$ is the expectation value of $\sigma_i$ without the contribution from the function node $a$, which is expected from the definition of the cavity operation. Hence, $m_{i\to a}$ can be rewritten as follows:

$$m_{i\to a} = \frac{\sum_{\boldsymbol{\sigma}} \sigma_i \exp(-\beta H_{i\to a}(\boldsymbol{\sigma}))}{\sum_{\boldsymbol{\sigma}} \exp(-\beta H_{i\to a}(\boldsymbol{\sigma}))}, \tag{2.22}$$

where $H_{i\to a}$ denotes the Hamiltonian without the interaction $a$, which reads

$$H_{i\to a} = H_{\text{cavity}} - \sum_{b\in\partial i\backslash a} J_b\sigma_i \prod_{j\in\partial b\backslash i}\sigma_j, \tag{2.23}$$

where $H_{\text{cavity}}$ is the Hamiltonian of the cavity system where the variable node $i$ together with its neighboring function nodes $b \in \partial i$ (except $a$) are all removed from the original system.

Similar to what we have done in Sect. 2.2, we can sum over all possible configurations of variable nodes not on the boundary of the cavity at first, and we, thus, get the marginal distribution of the boundary nodes in the cavity system. We then have

$$\begin{aligned} m_{i\to a} &= \frac{\frac{\sum_{\boldsymbol{\sigma}}\sigma_i \exp(-\beta H_{i\to a}(\boldsymbol{\sigma}))}{Z_{\text{cavity}}}}{\frac{\sum_{\boldsymbol{\sigma}}\exp(-\beta H_{i\to a}(\boldsymbol{\sigma}))}{Z_{\text{cavity}}}} \\ &= \frac{\sum_{\sigma_i}\sum_{\mathcal{B}}\sigma_i P_{\text{cavity}}(\mathcal{B})\exp(\sum_{b\in\partial i\backslash a}\beta J_b\sigma_i\prod_{j\in\partial b\backslash i}\sigma_j)}{\sum_{\sigma_i}\sum_{\mathcal{B}}P_{\text{cavity}}(\mathcal{B})\exp(\sum_{b\in\partial i\backslash a}\beta J_b\sigma_i\prod_{j\in\partial b\backslash i}\sigma_j)}, \end{aligned} \tag{2.24}$$

where $Z_{\text{cavity}}$ denotes the partition function related to $H_{\text{cavity}}$, and $\mathcal{B} \equiv \{\sigma_j | j \in \partial b\backslash i; b \in \partial i\backslash a\}$, which denotes the boundary of the cavity. Then we factorize the cavity probability according to the cavity approximation:

$$P_{\text{cavity}}(\mathcal{B}) \approx \prod_{b\in\partial i\backslash a}\prod_{j\in\partial b\backslash i} q_{j\to b}(\sigma_j). \tag{2.25}$$

Using the same techniques as in Eq. (2.17), we finally arrive at

$$m_{i\to a} = \frac{\prod_{b\in\partial i\backslash a}\Lambda^{+}_{b\to i} - \prod_{b\in\partial i\backslash a}\Lambda^{-}_{b\to i}}{\prod_{b\in\partial i\backslash a}\Lambda^{+}_{b\to i} + \prod_{b\in\partial i\backslash a}\Lambda^{-}_{b\to i}}. \tag{2.26}$$

If we define the conjugate cavity magnetization as

$$\hat{m}_{b\to j} \equiv \tanh(\beta J_b)\prod_{j\in\partial b\backslash i} m_{j\to b}, \tag{2.27}$$

we can then write Eq. (2.26) into the following form:

$$m_{i\to a} = \frac{\prod_{b\in\partial i\backslash a}(1+\hat{m}_{b\to i}) - \prod_{b\in\partial i\backslash a}(1-\hat{m}_{b\to i})}{\prod_{b\in\partial i\backslash a}(1+\hat{m}_{b\to i}) + \prod_{b\in\partial i\backslash a}(1-\hat{m}_{b\to i})}. \tag{2.28}$$

The above expression can be transformed into the language of cavity fields, e.g., a cavity local field $h_{i\to a}$ and cavity bias $u_{a\to i}$ as also defined in the seminal work [5]. We can then use these fields or biases to parameterize the cavity probability:

$$
\begin{aligned}
q_{i\to a}(\sigma_i) &\equiv \frac{\exp(\beta h_{i\to a}\sigma_i)}{2\cosh \beta h_{i\to a}},\\
p_{a\to i}(\sigma_i) &\equiv \frac{\exp(\beta u_{a\to i}\sigma_i)}{2\cosh \beta u_{a\to i}},
\end{aligned}
\tag{2.29}
$$

where $p_{a\to i}(\sigma_i) = \frac{1+\sigma_i \hat{m}_{a\to i}}{2}$. It then proceeds that

$$
\begin{aligned}
m_{i\to a} &= \tanh \beta h_{i\to a},\\
\hat{m}_{a\to i} &= \tanh \beta u_{a\to i}.
\end{aligned}
\tag{2.30}
$$

Therefore, Eqs. (2.27) and (2.28) turn out to be

$$
\begin{aligned}
h_{i\to a} &= \frac{1}{\beta}\left(\sum_{b\in \partial i\backslash a} \beta u_{b\to i}\right),\\
u_{a\to i} &= \frac{1}{\beta}\tanh^{-1}\left[\tanh(\beta J_a)\prod_{j\in\partial a\backslash i}\tanh(\beta h_{j\to a})\right].
\end{aligned}
\tag{2.31}
$$

Equation (2.31) is the very message passing equation in the Sourlas-code scenario. In essence, the cavity method is a probabilistic iterative method. One can iteratively solve these equations, until a fixed point of messages ($\{m_{i\to a}\}$) is reached. These messages are then used to evaluate the full magnetization $m_i$ as follows:

$$
m_i = \tanh\left(\sum_{b\in\partial i}\beta u_{b\to i}\right),
\tag{2.32}
$$

and the sent message can be decoded by the maximizer of the posterior marginal (MPM), i.e., $\sigma_i = \arg\max_{\sigma_i} P_i(\sigma_i)$, where $P_i(\sigma_i) = \frac{1+\sigma_i m_i}{2}$. In addition, the free energy and other thermodynamic quantities of interest can be evaluated according to the derived formulas. The computational complexity is clearly of the order of $O(N)$ for a sparsely connected factor graph and the order $O(N^2)$ in the case that all variable nodes connect to each function node. We remark that this procedure is quite general and can be adapted to learning problems of a variety of neural networks, which we shall introduce in the remaining chapters.

## References


1. M. Mézard, G. Parisi, M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987)
2. D. Sherrington, S. Kirkpatrick, Phys. Rev. Lett. **35**(26), 1792 (1975)
3. S.F. Edwards, P.W. Anderson, J. Phys. F: Metal Phys. **5**(5), 965 (1975)
4. L. Viana, A.J. Bray, J. Phys. C: Solid State Phys. **18**(15), 3037 (1985)

5. M. Mézard, G. Parisi, Eur. Phys. J. B **20**, 217 (2001)
6. M. Mézard, R. Zecchina, Phys. Rev. E **66**(5), 056126 (2002)
7. S. Franz, M. Leone, F. Ricci-Tersenghi, R. Zecchina, Phys. Rev. Lett. **87**(12), 127209 (2001)
8. M. Mézard, G. Parisi, M.A. Virasoro, Europhys. Lett. (EPL) **1**(2), 77 (1986)
9. M. Mezard, J. Phys. A **22**(12), 2181 (1989)
10. C.E. Shannon, Bell Syst. Tech. J. **27**(3), 379 (1948)
11. N. Sourlas, Nature **339**(6227), 693 (1989)
12. H. Huang, H. Zhou, Phys. Rev. E **80**, 056113 (2009)
13. F. Kschischang, B. Frey, H.A. Loeliger, IEEE Trans. Inf. Theory **47**(2), 498 (2001)

# Chapter 3
# Variational Mean-Field Theory and Belief Propagation


In the previous chapter, we have introduced the cavity method and its application to computing the approximate free energy of a multi-spin interaction model, and the approximation is equivalent to the Bethe approximation, which we shall provide an in-depth introduction in this chapter. In this chapter, we apply the variational method together with the mean-field approximation (MFA) and Bethe approximation (BA) to construct the free energy of the multi-spin interaction model. We show that the belief propagation (BP) algorithm in computer science can be derived under the variational framework, which is in fact equivalent to the cavity method in physics. Furthermore, we emphasize that BA is a more accurate approximation, which reduces to MFA when the coupling is relatively weak or when a high-temperature limit is performed. Finally, we give a brief introduction of the inverse Ising model, where model parameters (couplings and fields) can be learned by using the mean-field methods. Besides being a useful tool in statistical physics, the BP algorithm is also an efficient way to solve many important inference problems in areas of computer science, modern coding and learning in neural networks—one focus of this book.

## 3.1 Variational Method

The variational method is an important technique for statistical inference problems. With the target function we want to optimize and some constraints the problem should satisfy, we can apply the variation of model parameters on the target function. We take a simple example of the derivation of the Boltzmann distribution in statistical physics. The entropy of a system in statistical physics can be defined by

$$S = -k \sum_r P_r \ln P_r, \tag{3.1}$$

where $k$ indicates the Boltzmann constant, $r$ is the index of a thermodynamic state and $P_r$ is the to-be-determined distribution of the state $r$. According to the theory of



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_3

thermodynamics, the system is in equilibrium when the entropy reaches its maximum, and the distribution must meet the following two constraints:

$$\sum_r P_r = 1, \tag{3.2}$$

$$\sum_r E_r P_r = \mu, \tag{3.3}$$

which correspond to the normalization of a probability measure and a target mean energy level $\mu$ of the system, respectively. Hence, we can use the Lagrange multiplier method:

$$L = S + \lambda_1\Big(\sum_r P_r - 1\Big) + \lambda_2\Big(\sum_r E_r P_r - \mu\Big), \tag{3.4}$$

where $\lambda_1$ and $\lambda_2$ are the Lagrange multipliers for the two constraints, respectively. Then, the equilibrium requires that $\frac{\partial L}{\partial P_r} = 0$, and we finally arrive at

$$\begin{aligned} P_r &= \frac{e^{-\beta E_r}}{Z}, \\ Z &= \sum_r e^{-\beta E_r}, \end{aligned} \tag{3.5}$$

where the inverse temperature $\beta = \frac{1}{kT}$ can be deduced from the second law of thermodynamics, and $Z$ is the partition function, namely the normalization constant to enforce the first constraint. In the following, we assume $k = 1$ for optimization problems in a high-dimensional parameter space.

In sum, from the Lagrange multiplier method with a little knowledge from the equilibrium thermodynamics, we derive the well-known Boltzmann distribution, where the inverse temperature clearly tunes the energy level of the system [1].

## 3.2 Variational Free Energy

The behavior of the free energy contributes to the emergent behavior of a thermodynamic system. However, calculating the free energy in a brute-force way is intractable due to the $O(2^N)$ computational complexity. To overcome the barrier, the variational method provides an effective way to construct an approximate free energy. We take an example of the above-mentioned multi-spin interaction model which is captured by the following Boltzmann distribution:

$$\begin{aligned} P(\boldsymbol{x}) &= \frac{e^{-\beta E(\boldsymbol{x})}}{Z}, \\ Z &= \sum_{\boldsymbol{x}} e^{-\beta E(\boldsymbol{x})}, \end{aligned} \tag{3.6}$$

where $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ represents the state of $N$ spins in the system. The energy $E(\boldsymbol{x})$ is given by

$$E(\boldsymbol{x}) = -\sum_a J_a \prod_{i \in \partial a} x_i, \tag{3.7}$$

where $a$ is the index of the interaction, and $i \in \partial a$ specifies the set of spins that participate in the $a$th interaction where we use $\boldsymbol{x}_a$ to represent these spins. $J_a$ is the coupling strength of the $a$th interaction. The inverse temperature here can be set to an arbitrary value, and in an equivalent way, the temperature can be absorbed into the coupling $J_a$. We, thus, set $\beta = 1$ without loss of generality. We further define $f_a(\boldsymbol{x}_a) = e^{J_a \prod_{i \in \partial a} x_i}$, which denotes the contribution of the $a$th interaction to the Boltzmann measure. Thus, we can rewrite the distribution $P(\boldsymbol{x})$ and energy $E(\boldsymbol{x})$ into the following forms:

$$P(\boldsymbol{x}) = \frac{1}{Z} \prod_a f_a(\boldsymbol{x}_a), \tag{3.8}$$

$$E(\boldsymbol{x}) = -\sum_a \ln f_a(\boldsymbol{x}_a). \tag{3.9}$$

These expressions facilitate the following derivation of BP algorithm.

The Helmholtz free energy reads

$$F_H = -\ln Z. \tag{3.10}$$

As we mentioned above, an exact computation of the Helmholtz free energy is impossible for a large-size system. Instead, we introduce a *trial* probability distribution $b(\boldsymbol{x})$ and write the free energy, which is called the Gibbs free energy with some parameters (e.g., magnetizations) to be optimized:

$$F(b) = U(b) - H(b), \tag{3.11}$$

where we define $U(b)$ as the variational internal energy:

$$U(b) = \sum_{\boldsymbol{x}} b(\boldsymbol{x}) E(\boldsymbol{x}), \tag{3.12}$$

and $H(b)$ as the variational entropy:

$$H(b) = -\sum_{\boldsymbol{x}} b(\boldsymbol{x}) \ln b(\boldsymbol{x}). \tag{3.13}$$

It is then necessary to compute the difference between the Gibbs free energy and the Helmholtz free energy, as given by

$$\begin{aligned} F(b) - F_H &= \sum_{\boldsymbol{x}} b(\boldsymbol{x}) E(\boldsymbol{x}) + \sum_{\boldsymbol{x}} b(\boldsymbol{x}) \ln b(\boldsymbol{x}) + \ln Z \\ &= \sum_{\boldsymbol{x}} b(\boldsymbol{x})(-\ln Z - \ln P(\boldsymbol{x})) + \sum_{\boldsymbol{x}} b(\boldsymbol{x}) \ln b(\boldsymbol{x}) + \ln Z \\ &= \sum_{\boldsymbol{x}} b(\boldsymbol{x}) \ln \frac{b(\boldsymbol{x})}{P(\boldsymbol{x})} = D(b||P), \end{aligned} \tag{3.14}$$

where $D(b||P)$ is the Kullback–Leibler divergence between two probability distributions $b(\boldsymbol{x})$ and $P(\boldsymbol{x})$, which is always non-negative and is zero only if $b(\boldsymbol{x}) = P(\boldsymbol{x}), \forall \boldsymbol{x}$. Therefore, $F(b) \geq F_H$ and $F(b) = F_H$ only if $b(\boldsymbol{x}) = P(\boldsymbol{x}), \forall \boldsymbol{x}$.

The above analysis shows that the trial probability distribution $b(\boldsymbol{x})$ yielding a lower Gibbs free energy will have a smaller distance from the true distribution $P(\boldsymbol{x})$. That is to say, we transform the original free energy estimation problem to a (Gibbs) free energy minimization problem. To obtain a more accurate free energy, we must find a $b(\boldsymbol{x})$ to minimize the Gibbs free energy $F(b)$, which is exactly what the variational method wants to do. To proceed, we have to specify the trial probability $b(\boldsymbol{x})$ by introducing the so-called variational parameters, which can be physics-relevant quantities. In the next sections, we introduce two kinds of approximations for $b(\boldsymbol{x})$, which are mean-field and the Bethe approximations.

### 3.2.1 Mean-Field Approximation

The mean-field approximation for $b(\boldsymbol{x})$ is written in a factorized form:

$$b_{MF}(\boldsymbol{x}) = \prod_i b_i(x_i) = \prod_i \frac{1 + m_i x_i}{2}, \tag{3.15}$$

where $\boldsymbol{m}$ is the magnetization vector of spins $\boldsymbol{x}$. This approximation is the naive one that assumes each spin behaves independently of each other. Note that $x_i$ can only take two values $\pm 1$ (e.g., spin up and down, respectively). Given the form of $b_{MF}(\boldsymbol{x})$, we can compute the mean-field internal energy and mean-field entropy as follows:

$$\begin{aligned} U_{MF} &= \sum_{\boldsymbol{x}} b_{MF}(\boldsymbol{x}) E(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}} \prod_i b_i(x_i) \left( - \sum_a J_a \prod_{i \in \partial a} x_i \right) \\ &= - \sum_a J_a \left\langle \prod_{i \in \partial a} x_i \right\rangle \\ &= - \sum_a J_a \prod_{i \in \partial a} m_i, \end{aligned} \tag{3.16}$$

and

$$
\begin{aligned}
H_{MF} &= -\sum_{\boldsymbol{x}} b_{MF}(\boldsymbol{x}) \ln b_{MF}(\boldsymbol{x}) \\
&= -\sum_{\boldsymbol{x}} \prod_{i} \frac{1+m_i x_i}{2} \ln \prod_{i} \frac{1+m_i x_i}{2} \\
&= -\sum_{i} \sum_{\boldsymbol{x}} \prod_{j} \frac{1+m_j x_j}{2} \ln \frac{1+m_i x_i}{2} \\
&= -\sum_{i} \sum_{x_i} \sum_{\boldsymbol{x} \setminus x_i} \prod_{j(\neq i)} \frac{1+m_j x_j}{2} \frac{1+m_i x_i}{2} \ln \frac{1+m_i x_i}{2} \\
&= -\sum_{i} \sum_{x_i} \frac{1+m_i x_i}{2} \ln \frac{1+m_i x_i}{2} \\
&= \sum_{i} S_i,
\end{aligned}
\tag{3.17}
$$

where $S_i$ is defined as the entropy of spin $x_i$, and the symbol $\setminus$ indicates the operation of exclusion. Thus, the mean-field free energy can be derived as follows:

$$
\begin{aligned}
F_{MF} &= U_{MF} - H_{MF} \\
&= -\sum_{a} J_a \prod_{i \in \partial a} m_i + \sum_{i} \sum_{x_i} \frac{1+m_i x_i}{2} \ln \frac{1+m_i x_i}{2}.
\end{aligned}
\tag{3.18}
$$

The normalization constraint is automatically satisfied by the factorized form of the naive mean-field distribution [Eq. (3.15)]. The magnetization now becomes the variational parameter for the trial probability $b_{MF}(\boldsymbol{x})$. To minimize the upper bound of the Helmholtz free energy, we have to compute $\frac{\partial F_{MF}}{\partial m_i}$ and set the gradient to zero:

$$
\begin{aligned}
\frac{\partial F_{MF}}{\partial m_i} &= -\sum_{a} J_a \prod_{j \in \partial a \setminus i} m_j + \sum_{x_i} \frac{x_i}{2} \ln \frac{1+m_i x_i}{2} + \frac{x_i}{2} \\
&= -\sum_{a} J_a \prod_{j \in \partial a \setminus i} m_j + \frac{1}{2} \ln \frac{1+m_i}{1-m_i} = 0,
\end{aligned}
\tag{3.19}
$$

and finally, we derive the recursive-form of $m_i$:

$$
m_i = \tanh\left( \sum_{a \in \partial i} J_a \prod_{j \in \partial a \setminus i} m_j \right).
\tag{3.20}
$$

To obtain the fixed-point (equilibrium) values of $\boldsymbol{m}$, we can run these equations until a stationary point is reached. Using these equilibrium magnetizations, we can obtain the value of the Gibbs free energy [2].

However, the spin-independence assumption of the naive mean-field method may not be accurate, especially when a low-temperature thermodynamic phase is of interest. We need to consider the correlations among the spins in a short-range region of

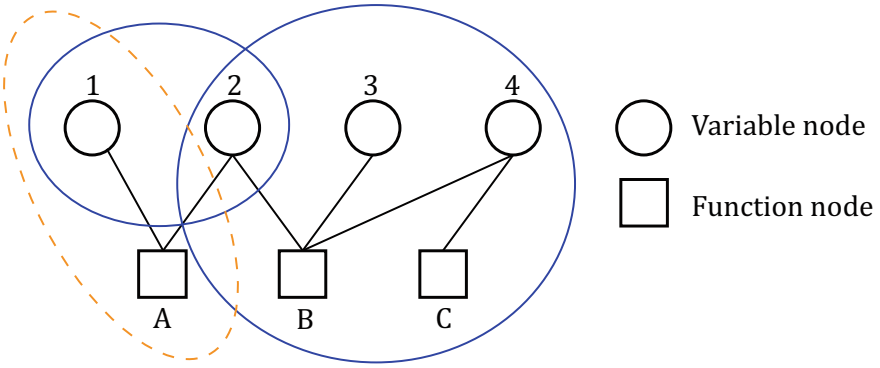



**Fig. 3.1** Regions in a factor graph. Solid circles are defined as regions, while the dashed circle is not a valid region

the factor graph, which is precisely the concept of the Bethe approximation, which we shall explore in the next section.

### *3.2.2 Bethe Approximation*

The Bethe approximation [3] is an extension of the classic mean-field method, taking into account correlations between nearest neighboring sites. To introduce the Bethe approximation, we first define the concept of region in the factor graph. As Fig. 3.1 shows, the region is defined by a set of function nodes and all the variable nodes connected to these functional nodes. Note that the function node set can be empty. Variable nodes and functional nodes represent spins and interactions in the multi-spin interaction model. In this setting, we can introduce the region energy $E_R(\boldsymbol{x}_R)$, the region internal energy $U_R(b_R)$, the region entropy $H_R(b_R)$ and the region free energy $F_R(b_R)$ as follows:

$$E_R(\boldsymbol{x}_R) = -\sum_{a \in R} \ln f_a(\boldsymbol{x}_a), \tag{3.21}$$

$$U_R(b_R) = \sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R) E_R(\boldsymbol{x}_R), \tag{3.22}$$

$$H_R(b_R) = -\sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R) \ln b_R(\boldsymbol{x}_R), \tag{3.23}$$

$$F_R(b_R) = U_R(b_R) - H_R(b_R), \tag{3.24}$$

where $\boldsymbol{x}_R$ are the variable nodes in the region $R$, and $b_R(\boldsymbol{x}_R)$ is the joint distribution of $\boldsymbol{x}_R$. The basic idea of a region-based free energy approximation is to break up the

factor graph into regions and then sum up their contributions to approximate the true free energy, where all the variable nodes and function nodes should be summed up only once. Because overlaps between different regions cannot be avoided in a non-naive approximation, counting numbers $C_R$ (an integer that may be zero or negative) must be introduced to avoid double calculation. Given a region set $\mathcal{R}$, the total internal energy $U_{\mathcal{R}}$ and entropy $H_{\mathcal{R}}$ can be written as

$$U_{\mathcal{R}} = \sum_{R \in \mathcal{R}} C_R U_R(b_R), \tag{3.25}$$

$$H_{\mathcal{R}} = \sum_{R \in \mathcal{R}} C_R H_R(b_R), \tag{3.26}$$

with the following two constraints for counting numbers:

$$\sum_{R \in \mathcal{R}} \mathbb{I}[a \in R] C_R = 1, \tag{3.27}$$

$$\sum_{R \in \mathcal{R}} \mathbb{I}[i \in R] C_R = 1, \tag{3.28}$$

where $\mathbb{I}[a \in R] = 1$ when the function node $a$ is in the region $R$, and takes zero otherwise. $\mathbb{I}[i \in R]$ has a similar meaning for variable nodes.

In the Bethe approximation, the factor graph is broken into two kinds of regions (see Fig. 3.2), which are a large region $R_L$ with one functional node and the variable nodes connected to it, and a small region $R_S$ with only one variable node. Under this division, counting numbers can be derived as $C_{R_L} = 1$ and $C_{R_S} = 1 - d_i$, where $d_i$ is the number of the function nodes connected to the variable node $i$ in the small region. These counting numbers can also be derived from the identity $C_R = 1 - \sum_{S \in \mathcal{S}(R)} C_S$, where $\mathcal{S}(R)$ denotes the region set that is the set of super-regions of $R$. If the set of variable and function nodes in $R_1$ are a subset of nodes in $R_2$, then $R_2$ is the super-region of $R_1$ [4]. Thus, we can compute the Bethe internal energy $U_{Bethe}$ and the Bethe entropy $H_{Bethe}$ as follows:

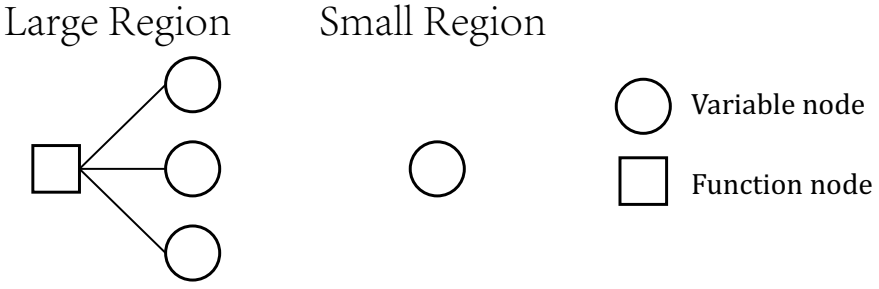



**Fig. 3.2** Region division in the Bethe approximation

$$U_{Bethe} = -\sum_a \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln f_a(\boldsymbol{x}_a), \tag{3.29}$$

$$H_{Bethe} = -\sum_a \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln b_a(\boldsymbol{x}_a) + \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \ln b_i(x_i), \tag{3.30}$$

where we replace $b_{R_L}(\boldsymbol{x}_{R_L})$ and $b_{R_S}(\boldsymbol{x}_{R_S})$ with $b_a(\boldsymbol{x}_a)$ and $b_i(x_i)$, respectively. The Bethe free energy is then given by

$$\begin{aligned} F_{Bethe} = &-\sum_a \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln f_a(\boldsymbol{x}_a) + \sum_a \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln b_a(\boldsymbol{x}_a) \\ &- \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \ln b_i(x_i), \end{aligned} \tag{3.31}$$

which is the target function to minimize later. By taking into account the nearest-neighbor correlations, the trial probability distribution can also be written in a compact form [4, 5]:

$$b_{BA}(\boldsymbol{x}) = \frac{\prod_a b_a(\boldsymbol{x}_a)}{\prod_i b_i(x_i)^{d_i - 1}}, \tag{3.32}$$

which is automatically normalized and exact when the factor graph is a tree, but still a good approximation when the factor graph is not tree-like. A rigorous proof is hard, but the approximation should be compared with simulations in practice. Inserting the form of $b_{BA}(\boldsymbol{x})$ into the Gibbs free energy, one can derive the same form as that in Eq. (3.31).

Before using the Lagrange multiplier method, we first formulate the probability constraints as follows:

$$\sum_{x_i} b_i(x_i) = 1, \ \forall i; \tag{3.33}$$

$$\sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) = 1, \ \forall a; \tag{3.34}$$

$$\sum_{\boldsymbol{x}_a \backslash x_i} b_a(\boldsymbol{x}_a) = b_i(x_i), \ \forall (i, a). \tag{3.35}$$

Finally, the Lagrange objective function reads

$$\begin{aligned} L = &-\sum_a \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln f_a(\boldsymbol{x}_a) + \sum_a \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln b_a(\boldsymbol{x}_a) \\ &- \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \ln b_i(x_i) + \sum_i \lambda_i \Big( \sum_{x_i} b_i(x_i) - 1 \Big) \\ &+ \sum_a \lambda_a \Big( \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) - 1 \Big) + \sum_{i,a} \sum_{x_i} \rho_{i,a}(x_i) \Big( \sum_{\boldsymbol{x}_a \backslash x_i} b_a(\boldsymbol{x}_a) - b_i(x_i) \Big). \end{aligned} \tag{3.36}$$

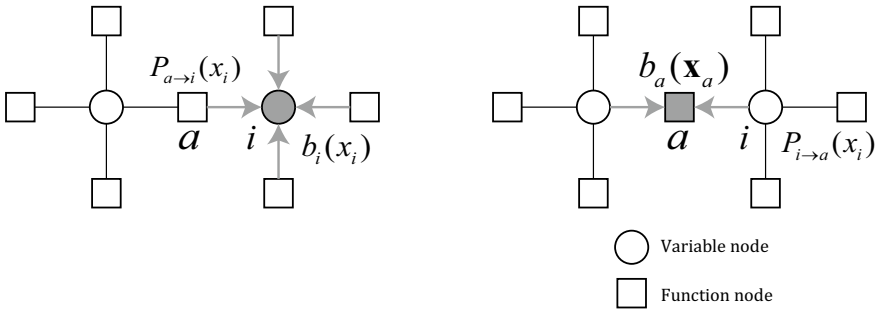



**Fig. 3.3** Message passing process in the BP algorithm. (Left Panel) cavity probabilities converge to a variable node; (Right panel) cavity probabilities converge to a function node

After performing the variation on $L$, we can obtain the form of the spin distribution $b_i(x_i)$ and joint distribution $b_a(\boldsymbol{x}_a)$ [4]:

$$b_i(x_i) = \frac{1}{Z_i} \prod_{a\in\partial i} P_{a\to i}(x_i), \tag{3.37a}$$

$$b_a(\boldsymbol{x}_a) = \frac{1}{Z_a} f_a(\boldsymbol{x}_a) \prod_{i\in\partial a} \prod_{b\in\partial i\setminus a} P_{b\to i}(x_i), \tag{3.37b}$$

where we define $P_{a\to i}(x_i)$ and $P_{i\to a}(x_i)$ as the messages passing between the functional nodes and variable nodes in two directions as illustrated in Fig. 3.3. These two messages obey the following iterative equations:

$$P_{a\to i}(x_i) = \sum_{\boldsymbol{x}_j : j\in\partial a\setminus i} f_a(\boldsymbol{x}_a) \prod_{j\in\partial a\setminus i} P_{j\to a}(x_j), \tag{3.38a}$$

$$P_{i\to a}(x_i) = \frac{1}{Z_{i\to a}} \prod_{b\in\partial i\setminus a} P_{b\to i}(x_i). \tag{3.38b}$$

Note that Eq. (3.38a) is compatible with the marginal probability constraint used to write the constrained Bethe free energy, while Eq. (3.38b) follows directly from the result of $b_i(x_i)$ by just excluding the function node $a$. Finally, the (joint) marginal probabilities $b_i(x_i), b_a(\boldsymbol{x}_a)$ can be interpreted as beliefs and written in an explicit form as

$$b_i(x_i) \propto \prod_{a\in\partial i} P_{a\to i}(x_i), \tag{3.39a}$$

$$b_a(\boldsymbol{x}_a) \propto f_a(\boldsymbol{x}_a) \prod_{i\in\partial a} P_{i\to a}(x_i), \tag{3.39b}$$

which is consistent with Eq. (3.37).

Equation (3.38) is also called the belief propagation (BP) algorithm in computer science, where we can perform the iteration of the messages $\{P_{a\to i}(x_i),\ P_{i\to a}(x_i)\}$ to their fixed point and calculate the beliefs $b_i(x_i)$ and $b_a(\boldsymbol{x}_a)$. The fixed points of the BP algorithm correspond to stationary points of the constrained Bethe free energy [6]. Depending on specific settings, the number of stationary points may be different, being finite or exponentially large. For example, if the factor graph is loopy, or the model has a complex low-temperature phase, the BP iteration may not converge, or oscillate among several solutions. Note that the cavity method allows an extension to handling the case of exponentially many states (in physics, corresponding to one-step replica symmetry breaking, see Chap. 9). The probability distributions of cavity fields over the states are then required to be introduced. We will provide an in-depth discussion about this point in Chap. 9.

The messages $P_{i\to a}(x_i)$ here can be interpreted as the probability distribution of the variable node $i$ with the removal of function node $a$, which is similar to the definition in the cavity method. Actually, we can prove that the BP equations are equivalent to the cavity equations as follows. First, we substitute the expression of $P_{b\to i}(x_i)$ into $P_{i\to a}(x_i)$ in the BP equation, and we obtain

$$\begin{aligned} P_{i\to a}(x_i) &= \frac{1}{Z_{i\to a}} \prod_{b\in\partial i\backslash a} \sum_{\boldsymbol{x}_j : j\in\partial b\backslash i} f_b(\boldsymbol{x}_b) \prod_{j\in\partial b\backslash i} P_{j\to b}(x_j) \\ &= \frac{1}{Z_{i\to a}} \prod_{b\in\partial i\backslash a} \sum_{\boldsymbol{x}_j : j\in\partial b\backslash i} e^{J_b \prod_{j\in\partial b} x_j} \prod_{j\in\partial b\backslash i} \frac{1+m_{j\to b}x_j}{2}. \end{aligned} \tag{3.40}$$

We then define $A_b^+ = \sum_{\boldsymbol{x}_j : j\in\partial b\backslash i} e^{J_b \prod_{j\in\partial b\backslash i} x_j} \prod_{j\in\partial b\backslash i} \frac{1+m_{j\to b}x_j}{2}$, where we take $x_i = +1$. $A_b^-$ follows the similar definition with $x_i = -1$. Thus, $Z_{i\to a} = \prod_{b\in\partial i\backslash a} A_b^+ + \prod_{b\in\partial i\backslash a} A_b^-$. After a few algebra operations, $A_b^+$ and $A_b^-$ can be written, respectively, as

$$A_b^+ = \cosh J_b \left(1 + \tanh J_b \prod_{j\in\partial b\backslash i} m_{j\to b}\right), \tag{3.41}$$

$$A_b^- = \cosh J_b \left(1 - \tanh J_b \prod_{j\in\partial b\backslash i} m_{j\to b}\right), \tag{3.42}$$

which is exactly the same as that derived by the cavity method in the previous chapter. According to the definition, we can then derive $m_{i\to a}$:

$$
\begin{aligned}
m_{i\to a} &= \sum_{x_i} P_{i\to a}(x_i) \\
&= \frac{\prod_{b\in\partial i\setminus a} A_b^+ - \prod_{b\in\partial i\setminus a} A_b^-}{\prod_{b\in\partial i\setminus a} A_b^+ + \prod_{b\in\partial i\setminus a} A_b^-} \\
&= \frac{\prod_{b\in\partial i\setminus a}(1+\tanh J_b \prod_{j\in\partial b\setminus i} m_{j\to b}) - \prod_{b\in\partial i\setminus a}(1-\tanh J_b \prod_{j\in\partial b\setminus i} m_{j\to b})}{\prod_{b\in\partial i\setminus a}(1+\tanh J_b \prod_{j\in\partial b\setminus i} m_{j\to b}) + \prod_{b\in\partial i\setminus a}(1-\tanh J_b \prod_{j\in\partial b\setminus i} m_{j\to b})}.
\end{aligned}
\tag{3.43}
$$

After introducing an auxiliary variable $u_{b\to i}$ through $\tanh u_{b\to i} = \tanh J_b \prod_{j\in\partial b\setminus i} m_{j\to b}$, we finally obtain

$$
m_{i\to a} = \tanh\left(\sum_{b\in\partial i\setminus a} u_{b\to i}\right), \tag{3.44a}
$$

$$
\tanh u_{b\to i} = \tanh J_b \prod_{j\in\partial b\setminus i} m_{j\to b}, \tag{3.44b}
$$

which is the standard cavity equation when $\beta = 1$.

The Bethe approximation is merely a pair approximation of a more general method—cluster variational method [5]. The cluster variational method is able to treat arbitrary large clusters of correlated sites, and yet, the computational complexity increases. Recent developments also include loop corrections for probabilistic inference on factor graphs [7, 8].

### 3.2.3 *From the Bethe to Naive Mean-Field Approximation*

In the naive mean-field approximation, we use a factorized form of the trial probability distribution that neglects the correlation among spins. In contrast, the Bethe approximation considers a short-range correlation among spins, where it is expected that in a high temperature, even these short-range correlations become unimportant, and thus, the naive mean-field approximation will be recovered. More precisely, we take an example of a two-body interaction model. Suppose our model is a two-body interaction model with inverse temperature $\beta$. In this setting, the mean-field iteration equations are given by

$$
m_i = \tanh\Big(\beta \sum_{j\in\partial i} J_{ij} m_j\Big). \tag{3.45}
$$

Next, we derive this equation from the Bethe approximation.

In the Bethe approximation, the cavity iteration equations are given by

$$m_{i\to a} = \tanh \sum_{b\in\partial i\setminus a} u_{b\to i}, \tag{3.46a}$$

$$\tanh u_{b\to i} = \tanh \beta J_b \prod_{j\in\partial b\setminus i} m_{j\to b}, \tag{3.46b}$$

where $m_{i\to a}$ can be derived as

$$\begin{aligned} m_{i\to a} &= \tanh\left(\sum_{b\in\partial i\setminus a} u_{b\to i}\right) \\ &= \tanh\left(\sum_{b\in\partial i} u_{b\to i} - u_{a\to i}\right) \\ &= \tanh(\tanh^{-1}(m_i) - u_{a\to i}) \\ &= \frac{m_i - \tanh\beta J_a \prod_{j\in\partial a\setminus i} m_{j\to a}}{1 - m_i \tanh\beta J_a \prod_{j\in\partial a\setminus i} m_{j\to a}}, \end{aligned} \tag{3.47}$$

where we have used the identity $\tanh(x+y) = \frac{\tanh x + \tanh y}{1+\tanh x \tanh y}$. Considering the two-body interaction, we have

$$m_{i\to j} = \frac{m_i - \tanh\beta J_{ij} m_{j\to i}}{1 - m_i \tanh\beta J_{ij} m_{j\to i}}, \tag{3.48a}$$

$$m_{j\to i} = \frac{m_j - \tanh\beta J_{ij} m_{i\to j}}{1 - m_j \tanh\beta J_{ij} m_{i\to j}}. \tag{3.48b}$$

We can then eliminate $m_{i\to j}$ and $m_{j\to i}$ in the cavity equation, by obtaining the non-cavity functions of $m_{j\to i}$ and $m_{i\to j}$ as a function of single magnetizations. We first have the following expressions based on Eq. (3.48) [9]:

$$m_{i\to j} = f(m_i, m_j, \tanh\beta J_{ij}), \tag{3.49}$$

$$m_{j\to i} = f(m_j, m_i, \tanh\beta J_{ij}), \tag{3.50}$$

$$f(a,b,t) = \frac{1 - t^2 - \sqrt{(1-t^2)^2 - 4t(a-bt)(b-at)}}{2t(b-at)}. \tag{3.51}$$

Thus, we can write a non-cavity version of $m_i$ as follows:

$$m_i = \tanh\Big(\sum_{j\in\partial i} \tanh^{-1}(f(m_j, m_i, \tanh\beta J_{ij})\tanh\beta J_{ij})\Big). \tag{3.52}$$

Since we assume $\beta J_{ij}$ is weak (e.g., in a high-temperature phase), we can perform the Taylor expansions like $\tanh^{-1} x \approx x$, $\tanh x \approx x$, $(1+x)^a \approx 1 + ax + \frac{1}{2}a(a-1)x^2$, when $x$ is a small quantity, and we finally get

$$m_i = \tanh\Big(\sum_{j\in\partial i}(\beta J_{ij}m_j - \beta^2 J_{ij}^2(1-m_j^2)m_i)\Big), \tag{3.53}$$

where the second term in the summation is called the Onsager reaction term, a characteristic of a high-temperature expansion solution of a spin glass model [10, 11], which we shall introduce in more details later. Neglecting the second-order term of couplings, one recovers the naive mean-field equation.

## 3.3 Mean-Field Inverse Ising Problem

In the previous sections, we describe how to find the statistical physics solutions of an equilibrium thermodynamic problem under some approximations, which is exactly a direct problem. However, if we acquire data samples from an unknown model, we can predict the model parameters, e.g., couplings and fields, from these raw data samples, which is called the inverse problem. The direct problem can provide insights into the inverse problem. Let us explain this in more details.

An Ising model considering only up to pairwise interactions is described by

$$H(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i, \tag{3.54a}$$

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}e^{\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i}. \tag{3.54b}$$

Note that $\beta$ has been absorbed into the model parameters in the current setting. Given measured magnetizations $m_i = \langle\sigma_i\rangle_{\text{data}}$ and correlation functions $C_{ij} = \langle\sigma_i\sigma_j\rangle_{\text{data}} - m_i m_j$, what we want to estimate is the coupling constants and external fields $\{J_{ij}, h_i\}$, which is a typical unsupervised learning problem. This is exactly the Boltzmann machine learning [12]. It starts from a set of initial parameters $\{J_{ij}, h_i\}$ and then updates the parameters by an increment:

$$\Delta J_{ij} = \eta(\langle\sigma_i\sigma_j\rangle_{\text{data}} - \langle\sigma_i\sigma_j\rangle_{\text{Ising}}), \tag{3.55a}$$

$$\Delta h_i = \eta(\langle\sigma_i\rangle_{\text{data}} - \langle\sigma_i\rangle_{\text{Ising}}), \tag{3.55b}$$

where $\eta$ is a predefined learning rate. The iteration runs until the model average and data average match with each other within a certain accuracy. The model average can be estimated by the Monte Carlo algorithms, which we shall introduce in the following chapter. However, when the system size is large, the mean-field method is relatively fast.

To carry out the inference, we first compute the magnetization:

$$m_i = \frac{\partial \log Z(J_{ij}^*, h_i^*)}{\partial h_i} = \sum_{\{\boldsymbol{\sigma}\}}\sigma_i\frac{e^{\sum_{i<j} J_{ij}^*\sigma_i\sigma_j + \sum_i h_i^*\sigma_i}}{Z}, \tag{3.56}$$

and then we apply the fluctuation-response theorem [13]:

$$\begin{aligned}\frac{\partial m_i}{\partial h_j} =& \sum_{\{\boldsymbol{\sigma}\}} \sigma_i \sigma_j \frac{e^{\sum_{i<j} J_{ij}^* \sigma_i \sigma_j + \sum_i h_i^* \sigma_i}}{Z} \\ &- \sum_{\{\boldsymbol{\sigma}\}} \sigma_i \frac{e^{\sum_{i<j} J_{ij}^* \sigma_i \sigma_j + \sum_i h_i^* \sigma_i}}{Z} \sum_{\{\boldsymbol{\sigma}\}} \sigma_j \frac{e^{\sum_{i<j} J_{ij}^* \sigma_i \sigma_j + \sum_i h_i^* \sigma_i}}{Z} \\ =& C_{ij} = \langle \sigma_i \sigma_j \rangle_{\text{data}} - m_i m_j .\end{aligned} \tag{3.57}$$

The symbol with the superscript $*$ indicates the current estimates of the model parameters. These steps amount to the expectation step of a standard Expectation-Maximization procedure [14]. The updating procedure in Eq. (3.55) corresponds to the M step.

Using the above relationship $C_{ij} = \frac{\partial m_i}{\partial h_j}$ and the naive mean-field equation $m_i = \tanh(h_i + \sum_{k \neq i} J_{ik} m_k)$, we get

$$\begin{aligned} C_{ij} &= (1 - m_i^2)\Big[\delta_{ij} + \sum_{k \neq i} J_{ik} C_{kj}\Big], \\ \mathbf{C} &= \mathbf{P} + \mathbf{PJC}, \end{aligned} \tag{3.58}$$

where $\mathbf{P}$ is a diagonal matrix with $\mathbf{P}_{ij} = (1 - m_i^2)\delta_{ij}$. Finally, we obtain the naive mean-field (nMF) solution of the inverse Ising problem:

$$J_{ij}^{\text{nMF}} = (\mathbf{P})_{ij}^{-1} - (\mathbf{C})_{ij}^{-1}. \tag{3.59}$$

The external fields can then be reconstructed based on the predicted couplings. The naive mean-field solution is the simplest one among other mean-field methods, including high-temperature expansion, small-correlation expansion and the Bethe approximation [2, 15].

## References


1. J.M. Yeomans, *Statistical Mechanics of Phase Transitions* (Oxford University Press, Oxford, 1992)
2. H. Huang, Y. Kabashima, Phys. Rev. E **87**, 062129 (2013)
3. H.A. Bethe, Proc. R. Soc. Lon. Ser. A-Math. Phys. Sci. **150**(871), 552 (1935)
4. J. Yedidia, W. Freeman, Y. Weiss, IEEE Trans. Inf. Theory **51**(7), 2282 (2005)
5. A. Pelizzola, J. Phys. A **38**(33), R309 (2005)
6. T. Heskes, Neural Comput. **16**(11), 2379 (2004)
7. J.M. Mooij, H.J. Kappen, J. Mach. Learn. Res. **8**(40), 1113 (2007)
8. J.Q. Xiao, H. Zhou, J. Phys. A: Math. Theor. **44**(42), 425001 (2011)
9. F. Ricci-Tersenghi, J. Stat. Mech.: Theory Exper. **2012**(8), 8015 (2012)
10. D.J. Thouless, P.W. Anderson, R.G. Palmer, Philos. Mag. **35**(3), 593 (1977)

11. T. Plefka, J. Phys. A **15**(6), 1971 (1982)
12. D.H. Ackley, G.E. Hinton, T.J. Sejnowski, Cognit. Sci. **9**(1), 147 (1985)
13. H.J. Kappen, F.B. Rodríguez, in *Advances in Neural Information Processing Systems* (1998), pp. 280–286
14. A.P. Dempster, N.M. Laird, D.B. Rubin, J. R. Stat. Soc. Ser. B **39**, 1 (1977)
15. H.C. Nguyen, R. Zecchina, J. Berg, Adv. Phys. **66**(3), 197 (2017)

# Chapter 4
# Monte Carlo Simulation Methods


A few systems in equilibrium physics can be analytically solved. It is, therefore, necessary to develop numerical techniques to estimate the equilibrium properties of a physics system. For example, given the Hamiltonian of the Ising model, it still requires $O(2^N)$ time complexity to directly compute expected energy, where $N$ is the number of spins. To either check how accurate a crude approximation is, e.g., mean-field approximation or the Bethe approximation, or estimate the typical energy level of a statistical mechanics model that cannot be analytically solved, we rely on the Monte Carlo simulation techniques, including their variants, which are widely used not only in the physics field itself but also in the machine learning community. For example, the Gibbs sampling is performed with the classical Monte Carlo methods or its variants with the help of importance sampling. In this chapter, we will introduce the basic knowledge about the sampling method and its applications to standard physics models.

## 4.1 Monte Carlo Method

The main idea of the Monte Carlo method is simple. For example, calculating a multi-dimensional integral can be carried out by drawing a set of samples according to a predefined distribution. We first introduce the standard steps to implement the Monte Carlo method:

- Transforming the original problem of interest to a statistical problem, like calculating the expectation of some random variables under a specific distribution.
- Sampling random variables from the specific distribution.
- Using the samples from the second step to compute any quantity of interest and obtaining the result of the problem.

We give here a representative example of estimating an integral or a sum:



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_4

$$\begin{aligned}\langle A\rangle &= \int A(\mathbf{x}) f(\mathbf{x}) d\mathbf{x},\\ \langle A\rangle &= \sum_{\mathbf{x}} A(\mathbf{x}) p(\mathbf{x}).\end{aligned} \tag{4.1}$$

To calculate the above expectations, one can sample random variables from the distribution $f(\mathbf{x})$ or $p(\mathbf{x})$ and then obtain a sample collection $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_M\}$ of the size $M$. Finally, $\{A(\mathbf{x}_1), A(\mathbf{x}_2), A(\mathbf{x}_3), \ldots, A(\mathbf{x}_M)\}$ can be obtained. As $A(\mathbf{x}_i)$ is independently estimated, the law of large number implies that

$$\lim_{M\to\infty} P\left(\left|\frac{1}{M}\sum_{i=1}^{M} A(\mathbf{x}_i) - \langle A\rangle\right| < \epsilon\right) = 1, \forall\epsilon > 0. \tag{4.2}$$

Thus, the expectation can be estimated as

$$\langle A\rangle \simeq \frac{1}{M}\sum_{\mathbf{x}_i} A(\mathbf{x}_i). \tag{4.3}$$

As the collected samples $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_M\}$ are independent and identically distributed, the statistical error due to the sampling is related to the variance of $A$ and can be estimated to be of the order of $O(M^{-\frac{1}{2}})$ [1]. Moreover, the Monte Carlo estimator is unbiased. Given a large number of the Monte Carlo samples, the empirical estimation converges to the true expectation we want to compute [2]. Interested readers can figure out the procedure as the above description to estimate the integral $\int_{-\infty}^{\infty} x^2 \mathrm{Dx}$, where Dx indicates the random variable $x$ is a standard Gaussian variable. The Monte Carlo estimation can be compared with the analytic result of 1. As the number of random samples increases, the estimation will approach the exact result. In the remaining chapters, we will also show this kind of method is also very effective and popular to solve the saddle-point equations of the replica method applied to solve a variety of neural network models.

## 4.2 Importance Sampling

In the Monte Carlo simulation, sampling a distribution is usually required, e.g., the Gaussian distribution as mentioned in the previous section. Unfortunately, most distributions are very challenging to sample, e.g., the Boltzmann distribution in statistical physics. Here, we shall introduce some basic strategies to generate random samples from distributions that are more complicated than the commonly used ones, such as uniform, Poisson and Gaussian distributions.

By introducing a simple trial distribution, say $q(\boldsymbol{x})$ that is easy to sample, we can recast Eq. (4.1) as

$$\begin{aligned}\langle A\rangle &= \int \frac{A(\mathbf{x})f(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})d\mathbf{x},\\ \langle A\rangle &= \sum_{\mathbf{x}} \frac{A(\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x}).\end{aligned} \tag{4.4}$$

The expectations $\langle A\rangle_f$ and $\langle A\rangle_p$ are then transformed to $\langle \frac{A(\mathbf{x})f(\mathbf{x})}{q(\mathbf{x})}\rangle_q$, and $\langle \frac{A(\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})}\rangle_q$. Therefore, we can first sample the distribution $q(\mathbf{x})$ to get samples $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_M\}$ and then calculate the expectations:

$$\langle A\rangle \simeq \frac{\sum_{i=1}^{M} A(\mathbf{x}_i)p(\mathbf{x}_i)/q(\mathbf{x}_i)}{M}, \tag{4.5}$$

where the factor $\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$ can be thought of as an importance weight of the sample $\mathbf{x}_i$ in computing the expectation. This estimation is, thus, called the importance sampling [2]. When $q(\mathbf{x}_i) = p(\mathbf{x}_i)$, the importance sampling turns out to be Eq. (4.3). Choosing a trial distribution is important; otherwise, the Monte Carlo estimation will become noisy with a large variance, being very slow to converge to a quantity of satisfied accuracy. An annealed importance sampling is introduced to build a suitable $q(\mathbf{x})$ starting from a trivial one $[p_0(\mathbf{x})]$. A common scheme specifying the intermediate distribution is given by

$$p_j(\mathbf{x}) = p_0(\mathbf{x})^{1-\beta_j} p_n(\mathbf{x})^{\beta_j}, \tag{4.6}$$

where $0 = \beta_0 < \beta_1 < \cdots < \beta_n = 1$. In other words, $p_j(\mathbf{x})$ interpolates between $p_0(\mathbf{x})$ and $p_n(\mathbf{x}) = p(\mathbf{x})$. The samples can then be sequentially generated by designing an appropriate transition probability of two states. Interested readers can find the original paper [3] for implementation details.

## 4.3 Markov Chain Sampling

To realize a sampling where a sequence of samples are generated, one can construct a Markov chain during sampling. The Markov property implies that the next state of a dynamics is only related to the current state, and the conditional probability can be written as

$$P\left[\mathbf{S}_{t+1}|\mathbf{S}_1, \ldots, \mathbf{S}_t\right] = P\left[\mathbf{S}_{t+1}|\mathbf{S}_t\right], \tag{4.7}$$

where $\mathbf{S}_t$ is the state at time $t$. A Markov chain obeys the Markov property for its dynamics. One can construct a time-homogeneous Markov chain by setting up an initial distribution $\pi(\mathbf{S}_0)$ together with the transition probability $W(\mathbf{S} \to \mathbf{S}')$. A stationary distribution $\pi(\mathbf{S}')$ can, thus, be identified by satisfying the following condition:

$$\pi(\mathbf{S}') = \sum_{\mathbf{S}} W(\mathbf{S} \to \mathbf{S}')\pi(\mathbf{S}). \tag{4.8}$$

The task of designing a Markov chain becomes simple if the detailed balance criterion is obeyed [1], i.e.,

$$W(\mathbf{S} \to \mathbf{S}')\pi(\mathbf{S}) = W(\mathbf{S}' \to \mathbf{S})\pi(\mathbf{S}'). \tag{4.9}$$

The detailed balance criterion guarantees that the designed Markov chain converges to the target distribution [Eq. (4.8)] [1].

## 4.4 Monte Carlo Simulations in Statistical Physics

In statistical physics, an equilibrium system is described by the Boltzmann distribution:

$$P_{\text{eq}}(\boldsymbol{s}) = \frac{1}{Z} e^{-\beta \mathcal{H}(\boldsymbol{s})}, \tag{4.10}$$

where the partition function $Z = \sum_s e^{-\beta \mathcal{H}(\boldsymbol{s})}$, and $\mathcal{H}(\boldsymbol{s})$ is the system's Hamiltonian. Then the expectation or thermal average of an observable $O(\boldsymbol{s})$ is given by

$$\langle O \rangle = \frac{1}{Z} \sum_{\boldsymbol{s}} O(\boldsymbol{s}) e^{-\beta \mathcal{H}(\boldsymbol{s})}. \tag{4.11}$$

The partition function is usually intractable, making an analytic estimation of thermodynamic quantities impossible. The Markov Chain Monte Carlo (MCMC) is then useful for estimating the quantities of interest. To illustrate the MCMC method, we simulate the SK model as an example. The SK model is a fully connected mean-field glass model, and the statistical mechanics properties were first studied analytically in the seminal work [4]. The Hamiltonian is given by

$$\mathcal{H} = -\frac{1}{2} \sum_{i \neq j} J_{ij} \sigma_i \sigma_j, \tag{4.12}$$

where the spin $\sigma_i = \pm 1$, and the couplings follow independently a Gaussian distribution of zero mean and variance $1/N$. The model has a paramagnetic-to-spin glass transition at the critical temperature $T = 1$. By using the MCMC method, we can acquire the equilibrium properties of the SK model, which can be compared with the theoretical analysis.

Next, we introduce two Monte Carlo techniques to numerically evaluate the model. But we emphasize that both methods are generally applicable to other similar models, for which an exact computation of relevant thermodynamic quantities may be impossible.

### 4.4.1 Metropolis Algorithm

The detailed balance condition of the Boltzmann distribution can be written as follows:

$$P_{\text{eq}}(\mathbf{s}_i)W(\mathbf{s}_i \to \mathbf{s}_j) = P_{\text{eq}}(\mathbf{s}_j)W(\mathbf{s}_j \to \mathbf{s}_i), \tag{4.13}$$

where $W(\mathbf{s}_i \to \mathbf{s}_j)$ is the transition probability from state $\mathbf{s}_i$ to state $\mathbf{s}_j$. The ratio between two transition probabilities can be rewritten as

$$\frac{W(\mathbf{s}_i \to \mathbf{s}_j)}{W(\mathbf{s}_j \to \mathbf{s}_i)} = e^{-\beta\Delta\mathcal{H}(\mathbf{s}_i,\mathbf{s}_j)}, \tag{4.14}$$

where $\Delta\mathcal{H}(\mathbf{s}_i, \mathbf{s}_j) = \mathcal{H}(\mathbf{s}_j) - \mathcal{H}(\mathbf{s}_i)$, and the Boltzmann distribution is used. Our purpose is to find a transition probability matrix satisfying the detailed balance condition. In fact, choosing the transition probability form is not unique, and there are two frequently used forms. One is the Metropolis algorithm:

$$W(\mathbf{s}_i \to \mathbf{s}_j) = \begin{cases} 1, & \Delta\mathcal{H}(\mathbf{s}_i, \mathbf{s}_j) < 0; \\ e^{-\beta\Delta\mathcal{H}(\mathbf{s}_i,\mathbf{s}_j)}, & \Delta\mathcal{H}(\mathbf{s}_i, \mathbf{s}_j) \geq 0, \end{cases} \tag{4.15}$$

which can be also recast into the form $W(\mathbf{s}_i \to \mathbf{s}_j) = \min(e^{-\beta\Delta\mathcal{H}(\mathbf{s}_i,\mathbf{s}_j)}, 1)$. A pseudocode is given in Algorithm 4.1. Another popular choice is the heat-bath algorithm:

$$W(\mathbf{s}_i \to \mathbf{s}_j) = \frac{e^{-\beta\Delta\mathcal{H}(\mathbf{s}_i,\mathbf{s}_j)}}{1 + e^{-\beta\Delta\mathcal{H}(\mathbf{s}_i,\mathbf{s}_j)}}. \tag{4.16}$$

It can be verified that the Metropolis dynamics is always more likely to accept an attempt of spin changes that leads to a small change of energy. In addition, if we define the transition probability as a function $F(e^{-\beta\Delta\mathcal{H}})$, it can be also verified that the above two choices satisfy $\frac{F(x)}{F(1/x)} = x$ for all $x$, compatible with the detailed balance criterion.

For a fast sampling, we can flip just one single spin (rather than a small group of spins) at each step of the Metropolis dynamics, and then, we can obtain the following transition rule:

$$W(\sigma_i \to -\sigma_i) = \frac{1}{2}[1 - \sigma_i \tanh \beta h_i], \tag{4.17}$$

where $h_i = \sum_{j\neq i} J_{ij}\sigma_j$ is the local field acting on the spin $\sigma_i$. This rule is derived from the heat-bath choice.

A random initial state is far from equilibrium with a high probability, and thus, a Markov chain dynamics requires a relaxation time for the system to reach the equilibrium state. This time scale is called the equilibration time $\tau_{eq}$. In practice, $\tau_{eq}$ is measured in the unit of the Monte Carlo sweep (MCS), in which one MCS equals to $N$ proposed single-spin-updates. To verify whether the system arrives at equilibrium, it is necessary in practice to check the evolution of some observables, e.g., energy.

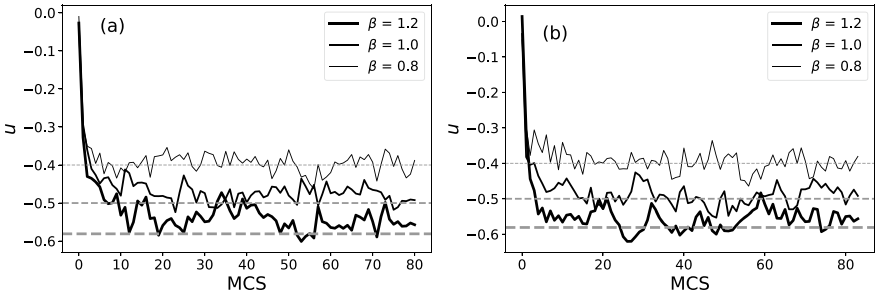



**Fig. 4.1** Evolutions of the energy density of the SK model with $N = 500$ and $J_{ij} \sim \mathcal{N}(0, 1/N)$. **a** Metropolis Monte Carlo simulation. **b** Parallel Tempering Monte Carlo. The dashed lines are the corresponding predictions of replica theory. The time step is a measure in the unit of the Monte Carlo step (MCS). Each step means a sweep of all spins for the proposed update. $\beta$ defines the inverse temperature

**Algorithm 4.1** Metropolis Algorithm

**Input:** The number of samples $M$, temperature $T$, $\tau_{eq}$, $\delta t$
**Output:** A collection of samples
1: Initialize configuration $S$ randomly;
2: Initialize i = 0;
3: Initialize counter = 0;
4: **while** (i< $M$) **do**
5: generate a trial state $\mathbf{S}'$;
6: compute $W(\mathbf{S}' \to \mathbf{S}|T)$;
7: **if** $W >$ rand(0, 1) **then**
8: $\mathbf{S} = \mathbf{S}'$
9: **if** [(counter $> \tau_{eq}$) and (counter % $\delta t$==0))] **then**
10: Append $\mathbf{S}$ to the sample collection.
11: i = i+1
12: counter = counter + 1
13: **return** the sample collection.

As shown in Fig. 4.1a, the energy of the SK model arrives at equilibrium at about $\tau_{eq}$ MCSs, which depends on the temperature. After the relaxation, the energy fluctuates around a typical value, which could be predicted by theory. Therefore, samples can be collected after $\tau_{eq}$ MCSs to estimate equilibrium values of thermodynamic quantities of interest.

Even if the dynamics reaches a steady state, an independent sampling of the equilibrium state requires a certain number of MCSs separating two consecutive samplings. Therefore, we need to compute a time-dependent autocorrelation function of any observable $O$:

$$C_O(t) = \frac{\langle O(t_0)\, O(t_0 + t)\rangle - \langle O(t_0)\rangle \langle O(t_0 + t)\rangle}{\langle O^2(t_0)\rangle - \langle O(t_0)\rangle^2}, \tag{4.18}$$

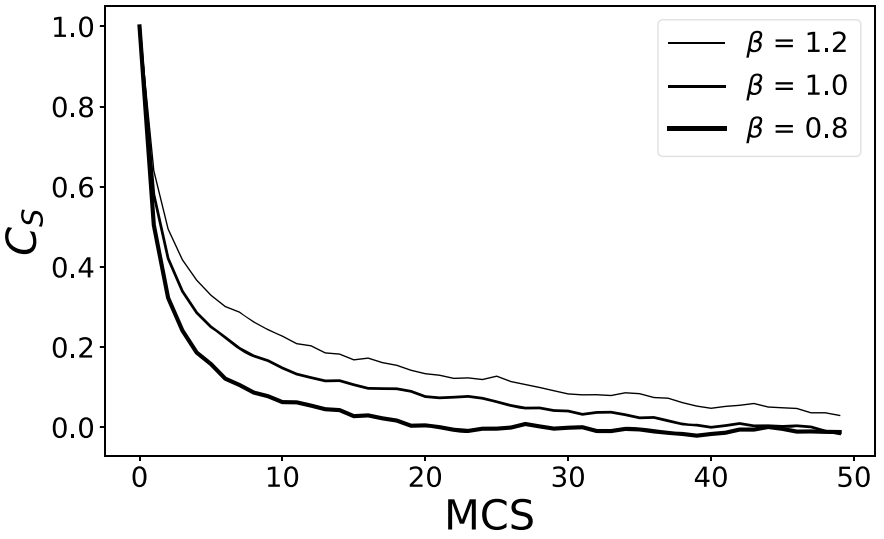



**Fig. 4.2** The relaxation dynamics of the autocorrelation function for the same SK model defined in Fig. 4.1

where $\langle \cdot \rangle$ indicates a thermal average, and $t_0$ denotes the starting time. In general, $C_O(t) \sim \exp(-t/\tau_{\text{auto}})$, and $\tau_{\text{auto}}$ is the corresponding time scale. The correlation length diverges at a continuous phase transition, while the autocorrelation time also diverges at the transition, which is also called the critical slowing down phenomenon. In glass physics, the Edwards–Anderson order parameter $q_{\text{EA}} = \frac{1}{N}\sum_i \langle \sigma_i \rangle^2$ [5], which can be treated as the long-time limit of the time-dependent autocorrelation function $q_{\text{EA}} = \lim_{t\to\infty} C(t)$, where $C(t) = \frac{1}{N}\sum_i \langle \sigma_i(0)\sigma_i(t)\rangle$. The Edwards–Anderson order parameter can also be used to detect ergodicity breaking. A typical example of the autocorrelation profile is shown in Fig. 4.2 for the SK model at different temperatures.

### *4.4.2 Parallel Tempering Monte Carlo*

When we are interested in a low-temperature phase for a spin glass model (e.g., the Sherrington–Kirkpatrick model, the Hopfield model, etc.), the Metropolis algorithm is easy to get trapped in a local minimum, once the Gibbs measure is decomposed into an exponential (in the number of degrees of freedom) number of metastable states. In general, there does not exist one efficient local dynamics method overcoming this challenging fair-sampling problem. However, there do exist a variety of sampling heuristics. One well-known example is the simulated annealing [6], where the starting temperature for the Metropolis sampling is much higher than the target low temperature, and the dynamics is run at each intermediate decreasing temper-

ature for a certain number of MCSs, and finally, a ground state of lower energy is expected to be reached by the annealing process.

The other more efficient one is the parallel tempering method [7], focusing on overcoming energy barriers by simulating several copies of the original system at different temperatures. In this method, $M$ replicas without interaction, which means replicas are independent, are used to construct an ensemble. The $m^{th}$ replica has the original Hamiltonian $\mathcal{H}(X_m)$ and obeys the Boltzmann distribution with an inverse temperature $\beta_m$. The corresponding inverse temperatures satisfy $\beta_m < \beta_{m+1}$ for convenience. Then the state of the ensemble can be described by an extended state $\{X\} = \{X_1, X_2, \ldots, X_M\}$, and the partition function of the ensemble is given by

$$\mathcal{Z} = \sum_{\{X\}} \exp\left(-\sum_{m=1}^{M} \beta_m \mathcal{H}(X_m)\right) = \prod_{m=1}^{M} Z(\beta_m), \tag{4.19}$$

where $Z(\beta_m)$ is the partition function of the original system with $\beta_m$. The probability of the extended state with a temperature set can be written as

$$P(\{X, \beta\}) = \prod_{m=1}^{M} P_{\text{eq}}(X_m, \beta_m) = \frac{1}{\mathcal{Z}} \exp\left(-\sum_{m=1}^{M} \beta_m \mathcal{H}(X_m)\right). \tag{4.20}$$

To construct the detailed balance condition, we only consider exchanging configurations between two replicas. For example, the extended state $\{\ldots; X, \beta_m; \ldots; X', \beta_n; \ldots\}$ changes to $\{\ldots; X', \beta_m; \ldots; X, \beta_n; \ldots\}$ with a transition probability $W(X', \beta_m; X, \beta_n | X, \beta_m; X', \beta_n)$. The detailed balance condition can, thus, be written as

$$\begin{aligned} &P(\{\ldots; X, \beta_m; \ldots; X', \beta_n; \ldots\}) W(X', \beta_m; X, \beta_n | X, \beta_m; X', \beta_n) \\ &= P(\{\ldots; X', \beta_m; \ldots; X, \beta_n; \ldots\}) W(X, \beta_m; X', \beta_n | X', \beta_m; X, \beta_n). \end{aligned} \tag{4.21}$$

It is then easy to derive the ratio between the two transition probabilities:

$$\frac{W(X', \beta_m; X, \beta_n | X, \beta_m; X', \beta_n)}{W(X, \beta_m; X', \beta_n | X', \beta_m; X, \beta_n)} = \exp(-\Delta), \tag{4.22}$$

where

$$\Delta = (\beta_n - \beta_m)\left(\mathcal{H}(X) - \mathcal{H}\left(X'\right)\right). \tag{4.23}$$

A reasonable choice of the transition probability can then be expressed as follows:

$$W(X', \beta_m; X, \beta_n | X, \beta_m; X', \beta_n) = \begin{cases} 1, & \text{for } \Delta < 0, \\ \exp(-\Delta), & \text{for } \Delta > 0. \end{cases} \tag{4.24}$$

In sum, the parallel tempering Monte Carlo can be implemented by the following procedure. First, using the conventional MCMC method to simulate each replica in the

ensemble for a certain number of MCSs. Then configurations of two neighboring temperatures are exchanged with the transition probability $W(X', \beta_m; X, \beta_{m+1}|X, \beta_m; X', \beta_{m+1})$. In general, arbitrary pairs of replicas (say, at two different temperatures $T_n$ and $T_m$) with associated microscopic configurations can undergo temperature switching [8]. We remark that the probability for the temperature exchange between nonadjacent replicas decreases exponentially, yet essential to speed up crossing the high energy barriers [8].

Finally, an expectation of any observable $O$ can be obtained:

$$\langle O \rangle_{\beta_m} = \frac{1}{M} \sum_{t=1}^{M} O\left(X_m(t)\right). \tag{4.25}$$

A pseudo-code for the parallel tempering method is shown in Algorithm 4.2.

**Algorithm 4.2** Parallel tempering Monte Carlo

**Input:** The number of samples $L$, $\beta_{max}$, $\beta_{min}$, and the number of temperatures $M$.
**Output:** Sample collection.
1: Initialize $\beta_1 = \beta_{min}$, $\beta_M = \beta_{max}$;
2: Linear initialization of $\beta$: $\beta_m = \beta_1 + (\beta_M - \beta_1)\ \frac{m-1}{M-1}$;
3: Initialize the extended state randomly: $\{X\} = \{X_1, X_2, \ldots, X_M\}$;
4: Initialize i = 0;
5: Initialize counter = 0;
6: **while** (i< $L$) **do**
7: Applying the MCMC (e.g., the Metropolis method) for each replica for a few MCSs
8: **for** $\beta_m$ in $\{\beta_1, \beta_2, \ldots, \beta_{M-1}\}$ **do**
9: compute $\Delta = (\beta_{m+1} - \beta_m)\left(\mathcal{H}(X_m) - \mathcal{H}\left(X_{m+1}\right)\right)$.
10: **if** $\exp(-\Delta) >$ rand(0, 1) **then**
11: Swap $X_m$ and $X_{m+1}$.
12: Append $\{X\}$ to the sample collection.
13: i = i+1
14: **return** the sample collection.

A high-temperature phase has a fast dynamics, while a low-temperature phase has a very slow dynamics, due to the potential rugged energy landscape. To ensure a proper acceptance ratio, the acceptance probability $e^{-\Delta}$ should be of order of one. According to Eq. (4.23), one has

$$-\Delta = \delta\left(\mathcal{H}\left(X_{n+1}\right) - \mathcal{H}\left(X_n\right)\right) \sim \delta^2 \frac{\mathrm{d}}{\mathrm{d}\beta} E, \tag{4.26}$$

where $\delta$ indicates the small inverse-temperature difference, and $E = \langle \mathcal{H} \rangle$ is the mean thermal energy and is an extensive quantity. To ensure the acceptance probability is of order one, the difference between neighboring temperatures $\delta$ should be of order $\frac{1}{\sqrt{N}}$, implying that a number of order $\sqrt{N}$ of replicas are required [7]. In essence, the

new configuration from the fast mixing chain allows the chains at a low temperature to sample the state space more efficiently, compared with a pure Metropolis local dynamics.

## References

1. U. von Toussaint, Rev. Mod. Phys. **83**(3), 943 (2011)
2. H.G. Katzgraber (2009). arXiv:0905.1629
3. R.M. Neal, Stat. Comput. **11**(2), 125 (2001)
4. D. Sherrington, S. Kirkpatrick, Phys. Rev. Lett. **35**(26), 1792 (1975)
5. S.F. Edwards, P.W. Anderson, J. Phys. F: Met. Phys. **5**(5), 965 (1975)
6. S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science **220**(4598), 671 (1983)
7. K. Hukushima, K. Nemoto, J. Phys. Soc. Jpn. **65**(6), 1604 (1996)
8. C.E. Fiore, M.G.E. da Luz, Phys. Rev. E **82**, 031104 (2010)

# Chapter 5
# High-Temperature Expansion


In this chapter, we introduce one important theoretical technique—high-temperature expansion, to derive the Thouless–Anderson–Palmer (TAP) equation, a seminal equation in standard spin glass theory (Thouless et al. in Phil. Mag. 35(3):593, 1977 [1]; Plefka in J. Phys. A 15(6):1971, 1982 [2]; Georges and Yedidia in J. Phys. A: Math. Gen. 24:2173, 1991 [3]). This technique is quite popular and useful even in machine learning community, acting as a perturbation analysis to derive efficient algorithms for inference and learning (Maillard et al. in J. Stat. Mech.: Theory Exper. 2019(11):113301, 2019 [4]).

## 5.1 Statistical Physics Setting

In statistical physics, given a Hamiltonian $H$, the corresponding partition function is defined as

$$Z = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma})}, \tag{5.1}$$

which is the normalization constant of the Boltzmann distribution. The inverse temperature $\beta = 1/T$, and $\boldsymbol{\sigma}$ denotes the configuration vector. The average of any thermodynamic quantity $A(\boldsymbol{\sigma})$ with respect to the Boltzmann distribution is given by

$$\langle A \rangle = \sum_{\boldsymbol{\sigma}} A(\boldsymbol{\sigma}) P(\boldsymbol{\sigma}), \tag{5.2}$$

where the Boltzmann distribution $P(\boldsymbol{\sigma}) = \frac{e^{-\beta H(\boldsymbol{\sigma})}}{Z}$. The internal energy $E(\beta)$ is, thus, defined as

$$E(\beta) = \langle H \rangle = \sum_{\boldsymbol{\sigma}} H(\boldsymbol{\sigma}) P(\boldsymbol{\sigma}). \tag{5.3}$$

According to the probabilistic interpretation, the entropy is defined as



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_5

$$S(\beta) = -\sum_{\sigma} P(\boldsymbol{\sigma}) \ln P(\boldsymbol{\sigma}). \tag{5.4}$$

The Helmholtz free energy is, thus, defined by

$$F(\beta) = E(\beta) - TS(\beta) = -\frac{1}{\beta} \ln Z(\beta). \tag{5.5}$$

In a complex system, like a neural network, the Boltzmann distribution is commonly hard to compute (including uniform sampling). However, the variational method approximates the intractable distribution $P(\boldsymbol{\sigma})$ by $Q(\boldsymbol{\sigma})$ which belongs to a family $\mathcal{M}$ of tractable distributions. The distribution $Q$ is chosen such that it minimizes a certain distance measure $D(Q, P)$ within the family $\mathcal{M}$. For example, $D(Q, P)$ can be chosen as the Kullback–Leibler (KL) divergence between $Q$ and $P$:

$$\mathrm{KL}(Q||P) = \sum_{\sigma} Q(\boldsymbol{\sigma}) \ln \frac{Q(\boldsymbol{\sigma})}{P(\boldsymbol{\sigma})} = \left\langle \ln \frac{Q}{P} \right\rangle_Q, \tag{5.6}$$

where $\langle \cdots \rangle_Q$ denotes an expectation with respect to $Q$. Inserting $P(\boldsymbol{\sigma}) = \dfrac{e^{-\beta H(\boldsymbol{\sigma})}}{Z}$ into Eq. (5.6), we get

$$\mathrm{KL}(Q||P) = \ln Z + \beta E[Q] - S[Q] = \ln Z + \beta F[Q], \tag{5.7}$$

where the variational energy is then defined by

$$E[Q] = \sum_{\sigma} Q(\boldsymbol{\sigma}) H(\boldsymbol{\sigma}), \tag{5.8}$$

and the entropy of the trial distribution $Q$ is given by

$$S[Q] = -\sum_{\sigma} Q(\boldsymbol{\sigma}) \ln Q(\boldsymbol{\sigma}). \tag{5.9}$$

The variational free energy is, thus, given by

$$F[Q] = E[Q] - TS[Q]. \tag{5.10}$$

We remark that $F[Q]$ constructs an upper bound to the Helmholtz free energy, due to the non-negativity of the KL divergence. The bound is tight once $Q = P$.

To proceed, we introduce the Gibbs free energy $G_\beta(\mathbf{m})$ under the distribution $Q$ as follows:

$$G_\beta(\mathbf{m}) = \min_{Q} \{F[Q] | \langle \boldsymbol{\sigma} \rangle_Q = \mathbf{m}\}. \tag{5.11}$$

The Helmholtz free energy is just a thermodynamic value equal to $E - TS$ at equilibrium, but the Gibbs free energy is a function that gives the value of $E - TS$ when some constraints (e.g., magnetizations) are applied. The advantage of working with a Gibbs free energy instead of a direct computation of the Helmholtz free energy is that it is much easier to apply intuitive approximations, as we explain below.

We then minimize the Gibbs free energy in the following steps. First, we constrain the minimization in the family of distributions satisfying $\langle \boldsymbol{\sigma} \rangle_Q = \mathbf{m}$ for fixed $\mathbf{m}$. By adding a Lagrange multiplier $\boldsymbol{\lambda}$, we obtain

$$\begin{aligned} G_\beta(\mathbf{m}, \boldsymbol{\lambda}) &= E[Q] - TS[Q] - \frac{1}{\beta} \sum_i \lambda_i (\langle \sigma_i \rangle_Q - m_i) \\ &= \sum_{\boldsymbol{\sigma}} Q(\boldsymbol{\sigma}) H(\boldsymbol{\sigma}) - TS[Q] - \frac{1}{\beta} \sum_{\boldsymbol{\sigma}} \sum_i \lambda_i \sigma_i Q(\boldsymbol{\sigma}) + \frac{1}{\beta} \sum_i \lambda_i m_i \\ &= \sum_{\boldsymbol{\sigma}} Q(\boldsymbol{\sigma}) [H(\boldsymbol{\sigma}) - \frac{1}{\beta} \sum_i \lambda_i \sigma_i] - TS[Q] + \frac{1}{\beta} \sum_i \lambda_i m_i . \end{aligned} \tag{5.12}$$

Equation (5.12) is of the form of the variational free energy [Eq. (5.10)], where $H(\boldsymbol{\sigma})$ is replaced by $H(\boldsymbol{\sigma}) - \sum_i \frac{\lambda_i}{\beta} \sigma_i$. Hence, the valid distribution is given by

$$Q_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}) = \frac{e^{-\beta H(\boldsymbol{\sigma}) + \sum_i \lambda_i \sigma_i}}{Z_{\boldsymbol{\lambda}}}, \tag{5.13}$$

where $Z_{\boldsymbol{\lambda}} = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma}) + \sum_i \lambda_i \sigma_i}$. This equation comes from the fact that the variational free energy takes a minimum when $Q$ is the Boltzmann distribution $P(\boldsymbol{\sigma})$. Inserting this distribution back into Eq. (5.12) yields

$$\begin{aligned} G_\beta(\mathbf{m}, \boldsymbol{\lambda}) &= -\frac{1}{\beta} \ln \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma}) + \sum_i \lambda_i \sigma_i} + \frac{1}{\beta} \sum_i \lambda_i m_i \\ &= -\frac{1}{\beta} \ln \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma}) + \sum_i \lambda_i \sigma_i - \sum_i \lambda_i m_i} . \end{aligned} \tag{5.14}$$

The constraint $\langle \boldsymbol{\sigma} \rangle_Q = \mathbf{m}$ has been enforced by the Lagrange multiplier $\boldsymbol{\lambda}$ that is determined by

$$\beta G_\beta(\mathbf{m}) = \max_{\boldsymbol{\lambda}} \left\{ - \ln \sum_{\boldsymbol{\sigma}} e^{-\beta H(\sigma) + \sum_i \lambda_i \sigma_i} + \sum_i \lambda_i m_i \right\}. \tag{5.15}$$

The max operation is related to the property of the Hessian matrix. By using the Lagrangian multiplier method, we carry out the derivatives:

$$\frac{\partial \left( -\beta G_\beta(\mathbf{m}, \boldsymbol{\lambda}) \right)}{\partial \lambda_i} = \langle \sigma_i \rangle_Q - m_i = 0 \quad \Rightarrow m_i = \langle \sigma_i \rangle_Q, \tag{5.16}$$

$$\begin{aligned}\frac{\partial\left(-\beta G_\beta(\mathbf{m},\boldsymbol{\lambda})\right)}{\partial m_i} &= \sum_j \frac{\partial}{\partial \lambda_j}\left[-\beta F_\beta(\boldsymbol{\lambda})\right]\frac{\partial \lambda_j}{\partial m_i} - \lambda_i - \sum_j \frac{\partial \lambda_j}{\partial m_i} m_j \\ &= \sum_j \frac{\partial \lambda_j}{\partial m_i}(\langle\sigma_j\rangle_{\boldsymbol{\lambda}} - m_j) - \lambda_i \\ &= -\lambda_i \\ &= 0,\end{aligned} \tag{5.17}$$

where $-\beta F_\beta(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \ln Z_{\boldsymbol{\lambda}}$. Finally, we obtain

$$\min_{\mathbf{m}} G_\beta(\mathbf{m}) = F[P] = -\frac{1}{\beta}\ln Z. \tag{5.18}$$

Note that $G_\beta(\mathbf{m})$ is a convex function with a unique minimum at $\mathbf{m}_{\text{eq}}$. In sum, the approximate computation of $G_\beta(\mathbf{m})$ can be used to get an approximation for the true free energy $F[P]$ as well.

## 5.2 High-Temperature Expansion

In this section, we apply the high-temperature expansion to approximate the true Helmholtz free energy. We first introduce the seminal Sherrington–Kirkpatrick (SK) model. This model was introduced in 1975 as a simple model of spin glass [5]. It is actually an Ising model with disordered couplings. For simplicity, we ignore external fields here. The Hamiltonian of the model is given by

$$H(\boldsymbol{\sigma}) = -\sum_{i<j}^{N} J_{ij}\sigma_i\sigma_j, \tag{5.19}$$

where couplings $J_{ij}$ are independent and are Gaussian random variables for $i < j$ with mean $J_0$ (here we just assume $J_0$ to be 0) and variance $J^2/N$. $J_{ij}$ acts as quenched disorder for the model.

The Gibbs free energy is unfortunately intractable, making an optimization in the magnetization space challenging as well. Therefore, we need to consider a perturbation analysis of the free energy, e.g., in terms of high temperatures. The approximation accuracy can be controlled by including higher orders of expansion.

We define a new partition function associated with the Hamiltonian as follows:

$$\tilde{Z}_\beta = \sum_{\boldsymbol{\sigma}} e^{-\beta\tilde{H}(\boldsymbol{\sigma})}, \tag{5.20}$$

where the modified Hamiltonian is given by

$$\tilde{H}(\boldsymbol{\sigma}) = H(\boldsymbol{\sigma}) - \sum_i \frac{\lambda_i(\beta)}{\beta}(\sigma_i - m_i) = -\frac{1}{2}\sum_{i\neq j} J_{ij}\sigma_i\sigma_j - \sum_i \frac{\lambda_i(\beta)}{\beta}(\sigma_i - m_i), \tag{5.21}$$

where we write $\lambda_i(\beta)$ as an explicit function of the temperature, because $\lambda_i$ is used to enforce the magnetization that depends on the temperature. The relation between the Gibbs free energy and the new partition function is

$$-\beta G_\beta(\mathbf{m}, \boldsymbol{\lambda}) = \ln \tilde{Z}_\beta. \tag{5.22}$$

We then carry out the Taylor expansion at $\beta = 0$:

$$-\beta G_\beta(\mathbf{m}) = \ln \tilde{Z}_\beta\Big|_{\beta=0} + \frac{\partial}{\partial\beta}\ln \tilde{Z}_\beta\Big|_{\beta=0}\beta + \frac{\partial^2}{\partial\beta^2}\ln \tilde{Z}_\beta\Big|_{\beta=0}\frac{\beta^2}{2} + \cdots. \tag{5.23}$$

At $\beta = 0$, we obtain

$$\begin{aligned}\tilde{Z}_\beta\Big|_{\beta=0} &= \sum_{\boldsymbol{\sigma}} e^{\sum_i \lambda_i(\sigma_i - m_i)} \\ &= \prod_i \sum_{\sigma_i} e^{\lambda_i(\sigma_i - m_i)} \\ &= \prod_i e^{-\lambda_i m_i} \prod_i (2\cosh\lambda_i).\end{aligned} \tag{5.24}$$

Because

$$\frac{\partial \ln \tilde{Z}_{\beta=0}}{\partial\lambda_i} = -m_i + \tanh(\lambda_i) = 0 \Rightarrow \begin{cases} m_i = \tanh(\lambda_i) \\ \lambda_i = \operatorname{atanh}(m_i) \end{cases}, \tag{5.25}$$

then we can calculate the first term:

$$\begin{aligned}\ln \tilde{Z}_\beta|_{\beta=0} &= -\sum_i \operatorname{atanh}(m_i)\, m_i + \sum_i \ln(2\cosh(\operatorname{atanh} m_i)) \\ &= -\sum_i \frac{1}{2} m_i \ln\frac{1+m_i}{1-m_i} + \sum_i \ln\left[e^{-\frac{1}{2}\ln\frac{1+m_i}{1-m_i}} + e^{\frac{1}{2}\ln\frac{1+m_i}{1-m_i}}\right] \\ &= -\sum_i \frac{1}{2} m_i \ln\frac{1+m_i}{1-m_i} + \sum_i \ln\left(\left(\frac{1+m_i}{1-m_i}+1\right)\sqrt{\frac{1-m_i}{1+m_i}}\right) \\ &= \sum_i \left(-\frac{m_i}{2}\ln\frac{1+m_i}{2} + \frac{m_i}{2}\ln\frac{1-m_i}{2}\right) + \sum_i \ln\frac{2}{\sqrt{(1-m_i)(1+m_i)}} \\ &= -\sum_i \left(\frac{1+m_i}{2}\ln\frac{1+m_i}{2} + \frac{1-m_i}{2}\ln\frac{1-m_i}{2}\right).\end{aligned} \tag{5.26}$$

Here, we have used the mathematical identity: $\operatorname{atanh} m_i = \frac{1}{2} \ln \frac{1+m_i}{1-m_i}$. The second term is given by

$$\begin{aligned} \left. \frac{\partial \ln \tilde{Z}_\beta}{\partial \beta} \right|_{\beta=0} &= \frac{1}{\tilde{Z}_\beta} \sum_{\boldsymbol{\sigma}} \left[ (-H) e^{-\beta \tilde{H}} + \sum_i \frac{\partial \lambda_i}{\partial \beta} (\sigma_i - m_i) e^{-\beta \tilde{H}} \right] \\ &= - \langle H \rangle |_{\beta=0} \\ &= \frac{1}{2} \sum_{i \neq j} J_{ij} m_i m_j . \end{aligned} \tag{5.27}$$

Note that the thermal average is carried out under the Boltzmann measure of $\tilde{H}(\boldsymbol{\sigma})$, and the correlation between two spins is negligible in the high-temperature limit. The third term is given by

$$\begin{aligned} \frac{\partial^2}{\partial \beta^2} \ln \tilde{Z}_\beta &= - \frac{\partial \langle H \rangle}{\partial \beta} \\ &= - \frac{\partial}{\partial \beta} \left[ \sum_{\boldsymbol{\sigma}} \frac{H e^{-\beta \tilde{H}}}{\tilde{Z}_\beta} \right] \\ &= - \left[ \sum_{\boldsymbol{\sigma}} \frac{e^{-\beta \tilde{H}}}{\tilde{Z}_\beta} H \left( -H + \sum_i \frac{\partial \lambda_i}{\partial \beta} (\sigma_i - m_i) \right) - \sum_{\boldsymbol{\sigma}} \frac{H e^{-\beta \tilde{H}}}{\tilde{Z}_\beta} \cdot \frac{\partial \ln \tilde{Z}_\beta}{\partial \beta} \right] \\ &= - \left\langle H \left( -H + \sum_i \frac{\partial \lambda_i}{\partial \beta} (\sigma_i - m_i) \right) - H(-\langle H \rangle) \right\rangle \\ &= \left\langle H \left( H - \langle H \rangle - \sum_i \frac{\partial \lambda_i}{\partial \beta} (\sigma_i - m_i) \right) \right\rangle \\ &:= \langle u H \rangle . \end{aligned} \tag{5.28}$$

Here, we have introduced a very useful operator $u$ as follows [3]:

$$u := H - \langle H \rangle - \sum_i \frac{\partial \lambda_i}{\partial \beta} (\sigma_i - m_i) = H - \langle H \rangle - k \tag{5.29}$$

where $k := \sum_i \frac{\partial \lambda_i}{\partial \beta} (\sigma_i - m_i)$. Because

$$\frac{\partial \ln \tilde{Z}_\beta}{\partial m_i} = -\lambda_i \frac{\tilde{Z}_\beta}{\tilde{Z}_\beta} = -\lambda_i , \tag{5.30}$$

we have the following result:

$$\begin{aligned}\left.\frac{\partial\lambda_i}{\partial\beta}\right|_{\beta=0} &= -\frac{\partial}{\partial\beta}\frac{\partial\ln\tilde{Z}_\beta}{\partial m_i} = -\frac{\partial}{\partial m_i}\frac{\partial\ln\tilde{Z}_\beta}{\partial\beta} \\ &= -\frac{1}{2}\frac{\partial}{\partial m_i}\sum_{i\neq j}J_{ij}m_i m_j = -\sum_{j(\neq i)}J_{ij}m_j.\end{aligned} \tag{5.31}$$

We, thus, conclude that

$$\begin{aligned}u|_{\beta=0} &= -\frac{1}{2}\sum_{i\neq j}J_{ij}\sigma_i\sigma_j + \frac{1}{2}\sum_{i\neq j}J_{ij}m_i m_j + \sum_i\sum_{j(j\neq i)}J_{ij}m_j(\sigma_i - m_i) \\ &= -\frac{1}{2}\sum_{i\neq j}J_{ij}\left(\sigma_i - m_i\right)\left(\sigma_j - m_j\right).\end{aligned} \tag{5.32}$$

To proceed, we should first calculate the mean and variance of $u$:

$$\langle u\rangle = 0, \tag{5.33}$$

and

$$\left\langle u^2\right\rangle = \langle u\left(H - \langle H\rangle - k\right)\rangle = \langle uH\rangle - \langle u\rangle\langle H\rangle - \langle ku\rangle = \langle uH\rangle. \tag{5.34}$$

We can prove above Eqs. (5.33) and (5.34) by using the following identity:

$$\begin{aligned}\frac{d}{d\beta}\langle O\rangle &= \frac{1}{\tilde{Z}_\beta}\sum_{\boldsymbol{\sigma}}Oe^{-\beta\tilde{H}}\left(-H + \sum_i\frac{\partial\lambda_i(\beta)}{\partial\beta}(\sigma_i - m_i)\right) \\ &\quad + \frac{\sum_{\boldsymbol{\sigma}}Oe^{-\beta\tilde{H}}}{\tilde{Z}_\beta}\left[-\frac{\partial\ln\tilde{Z}_\beta}{\partial\beta}\right] + \frac{\sum_{\boldsymbol{\sigma}}\frac{\partial O}{\partial\beta}e^{-\beta\tilde{H}}}{\tilde{Z}_\beta} \\ &= \left\langle\frac{\partial O}{\partial\beta}\right\rangle - \langle Ou\rangle,\end{aligned} \tag{5.35}$$

where $O$ is any observable, and

$$\begin{aligned}&\frac{d}{d\beta}\langle\sigma_i\rangle = 0 = \left\langle\frac{\partial\sigma_i}{\partial\beta}\right\rangle - \langle\sigma_i u\rangle = -\langle\sigma_i u\rangle, \\ &\langle\left(\sigma_i - m_i\right)u\rangle = 0, \\ &\langle ku\rangle = 0.\end{aligned} \tag{5.36}$$

Note that the full derivative vanishes due to the constrained magnetization [i.e., as a constant, see also Eq. (5.11)]. Therefore, we can obtain $\left.\frac{\partial^2}{\partial\beta^2}\ln\tilde{Z}_\beta\right|_{\beta=0}$ by calculating $\left\langle u^2\right\rangle|_{\beta=0}$:

$$\begin{aligned}\left.\frac{\partial^2}{\partial\beta^2}\ln\tilde{Z}_\beta\right|_{\beta=0} &= \langle u^2\rangle|_{\beta=0} \\ &= \frac{1}{4}\sum_{i\neq j,k\neq l} J_{ij}J_{kl}\left\langle(\sigma_i-m_i)\left(\sigma_j-m_j\right)(\sigma_k-m_k)(\sigma_l-m_l)\right\rangle \\ &\simeq \frac{1}{2}\sum_{i\neq j} J_{ij}^2\left\langle(\sigma_i-m_i)^2\left(\sigma_j-m_j\right)^2\right\rangle \\ &= \frac{1}{2}\sum_{i\neq j} J_{ij}^2\left(1-m_i^2\right)\left(1-m_j^2\right),\end{aligned} \tag{5.37}$$

where we have used the formula: $\left\langle(\sigma_i-m_i)^2\right\rangle = 1-2m_i^2+m_i^2 = 1-m_i^2$. Finally, we obtain

$$\begin{aligned}-\beta G_\beta(\mathbf{m}) = &-\sum_i\left(\frac{1+m_i}{2}\ln\frac{1+m_i}{2}+\frac{1-m_i}{2}\ln\frac{1-m_i}{2}\right) \\ &+\frac{1}{2}\beta\sum_{i\neq j}J_{ij}m_im_j+\frac{\beta^2}{4}\sum_{i\neq j}J_{ij}^2\left(1-m_i^2\right)\left(1-m_j^2\right)+O(\beta^3).\end{aligned} \tag{5.38}$$

The first term on the right side of the above equation is called the mean-field variational entropy. The second term is the mean-field variational energy. The third term corresponds to the Onsager reaction correction. All three terms construct the TAP free energy for the SK model.

To minimize the free energy, we carry out the differentiation with respect to $\{m_i\}$,

$$\frac{\partial\left(-\beta G_\beta(\mathbf{m})\right)}{\partial m_i} = -\operatorname{atanh}(m_i)+\beta\sum_{j(\neq i)}J_{ij}m_j+\frac{\beta^2}{2}\sum_{j(\neq i)}J_{ij}^2\left(1-m_j^2\right)(-2m_i)=0, \tag{5.39}$$

and finally obtain the self-consistent equation (the so-called TAP equation):

$$m_i = \tanh\left(\beta\sum_{j(\neq i)}J_{ij}m_j-\beta^2\sum_{j(\neq i)}J_{ij}^2\left(1-m_j^2\right)m_i\right). \tag{5.40}$$

The first term on the right-hand side represents the standard mean-field approximation of local fields. The second term is called the Onsager reaction field added to remove the effects of self-response [6]. If we consider the external fields $\{h_i\}$, the TAP equation becomes

$$m_i^{t+1} = \tanh\left(\beta h_i+\beta\sum_{j(\neq i)}J_{ij}m_j^t-\beta^2\sum_{j(\neq i)}J_{ij}^2\left(1-(m_j^t)^2\right)m_i^{t-1}\right), \tag{5.41}$$

where we have put the correct time indexes for iteration [7]. In the thermodynamic limit, the TAP approximation becomes exact for the SK model, as the terms $O(\beta^3)$ vanish. The fixed points of TAP are the stationary points of the TAP free energy. At low temperatures, the TAP equation have many solutions with $m_i \neq 0$, which can be interpreted as stable or metastable thermodynamic states [8, 9].

## 5.3 Properties of the TAP Equation

In this section, we study the behavior of the solution of the TAP equation [Eq. (5.40) where external fields are added] around the spin glass transition point [6]. Because $J_{ij}$ are assumed to be independent random variables (for $i < j$) with zero mean and the variance $J^2/N$. The Onsager term of the TAP equation becomes

$$\beta^2 \sum_{j(\neq i)}^{N} J_{ij}^2 \left(1 - m_j^2\right) m_i = \beta^2 J^2 m_i - \beta^2 \sum_{j(\neq i)}^{N} J_{ij}^2 m_j^2 m_i, \tag{5.42}$$

when $N \to \infty$ (the law of large numbers applies). Around the spin glass transition point, we assume that the magnetizations $\{m_i\}$ are small, expand the right-hand side of the TAP equation to the first order in $\mathbf{m}$ and finally arrive at

$$m_i = \beta \sum_j J_{ij} m_j + \beta h_i - \beta^2 J^2 m_i. \tag{5.43}$$

We also assume that $h_i$ is not dominant. For the symmetric matrix $\mathbf{J}$, we have $\mathbf{J} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q}$ is the orthogonal matrix, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ in which $\{\lambda_i\}$ are eigenvalues of the interaction matrix $\mathbf{J}$. Let us write $J_{ij}$ in the following form:

$$J_{ij} = \sum_n Q_{in} Q_{jn} \lambda_n. \tag{5.44}$$

To proceed, we define the $\lambda$-magnetization and $\lambda$-field by [6]

$$m_{\lambda_n} = \sum_i Q_{in} m_i, \quad h_{\lambda_n} = \sum_i Q_{in} h_i. \tag{5.45}$$

Then we have the following result:

$$\begin{aligned}\beta \sum_i Q_{in} \sum_j J_{ij} m_j &= \beta \sum_i Q_{in} \sum_j \sum_m Q_{im} Q_{jm} \lambda_m m_j \\ &= \beta \sum_m \lambda_m \sum_i Q_{in} Q_{im} \sum_j Q_{jm} m_j \\ &= \beta \lambda_n \sum_j Q_{jn} m_j \\ &= \beta \lambda_n m_{\lambda_n},\end{aligned} \tag{5.46}$$

where we have used the orthogonal condition: $\sum_i Q_{in} Q_{im} = \delta_{nm}$. Then we can rewrite Eq. (5.43) as

$$m_\lambda = \beta m_\lambda \lambda + \beta h_\lambda - \beta^2 J^2 m_\lambda. \tag{5.47}$$

We can, thus, conclude that the $\lambda$-susceptibility can be expressed as [10]

$$\chi_\lambda = \frac{\partial m_\lambda}{\partial h_\lambda} = \frac{\beta}{1 - \beta\lambda + (\beta J)^2}. \tag{5.48}$$

In addition, the eigenvalues of the random matrix $\mathbf{J}$ follow the well-known semi-circle law[1] [11]:

$$\rho(\lambda) = \frac{\sqrt{4J^2 - \lambda^2}}{2\pi J^2}. \tag{5.49}$$

It is easy to derive from Eq. (5.48) that the susceptibility corresponding to the largest eigenvalue $\lambda = 2J$ diverges at $T_g = J$, suggesting a continuous phase transition. The location of this transition agrees exactly with that obtained from the replica result [5].

An alternative way to see the stability condition of the paramagnetic phase is to compute the Hessian matrix:

$$H_{ij} = \left.\frac{\partial^2 (\beta G_\beta(\mathbf{m}))}{\partial m_i \partial m_j}\right|_{\mathbf{m}=0} = -\beta J_{ij} + (\beta^2 J^2 + 1)\delta_{ij}. \tag{5.50}$$

The stability condition is that all the eigenvalues of the Hessian matrix should be positive, leading to the same result as above. The susceptibility matrix $\chi_{ij} = \frac{\partial m_i}{\partial h_j}$ is related to the Hessian matrix as $(\mathbf{H}^{-1})_{ij} = \beta^{-1}\chi_{ij} = \langle \sigma_i \sigma_j \rangle_c$ followed from the linear response theory. The subscript c denotes the connected two-point correlation.

## References

1. D.J. Thouless, P.W. Anderson, R.G. Palmer, Phil. Mag. **35**(3), 593 (1977)
2. T. Plefka, J. Phys. A **15**(6), 1971 (1982)

[1] We will derive this law in Chap. 17.

3. A. Georges, J. Yedidia, J. Phys. A: Math. Gen. **24**, 2173 (1991)
4. A. Maillard, L. Foini, A.L. Castellanos, F. Krzakala, M. Mézard, L. Zdeborova, J. Stat. Mech.: Theory Exper. **2019**(11), 113301 (2019)
5. D. Sherrington, S. Kirkpatrick, Phys. Rev. Lett. **35**(26), 1792 (1975)
6. H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, Oxford, 2001)
7. E. Bolthausen, Commun. Math. Phys. **325**(1), 333 (2014)
8. A. Crisanti, L. Leuzzi, G. Parisi, T. Rizzo, Phys. Rev. Lett. **92**(12), 127203 (2004)
9. T. Aspelmeier, A.J. Bray, M.A. Moore, Phys. Rev. Lett. **92**(8), 87203 (2004)
10. A.J. Bray, M.A. Moore, J. Phys. C: Solid State Phys. **12**(11), L441 (1979)
11. M.L. Mehta, *Random Matrices* (Academic, San Diego, 2004)

# Chapter 6
# Nishimori Line


In this chapter, we introduce the Nishimori line as an important concept, i.e., Nishimori temperature or constraint, on spin glass models of broad contexts. This concept was first discovered in the traditional two-body interaction spin glass model [1, 2], which demonstrated that on a special temperature, the model energy of a complex glass model is analytic, and the replica symmetry breaking (RSB) phase (introduced in Chap. 9) is not dominant for ground states, and thus, the underlying physics is greatly simplified. The concept was later connected to the Bayes optimal setting of statistical inference problems [3–6]. Thus, this concept is an important theoretical perspective to understand the Bayesian learning process, one of the most popular paradigms in the deep learning era. Here, we introduce the basic knowledge about this concept first, and we leave more applications to later chapters of learning theory.

## 6.1 Model Setting

The original model Hidetoshi Nishimori used to derive the special temperature is defined as follows:

$$H = -\sum_{i<j} J_{ij}\sigma_i\sigma_j, \tag{6.1}$$

where $J_{ij}$ acts as a quenched disorder. The coupling distribution function is specified as follows:

$$P(J_{ij}) = p\delta(J_{ij} - J) + (1-p)\delta(J_{ij} + J), \tag{6.2}$$

where $p$ denotes a ferromagnetic bias for the coupling, and $J$ is a positive constant. Each coupling is generated independently from this binomial distribution.

Let $J_{ij} = J\tau_{ij}$, where $\tau_{ij} = \pm 1$. For the sake of convenience, we then introduce an auxiliary temperature $\beta_p$ to parameterize the original distribution $P(J_{ij})$:



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_6

$$P(J_{ij}) = P(\tau_{ij}) = \frac{e^{\beta_p \tau_{ij}}}{2\cosh \beta_p}, \tag{6.3a}$$

$$\beta_p = \frac{1}{2} \ln\Big(\frac{1-p}{p}\Big). \tag{6.3b}$$

The form of $\beta_p$ ensures that the two forms of the coupling distribution are equivalent. Readers can easily verify this point by considering both possible values of the coupling.

## 6.2 Exact Result for Internal Energy

According to the model definition, the internal energy can be written as follows:

$$\begin{aligned} \langle H \rangle_{\boldsymbol{\tau},\boldsymbol{\sigma}} &= \sum_{\boldsymbol{\tau}} P(\boldsymbol{\tau}) \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \left( -J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j \right) \\ &= \sum_{\boldsymbol{\tau}} \frac{e^{\beta_p \sum_{i<j} \tau_{ij}}}{(2\cosh\beta_p)^{N_B}} \sum_{\boldsymbol{\sigma}} \frac{e^{\beta J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j}}{\sum_{\boldsymbol{\sigma}} e^{\beta J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j}} \left( -J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j \right), \end{aligned} \tag{6.4}$$

where $N_B$ is the number of interactions (also called bonds in a lattice model). Note that $P(\boldsymbol{\tau})$ is factorized, as the $\{\tau_{ij}\}$ are independent. We further remark that the Hamiltonian of the model is invariant under the following gauge transformation:

$$\tau_{ij} \to \tau_{ij} s_i s_j, \tag{6.5}$$

$$\sigma_i \to \sigma_i s_i. \tag{6.6}$$

Note that $\{s_i\}$ is also an Ising-valued configuration. Therefore, we apply this transformation to the model internal energy as follows:

$$\begin{aligned} \langle H \rangle_{\boldsymbol{\tau},\boldsymbol{\sigma}} &= -\sum_{\boldsymbol{\tau}} \frac{e^{\beta_p \sum_{i<j} \tau_{ij}s_is_j}}{(2\cosh\beta_p)^{N_B}} \frac{\sum_{\boldsymbol{\sigma}} e^{\beta J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j} J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j}{\sum_{\boldsymbol{\sigma}} e^{\beta J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j}} \\ &= -\frac{1}{2^N} \sum_{\boldsymbol{\tau}} \frac{\sum_{\boldsymbol{s}} e^{\beta_p \sum_{i<j} \tau_{ij}s_is_j}}{(2\cosh\beta_p)^{N_B}} \frac{\sum_{\boldsymbol{\sigma}} e^{\beta J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j} J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j}{\sum_{\boldsymbol{\sigma}} e^{\beta J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j}} \\ &= -\frac{1}{2^N} \sum_{\boldsymbol{\tau}} \frac{Z_s}{(2\cosh\beta_p)^{N_B}} \frac{\sum_{\boldsymbol{\sigma}} e^{\beta J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j} J \sum_{i<j} \tau_{ij}\sigma_i\sigma_j}{Z_{\boldsymbol{\sigma}}}, \end{aligned} \tag{6.7}$$

where $2^N$ is introduced to cancel the sum operation $\sum_{\boldsymbol{s}} \bullet$. Clearly, when $\beta J = \beta_p$, the partition functions $Z_{\mathbf{s}}$ and $Z_{\boldsymbol{\sigma}}$ cancel with each other. Then, we have

$$\begin{aligned}
\langle H\rangle_{\boldsymbol{\tau},\boldsymbol{\sigma}} &= -\frac{1}{2^N}\frac{1}{(2\cosh\beta_p)^{N_B}}\sum_{\boldsymbol{\tau}}\sum_{\boldsymbol{\sigma}} e^{\beta J\sum_{i<j}\tau_{ij}\sigma_i\sigma_j} J\sum_{i<j}\tau_{ij}\sigma_i\sigma_j \\
&= -\frac{1}{2^N}\frac{1}{(2\cosh\beta_p)^{N_B}}\frac{\partial}{\partial\beta}\sum_{\boldsymbol{\tau}}\sum_{\boldsymbol{\sigma}} e^{\beta J\sum_{i<j}\tau_{ij}\sigma_i\sigma_j} \\
&= -\frac{1}{2^N}\frac{1}{(2\cosh\beta_p)^{N_B}}\frac{\partial}{\partial\beta}\sum_{\boldsymbol{\sigma}}\prod_{i<j}\sum_{\tau_{ij}} e^{\beta J\tau_{ij}\sigma_i\sigma_j} \\
&= -\frac{1}{2^N}\frac{1}{(2\cosh\beta_p)^{N_B}}\frac{\partial}{\partial\beta}\sum_{\boldsymbol{\sigma}}(2\cosh\beta_p)^{N_B} \\
&= -N_B J\tanh\beta_p.
\end{aligned} \tag{6.8}$$

Therefore, under the Nishimori temperature $\beta J = \beta_p$, the internal energy for the model has an analytical expression. In general, the internal energy does not have a closed-form expression.

## 6.3 Proof of No RSB Effects on the Nishimori Line

In this section, we will prove that, using the gauge transformation, the distribution of spin glass order parameters does not have a complex structure on the Nishimori line ($\beta_p$) and coincides exactly with the distribution of magnetizations.

The magnetization distribution is defined as follows:

$$P_m(x;k) = \sum_{\boldsymbol{\tau}}\frac{e^{k_p\sum_{i<j}\tau_{ij}}}{(2\cosh k_p)^{N_B}}\sum_{\boldsymbol{\sigma}}\frac{e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}{\sum_{\boldsymbol{\sigma}} e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}\delta\left(x-\frac{1}{N}\sum_i\sigma_i\right), \tag{6.9}$$

where we have defined $k = \beta J$ and $k_p = \beta_p$. Double averages are performed in the definition of the magnetization distribution: the one over $\boldsymbol{\sigma}$ is the thermal average, and the other over $\boldsymbol{\tau}$ is the disorder average. Both averages are standard thermodynamic operations in the spin glass theory. The disorder average is usually challenging.

Next, we apply the following gauge transformation:

$$\tau_{ij} \to \tau_{ij}s_i s_j, \tag{6.10}$$

$$\sigma_i \to \sigma_i s_i. \tag{6.11}$$

Then, $P_m(x;k)$ changes to

$$P_m(x;k) = \frac{1}{2^N}\sum_{\boldsymbol{\tau}}\sum_{\boldsymbol{s}}\frac{e^{k_p\sum_{i<j}\tau_{ij}s_i s_j}}{(2\cosh k_p)^{N_B}}\sum_{\boldsymbol{\sigma}}\frac{e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}{\sum_{\boldsymbol{\sigma}} e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}\delta\left(x-\frac{1}{N}\sum_i\sigma_i s_i\right). \tag{6.12}$$

This form of $P_m(x;k)$ can be further simplified to make the underlying physics more transparent. A simple algebraic manipulation leads to

$$P_m(x;k) = \frac{1}{2^N}\sum_{\boldsymbol{\tau}}\sum_{\boldsymbol{s}}\frac{e^{k_p\sum_{i<j}\tau_{ij}s_is_j}}{(2\cosh k_p)^{N_B}}\sum_{\boldsymbol{\sigma}}\frac{e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}{\sum_{\boldsymbol{\sigma}}e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}\delta\left(x-\frac{1}{N}\sum_i\sigma_is_i\right) \times \frac{\sum_{\boldsymbol{s}'}e^{k_p\sum_{i<j}\tau_{ij}s_i's_j'}}{\sum_{\boldsymbol{s}}e^{k_p\sum_{i<j}\tau_{ij}s_is_j}}.$$

We then perform the second gauge transformation: $\tau_{ij}\to\tau_{ij}s_i's_j', \sigma_i\to\sigma_is_i',$ *and* $s_i\to s_is_i'$, resulting in

$$\begin{aligned}P_m(x;k) &= \sum_{\boldsymbol{\tau}}\frac{e^{k_p\sum_{i<j}\tau_{ij}}}{(2\cosh k_p)^{N_B}}\sum_{\boldsymbol{s}}\frac{e^{k_p\sum_{i<j}\tau_{ij}s_is_j}}{\sum_{\boldsymbol{s}}e^{k_p\sum_{i<j}\tau_{ij}s_is_j}} \\ &\quad\times\sum_{\boldsymbol{\sigma}}\frac{e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}{\sum_{\boldsymbol{\sigma}}e^{k\sum_{i<j}\tau_{ij}\sigma_i\sigma_j}}\delta\left(x-\frac{1}{N}\sum_i\sigma_is_i\right) \\ &= \sum_{\boldsymbol{\tau}}P(\boldsymbol{\tau})\sum_{\boldsymbol{\sigma}}P(\boldsymbol{\sigma})\sum_{\boldsymbol{s}}P(\boldsymbol{s})\delta\left(x-\frac{1}{N}\sum_i\sigma_is_i\right) \\ &= P_q(x;k,k_p),\end{aligned} \tag{6.13}$$

where $P(\boldsymbol{\sigma})$ and $P(\mathbf{s})$ are the Boltzmann measures with (rescaled) inverse temperature $k$ and $k_p$, respectively.

Under the Nishimori temperature, $P_m(x;k_p)=P_q(x;k_p,k_p)=P_q(x;k_p)$. We, thus, conclude that the distribution of spin glass order parameter (overlap $q=\frac{1}{N}\sum_i\sigma_is_i$) shares the same form as the magnetization distribution. It is well known that the magnetization distribution in statistical physics is simple, while the overlap distribution can be very complex (e.g., when replica symmetry breaking effects dominate the phase space, like in the SK model). The two equivalent distributions on the Nishimori line suggest an absence of spin glass phase for the ground states. However, RSB may be needed to describe the metastable (out of equilibrium) states of the system (e.g., in the study [7]). Altogether, on the Nishimori line, the system never enters the glassy phase and the dominant thermodynamic phase is always a RS type.

## References

1. H. Nishimori, J. Phys. C: Solid State Phys. **13**(21), 4071 (1980)
2. H. Nishimori, Prog. Theor. Phys. **66**(4), 1169 (1981)
3. Y. Iba, J. Phys. A: Math. Gen. **32**, 3875 (1999)
4. L. Zdeborova, F. Krzakala, Adv. Phys. **65**(5), 453 (2016)
5. H. Huang, J. Stat. Mech.: Theory Exper. **2017**(5), 053302 (2017)
6. T. Hou, H. Huang, Phys. Rev. Lett. **124**, 248302 (2020)
7. M. Yoshida, T. Uezu, T. Tanaka, M. Okada, J. Phys. Soc. Jpn. **76**(5), 54003 (2007)

# Chapter 7
# Random Energy Model


In this chapter, we briefly introduce the well-known random energy model (Derrida in Phys. Rev. Lett. 45:79, 1980 [1]; Derrida in Phys. Rev. B 24(5):2613, 1981 [2]), which is the infinite-body interaction limit of $p$-spin interaction models, but still captures characteristics of spin glasses (Gross and Mezard in Nuclear Phys. 240(4):431, 1984 [3]). Here, we focus on basic concepts and their connections to frozen phases commonly observed in other constraint satisfaction problems, e.g., binary Perceptron (introduced in Chap. 13).

## 7.1 Model Setting

We consider Ising-type spins, whose interaction follows the Hamiltonian:

$$\mathcal{H}(\sigma) = -\sum_{1\le i_1\dots i_p\le N} J_{i_1\dots i_p}\sigma_{i_1}\cdots\sigma_{i_p}, \tag{7.1}$$

where the coupling follows the Gaussian distribution defined by

$$P(J_{i_1\dots i_p}) = \sqrt{\frac{N^{p-1}}{\pi J^2 p!}}\exp\left[-\frac{J^2_{i_1\dots i_p}N^{p-1}}{J^2 p!}\right], \tag{7.2}$$

where $J$ is positive, and the scaling of variance ensures that extensive energy is well-defined. A generalized Hopfied model with multi-body interactions can also be included in this class of models [4]. In this scaling, it is easy to verify that $p = 2$ corresponds to the standard Sherrington–Kirkpatrick model.

We are interested in the distribution of the energy level $E$, to see if this distribution becomes simple in the limit $p \to \infty$. In general, the distribution can be very complex. According to the definition, we have



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_7

$$
\begin{aligned}
P(E) &= \overline{\delta(E - \mathcal{H}(\sigma))} \\
&= \int \frac{d\hat{E}}{2\pi} \overline{\exp\left[ \mathrm{i}\,\hat{E}E + \mathrm{i}\,\hat{E} \sum_{1 \le i_1 \ldots i_p \le N} J_{i_1 \ldots i_p} \sigma_{i_1} \cdots \sigma_{i_p} \right]} \\
&= \int \frac{d\hat{E}}{2\pi} e^{\mathrm{i}\,\hat{E}E} \prod_{1 \le i_1 \ldots i_p \le N} \overline{e^{\mathrm{i}\,\hat{E} J_{i_1 \ldots i_p} \sigma_{i_1} \cdots \sigma_{i_p}}},
\end{aligned}
\tag{7.3}
$$

where the quenched-disorder average (indicated by the over-bar) can be explicitly calculated out as follows:

$$
\begin{aligned}
\overline{e^{\mathrm{i}\,\hat{E} J_{i_1 \ldots i_p} \sigma_{i_1} \cdots \sigma_{i_p}}} &= \int P(J_{i_1 \ldots i_p}) dJ_{i_1 \ldots i_p} e^{\mathrm{i}\,\hat{E} J_{i_1 \ldots i_p} \sigma_{i_1} \cdots \sigma_{i_p}} \\
&= \exp\left[ \frac{(\mathrm{i}\,\hat{E}\sigma_{i_1} \cdots \sigma_{i_p})^2 J^2 p!}{4N^{p-1}} \right].
\end{aligned}
\tag{7.4}
$$

Note that the total number of the products in Eq. (7.3) can be approximated by $\frac{N^p}{p!}$ when $N \to \infty$. Therefore, we finally arrive at

$$
\begin{aligned}
P(E) &= \int \frac{d\hat{E}}{2\pi} e^{\mathrm{i}\,\hat{E}E + \frac{(\mathrm{i}\,\hat{E})^2 N J^2}{4}} \\
&= \frac{1}{\sqrt{N\pi J^2}} e^{-\frac{E^2}{NJ^2}},
\end{aligned}
\tag{7.5}
$$

which is exactly a Gaussian distribution with zero mean and a fluctuation of the order $\mathcal{O}(\sqrt{N})$.

The Gaussian distribution of energy levels in $p$-spin interaction models does not imply any information about whether the energy levels are correlated or not. To address this question, we derive the joint distribution of two energy levels, say $E_1$ and $E_2$, as follows:

$$
\begin{aligned}
&P(E_1, E_2, q) = \overline{\delta(E_1 - \mathcal{H}(\sigma^1))\delta(E_2 - \mathcal{H}(\sigma^2))} \\
&= \iint \frac{d\hat{E}_1 \hat{E}_2}{4\pi^2} e^{\mathrm{i}(\hat{E}_1 E_1 + \hat{E}_2 E_2)} \overline{\exp\left[ \mathrm{i} \left( \hat{E}_1 \sum_{i_1 < \cdots < i_p} J_{i_1 \ldots i_p} \sigma^1_{i_1} \cdots \sigma^1_{i_p} + \hat{E}_2 \sum_{i_1 < \cdots < i_p} J_{i_1 \ldots i_p} \sigma^2_{i_1} \cdots \sigma^2_{i_p} \right) \right]} \\
&= \iint \frac{d\hat{E}_1 \hat{E}_2}{4\pi^2} e^{\mathrm{i}(\hat{E}_1 E_1 + \hat{E}_2 E_2)} \prod_{i_1 < \cdots < i_p} \overline{\exp\left( \mathrm{i}\,\hat{E}_1 J_{i_1 \ldots i_p} \sigma^1_{i_1} \cdots \sigma^1_{i_p} + \mathrm{i}\,\hat{E}_2 J_{i_1 \ldots i_p} \sigma^2_{i_1} \cdots \sigma^2_{i_p} \right)},
\end{aligned}
\tag{7.6}
$$

where we have defined the overlap between two configurations as $q = \frac{1}{N} \sum_i \sigma^1_i \sigma^2_i$. To proceed, we must calculate the disorder average in the above expression of $P(E_1, E_2, q)$. The disorder average is carried out as follows:

$$\begin{aligned}
&\overline{\exp\left[\mathrm{i}\,\hat{E}_1 J_{i_1\ldots i_p}\sigma_{i_1}^1\cdots\sigma_{i_p}^1+\mathrm{i}\,\hat{E}_2 J_{i_1\ldots i_p}\sigma_{i_1}^2\cdots\sigma_{i_p}^2\right]}\\
&=\int dJ_{i_1\ldots i_p}P(J_{i_1\ldots i_p})\exp\left(\mathrm{i}\,\hat{E}_1 J_{i_1\ldots i_p}\sigma_{i_1}^1\cdots\sigma_{i_p}^1+\mathrm{i}\,\hat{E}_2 J_{i_1\ldots i_p}\sigma_{i_1}^2\cdots\sigma_{i_p}^2\right)\\
&=\exp\left[(\mathrm{i}\,\hat{E}_1)^2+(\mathrm{i}\,\hat{E}_2)^2+\frac{2J^2p!(\mathrm{i}\,\hat{E}_1)(\mathrm{i}\,\hat{E}_2)}{4N^{p-1}}(\sigma_{i_1}^1\cdots\sigma_{i_p}^1\sigma_{i_1}^2\cdots\sigma_{i_p}^2)\right],
\end{aligned}\tag{7.7}$$

where we have used the fact that spin takes a binary value $\pm1$. Inserting the disorder average into Eq. (7.6), we obtain

$$\begin{aligned}
P(E_1,E_2,q)&=\iint\frac{d\hat{E}_1\hat{E}_2}{4\pi^2}e^{\mathrm{i}(\hat{E}_1E_1+\hat{E}_2E_2)}\\
&\quad\times\prod_{i_1<\cdots<i_p}\exp\left[(\mathrm{i}\,\hat{E}_1)^2+(\mathrm{i}\,\hat{E}_2)^2+\frac{2J^2p!(\mathrm{i}\,\hat{E}_1)(\mathrm{i}\,\hat{E}_2)}{4N^{p-1}}(\sigma_{i_1}^1\cdots\sigma_{i_p}^1\sigma_{i_1}^2\cdots\sigma_{i_p}^2)\right]\\
&=\iint\frac{d\hat{E}_1\hat{E}_2}{4\pi^2}e^{\mathrm{i}(\hat{E}_1E_1+\hat{E}_2E_2)}\exp\left[\frac{J^2N}{4}\left((\mathrm{i}\,\hat{E}_1)^2+(\mathrm{i}\,\hat{E}_2)^2+2q^p(\mathrm{i}\,\hat{E}_1)(\mathrm{i}\,\hat{E}_2)\right)\right].
\end{aligned}\tag{7.8}$$

To arrive at the last equality, we have used the relationship $p!\sum_{i_1<i_2<\cdots<i_p}^{N}\bullet\simeq\sum_{i_1,i_2,\ldots,i_p}\bullet$ for large $N$, together with the definition of the overlap $q$. Finally, calculating the Gaussian integral out in Eq. (7.8), we conclude that the joint distribution parameterized by $q$ and $J$ is given by

$$\begin{aligned}
P(E_1,E_2,q)&=\frac{1}{\pi J^2N\sqrt{1-q^{2p}}}\exp\left[\frac{2E_1E_2q^p-E_1^2-E_2^2}{J^2N(1-q^{2p})}\right]\\
&=\left[N\pi J^2(1+q^p)N\pi J^2(1-q^p)\right]^{-1/2}\exp\left[-\frac{(E_1+E_2)^2}{2J^2N(1+q^p)}-\frac{(E_1-E_2)^2}{2J^2N(1-q^p)}\right].
\end{aligned}\tag{7.9}$$

Supposed that $|q|<1$, we immediately have $P(E_1,E_2,q)\xrightarrow{p\to\infty}P(E_1)P(E_2)$, where $P(E_1)$ and $P(E_2)$ are the Gaussian distributions derived before. This implies that the energy levels are uncorrelated, and each of them follows exactly the Gaussian distribution.

## 7.2 Phase Diagram

The above mathematical results draw concise physics pictures of the infinite-body interaction model. We can then easily compute the typical number of configurations with predefined energy level $E$,

$$\langle n(E)\rangle=2^NP(E)=\frac{1}{\sqrt{\pi NJ^2}}e^{N\left(\ln 2-\left(\frac{E}{NJ}\right)^2\right)}.\tag{7.10}$$

One can then derive a critical energy level $E_0 = NJ\sqrt{\ln 2}$, above which (in the absolute value) no configurations exist. However, for $|E| < E_0$, there are exponentially many configurations at the corresponding energy level. In the thermodynamic limit, the entropy density (per spin) below the critical energy level is given by $s(E) = \lim_{N\to\infty} \frac{\ln\langle n(E)\rangle}{N} = \ln 2 - \left(\frac{\epsilon}{J}\right)^2$, where $\epsilon$ denotes the energy density. According to the thermodynamic relationship $\frac{dS}{dE} = \frac{1}{T}$, one can also obtain the expression for the energy level $\epsilon = -\frac{J^2}{2T}$, which also determines the critical temperature $T_c = \frac{J}{2\sqrt{\ln 2}}$ where the entropy vanishes.

Finally, the equilibrium property of the random energy model is summarized by the free energy profile ($F = E - TS$):

$$F/N = \begin{cases} -T\ln 2 - \frac{J^2}{4T} & T > T_c \\ -J\sqrt{\ln 2} & T < T_c \end{cases}. \tag{7.11}$$

This implies that below the critical temperature, the free energy of the system does not depend on the temperature, due to the vanishing entropy for a system of discrete degrees of freedom. The vanishing entropy suggests that the system enter a frozen glassy phase—the transition is continuous in the thermodynamic sense (no latent heat). This frozen glassy phase is also discovered in the Gallager codes [5, 6] and binary Perceptron [7–9]. We finally remark that the one-step replica symmetry breaking (see Chap. 9) was confirmed to be exact for the random energy model [3].

## References

1. B. Derrida, Phys. Rev. Lett. **45**, 79 (1980)
2. B. Derrida, Phys. Rev. B **24**(5), 2613 (1981)
3. D. Gross, M. Mezard, Nuclear Phys. **240**(4), 431 (1984)
4. E. Gardner, J. Phys. A **20**(11), 3453 (1987)
5. A. Montanari, Eur. Phys. J. B **23**(1), 121 (2001)
6. H. Huang, Commun. Theor. Phys. **63**(1), 115 (2015)
7. W. Krauth, M. Mezard, J. De Phys. **50**(20), 3057 (1989)
8. H. Huang, Y. Kabashima, Phys. Rev. E **90**, 052813 (2014)
9. H. Huang, K.Y.M. Wong, Y. Kabashima, J. Phys. A: Math. Theor. **46**, 375002 (2013)

# Chapter 8
# Statistical Mechanical Theory of Hopfield Model


Hopfield model is a well-known abstract model of associative memory in the brain (Amari in Biolog. cybern. 26:175, 1977 [1]; Hopfield in Proc. Natl. Acad. Sci. USA 79:2554, 1982 [2]). Its equilibrium properties were first analyzed in the seminal paper (Amit et al. in Phys. Rev. Lett. 55(14):1530, 1985 [3]) by Amit, Gutfreund and Sompolinsky. To obtain the phase diagram, the replica method developed originally in spin glass theory was used and then became popular in neural network research. This work also opened a new discipline—computational/theoretical neuroscience, being an important branch of worldwide brain projects in this new century. In this chapter, we will introduce in detail physics of this model, including phase transitions in associative memory, by an in-depth application of the replica trick (Mézard et al. in Spin Glass Theory and Beyond. World Scientific, Singapore, 1987 [4]).

## 8.1 Hopfield Model

In the Hopfield network, all neurons are connected with each other by real-valued weights (see Fig. 8.1). Randomly generated patterns can be stored in this network by assigning the weights $w_{ij}$ in a Hebbian way (i.e., cells that fire together, wire together). After assigning all the weights, if one feeds a distorted pattern to the network, the network dynamics can converge to the correct undistorted pattern by locally updating the neural state.

In the Hopfield model, the state of neuron $i$ at time step $t$ takes binary values ($\pm 1$)

$$S_i(t) = \begin{cases} -1 & \text{inactive} \\ 1 & \text{active} \end{cases}. \tag{8.1}$$

The update rule takes the form



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_8

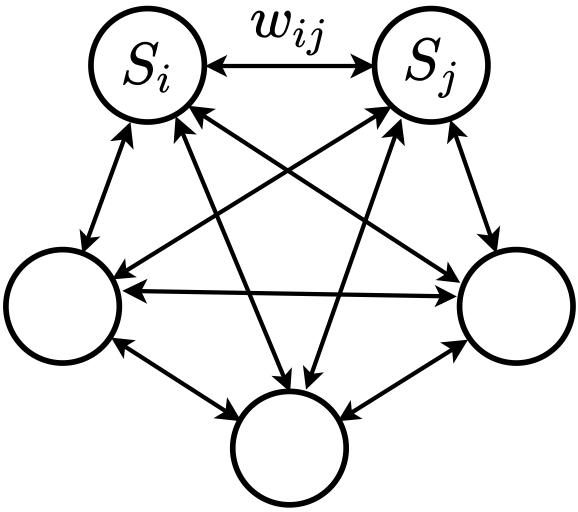



**Fig. 8.1** Typical structure of a Hopfield network. The circles represent neurons, and the lines with arrows represent symmetric weights between two neurons ($w_{ij} = w_{ji}$). Every neuron is connected to all other neurons

$$S_i(t+1) \leftarrow \mathrm{sgn}\left(\sum_j w_{ij} S_j(t) - \theta_i\right), \tag{8.2}$$

where $\mathrm{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$, and $\theta_i$ is the firing bias of the neuron $S_i$. In fact, this rule is a zero-temperature Monte Carlo dynamics of the model.

Now we need to choose the right weights $\{w_{ij}\}$ to ensure that the binary patterns $\{\boldsymbol{\xi}^{(\mu)}\}$ are attractors. If one feeds an input $\mathbf{S}(t=0)$ close to one of stored patterns (say $\boldsymbol{\xi}^{(\nu)}$) to the network, the network is expected to converge to $\boldsymbol{\xi}^{(\nu)}$.

We consider a simple setting for the network, namely storing just one pattern, say $\boldsymbol{\xi}^1$. We can choose the weights according to the following Hebbian rule:

$$w_{ij} = \frac{1}{N} \xi_i^{(1)} \xi_j^{(1)}, \tag{8.3}$$

for $i \neq j$, and $\theta_i = 0$. Usually we set $w_{ii} = 0$ for all $i$. To check this rule, we feed the pattern $\boldsymbol{\xi}^{(1)}$ to the network

$$\sum_{j=1}^{N} w_{ij} \xi_j^{(1)} = \frac{1}{N} \sum_{j=1}^{N} \xi_i^{(1)} \xi_j^{(1)} \xi_j^{(1)} = \frac{1}{N} \sum_{j=1}^{N} \xi_i^{(1)} = \xi_i^{(1)}. \tag{8.4}$$

Therefore

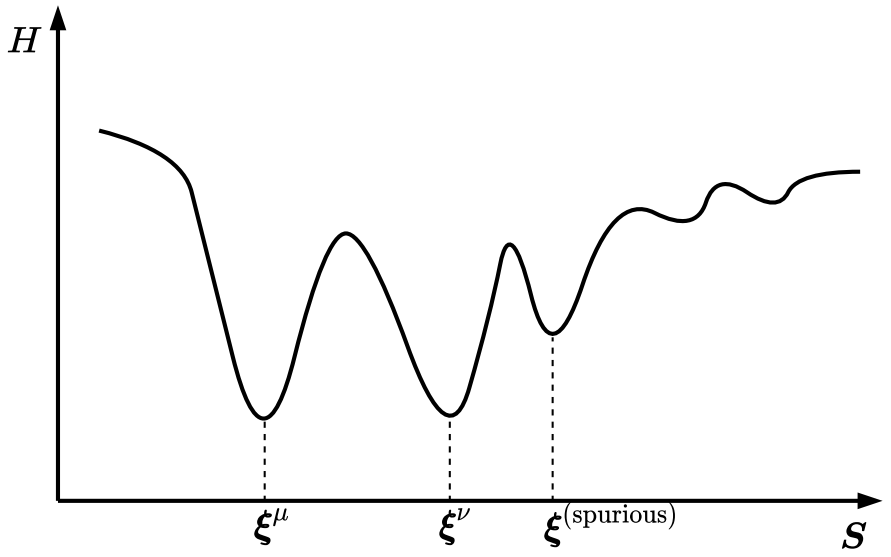



**Fig. 8.2** The energy landscape of the Hopfield network. Minima in the energy function are attractors in the state space. But not every attractor corresponds to a stored pattern. These metastable states are referred to as spurious memories (e.g., a linear combination of several stored patterns [5])

$$\operatorname{sgn}\left(\sum_{j=1}^{N} w_{ij}\xi_j^{(1)}\right) = \xi_i^{(1)} \quad \Rightarrow \quad \mathbf{S}(t>0) = \boldsymbol{\xi}^{(1)}. \tag{8.5}$$

If we feed the reversed pattern $-\boldsymbol{\xi}^{(1)}$ to the network

$$\operatorname{sgn}\left(-\sum_{j=1}^{N} w_{ij}\xi_j^{(1)}\right) = -\xi_i^{(1)} \quad \Rightarrow \quad \mathbf{S}(t>0) = -\boldsymbol{\xi}^{(1)}\,. \tag{8.6}$$

Therefore, if $\boldsymbol{\xi}^{(1)}$ is an attractor, then $-\boldsymbol{\xi}^{(1)}$ is an attractor as well. This is a general property of the Hopfield model, as we shall show by writing down the Hamiltonian.

In equilibrium statistical physics, the Hamiltonian (the energy function) is defined as

$$H = -\frac{1}{2}\sum_{i,j}^{N} w_{ij} S_i S_j, \tag{8.7}$$

where $w_{ij} = 1/N \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu$, which is symmetric, ensuring that an equilibrium state exists. Note that the pattern entries are independently selected as $P(\xi_i^\mu = \pm 1) = 1/2$. Under the zero-temperature dynamics of the model, the Hamiltonian $H$ remains unchanged or decrease. To show this, neglecting the firing bias, we consider the

update

$$S'_k = \mathrm{sgn}\left(\sum_j w_{kj} S_j\right), \tag{8.8}$$

and thus either $S'_k = S_k$ or $S'_k = -S_k$. In the first case, $H$ remains unchanged. In the other case,

$$H' - H = \sum_{j(\neq k)} w_{kj} S_k S_j + \sum_{i(\neq k)} w_{ik} S_i S_k = 2 \sum_{j(\neq k)} w_{kj} S_k S_j. \tag{8.9}$$

Because the sign of $\sum_j w_{kj} S_j$ is the same as $S'_k$ and $S'_k = -S_k$, it then follows that

$$H' - H < 0. \tag{8.10}$$

Hence, either $H$ remains unchanged or its value decreases in one update step. After a sufficient number of updates, the energy function falls into a certain minimum, which is expected to correspond to a stored pattern (Fig. 8.2). This derivation can be cross-checked by implementing a zero-temperature Monte Carlo sampling on the Hamiltonian of Hopfield model.

## 8.2 Replica Method

In the thermodynamic limit, the free energy has the self-averaging property, i.e., $-\beta f = \langle \ln Z \rangle$, where $Z$ is the partition function. As the number of degrees of freedom grows, the single-sample value of the free energy will converge sharply to the quenched average value. However, the expression $\langle \ln Z \rangle$, namely the quenched average, is difficult to calculate in a direct way, whereas $\langle Z \rangle$, namely the annealed average, is much easier to calculate. However, in most contexts of interest, $\langle \ln Z \rangle \neq \ln \langle Z \rangle$. In fact, the annealed average provides an upper bound to the quenched average, due to the Jensen's inequality. The replica trick can be used to make a transformation of this calculation by introducing many copies of the original systems. Then the original interaction system can be decoupled to an equivalent system where correlations among replicas are considered, which greatly simplifies the original challenging computation.

In mathematics, we have

$$\ln Z = \lim_{n \to 0} \frac{Z^n - 1}{n}. \tag{8.11}$$

Then we calculate the expectation

$$\langle \ln Z \rangle = \lim_{n\to 0} \frac{\langle Z^n \rangle - 1}{n} = \lim_{n\to 0} \frac{\ln \langle Z^n \rangle}{n}, \tag{8.12}$$

where $\langle \cdot \rangle$ is the disorder average over $\boldsymbol{\xi}$. Since $Z^n \simeq 1 + n \ln Z + \cdots$, we have $\langle Z^n \rangle \simeq 1 + n \langle \ln Z \rangle \cdots$. Therefore

$$\lim_{n\to 0} \frac{\ln \langle Z^n \rangle}{n} = \lim_{n\to 0} \frac{\ln(1 + n\langle \ln Z \rangle)}{n} = \lim_{n\to 0} \frac{n\langle \ln Z \rangle}{n} = \langle \ln Z \rangle \,, \tag{8.13}$$

where when $n$ is small enough, we can take the expansion like $Z^n = e^{n \ln Z} = 1 + n \ln Z + \cdots$. The averaged free energy per spin can thus be calculated by

$$f = \lim_{n\to 0} \lim_{N\to\infty} \frac{-\ln \langle Z^n \rangle}{\beta n N}. \tag{8.14}$$

We first assume that $n$ is an integer (for the power), and after the calculation of $\langle Z^n \rangle$, we carry out the limit of $\langle \ln Z \rangle$ as $n$ approaches 0. This seems hard to understand in physics; whereas the results must be compared with physics simulations of the model. In this sense, the cavity approximation is more physically transparent than the replica trick, although in most (we are not sure if all is suitable) cases, both methods yields the same result. We remark that the order of the two limits ($n \to 0$ and $N \to \infty$) has been exchanged for the purpose of applying the Laplace method in the thermodynamic limit. This operation is also not mathematically rigorous. But the final result is usually in consistent with physics intuition and numerical simulations.

Next, we suppose that the network is able to store $P$ random patterns ($P = \alpha N$, and $\alpha$ denotes the memory load). Note that $H = -\frac{1}{2}\sum_{i,j}^{N} w_{ij} S_i S_j$ and $w_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu$. Therefore

$$\begin{aligned}
\langle Z^n \rangle &= \left\langle \mathrm{Tr} \exp\left[ \frac{\beta}{2N} \sum_{i,j}^{N} \sum_{\mu=1}^{P} \sum_{\rho=1}^{n} \xi_i^\mu \xi_j^\mu S_i^\rho S_j^\rho \right] \right\rangle \\
&= \left\langle \mathrm{Tr} \exp\left[ \frac{\beta}{2N} \sum_{\rho,\mu} \left( \sum_i \xi_i^\mu S_i^\rho \right) \left( \sum_j \xi_j^\mu S_j^\rho \right) \right] \right\rangle \\
&= \left\langle \mathrm{Tr} \exp\left[ \frac{\beta N}{2} \sum_{\rho,\mu} \left( \frac{1}{N} \sum_i \xi_i^\mu S_i^\rho \right)^2 \right] \right\rangle \\
&= \left\langle \mathrm{Tr} \prod_{\rho,\mu} \exp\left[ \frac{\beta N}{2} \left( \frac{1}{N} \sum_i \xi_i^\mu S_i^\rho \right)^2 \right] \right\rangle ,
\end{aligned} \tag{8.15}$$

where Tr means the summation over all configurations $\{\mathbf{S}\}$, and $\langle \cdot \rangle$ means the quenched disorder average over the random patterns.

To linearize the quadratic term, we apply the following Gaussian integral:

$$e^{ab^2} = \sqrt{\frac{a}{\pi}} \int e^{-ax^2+2abx} dx, \tag{8.16}$$

by carrying out the following substitutions:

$$\begin{cases} b \to \frac{1}{N} \sum_i \xi_i^\mu S_i^\rho \\ x \to m_\rho^\mu \\ a \to \frac{\beta N}{2} \end{cases} . \tag{8.17}$$

It is then natural to introduce integrals over $m_\rho^\mu$

$$\begin{aligned} \langle Z^n \rangle &= \left\langle \mathrm{Tr} \int \prod_{\rho,\mu} \sqrt{\frac{\beta N}{2\pi}} dm_\rho^\mu \exp\left[ -\frac{\beta N}{2} \left(m_\rho^\mu\right)^2 + \beta m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho \right] \right\rangle \\ &= \left\langle \mathrm{Tr} \int \prod_{\rho,\mu} \sqrt{\frac{\beta N}{2\pi}} dm_\rho^\mu \exp\left[ -\frac{\beta N}{2} \sum_{\rho,\mu} \left(m_\rho^\mu\right)^2 + \beta \sum_{\rho,\mu} m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho \right] \right\rangle \\ &= \left\langle \mathrm{Tr} \int \prod_{\rho,\mu} \sqrt{\frac{\beta N}{2\pi}} dm_\rho^\mu \exp\left[ -\frac{\beta N}{2} \sum_{\mu\geq 2} \sum_\rho \left(m_\rho^\mu\right)^2 + \right. \right. \\ &\left. \left. \beta \sum_{\mu\geq 2} \sum_\rho m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho \; - \frac{\beta N}{2} \sum_\rho \left(m_\rho^1\right)^2 + \beta \sum_\rho m_\rho^1 \sum_i \xi_i^1 S_i^\rho \right] \right\rangle . \end{aligned} \tag{8.18}$$

In the above equation, we have separated the first pattern from other patterns. We further assume that only the first pattern ($\mu = 1$) is retrieved, and thus the overlap $m_\rho^\mu \sim O(1)$ (the definition of the overlap will become clear in the following analysis). We next consider those non-retrieved patterns ($\mu \geq 2$). Because $\langle \sum_i \xi_i^\mu S_i^\rho \rangle_{\boldsymbol{\xi}} = 0$ and $\left\langle \left(\sum_i \xi_i^\mu S_i^\rho\right)^2 \right\rangle_{\boldsymbol{\xi}} = N + \langle \sum_{i\neq j} \xi_i^\mu \xi_j^\mu S_i^\rho S_j^\rho \rangle_{\boldsymbol{\xi}} = N$, the order of $m_\rho^\mu$ ($\mu \geq 2$) is given by

$$m_\rho^\mu = \frac{1}{N} \sum_i \xi_i^\mu S_i^\rho \approx O\left(\frac{1}{\sqrt{N}}\right). \tag{8.19}$$

To use an $m_\rho^\mu$ of $O(1)$, we rescale the original $m_\rho^\mu \to \frac{m_\rho^\mu}{\sqrt{\beta N}}$. Then we get

$$\left\langle Z^n \right\rangle = \left(\frac{1}{\sqrt{2\pi}}\right)^{n(P-1)} \left\langle \mathrm{Tr} \int \prod_{\rho,\mu>1} dm_\rho^\mu \prod_\rho \sqrt{\beta N} dm_\rho^1 \exp\left[ -\frac{1}{2} \sum_{\mu\geq 2} \sum_\rho \left(m_\rho^\mu\right)^2 + \sqrt{\frac{\beta}{N}} \sum_{\mu\geq 2} \sum_\rho m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho - \frac{\beta N}{2} \sum_\rho \left(m_\rho^1\right)^2 + \beta \sum_\rho m_\rho^1 \sum_i \xi_i^1 S_i^\rho \right] \right\rangle . \tag{8.20}$$

For the part of $\mu \geq 2$ involving in non-condensed patterns, we have

$$\begin{aligned} &\left\langle \exp\left[ \sqrt{\frac{\beta}{N}} \sum_{\mu\geq 2} \sum_\rho m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho \right] \right\rangle_{\xi_i^\mu:\mu>1} \\ &\propto \exp\left[ \sum_{\mu\geq 2,i} \ln\cosh\left( \sqrt{\frac{\beta}{N}} \sum_\rho m_\rho^\mu S_i^\rho \right) \right] \\ &\cong \exp\left[ \sum_{\mu\geq 2} \sum_i \frac{\beta}{2N} \left( \sum_\rho m_\rho^\mu S_i^\rho \right)^2 \right] , \end{aligned} \tag{8.21}$$

where we have used the formula $\langle \exp(A\xi) \rangle_{\{\xi=\pm 1\}} = \exp(-A) + \exp(A) = 2\cosh(A) \propto \exp(\ln\cosh(A))$, and taken the approximation $\ln\cosh x = \frac{x^2}{2} + \cdots$ as $x \to 0$.

We can then write down the following expressions:

$$\sum_\rho \left(m_\rho^\mu\right)^2 = \sum_{\rho,\sigma} m_\rho^\mu \delta_{\rho\sigma} m_\sigma^\mu , \tag{8.22}$$

and

$$\begin{aligned} \frac{1}{N} \sum_i \left( \sum_\rho m_\rho^\mu S_i^\rho \right)^2 &= \frac{1}{N} \sum_i \sum_\rho m_\rho^\mu S_i^\rho \sum_\sigma m_\sigma^\mu S_i^\sigma \\ &= \sum_{\rho,\sigma} m_\rho^\mu \frac{1}{N} \sum_i S_i^\rho S_i^\sigma m_\sigma^\mu \\ &:= \sum_{\rho,\sigma} m_\rho^\mu q_{\rho\sigma} m_\sigma^\mu . \end{aligned} \tag{8.23}$$

To further simplify the formulas, we define

$$\kappa_{\rho\sigma} = \delta_{\rho\sigma} - \frac{\beta}{N} \sum_i S_i^\rho S_i^\sigma := \delta_{\rho\sigma} - \beta q_{\rho\sigma}, \tag{8.24}$$

and in the matrix form

$$\mathbf{K} = \mathbf{I} - \beta \mathbf{Q}, \tag{8.25}$$

where

$$q_{\rho\sigma} = \begin{cases} \frac{1}{N} \sum_i S_i^\rho S_i^\sigma & \rho \neq \sigma \\ 1 & \rho = \sigma \end{cases}, \tag{8.26}$$

and $\mathbf{K}$, $\mathbf{Q}$ are symmetric $n \times n$ matrices with elements $\kappa_{\rho\sigma}$ and $q_{\rho\sigma}$, respectively. $\mathbf{I}$ is an identity matrix.

Thus, we need to introduce $q_{\rho\sigma}$ by an integral of a Dirac delta function, and obtain

$$\begin{aligned} \langle Z^n \rangle \propto \mathrm{Tr} \int \prod_{\rho,\sigma} dq_{\rho\sigma} \delta \left( q_{\rho\sigma} - \frac{1}{N} \sum_i S_i^\rho S_i^\sigma \right) \\ \times \prod_{\mu \geq 2, \rho} dm_\rho^\mu \exp \left[ -\frac{1}{2} \sum_{\mu \geq 2} \sum_{\rho,\sigma} m_\rho^\mu \kappa_{\rho\sigma} m_\sigma^\mu \right] \\ \times \left\langle \int \prod_\rho dm_\rho^1 \exp \left[ -\frac{\beta N}{2} \sum_\rho \left( m_\rho^1 \right)^2 + \frac{\beta N}{N} \sum_\rho m_\rho^1 \sum_i \xi_i^1 S_i^\rho \right] \right\rangle_{\xi^1}, \end{aligned} \tag{8.27}$$

where we have neglected irrelevant prefactors. By using the multivariate Gaussian integral

$$\int_{R^n} d\mathbf{m} e^{-\mathbf{M}^\mathrm{T} \mathbf{K} \mathbf{M}} = \sqrt{\frac{\pi^n}{\det(\mathbf{K})}}, \tag{8.28}$$

we get

$$\int \prod_{\mu \geq 2, \rho} dm_\rho^\mu \exp \left[ -\frac{1}{2} \sum_{\mu \geq 2} \sum_{\rho,\sigma} m_\rho^\mu \kappa_{\rho\sigma} m_\sigma^\mu \right] = \frac{C}{(\det \mathbf{K})^{\frac{P-1}{2}}}, \tag{8.29}$$

where $C$ is a constant. Because $\det(e^{\mathbf{K}}) = e^{\mathrm{Tr}\,\mathbf{K}}$, and $\det \mathbf{K} = e^{\mathrm{Tr} \ln \mathbf{K}}$, we have

$$(\det \mathbf{K})^{-\frac{P-1}{2}} = e^{-\frac{P-1}{2} \mathrm{Tr} \ln \mathbf{K}} = e^{-\frac{P-1}{2} \mathrm{Tr} \ln [\mathbf{I} - \beta \mathbf{Q}]}. \tag{8.30}$$

By using the Fourier representation of the Dirac delta function

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-\mathrm{i}kx} dk, \tag{8.31}$$

we obtain

$$\begin{aligned}
&\mathrm{Tr}\int\prod_{\rho}dm_{\rho}^{1}\prod_{\rho,\sigma}dq_{\rho\sigma}\,\delta\left(q_{\rho\sigma}-\frac{1}{N}\sum_{i}S_{i}^{\rho}S_{i}^{\sigma}\right)\cdot\mathcal{I}\\
&\propto\mathrm{Tr}\int\prod_{\rho}dm_{\rho}^{1}\prod_{\rho,\sigma}dq_{\rho\sigma}dr_{\rho\sigma}\exp\left[-\frac{N\alpha\beta^{2}}{2}\sum_{\rho,\sigma}r_{\rho\sigma}q_{\rho\sigma}+\frac{\alpha\beta^{2}}{2}\sum_{i,\rho,\sigma}r_{\rho\sigma}S_{i}^{\rho}S_{i}^{\sigma}\right]\cdot\mathcal{I}\,,
\end{aligned}\tag{8.32}$$

where the symbol $\mathcal{I}$ represents the other non-shown parts in Eq. (8.27), and we have rescaled $r_{\rho\sigma}\to-\frac{\mathrm{i}N\alpha\beta^{2}}{2}r_{\rho\sigma}$ (after the transformation, $r_{\rho\sigma}\sim O(1)$), and used $\alpha=P/N$.

Then we define the $S_i$-dependent part as

$$\begin{aligned}
&\left\langle\mathrm{Tr}\exp\left[\beta\sum_{\rho}m_{\rho}^{1}\sum_{i}\xi_{i}^{1}S_{i}^{\rho}+\frac{\alpha\beta^{2}}{2}\sum_{i,\rho,\sigma}r_{\rho\sigma}S_{i}^{\rho}S_{i}^{\sigma}\right]\right\rangle_{\boldsymbol{\xi}^{1}}\\
&=\left\langle\exp\left\{\sum_{i}\ln\mathrm{Tr}\exp\left(\beta\sum_{\rho}m_{\rho}^{1}\xi_{i}^{1}S^{\rho}+\frac{\alpha\beta^{2}}{2}\sum_{\rho,\sigma}r_{\rho\sigma}S^{\rho}S^{\sigma}\right)\right\}\right\rangle_{\boldsymbol{\xi}^{1}}\\
&=\exp\left\{N\left\langle\ln\mathrm{Tr}\exp\left(\beta\sum_{\rho}m_{\rho}^{1}\xi^{1}S^{\rho}+\frac{\alpha\beta^{2}}{2}\sum_{\rho,\sigma}r_{\rho\sigma}S^{\rho}S^{\sigma}\right)\right\rangle_{\xi^{1}}\right\}\\
&:=\exp\left\{N\left\langle\ln\mathrm{Tr}\exp(\beta H_{\xi^{1}})\right\rangle_{\xi^{1}}\right\}\,,
\end{aligned}\tag{8.33}$$

where we have used the fact that the sum over $i$ is equivalent to taking the average over the pattern configuration because of i.i.d properties of the random pattern, and we have defined

$$\beta H_{\xi^{1}}=\frac{1}{2}\alpha\beta^{2}\sum_{\rho,\sigma}r_{\rho\sigma}S^{\rho}S^{\sigma}+\beta\sum_{\rho}m_{\rho}^{1}\xi^{1}S^{\rho}\,,\tag{8.34}$$

where $\xi^1$ is just a typical entry of the random pattern vector.

Finally, we obtain

$$\begin{aligned}
\left\langle Z^{n}\right\rangle\propto&\int\prod_{\rho}dm_{\rho}^{1}\prod_{\rho,\sigma}dq_{\rho\sigma}dr_{\rho\sigma}\exp\left[-\frac{N}{2}\alpha\beta^{2}\sum_{\rho,\sigma}r_{\rho\sigma}q_{\rho\sigma}\right]\\
&\times\exp\left[-\frac{P-1}{2}\mathrm{Tr}\ln[\mathbf{I}-\beta\mathbf{Q}]\right]\exp\left[-\frac{\beta N}{2}\sum_{\rho}\left(m_{\rho}^{1}\right)^{2}+N\left\langle\ln\mathrm{Tr}\,e^{\beta H_{\xi^{1}}}\right\rangle_{\xi^{1}}\right]\,.
\end{aligned}\tag{8.35}$$

Because we assume that $N$ is large enough, we can use the Laplace's method, which is

$$\int_{a}^{b}e^{Nf(z)}dz\approx\sqrt{\frac{2\pi}{-Nf''(z_{0})}}e^{Nf(z_{0})}.\tag{8.36}$$

where $z_0$ is the maximum point. Thus, we can perform the approximation $\langle Z^n \rangle \sim e^{NF(\theta^*)}$, where $F(\theta^*) = \max_\theta F(\theta)$. Here, we use $\theta$ to indicate the order parameter set of the model.

Then the quenched disorder averaged free energy becomes

$$\langle \ln Z \rangle = \lim_{n\to 0} \frac{\ln \langle Z^n \rangle}{n} = \lim_{n\to 0} \frac{\ln e^{NF(\theta^*)}}{n} = N \lim_{n\to 0} \frac{F(\theta^*)}{n} , \tag{8.37}$$

where

$$\begin{aligned} F\left(r_{\rho\sigma}, q_{\rho\sigma}, m_\rho^1\right) = &-\frac{\alpha\beta^2}{2} \sum_{\rho,\sigma} r_{\rho\sigma} q_{\rho\sigma} - \frac{\alpha}{2} \operatorname{Tr} \ln[\mathbf{I} - \beta\mathbf{Q}] \\ &- \frac{\beta}{2} \sum_\rho \left(m_\rho^1\right)^2 + \left\langle \ln \operatorname{Tr} e^{\beta H_{\xi^1}} \right\rangle_{\xi^1} . \end{aligned} \tag{8.38}$$

We have taken the approximation $P - 1 \simeq P = \alpha N$ as $N$ is large enough. Note that $(r_{\rho\sigma}, q_{\rho\sigma}, m_\rho^1)$ is the order parameter set of the model. Their physical meanings will be clear in the following analysis.

To calculate the maximum of $F\left(r_{\rho\sigma}, q_{\rho\sigma}, m_\rho^1\right)$, we first calculate the derivatives of $F\left(r_{\rho\sigma}, q_{\rho\sigma}, m_\rho^1\right)$ with respect to the order parameters.

First, we take a derivative with respect to $q_{\rho\sigma}$,

$$\frac{\partial F}{\partial q_{\rho\sigma}} = 0 \Rightarrow \frac{\partial}{\partial q_{\rho\sigma}} \left[ -\frac{N\alpha\beta^2}{2} \sum_{\rho,\sigma} r_{\rho\sigma} q_{\rho\sigma} + \frac{\beta}{2} (\sqrt{\beta N})^2 \sum_{\mu\geq 2} \sum_{\rho,\sigma} m_\rho^\mu q_{\rho\sigma} m_\sigma^\mu \right] = 0 , \tag{8.39}$$

where the second term inside the bracket comes from the original formula [Eq. (8.27)] in which the integral over $\{m_\rho^\mu\}$ is kept. Note that the magnetization is rescaled back. We then obtain the conjugated order parameter

$$r_{\rho\sigma} = \frac{1}{\alpha} \sum_{\mu\geq 2} m_\rho^\mu m_\sigma^\mu , \tag{8.40}$$

where we need to use the rescaling $m_\rho^\mu \to \frac{m_\rho^\mu}{\sqrt{\beta N}}$ that is done before. $r_{\rho\sigma}$ is thus understood as the sum of effects of non-condensed patterns (only one retrieved pattern here).

Second, we take a derivative with respect to $m_\rho^\mu$ [see the original formula Eq. (8.20)]

$$\frac{\partial F}{\partial m_\rho^\mu} = 0 \Rightarrow \frac{\partial}{\partial m_\rho^\mu}\left[-\frac{\beta N}{2}\left(m_\rho^\mu\right)^2 + \beta m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho\right] = 0\,, \tag{8.41}$$

and obtain

$$m_\rho^\mu = \frac{1}{N}\sum_i \xi_i^\mu S_i^\rho\,. \tag{8.42}$$

The parameter $m_\rho^\mu$ is exactly the overlap between the state of the system and the $\mu$th pattern, characterizing the quality of memory retrieval.

Finally, from the requirement of a stationary free energy [see Eq. (8.32)]

$$\frac{\partial F}{\partial r_{\rho\sigma}} = 0 \Rightarrow \frac{\partial}{\partial r_{\rho\sigma}}\left[-\frac{N\alpha\beta^2}{2}\sum_{\rho,\sigma} r_{\rho\sigma}q_{\rho\sigma} + \frac{\alpha\beta^2}{2}\sum_{i,\rho,\sigma} r_{\rho\sigma}S_i^\rho S_i^\sigma\right] = 0\,, \tag{8.43}$$

we obtain the Edwards–Anderson order parameter

$$q_{\rho\sigma} = \frac{1}{N}\sum_i S_i^\rho S_i^\sigma\,. \tag{8.44}$$

$q_{\rho\sigma}$ is understood as the mutual overlap of two pure states in general. If a single state dominates the phase space, the Edwards–Anderson order parameter characterizes the size of that state.

### 8.2.1 Replica-Symmetric Ansätz

To proceed, we need to make an approximation about the overlap matrix, i.e., considering the simplest form—the overlap is invariant under permutation of replica indexes. This is called the replica symmetry (RS) ansätz

$$\begin{cases} r_{\rho\sigma} = r, & \forall \rho, \sigma \\ m_\rho^1 = m, & \forall \rho \\ q_{\rho\sigma} = q, & \forall \rho \neq \sigma \end{cases}\,. \tag{8.45}$$

Then we have

$$\begin{aligned} F(r,q,m) = &-\frac{\alpha\beta^2}{2}rq\left(n^2-n\right) - \frac{\alpha\beta^2}{2}nr - \frac{\alpha}{2}\operatorname{Tr}\ln[\mathbf{I}-\beta\mathbf{Q}] \\ &-\frac{\beta}{2}nm^2 + \left\langle \ln \operatorname{Tr} e^{\beta H_{\xi^1}}\right\rangle\,, \end{aligned} \tag{8.46}$$

and

$$\begin{aligned}\langle \ln Z \rangle = & \frac{N\alpha\beta^2 rq}{2} - \frac{N\alpha\beta^2 r}{2} - \frac{\alpha N}{2} \lim_{n\to 0} \frac{\mathrm{Tr}\ln[\mathbf{I} - \beta\mathbf{Q}]}{n} - \frac{\beta N m^2}{2} \\ & + N \lim_{n\to 0} \frac{\left\langle \ln \mathrm{Tr}\, e^{\beta H_{\xi^1}} \right\rangle}{n} ,\end{aligned} \tag{8.47}$$

where

$$\beta H_{\xi^1} = \beta m \xi^1 \sum_\rho S^\rho + \frac{1}{2}\alpha\beta^2 r \sum_{\rho,\sigma} S^\rho S^\sigma \ . \tag{8.48}$$

First, we calculate the last term of $\langle \ln Z \rangle$.

$$\begin{aligned}\mathrm{Tr}\, e^{\beta H_{\xi^1}} &= \mathrm{Tr}\, e^{\beta m \xi^1 \Sigma_\rho S^\rho + \frac{1}{2}\alpha\beta^2 r \left(\sum_\rho S^\rho\right)^2} \\ &:= \mathrm{Tr}\, e^{A\left(\Sigma_\rho S^\rho\right)^2 + B\Sigma_\rho S^\rho} \\ &= \mathrm{Tr}\sqrt{\frac{A}{\pi}} \int dz\, e^{-Az^2 + 2Az\sum_\rho S^\rho + B\Sigma_\rho S^\rho} \\ &= \sqrt{\frac{A}{\pi}} \int dz\, e^{-Az^2}\, \mathrm{Tr} \prod_\rho e^{(2Az+B)S^\rho} \\ &= \sqrt{\frac{\alpha\beta^2 r}{2\pi}} \int dz\, e^{-\frac{1}{2}\alpha\beta^2 r z^2} \left[2\cosh\left(\alpha\beta^2 r z + \beta m \xi^1\right)\right]^n \\ &= \sqrt{\frac{\alpha\beta^2 r}{2\pi}} \int dz\, e^{-\frac{1}{2}\alpha\beta^2 r z^2 + n \ln\left[2\cosh\left(\alpha\beta^2 r z + \beta m \xi^1\right)\right]} \\ &= \sqrt{\frac{1}{2\pi}} \int dz\, e^{-\frac{1}{2}z^2 + n \ln\left[2\cosh\left(\beta\sqrt{\alpha r} z + \beta m \xi^1\right)\right]} \ .\end{aligned} \tag{8.49}$$

Note that $A$ and $B$ are auxiliary variables in intermediate computations. The limit of the above term is clearly given by

$$\lim_{n\to 0} \mathrm{Tr}\, e^{\beta H_{\xi^1}} = \sqrt{\frac{1}{2\pi}} \int dz\, e^{-\frac{1}{2}z^2} = 1 \ . \tag{8.50}$$

Thus, we can obtain the limit by the derivative with respect to $n$

$$
\begin{aligned}
&\lim_{n\to 0}\frac{\left\langle \ln \operatorname{Tr} e^{\beta H_{\xi^1}}\right\rangle}{n} \\
&=\left\langle \lim_{n\to 0}\frac{\frac{d}{dn}\operatorname{Tr} e^{\beta H_{\xi^1}}}{\operatorname{Tr} e^{\beta H_{\xi^1}}}\right\rangle \\
&=\left\langle \sqrt{\frac{1}{2\pi}}\lim_{n\to 0}\frac{d}{dn}\int dz\, e^{-\frac{1}{2}z^2+n\ln\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]}\right\rangle \\
&=\left\langle \sqrt{\frac{1}{2\pi}}\lim_{n\to 0}\int dz\, e^{-\frac{1}{2}z^2}\frac{d}{dn}\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]^n\right\rangle \\
&=\left\langle \sqrt{\frac{1}{2\pi}}\lim_{n\to 0}\int dz\, e^{-\frac{1}{2}z^2}\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]^n\right. \\
&\quad \left.\ln\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]\right\rangle \\
&=\left\langle \sqrt{\frac{1}{2\pi}}\int dz\, e^{-\frac{1}{2}z^2}\lim_{n\to 0}\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]^n\right. \\
&\quad \left.\ln\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]\right\rangle \\
&=\left\langle \sqrt{\frac{1}{2\pi}}\int dz\, e^{-\frac{1}{2}z^2}\ln\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]\right\rangle \\
&=\int Dz\left\langle \ln\left[2\cosh\left(\beta\sqrt{\alpha r}z+\beta m\xi^1\right)\right]\right\rangle .
\end{aligned}
\tag{8.51}
$$

Then we calculate the third term of $\langle \ln Z\rangle$. Since $\mathbf{Q}$ is a symmetric matrix, we can diagonalize this matrix and get

$$
\mathbf{AQA}^{-1}=\Lambda=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n) . \tag{8.52}
$$

We can thus expand $\ln[\mathbf{I}-\beta\mathbf{Q}]$ to a power series with respect to $\mathbf{Q}$ (here we take the formula $\ln(1-x)=-\sum_{n=1}^{\infty}\frac{x^n}{n}$) and obtain

$$
\begin{aligned}
\operatorname{Tr}\ln[\mathbf{I}-\beta\mathbf{Q}] &= \operatorname{Tr}\left\{\mathbf{A}\cdot\ln[\mathbf{I}-\beta\mathbf{Q}]\cdot\mathbf{A}^{-1}\right\} \\
&= -\operatorname{Tr}\left\{\sum_{l=1}^{\infty}\frac{\beta^l\left(\mathbf{AQA}^{-1}\right)^l}{l}\right\} \\
&= -\operatorname{Tr}\left\{\sum_{l=1}^{\infty}\frac{\beta^l(\Lambda)^l}{l}\right\} \\
&= -\sum_{l=1}^{\infty}\frac{\beta^l}{l}\sum_{i=1}^{n}\lambda_i^l=\sum_{i=1}^{n}\ln\left[1-\beta\lambda_i\right] .
\end{aligned}
\tag{8.53}
$$

This result is equivalent to the matrix identity: $\mathrm{Tr}\ln\mathbf{K} = \ln\det\mathbf{K}$ for a positive definite matrix.

Then we calculate the eigenvalues of $\mathbf{Q}$ by

$$
\begin{aligned}
&\begin{vmatrix} 1-\lambda & q & \cdots & q \\ q & 1-\lambda & \cdots & q \\ \vdots & \vdots & & \vdots \\ q & q & \cdots & 1-\lambda \end{vmatrix} \\
&= \begin{vmatrix} 1-\lambda+(n-1)q & 1-\lambda+(n-1)q & \cdots & 1-\lambda+(n-1)q \\ q & 1-\lambda & \cdots & q \\ \vdots & \vdots & & \vdots \\ q & q & \cdots & 1-\lambda \end{vmatrix} \\
&= [1-\lambda+(n-1)q] \begin{vmatrix} 1 & 1 & \cdots & 1 \\ q & 1-\lambda & \cdots & q \\ \vdots & \vdots & & \vdots \\ q & q & \cdots & 1-\lambda \end{vmatrix} \\
&= [1-\lambda+(n-1)q] \begin{vmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1-\lambda-q & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1-\lambda-q \end{vmatrix} \\
&= [1-\lambda+(n-1)q](1-q-\lambda)^{n-1} = 0\,.
\end{aligned}
\tag{8.54}
$$

Thus, $\mathbf{Q}$ have one eigenvalue with the value $(1+(n-1)q)$ and $(n-1)$ eigenvalues with values $(1-q)$. Then the trace turns out to be

$$
\mathrm{Tr}\ln[\mathbf{I}-\beta\mathbf{Q}] = \ln(1-\beta+\beta q-n\beta q) + (n-1)\ln(1-\beta+\beta q)\,, \tag{8.55}
$$

and

$$
\begin{aligned}
\lim_{n\to 0}\frac{\mathrm{Tr}\ln[\mathbf{I}-\beta\mathbf{Q}]}{n} &= \lim_{n\to 0}\left[\frac{\ln(\frac{1-\beta+\beta q-n\beta q}{1-\beta+\beta q})}{n} + \ln(1-\beta+\beta q)\right] \\
&= -\frac{\beta q}{1-\beta+\beta q} + \ln(1-\beta+\beta q)\,,
\end{aligned}
\tag{8.56}
$$

where we calculate the limit by the L′Hospital's rule.

Taken all together, the free energy of the Hopfield model can be written as

$$
\begin{aligned}
-\beta f &= \frac{1}{N}\langle \ln Z\rangle \\
&= \frac{\alpha\beta^2}{2} r(q-1) - \frac{\alpha}{2}\left[\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\right] - \frac{\beta}{2}m^2 \\
&+ \int Dz \left\langle \ln\left[2\cosh\left(\beta\sqrt{\alpha r}z + \beta m \xi^1\right)\right]\right\rangle \ .
\end{aligned} \tag{8.57}
$$

To complete the Laplace method, we finally derive the saddle-point equations for all order parameters in the RS ansätz. More precisely, we take derivatives of the free energy with respect to all the order parameters

$$
\begin{cases}
\frac{\partial(-\beta f)}{\partial r} = 0 \\
\frac{\partial(-\beta f)}{\partial m} = 0 \\
\frac{\partial(-\beta f)}{\partial q} = 0
\end{cases} , \tag{8.58}
$$

and get

$$
\begin{aligned}
q &= -\frac{1}{\beta\sqrt{2\pi\alpha r}} \int dz e^{-\frac{1}{2}z^2} z \left\langle \tanh\left(\beta\sqrt{\alpha r}z + \beta m\xi^1\right)\right\rangle + 1 \\
&= \frac{1}{\beta\sqrt{2\pi\alpha r}} \int dz \frac{de^{-\frac{1}{2}z^2}}{dz} \left\langle \tanh\left(\beta\sqrt{\alpha r}z + \beta m\xi^1\right)\right\rangle + 1 \\
&= \frac{1}{\beta\sqrt{2\pi\alpha r}} e^{-\frac{1}{2}z^2} \left\langle \tanh\left(\beta\sqrt{\alpha r}z + \beta m\xi^1\right)\right\rangle \Big|_{-\infty}^{+\infty} \\
&\quad - \int Dz \left\langle 1 - \tanh^2\left(\beta\sqrt{\alpha r}z + \beta m\xi^1\right)\right\rangle + 1 \\
&= \int Dz \left\langle \tanh^2\left(\beta\sqrt{\alpha r}z + \beta m\xi^1\right)\right\rangle \\
&= \int Dz \tanh^2 \beta(\sqrt{\alpha r}z + m) \ .
\end{aligned} \tag{8.59}
$$

Moreover, $r$ and $m$ can be analogously computed, which leads to the following saddle-point equations for the associative memory model.

$$
q = \int Dz \tanh^2 \beta(\sqrt{\alpha r}z + m) \ , \tag{8.60}
$$

$$
m = \int Dz \langle \xi \tanh \beta(\sqrt{\alpha r}z + m\xi)\rangle = \int Dz \tanh \beta(\sqrt{\alpha r}z + m) \ , \tag{8.61}
$$

$$
r = \frac{q}{(1-\beta+\beta q)^2} \ . \tag{8.62}
$$

Phase transitions can be deduced from an analysis of the behavior of these equations and the corresponding free energy function.

### 8.2.2 Zero-Temperature Limit

Under the replica-symmetric assumption, as $T \to 0$ ($\beta \to \infty$), we have

$$\tanh(\beta x) \to \mathrm{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}, \tag{8.63}$$

Equation (8.61) becomes

$$\begin{aligned} m &= \int Dz \, \mathrm{sign}(\sqrt{\alpha r} z + m) + O(T) \\ &= \mathrm{erf}\left( \frac{m}{\sqrt{2\alpha r}} \right) + O(T) \, . \end{aligned} \tag{8.64}$$

On the other hand, as $\beta \to \infty$

$$\begin{aligned} 1 - q &= \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \left( 1 - \tanh^2 \beta(\sqrt{\alpha r} z + m) \right) \\ &\simeq \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \Big|_{\tanh^2 \beta(\sqrt{\alpha r} z + m) = 0} \int dz \left( 1 - \tanh^2 \beta(\sqrt{\alpha r} z + m) \right) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{m^2}{2\alpha r}} \frac{1}{\beta \sqrt{\alpha r}} \int dz \frac{\partial}{\partial z} \tanh \beta(\sqrt{\alpha r} z + m) \\ &= \frac{2}{\sqrt{2\pi}} \frac{1}{\beta \sqrt{\alpha r}} e^{-\frac{m^2}{2\alpha r}} . \end{aligned} \tag{8.65}$$

Equation (8.60) thus yields $q = 1 - CT$, where

$$C \stackrel{\mathrm{def}}{=} \sqrt{\frac{2}{\pi r \alpha}} \, e^{-\frac{m^2}{2\alpha r}} \, . \tag{8.66}$$

Using these intermediate results, Eq. (8.62) becomes $r = (1 - C)^{-2}$.

The equations of $m$ and $r$ can be reduced to one equation, by defining an auxiliary variable $y = m/\sqrt{2\alpha r}$. We then have

$$\mathrm{erf}(y) = y \left( \sqrt{2\alpha} + \frac{2}{\sqrt{\pi}} e^{-y^2} \right) \, . \tag{8.67}$$

One solution is given by $y = m = 0$, which is a spin glass (SG) solution. For $\alpha \geq \alpha_c = 0.138$, this is the unique solution. For $a < \alpha_c$, Ferromagnetic solutions $m \neq 0$ appear ($2P$ such solutions, due to the model symmetry). At $\alpha = \alpha_c$, the overlap $m$ takes the value $m = 0.967$ [6].

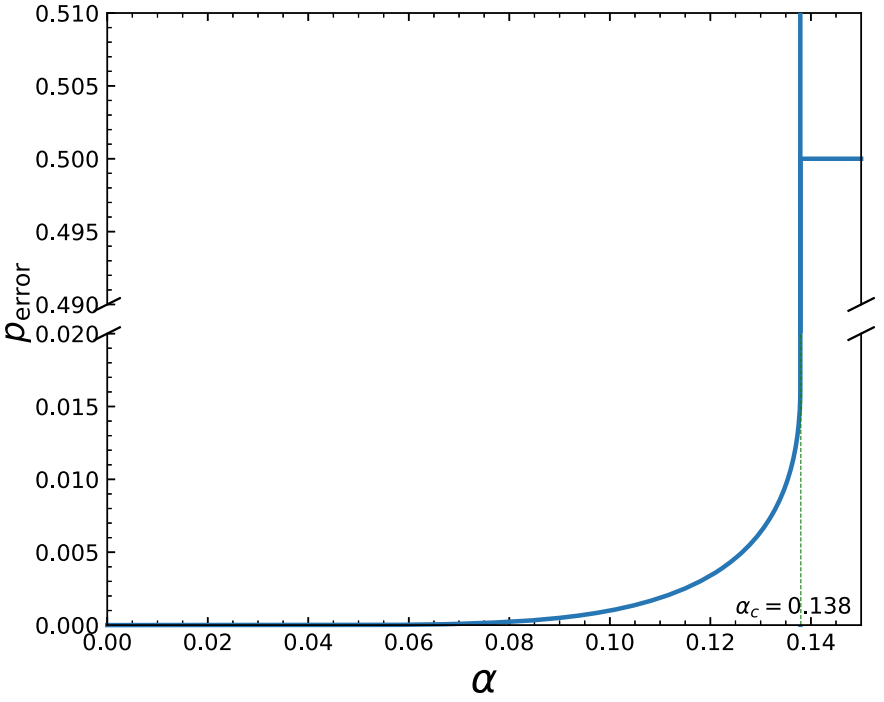



**Fig. 8.3** The error probability as a function of $\alpha$ at $T = 0$

Equation (8.67) can be solved numerically. By using the relation $m = \text{erf}(y)$, we can obtain the values of $m$. The error probability is given by $P_{error} = (1 - m)/2$, which is shown in Fig. 8.3. From the plot, we can see that there is a critical value $\alpha_c = 0.138$ where the error probability jumps to $1/2$, indicating a discontinuous transition to a spin glass phase. When $\alpha < \alpha_c$, the error probability is quite low, which means that the network can reliably retrieve one of the stored patterns. When $\alpha > \alpha_c$, the error probability is $1/2$, suggesting the network could not have a significant memory.

## 8.3 Phase Diagram

By solving Eqs. (8.61), (8.60) and (8.62) numerically, we can obtain the phase diagram of the Hopfield network (Fig. 8.4) [3, 6]. At a very high temperature, the thermal noise impairs the retrieval process, therefore $m = 0, q = 0$ and $r = 0$. Interesting, from an inverse Ising perspective, given the configurations from this phase, the couplings of the model can be easily inferred by a reverse engineering process [7, 8]. As the temperature is lowered down, the paramagnetic phase becomes unsta-

ble at a critical temperature-load line ($T_g(\alpha)$), which can be obtained analytically through a linear stability analysis of Eq. (8.60), i.e., $T_g = 1 + \sqrt{\alpha}$, where $\alpha$ is the memory load.

On the other hand, with decreasing memory load, the spin glass phase becomes metastable at a critical line $T_M(\alpha)$, where the retrieval phase becomes locally stable. This transition is thus a first-order phase transition. In this phase, spurious states (i.e., a linear combination of several stored patterns) also emerge as metastable states. Once $\alpha < 0.051$, the retrieval phase becomes globally stable when a critical temperature line $T_c$ is crossed. The discontinuous transition point can be obtained by analyzing the saddle-point equation, and equaling the free energies of two competing phases. $T_M \simeq 1 - 1.95\sqrt{\alpha}$, and $T_c \simeq 1 - 2.6\sqrt{\alpha}$ [6].

At $T = 0$, the entropy per spin $S = -\left.\frac{\partial f}{\partial T}\right|_{T\to 0} = -\frac{1}{2}\alpha[\ln(1-C) + C/(1-C)]$ with $C = \beta(1-q)$ is negative for all replica-symmetric solutions, which is unphysical. Below the dashed line (so-called AT line in spin glass theory; see Chap. 9) in Fig. 8.4, the retrieval states become unstable, the replica symmetry breaking (RSB) effects should be considered (a general introduction of RSB will be presented in

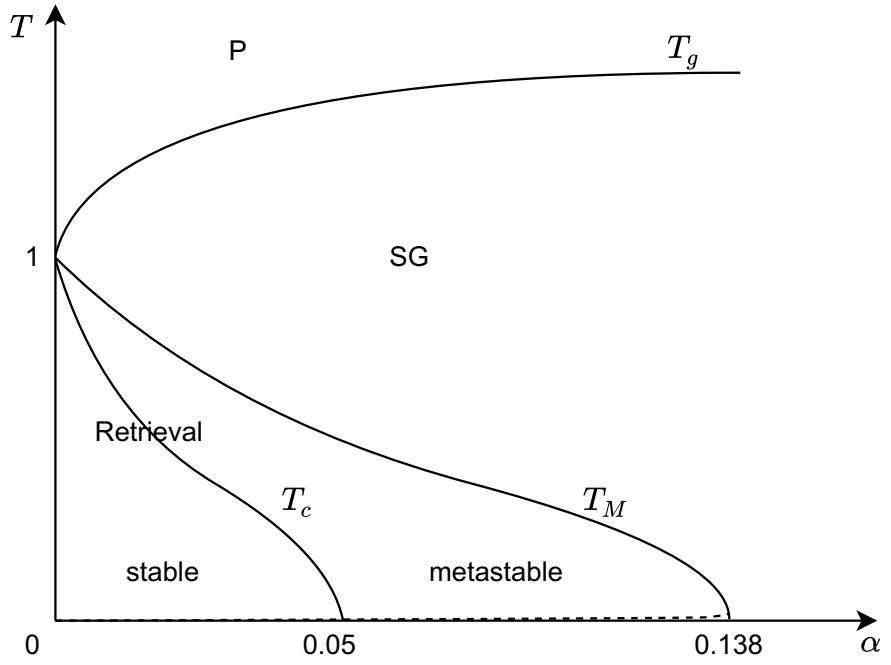



**Fig. 8.4** The phase diagram of Hopfield model (adapted from Ref. [3]). Three phases (paramagnetic, spin glass and retrieval) exist. The paramagnetic phase is separated by a continuous transition to the spin glass phase ($T_g$ line). The phase transition from retrieval phase to spin glass phase on the $T_M$ is discontinuous. Below $T_c$ line, the retrieval phase becomes globally stable. Below the dash line ($T_R$), the replica-symmetric solution becomes unstable

Chap. 9). In physics, this implies that the permutation symmetry of replica indexes in the overlap matrix does not hold, requiring that a higher level of approximation should be taken. However, as shown in the Fig. 8.4, the RSB effect in the retrieval phase is very weak. As $\alpha \to \infty$, the Hopfield model reduces to the well-known SK model.

## 8.4 Hopfield Model with Arbitrary Hebbian Length

In this section, we generalize the standard Hopfield model to the case of arbitrary Hebbian length. This is inspired by the Monkey experiments where the monkey is trained to recognize and match visual stimuli, the temporal order of the stimulus presentations is maintained during training. The experiments revealed that the monkey's temporal cortex is able to convert the temporal association of stimuli into a spatial correlation in the patterns of sustained activities [9, 10]. This experimental result was first modeled by Griniasty et.al. [11], who takes one Hebbian length into the construction of the coupling matrix, i.e., the neighboring patterns in the sequence of presentation contribute to Hebbian learning. In this model, a novel phase of correlated- attractors emerges due to this revised Hebbian rule. The correlated attractor triggered by one stimulus pattern becomes correlated with neighboring patterns around the stimulus, although the patterns themselves are all independent.

Motivated by the observation that Hebbian learning can occur in a wider learning window [12, 13], we propose to extend the Hebbian length to an arbitrary value [14], and thus define the following coupling matrix of neurons:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \left[ c\xi_i^\mu \xi_j^\mu + \gamma \sum_{r=1}^{d} \left( \xi_i^\mu \xi_j^{\mu+r} + \xi_i^{\mu+r} \xi_j^\mu \right) \right], \tag{8.68}$$

where $c$ specifies the standard Hebbian strength, $\gamma$ specifies the coupling strength between $r$-separated patterns, and $d$ is thus the Hebbian length of our model. The case of $d = 1$ has been studied by previous works [11, 15], while $d = 0$ recovers the standard Hopfield model [1–3]. $\xi_i^\mu$ follows independently a binomial distribution, i.e., $p(\xi_i^\mu = \pm 1) = \frac{1}{2}\delta(\xi_i^\mu + 1) + \frac{1}{2}\delta(\xi_i^\mu - 1)$. We are interested in the limit of large values of $P$ and $N$, thereby defining $\alpha = \frac{P}{N}$. $\alpha$ is also called the memory load of the associative memory model.

### *8.4.1 Computation of the Disorder-Averaged Free Energy*

The matrix $\mathbf{J}$ can be recast into the form

$$\mathbf{J} = \frac{1}{N}\boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}\boldsymbol{\xi} \ , \tag{8.69}$$

where $\mathbf{X}$ is a $P \times P$ circulant matrix, a special form of Toeplitz matrix with elements

$$\begin{aligned} X_{\mu\eta} &= c\delta_{\mu\eta} + \gamma \sum_{r=1}^{d} \left(\delta_{\mu,(\eta+r) \bmod P} + \delta_{\mu,(\eta-r) \bmod P}\right) \\ &= (c-\gamma)\delta_{\mu\eta} + \gamma \sum_{r=-d}^{d} \delta_{\mu,(\eta+r) \bmod P} \ . \end{aligned} \tag{8.70}$$

The $m$th eigenvalue of $\mathbf{X}$ is given by [16]

$$\begin{aligned} \lambda_m &= \sum_{k=0}^{P-1} X_{1(k+1)} e^{-2\pi imk/P} \\ &= \sum_{k=0}^{P-1} X_{1(k+1)} \cos\left(2\pi \frac{mk}{P}\right) \\ &= \sum_{k=0}^{P-1} \left[ c\delta_{0k} + \gamma \sum_{r=1}^{d} (\delta_{0,(k+r) \bmod P} + \delta_{0,(k-r) \bmod P}) \right] \cos\left(2\pi \frac{mk}{P}\right) \\ &= c + \gamma \sum_{r=1}^{d} \left[ \cos\left(-2\pi \frac{mr}{P}\right) + \cos\left(2\pi \frac{mr}{P}\right) \right] \\ &= c + 2\gamma \sum_{r=1}^{d} \cos\left(2\pi \frac{mr}{P}\right) \ , \end{aligned} \tag{8.71}$$

for $m = 0, 1, \ldots, P-1$.

The Hamiltonian of the model is defined by

$$\mathcal{H}(\mathbf{s}) = -\frac{1}{2} \sum_{i \neq j} J_{ij} s_i s_j \ . \tag{8.72}$$

The partition function is thus given by

$$Z = \mathrm{Tr} \exp\left[ \frac{\beta}{2N} \mathbf{s}^{\mathrm{T}} \boldsymbol{\xi}^{\mathrm{T}} \mathbf{X} \boldsymbol{\xi} \mathbf{s} \right] \ , \tag{8.73}$$

where Tr indicates the summation over all discrete states $\mathbf{s}$. In general, to compute a disorder averaged free energy ($\langle -T \ln Z \rangle$) is a computationally hard task. However, the well-known replica trick developed in spin glass theory [4] can be used to get around the difficulty, but assumptions on the replica matrix are required (detailed below). The replica method uses the mathematical identity

$$\langle \ln Z \rangle = \lim_{n\to 0} \frac{\ln \langle Z^n \rangle}{n} , \tag{8.74}$$

where $\langle \cdot \rangle$ denotes the expectation over the distribution of $\boldsymbol{\xi}$. To proceed, we have to compute an integer-power of the partition function

$$Z^n = \mathrm{Tr} \exp\left[ \frac{\beta}{2N} \sum_{a=1}^{n} \left(\mathbf{s}^a\right)^{\mathrm{T}} \boldsymbol{\xi}^{\mathrm{T}} \mathbf{X} \boldsymbol{\xi} \mathbf{s}^a \right] . \tag{8.75}$$

We consider the situation where there are $S$ condensed (or foreground) patterns and $P - S$ non-condensed (or background) patterns, which is reasonable in our current setting. The choice of $S$ can be justified a posterior, e.g., through solving the mean-field dynamics or saddle-point equations. Thus, we can reorganize the matrix $\mathbf{X}$ as a block matrix, i.e.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{FF} & \mathbf{X}_{FB} \\ \mathbf{X}_{BF} & \mathbf{X}_{BB} \end{bmatrix} , \tag{8.76}$$

where $\mathbf{X}_{FF} \in \mathbb{R}^{S\times S}$, $\mathbf{X}_{BF}^{\mathrm{T}} = \mathbf{X}_{FB} \in \mathbb{R}^{S\times(P-S)}$ and $\mathbf{X}_{BB} \in \mathbb{R}^{(P-S)\times(P-S)}$.

It then follows that

$$\begin{aligned} Z^n = \mathrm{Tr} \exp\Bigg[ & \frac{\beta}{2N} \sum_{a,i,j,\mu\in B,\nu\in B} s_i^a \xi_i^\mu X_{\mu\nu} \xi_j^\nu s_j^a + \frac{\beta}{N} \sum_{a,i,j,\mu\in B,\nu\in F} s_i^a \xi_i^\mu X_{\mu\nu} \xi_j^\nu s_j^a \\ & + \frac{\beta}{2N} \sum_{a,i,j,\mu\in F,\nu\in F} s_i^a \xi_i^\mu X_{\mu\nu} \xi_j^\nu s_j^a \Bigg] . \end{aligned} \tag{8.77}$$

We then diagonalize the submatrix $\mathbf{X}_{BB}$ as $\mathbf{X}_{BB}^{\mu\nu} = \sum_\sigma \lambda_\sigma \eta_\mu^\sigma \eta_\nu^\sigma$, where $\lambda_\sigma$ and $\eta_\mu^\sigma$ are denoted as its eigenvalues and eigenvectors, respectively. We thus obtain

$$\begin{aligned} Z^n &= \mathrm{Tr} \exp\left[ \frac{\beta}{2N} \sum_{a,\sigma} \lambda_\sigma \left( \sum_{i,\mu\in B} s_i^a \xi_i^\mu \eta_\mu^\sigma \right)^2 + \frac{\beta}{N} \sum_{a,i,j,\mu\in B,\nu\in F} s_i^a \xi_i^\mu X_{\mu\nu} \xi_j^\nu s_j^a \right. \\ &\qquad \left. + \frac{\beta}{2N} \sum_{a,i,j,\mu\in F,\nu\in F} s_i^a \xi_i^\mu X_{\mu\nu} \xi_j^\nu s_j^a \right] \\ &= \mathrm{Tr} \prod_{a,\sigma} \int Dx_\sigma^a \exp\left[ \sum_{i,\mu\in B} \frac{\xi_i^\mu}{\sqrt{N}} \left( \sum_{a,\sigma} s_i^a \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \frac{\beta}{\sqrt{N}} \sum_{a,j,\nu\in F} s_i^a X_{\mu\nu} \xi_j^\nu s_j^a \right) \right. \\ &\qquad \left. + \frac{\beta}{2N} \sum_{a,i,j,\mu\in F,\nu\in F} s_i^a \xi_i^\mu X_{\mu\nu} \xi_j^\nu s_j^a \right] , \end{aligned} \tag{8.78}$$

where we have used the Hubbard–Stratonovich transformation, i.e., $\exp\left[\frac{1}{2}b^2\right] = \int Dx \exp\left[\pm bx\right]$, where $Dx = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$.

We then define

$$\Phi_B = \exp\left[ \sum_{i,\mu\in B} \frac{\xi_i^\mu}{\sqrt{N}} \left( \sum_{a,\sigma} s_i^a \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \frac{\beta}{\sqrt{N}} \sum_{a,j,\nu\in F} s_i^a X_{\mu\nu} \xi_j^\nu s_j^a \right) \right] , \quad (8.79)$$

and

$$\Phi_F = \exp\left[ \frac{\beta}{2N} \sum_{a,i,j,\mu\in F,\nu\in F} s_i^a \xi_i^\mu X_{\mu\nu} \xi_j^\nu s_j^a \right] . \quad (8.80)$$

Taking the disorder average over $\{\xi_i^\mu\}$, we write the result as

$$\left\langle Z^n \right\rangle = \left\langle \mathrm{Tr} \prod_{a,\sigma} \int Dx_\sigma^a \Phi_B \Phi_F \right\rangle . \quad (8.81)$$

We first carry out the average over the distribution of background patterns, which yields

$$\langle \Phi_B \rangle = \exp\left\{ \frac{1}{2N} \sum_{i,\mu\in B} \left[ \sum_a s_i^a \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \frac{\beta}{\sqrt{N}} \sum_{j,\nu\in F} X_{\mu\nu} \xi_j^\nu s_j^a \right) \right]^2 \right\} . \quad (8.82)$$

Introducing the state overlap as one order parameter: $q_{ab} = \frac{1}{N} \sum_i^N s_i^a s_i^b$ for $a \neq b$, and $m_\mu^a = \frac{1}{N} \sum_i \xi_i^\mu s_i^a$ as another order parameter, we have

$$\begin{aligned} \langle \Phi_B \rangle = & \int \prod_{a\neq b} \frac{dq_{ab} d\hat{q}_{ab}}{2\pi/N} \prod_{a,\mu\in F} \frac{dm_\mu^a d\hat{m}_\mu^a}{2\pi/N} \\ & \times \exp\left[ -\frac{1}{2} N \sum_{a\neq b} \hat{q}_{ab} q_{ab} + \frac{1}{2} \sum_{a\neq b} \hat{q}_{ab} \sum_i s_i^a s_i^b - N \sum_{a,\mu\in F} m_\mu^a \hat{m}_\mu^a + \sum_{a,\mu\in F} \hat{m}_\mu^a \sum_i \xi_i^\mu s_i^a \right] \\ & \times \exp\left[ \frac{1}{2} \sum_{\mu\in B} \sum_a \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu^a \right)^2 \right] \\ & \times \exp\left[ \frac{1}{2} \sum_{\mu\in B} \sum_{a\neq b} q_{ab} \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu^a \right) \right. \\ & \left. \times \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^b + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu^b \right) \right] . \end{aligned} \quad (8.83)$$

In the above derivations, we have inserted Dirac delta functions for defining those order parameters, and then applied the integral representations of these delta functions. The hatted order parameters are the byproducts of conjugated counterparts.

Under the replica symmetric ansätz with $q_{ab} = q$ and $\hat{q}_{ab} = \hat{q}$ for $a \neq b$, $m_\mu^a = m_\mu$ and $\hat{m}_\mu^a = \hat{m}_\mu$, we arrive at

$$\begin{aligned}
\langle \Phi_B \rangle &= \int \frac{dqd\hat{q}}{(2\pi/N)^{n(n-1)}} \frac{dmd\hat{m}}{(2\pi/N)^{nS}} - Nn \sum_{\mu\in F} m_\mu \hat{m}_\mu \exp\Bigg[ -\frac{1}{2} Nn(n-1)\hat{q}q \\
&\quad + \frac{1}{2}\hat{q} \sum_{a\neq b} \sum_i s_i^a s_i^b + \sum_{a,\mu\in F} \hat{m}_\mu \sum_i \xi_i^\mu s_i^a \Bigg] \times \exp\left[ \frac{1}{2} \sum_{\mu\in B} \sum_a \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a \right.\right. \\
&\quad \left.\left. + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu \right)^2 \right] \times \exp\left[ \frac{q}{2} \sum_{\mu\in B} \sum_{a\neq b} \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu \right) \right. \\
&\quad \left. \times \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^b + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu \right) \right] \\
&= \int \frac{dqd\hat{q}}{(2\pi/N)^{n(n-1)}} \frac{dmd\hat{m}}{(2\pi/N)^{nS}} \exp\left[ -\frac{1}{2} Nn(n-1)\hat{q}q + \frac{1}{2}\hat{q} \sum_{a\neq b} \sum_i s_i^a s_i^b - Nn \sum_{\mu\in F} m_\mu \hat{m}_\mu \right. \\
&\quad \left. + \sum_{a,\mu\in F} \hat{m}_\mu \sum_i \xi_i^\mu s_i^a \right] \times \exp\left[ \frac{1-q}{2} \sum_{\mu\in B} \sum_a \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu \right)^2 \right] \\
&\quad \times \exp\left[ \frac{q}{2} \sum_{\mu\in B} \left( \sum_{a,\sigma} \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \beta n\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu \right)^2 \right] .
\end{aligned} \tag{8.84}$$

We apply the Hubbard–Stratonovich transformation once again, and obtain

$$\begin{aligned}
\langle \Phi_B \rangle &= \int \frac{dqd\hat{q}}{(2\pi/N)^{n(n-1)}} \frac{dmd\hat{m}}{(2\pi/N)^{nS}} \prod_{\mu,a} Dy_\mu^a \prod_\mu Dz_\mu \\
&\quad \times \exp\left[ -\frac{1}{2} Nn(n-1)\hat{q}q + \frac{1}{2}\hat{q} \sum_{a\neq b} \sum_i s_i^a s_i^b - Nn \sum_{\mu\in F} m_\mu \hat{m}_\mu + \sum_{a,\mu\in F} \hat{m}_\mu \sum_i \xi_i^\mu s_i^a \right] \\
&\quad \times \exp\left[ \sqrt{1-q} \sum_{\mu\in B} \sum_a \left( \sum_\sigma \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \beta\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu \right) y_\mu^a \right] \\
&\quad \times \exp\left[ \sqrt{q} \sum_{\mu\in B} \left( \sum_{a,\sigma} \eta_\mu^\sigma \sqrt{\beta\lambda_\sigma} x_\sigma^a + \beta n\sqrt{N} \sum_{\nu\in F} X_{\mu\nu} m_\nu \right) z_\mu \right] .
\end{aligned} \tag{8.85}$$

By collecting terms containing $x_\sigma^a$, we have

$$\begin{aligned}\langle \Phi_B \rangle = \int \frac{dqd\hat{q}}{(2\pi/N)^{n(n-1)}} \frac{dmd\hat{m}}{(2\pi/N)^{nS}} \prod_{\mu,a} Dy_\mu^a \prod_\mu Dz_\mu \\ \times \exp\left[ -\frac{1}{2} Nn(n-1)\hat{q}q + \frac{1}{2}\hat{q} \sum_{a\neq b} \sum_i s_i^a s_i^b - Nn \sum_{\mu\in F} m_\mu \hat{m}_\mu + \sum_{a,\mu\in F} \hat{m}_\mu \sum_i \xi_i^\mu s_i^a \right] \\ \times \exp\left[ \sum_{a,\sigma} x_\sigma^a \sqrt{\beta\lambda_\sigma} \sum_{\mu\in B} \eta_\mu^\sigma \left( \sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu \right) \right] \\ \times \exp\left[ \beta\sqrt{N} \sum_{a,\mu\in B} \sum_{\nu\in F} X_{\mu\nu} m_\nu \left( \sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu \right) \right] .\end{aligned} \tag{8.86}$$

According to the definition of the overlap, $\Phi_F$ can be written as

$$\Phi_F = \exp\left[ \frac{\beta n N}{2} \sum_{\mu\in F, \nu\in F} m_\mu X_{\mu\nu} m_\nu \right] . \tag{8.87}$$

Collecting all the results derived above, we have

$$\begin{aligned}\langle Z^n \rangle = \mathrm{Tr} \int \prod_{a,\sigma} Dx_\sigma^a \frac{dqd\hat{q}}{(2\pi/N)^{n(n-1)}} \frac{dmd\hat{m}}{(2\pi/N)^{nS}} \prod_{\mu,a} Dy_\mu^a \prod_\mu Dz_\mu \\ \times \exp\left[ -\frac{1}{2} Nn(n-1)\hat{q}q + \frac{1}{2}\hat{q} \sum_{a\neq b} \sum_i s_i^a s_i^b - Nn \sum_{\mu\in F} m_\mu \hat{m}_\mu \right] \\ \times \left\langle \exp\left[ \sum_{a,\mu\in F} \hat{m}_\mu \sum_i \xi_i^\mu s_i^a \right] \right\rangle \times \exp\left[ \sum_{a,\sigma} x_\sigma^a \sqrt{\beta\lambda_\sigma} \sum_{\mu\in B} \eta_\mu^\sigma \left( \sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu \right) \right] \\ \times \exp\left[ \beta\sqrt{N} \sum_{a,\mu\in B} \sum_{\nu\in F} X_{\mu\nu} m_\nu \left( \sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu \right) \right] \\ \times \exp\left[ \frac{\beta n N}{2} \sum_{\mu\in F, \nu\in F} m_\mu X_{\mu\nu} m_\nu \right] .\end{aligned} \tag{8.88}$$

We define the term summing over $\{s_i^a\}$ as

$$\begin{aligned}\langle \Phi_S \rangle &= \left\langle \mathrm{Tr} \exp\left[ \frac{1}{2}\hat{q} \sum_{a\neq b} \sum_i s_i^a s_i^b + \sum_{a,\mu\in F} \hat{m}_\mu \sum_i \xi_i^\mu s_i^a \right] \right\rangle \\ &= \exp\left[ -\frac{nN}{2}\hat{q} \right] \mathrm{Tr} \left\langle \prod_i \exp\left[ \frac{1}{2}\hat{q} \left( \sum_a s_i^a \right)^2 + \sum_{a,\mu\in F} \hat{m}_\mu \xi_i^\mu s_i^a \right] \right\rangle \\ &= \exp\left[ -\frac{nN}{2}\hat{q} \right] \left\{ \left\langle \mathrm{Tr} \exp\left[ \frac{1}{2}\hat{q} \left( \sum_a s^a \right)^2 + \sum_{a,\mu\in F} \hat{m}_\mu \xi^\mu s^a \right] \right\rangle \right\}^N . \end{aligned} \tag{8.89}$$

Applying the Hubbard–Stratonovich transformation, we obtain

$$\begin{aligned}\langle \Phi_S \rangle &= \exp\left[ -\frac{nN}{2}\hat{q} \right] \left\{ \left\langle \int Dz \prod_a \mathrm{Tr} \exp\left[ \sqrt{\hat{q}} s^a z + \sum_{\mu\in F} \hat{m}_\mu \xi^\mu s^a \right] \right\rangle \right\}^N \\ &= \exp\left[ -\frac{nN}{2}\hat{q} \right] \left\{ \left\langle \int Dz \prod_a 2\cosh\left[ \sqrt{\hat{q}} z + \sum_{\mu\in F} \hat{m}_\mu \xi^\mu \right] \right\rangle \right\}^N \\ &= \exp\left[ -\frac{nN}{2}\hat{q} \right] \exp\left\{ N \ln\left[ \left\langle \int Dz\, 2^n \cosh^n\left( \sqrt{\hat{q}} z + \sum_{\mu\in F} \hat{m}_\mu \xi^\mu \right) \right\rangle \right] \right\} . \end{aligned} \tag{8.90}$$

In the limit $n \to 0$,

$$\langle \Phi_S \rangle = \exp\left[ -\frac{nN}{2}\hat{q} \right] \exp\left\{ nN \left\langle \int Dz\, \ln\left[ 2\cosh\left( \sqrt{\hat{q}} z + \sum_{\mu\in F} \hat{m}_\mu \xi^\mu \right) \right] \right\rangle \right\} . \tag{8.91}$$

Taken together, we have

$$\begin{aligned}\langle Z^n\rangle = &\int \prod_{a,\sigma} Dx_\sigma^a \frac{dqd\hat{q}}{(2\pi/N)^{n(n-1)}} \frac{dmd\hat{m}}{(2\pi/N)^{nS}} \prod_{\mu,a} Dy_\mu^a \prod_\mu Dz_\mu \\ &\times \exp\left[-\frac{1}{2}Nn(n-1)\hat{q}q - \frac{nN}{2}\hat{q} - Nn\sum_{\mu\in F} m_\mu \hat{m}_\mu + \frac{\beta nN}{2}\sum_{\mu\in F,\nu\in F} m_\mu X_{\mu\nu} m_\nu\right] \\ &\times \exp\left[\sum_{a,\sigma} x_\sigma^a \sqrt{\beta\lambda_\sigma} \sum_{\mu\in B} \eta_\mu^\sigma \left(\sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu\right)\right] \\ &\times \exp\left[\beta\sqrt{N} \sum_{a,\mu\in B}\sum_{\nu\in F} X_{\mu\nu} m_\nu \left(\sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu\right)\right] \\ &\times \exp\left\{nN\left\langle \int Dz\, \ln\left[2\cosh\left(\sqrt{\hat{q}} z + \sum_{\mu\in F} \hat{m}_\mu \xi^\mu\right)\right]\right\rangle\right\}. \end{aligned} \tag{8.92}$$

To proceed, we first denote the vectors $\mathbf{y}^a = [y_\mu^a; \mu \in B]^{\mathrm{T}}$, $\mathbf{z} = [z_\mu; \mu \in B]^{\mathrm{T}}$, $\mathbf{m} = [m_\mu; \mu \in F]^{\mathrm{T}}$, $\hat{\mathbf{m}} = [\hat{m}_\mu; \mu \in F]^{\mathrm{T}}$ and $\boldsymbol{\xi}_F = [\xi^\mu; \mu \in F]^{\mathrm{T}}$. Integrating out $\{x_\sigma^a\}$, we get

$$\begin{aligned} &\int \prod_{a,\sigma} Dx_\sigma^a \exp\left[\sum_{a,\sigma} x_\sigma^a \sqrt{\beta\lambda_\sigma} \sum_{\mu\in B} \eta_\mu^\sigma \left(\sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu\right)\right] \\ &= \exp\left[\frac{1}{2}\beta \sum_{a,\sigma,\mu\in B,\nu\in B} \lambda_\sigma \eta_\mu^\sigma \eta_\nu^\sigma (\sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu)(\sqrt{1-q}\, y_\nu^a + \sqrt{q} z_\nu)\right] \\ &= \exp\left[\frac{1}{2}\beta \sum_{a,\mu\in B,\nu\in B} X_{\mu\nu} (\sqrt{1-q}\, y_\mu^a + \sqrt{q} z_\mu)(\sqrt{1-q}\, y_\nu^a + \sqrt{q} z_\nu)\right] \\ &= \exp\left[\frac{1}{2}\beta(1-q) \sum_{a,\mu\in B,\nu\in B} y_\mu^a X_{\mu\nu} y_\nu^a + \beta\sqrt{(1-q)q}\right. \\ &\quad \left. \times \sum_{a,\mu\in B,\nu\in B} z_\mu X_{\mu\nu} y_\nu^a + \frac{1}{2} n\beta q \sum_{\mu\in B,\nu\in B} z_\mu X_{\mu\nu} z_\nu\right] \\ &= \exp\left[\frac{1}{2}\beta(1-q)\sum_a (\mathbf{y}^a)^{\mathrm{T}} \mathbf{X}_{BB} \mathbf{y}^a + \beta\sqrt{(1-q)q} \sum_a \mathbf{z}^{\mathrm{T}} \mathbf{X}_{BB} \mathbf{y}^a + \frac{1}{2} n\beta q \mathbf{z}^{\mathrm{T}} \mathbf{X}_{BB} \mathbf{z}\right]. \end{aligned} \tag{8.93}$$

Collecting all terms containing $\{y_\mu^a\}$, we get

$$
\begin{aligned}
&\int \prod_{\mu,a} \frac{dy_\mu^a}{\sqrt{2\pi}} \prod_a \exp\left[ -\frac{1}{2} \sum_{\mu\in B,\nu\in B} y_\mu^a \left(\delta_{\mu\nu} - \beta(1-q)X_{\mu\nu}\right) y_\nu^a \right] \\
&\quad \times \exp\left[ \beta\sqrt{1-q} \sum_{\nu\in B} \left( \sum_{\mu\in F} \sqrt{N} X_{\nu\mu} m_\mu + \sqrt{q} \sum_{\mu\in B} z_\mu X_{\mu\nu} \right) y_\nu^a \right] \\
&= \int \prod_{\mu,a} \frac{dy_\mu^a}{\sqrt{2\pi}} \prod_a \exp\left[ -\frac{1}{2} (\mathbf{y}^a)^{\mathrm{T}} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right) \mathbf{y}^a \right] \\
&\quad \times \exp\left[ \beta\sqrt{1-q} \left( \sqrt{N}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB} + \sqrt{q}\mathbf{z}^{\mathrm{T}}\mathbf{X}_{BB} \right) \mathbf{y}^a \right] \\
&= \frac{1}{\sqrt{[\det\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)]^n}} \exp\left[ \frac{1}{2} n\beta^2(1-q) \left( \sqrt{N}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB} + \sqrt{q}\mathbf{z}^{\mathrm{T}}\mathbf{X}_{BB} \right) \right. \\
&\quad \left. \cdot \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \left( \sqrt{N}\mathbf{X}_{BF}\mathbf{m} + \sqrt{q}\mathbf{X}_{BB}\mathbf{z} \right) \right] \\
&= \frac{1}{\sqrt{[\det\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)]^n}} \exp\left[ \frac{1}{2} nN\beta^2(1-q)\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BF}\mathbf{m} \right] \\
&\quad \times \exp\left[ \frac{1}{2} n\beta^2(1-q)q\mathbf{z}^{\mathrm{T}}\mathbf{X}_{BB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BB}\mathbf{z} \right] \\
&\quad \times \exp\left[ n\beta^2(1-q)\sqrt{Nq}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BB}\mathbf{z} \right] ,
\end{aligned}
\tag{8.94}
$$

where $\mathbb{I}$ indicates an identity matrix.

We then collect all terms containing $\{z_\mu\}$, integrate out $\{z_\mu\}$ in the limit $n \to 0$, and finally obtain

$$
\begin{aligned}
&\int \prod_\mu \frac{dz_\mu}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\mathbf{z}^{\mathrm{T}} \left[ \mathbb{I} - n\beta q\mathbf{X}_{BB} - n\beta^2(1-q)q\mathbf{X}_{BB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BB} \right] \mathbf{z} \right\} \\
&\times \exp\left\{ \beta n\sqrt{qN} \left[ \mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB} + \beta(1-q)\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BB} \right] \mathbf{z} \right\} \\
&= \exp\left\{ -\frac{1}{2} \ln\det\left[ \mathbb{I} - n\beta q\mathbf{X}_{BB} - n\beta^2(1-q)q\mathbf{X}_{BB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BB} \right] \right\} ,
\end{aligned}
\tag{8.95}
$$

where to arrive at the last equality, we consider the limit of $n \to 0$ (i.e., neglecting terms involving $O(n^2)$).

To sum up, we rewrite $\langle Z^n \rangle$ as

$$
\begin{aligned}
\langle Z^n \rangle = &\int \frac{dqd\hat{q}}{(2\pi/N)^{n(n-1)}} \prod_\mu \frac{dm_\mu d\hat{m}_\mu}{(2\pi/N)^{nS}} \times \exp\left\{ nN \left\langle \int Dz \, \ln\left[ 2\cosh\left( \sqrt{\hat{q}}z + \hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F \right) \right] \right\rangle \right\} \\
&\times \exp\left[ -\frac{1}{2} Nn(n-1)\hat{q}q - \frac{nN}{2}\hat{q} - Nn\mathbf{m}^{\mathrm{T}}\hat{\mathbf{m}} + \frac{\beta nN}{2}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FF}\mathbf{m} \right] \\
&\times \exp\left[ -\frac{n}{2} \ln\det\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right) \right] \\
&\times \exp\left[ \frac{nN\beta^2(1-q)}{2} \mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BF}\mathbf{m} \right] \\
&\times \exp\left\{ -\frac{1}{2} \ln\det\left[ \mathbb{I} - n\beta q\mathbf{X}_{BB} - n\beta^2(1-q)q\mathbf{X}_{BB} \left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1} \mathbf{X}_{BB} \right] \right\} .
\end{aligned}
\tag{8.96}
$$

By applying the Laplace's method, we get the averaged free energy as

$$
\begin{aligned}
-\beta f \equiv \frac{1}{N}\langle \ln Z\rangle = & \left\langle \int Dz\, \ln\left[2\cosh\left(\sqrt{\hat{q}}z + \hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F\right)\right]\right\rangle + \frac{1}{2}\hat{q}q - \frac{1}{2}\hat{q} - \mathbf{m}^{\mathrm{T}}\hat{\mathbf{m}} + \frac{\beta}{2}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FF}\mathbf{m} \\
& - \frac{1}{2N}\ln\det\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right) + \frac{\beta^2(1-q)}{2}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1}\mathbf{X}_{BF}\mathbf{m} \\
& - \lim_{n\to 0}\frac{1}{2nN}\ln\det\left[\mathbb{I} - n\beta q\mathbf{X}_{BB} - n\beta^2(1-q)q\mathbf{X}_{BB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1}\mathbf{X}_{BB}\right],
\end{aligned}
\tag{8.97}
$$

where the last two terms can be further simplified as follows:

$$
-\frac{1}{2N}\ln\det\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right) = -\frac{1}{2N}\sum_{\sigma}\ln\left[1 - \beta(1-q)\lambda_\sigma\right]\ ; \tag{8.98}
$$

and

$$
\begin{aligned}
& -\lim_{n\to 0}\frac{1}{2nN}\ln\det\left[\mathbb{I} - n\beta q\mathbf{X}_{BB} - n\beta^2(1-q)q\mathbf{X}_{BB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1}\mathbf{X}_{BB}\right] \\
& = -\lim_{n\to 0}\frac{1}{2nN}\sum_{\sigma}\ln\left[1 - n\beta q\lambda_\sigma - \frac{n\beta^2(1-q)q\lambda_\sigma^2}{1-\beta(1-q)\lambda_\sigma}\right] \\
& = \frac{1}{2N}\sum_{\sigma}\frac{\beta q\lambda_\sigma}{1-\beta(1-q)\lambda_\sigma}\ .
\end{aligned}
\tag{8.99}
$$

Finally, the averaged free energy is given by

$$
\begin{aligned}
\frac{1}{N}\langle \ln Z\rangle = & \left\langle \int Dz\, \ln\left[2\cosh\left(\sqrt{\hat{q}}z + \hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F\right)\right]\right\rangle + \frac{1}{2}\hat{q}q - \frac{1}{2}\hat{q} - \mathbf{m}^{\mathrm{T}}\hat{\mathbf{m}} + \frac{\beta}{2}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FF}\mathbf{m} \\
& - \frac{1}{2N}\sum_{\sigma}\ln\left[1 - \beta(1-q)\lambda_\sigma\right] + \frac{\beta^2(1-q)}{2}\mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1}\mathbf{X}_{BF}\mathbf{m} \\
& + \frac{1}{2N}\sum_{\sigma}\frac{\beta q\lambda_\sigma}{1-\beta(1-q)\lambda_\sigma}\ .
\end{aligned}
\tag{8.100}
$$

We rescale $\hat{q}$ by $\beta^2\hat{q}$, and $\hat{\mathbf{m}}$ by $\beta\hat{\mathbf{m}}$. We then define

$$
\mathbf{K} = \mathbf{X}_{FF} + \beta(1-q)\mathbf{X}_{FB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1}\mathbf{X}_{BF}. \tag{8.101}
$$

The stationary condition of the free energy with respect to $\mathbf{m}$ implies that $\hat{\mathbf{m}} = \mathbf{K}\mathbf{m}$. Therefore, the free energy can be reorganized as follows:

$$-\beta f = \frac{\beta^2 \hat{q}}{2}(q-1) - \frac{\beta}{2}\mathbf{m}^{\mathrm{T}}\mathbf{K}\mathbf{m} - \frac{\alpha}{2}\int_0^1 du \ln\left[1-\beta(1-q)\Lambda(u)\right] + \frac{\alpha\beta q}{2}\int_0^1 du \frac{\Lambda(u)}{1-\beta(1-q)\Lambda(u)} + \left\langle \int Dz \ln\left[2\cosh\left(\beta\sqrt{\hat{q}}z + \beta\hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F\right)\right]\right\rangle, \tag{8.102}$$

where $\Lambda(u) = c + 2\gamma\sum_{r=1}^{d}\cos(2\pi r u)$. In the limit $P \to \infty$, it can be proved that $X_{BB}$ is asymptotically equivalent to $\mathbf{X}$ [16]. Therefore, the summation over $\sigma$ can be replaced by an integral using the eigenvalue of the circulant matrix $\mathbf{X}$.

### 8.4.2 Derivation of Saddle-Point Equations

The order parameter should take values optimizing the free energy function, leading to the saddle-point equations (SDE). The saddle-point equation of $q$ is given by

$$q - 1 + \frac{1}{\beta\sqrt{\hat{q}}}\left\langle \int Dz\, z \tanh\left(\beta\sqrt{\hat{q}}z + \beta\hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F\right)\right\rangle = 0\,; \tag{8.103}$$

$$q - 1 + \left\langle \int Dz \left[1 - \tanh^2\left(\beta\sqrt{\hat{q}}z + \beta\hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F\right)\right]\right\rangle = 0\,; \tag{8.104}$$

$$q = \left\langle \int Dz \tanh^2\left(\beta\sqrt{\hat{q}}z + \beta\hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F\right)\right\rangle. \tag{8.105}$$

The saddle-point equation of $\mathbf{m}$ is given by

$$\mathbf{m} = \left\langle \boldsymbol{\xi}_F \int Dz \tanh\left(\beta\sqrt{\hat{q}}z + \beta\hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F\right)\right\rangle. \tag{8.106}$$

The saddle-point equation of $\hat{\mathbf{m}}$ is given by

$$\hat{\mathbf{m}} = \mathbf{X}_{FF}\mathbf{m} + \beta(1-q)\mathbf{X}_{FB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1}\mathbf{X}_{BF}\mathbf{m} := \mathbf{K}\mathbf{m}\,, \tag{8.107}$$

where $\mathbf{K} = \mathbf{X}_{FF} + \beta(1-q)\mathbf{X}_{FB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-1}\mathbf{X}_{BF}$, as derived at the end of the previous section. The saddle-point equation of $\hat{q}$ is given by

$$\begin{aligned}\hat{q} &= \frac{1}{N}\sum_{\sigma}\frac{q\lambda_\sigma^2}{\left[1-\beta(1-q)\lambda_\sigma\right]^2} + \mathbf{m}^{\mathrm{T}}\mathbf{X}_{FB}\left(\mathbb{I} - \beta(1-q)\mathbf{X}_{BB}\right)^{-2}\mathbf{X}_{BF}\mathbf{m} \\ &= \alpha q\int_0^1 \frac{\Lambda^2(u)du}{\left(1-\beta(1-q)\Lambda(u)\right)^2} - \beta^{-1}\mathbf{m}^{\mathrm{T}}\frac{\partial\mathbf{K}}{\partial q}\mathbf{m}\,.\end{aligned} \tag{8.108}$$

Finally, the saddle-point equations are summarized as follows:

$$\hat{\mathbf{m}} = \mathbf{K}\mathbf{m} \,, \tag{8.109a}$$

$$q = \left\langle \int Dz \, \tanh^2 \left( \beta \sqrt{\hat{q}} z + \beta \hat{\mathbf{m}}^{\mathrm{T}} \boldsymbol{\xi}_F \right) \right\rangle \,, \tag{8.109b}$$

$$\hat{q} = \alpha q \int_0^1 du \frac{\Lambda^2(u)}{(1 - \beta(1-q)\Lambda(u))^2} - \beta^{-1} \mathbf{m}^{\mathrm{T}} \frac{\partial \mathbf{K}}{\partial q} \mathbf{m} \,, \tag{8.109c}$$

$$\mathbf{m} = \left\langle \boldsymbol{\xi}_F \int Dz \, \tanh \left( \beta \sqrt{\hat{q}} z + \beta \hat{\mathbf{m}}^{\mathrm{T}} \boldsymbol{\xi}_F \right) \right\rangle \,. \tag{8.109d}$$

We next determine the critical temperature between the paramagnetic phase and spin glass phase. In the spin glass phase, $q \neq 0$ but $\mathbf{m} = 0$. Expanding $q = \left\langle \int Dz \tanh^2 \left( \beta \sqrt{\hat{q}} z + \beta \hat{\mathbf{m}}^{\mathrm{T}} \boldsymbol{\xi}_F \right) \right\rangle$, and $\hat{q} = \alpha q \int_0^1 du \frac{\Lambda^2(u)}{(1-\beta(1-q)\Lambda(u))^2} + \mathbf{m}^{\mathrm{T}} \frac{\partial \mathbf{K}}{\partial C} \mathbf{m}$ [$C \equiv \beta(1-q)$] in powers of $q$ and $\hat{q}$, we have

$$q \simeq \beta^2 \hat{q} \simeq \beta^2 \alpha q \int_0^1 du \frac{\Lambda^2(u)}{(1 - \beta\Lambda(u))^2} + O(q^2) \,. \tag{8.110}$$

$T_g$ can then be obtained by solving

$$1 = \alpha \int_0^1 du \frac{\Lambda^2(u)}{\left(T_{\mathrm{g}} - \Lambda(u)\right)^2} \,. \tag{8.111}$$

For the standard Hopfield model, Eq. (8.111) can be analytically solved with the result $T_{\mathrm{g}} = 1 + \sqrt{\alpha}$.

### 8.4.3 Computation Transformation to Solve the SDE

To solve the SDE numerically is challenging, due to the computation of $\mathbf{K}$, which involves the block structure of $\mathbf{X}$. To get rid of dependence on $N$ and $P$ (we are only interested in the large $N$ and $P$ limit), we propose the following numerical technique. We first define $C = \beta(1-q)$.

Note that if $C = 0$, $\mathbf{K} = \mathbf{X}_{FF}$, $\frac{\partial \mathbf{K}}{\partial C} = \mathbf{X}_{FB}\mathbf{X}_{BF}$. Let

$$\mathbf{X}\mathbf{X}^{\top} = \begin{bmatrix} \mathbf{H} & \cdots \\ \cdots & \cdots \end{bmatrix}, \tag{8.112}$$

where $\mathbf{H}$ is an $S \times S$ symmetric matrix. Then, we have

$$\mathbf{H} = \mathbf{X}_{FF}\mathbf{X}_{FF}^{\top} + \mathbf{X}_{FB}\mathbf{X}_{BF} = \mathbf{X}_{FF}\mathbf{X}_{FF}^{\top} + \left.\frac{\partial \mathbf{K}}{\partial C}\right|_{C=0} . \tag{8.113}$$

The matrix $\mathbf{H}$ can be computed as follows:

$$\mathbf{H} = \begin{bmatrix} h_0 & h_1 & \cdots & h_{S-1} \\ h_1 & h_0 & \cdots & h_{S-2} \\ \vdots & \vdots & & \vdots \\ h_{S-1} & h_{S-2} & \cdots & h_0 \end{bmatrix} , \tag{8.114}$$

where

$$\begin{aligned} h_l &= \frac{1}{P}\sum_{m=0}^{P-1}\left[c + 2\gamma\sum_{r=1}^{d}\cos\left(\frac{2\pi rm}{P}\right)\right]^2 \exp\left(\frac{2\pi \mathrm{i} ml}{P}\right) \\ &= \int_0^1 dx\left[c - \gamma + \gamma\sum_{r=-d}^{d}\exp\left(2\pi \mathrm{i} rx\right)\right]^2 \exp\left(2\pi \mathrm{i} lx\right) \\ &= \int_0^1 dx\left[c + 2\gamma\sum_{r=1}^{d}\cos\left(2\pi rx\right)\right]^2 \cos\left(2\pi lx\right) . \end{aligned} \tag{8.115}$$

Finally, we arrive at

$$\left.\frac{\partial \mathbf{K}}{\partial C}\right|_{C=0} = \mathbf{H} - \mathbf{X}_{FF}\mathbf{X}_{FF}^{\top} = \mathbf{H} - \left(\mathbf{K}|_{C=0}\right)^2 . \tag{8.116}$$

If $C \neq 0$, we have $\mathbf{K} = \mathbf{X}_{FF} - \mathbf{X}_{FB}\frac{1}{\mathbf{X}_{BB} - C^{-1}\mathbb{I}}\mathbf{X}_{BF}$. To calculate $\mathbf{K}$ numerically in the large $P$ limit, we notice that

$$\left(\mathbf{X} - C^{-1}\mathbb{I}\right)^{-1} = \begin{bmatrix} \mathbf{F}_1^{-1} & \cdots \\ \cdots & \cdots \end{bmatrix} , \tag{8.117}$$

where $\mathbf{F}_1^{-1} \in \mathbb{R}^{S\times S}$, and is a submatrix of $(\mathbf{X} - C^{-1}\mathbb{I})^{-1}$. Since $\mathbf{X} - C^{-1}\mathbb{I}$ is a circulant matrix, its inverse matrix can be calculated by $(\mathbf{X} - C^{-1}\mathbb{I})^{-1} = \mathrm{Circ}(w_0, w_1, \ldots, w_{P-1})$, where

$$w_k = \int_0^1 dx \frac{\cos(2\pi kx)}{c - C^{-1} + 2\gamma\sum_{r=1}^{d}\cos(2\pi rx)} , \tag{8.118}$$

for $k = 0, 1, \ldots, P-1$ in the limit $P \to \infty$. Thus, $\mathbf{F}_1^{-1}$ can be written as

$$\mathbf{F}_1^{-1} = \begin{bmatrix} w_0 & w_1 & \cdots & w_{S-1} \\ w_1 & w_0 & \cdots & w_{S-2} \\ \vdots & \vdots & & \vdots \\ w_{S-1} & w_{S-2} & \cdots & w_0 \end{bmatrix} . \tag{8.119}$$

By using the matrix formula for the inverse of a block matrix, we can prove that $\mathbf{K}$ can be expressed as

$$\mathbf{K} = \mathbf{F}_1 + C^{-1}\mathbb{I} . \tag{8.120}$$

Hence, to calculate $\mathbf{K}$ numerically, we first calculate $w_k$ for $k = 0, 1, \ldots, S-1$ to get $\mathbf{F}_1^{-1}$, and then calculate its inverse matrix $\mathbf{F}_1$, and finally add the matrix $C^{-1}\mathbb{I}$ to $\mathbf{F}_1$.

The term $\frac{\partial \mathbf{K}}{\partial C} = -\frac{1}{\beta}\frac{\partial \mathbf{K}}{\partial q}$ can be calculated as follows:

$$\frac{\partial \mathbf{K}}{\partial C} = \frac{\partial \mathbf{F}_1}{\partial C} - \frac{1}{C^2}\mathbb{I} = -\mathbf{F}_1 \frac{\partial \mathbf{F}_1^{-1}}{\partial C}\mathbf{F}_1 - \frac{1}{C^2}\mathbb{I} , \tag{8.121}$$

where the entry of $\frac{\partial \mathbf{F}_1^{-1}}{\partial C}$ is computed as

$$\frac{\partial w_k}{\partial C} = -\int_0^1 dx \frac{C^{-2}\cos(2\pi k x)}{\left[c - C^{-1} + 2\gamma \sum_{r=1}^{d} \cos(2\pi r x)\right]^2} , \tag{8.122}$$

for $k = 0, 1, \ldots, S-1$.

### 8.4.4 Zero-Temperature Limit

In the limit $T \to 0$ ($\beta \to \infty$), it is easy to derive that

$$\begin{aligned} \int Dz \, \tanh(\beta(\sqrt{\hat{q}}z + x)) &= \sqrt{\frac{2}{\pi}} \int_0^{\frac{1}{\sqrt{\hat{q}}}x} dz \exp\left(-\frac{1}{2}z^2\right) + O(T) \\ &\equiv \operatorname{erf}\left(\frac{1}{\sqrt{2\hat{q}}}x\right) + O(T) , \end{aligned} \tag{8.123}$$

and

$$\begin{aligned}
&\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}\left(1-\tanh^2\beta[az+b]\right) \\
&\simeq \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\Big|_{\tanh^2\beta(az+b)=0} \times \int dz\left(1-\tanh^2\beta[az+b]\right) \\
&= \frac{1}{\sqrt{2\pi}} e^{-b^2/2a^2}\frac{1}{a\beta}\int dz \frac{\partial}{\partial z}\tanh\beta[az+b] \\
&= \sqrt{\frac{2}{\pi}}\frac{1}{a\beta}e^{-b^2/2a^2}\,.
\end{aligned} \tag{8.124}$$

We thus obtain

$$\mathbf{m} = \left\langle \boldsymbol{\xi}_F \,\mathrm{erf}\left[\frac{1}{\sqrt{2\hat{q}}}\boldsymbol{\xi}_F^{\mathrm{T}}\mathbf{K}\mathbf{m}\right]\right\rangle . \tag{8.125}$$

In the limit $T \to 0$, we also have

$$\begin{aligned}
\beta(1-q) &= \beta\int Dz\left\langle 1-\tanh^2\left[\beta\sqrt{\hat{q}}z+\beta\boldsymbol{\xi}_F^{\mathrm{T}}\mathbf{K}\mathbf{m}\right]\right\rangle \\
&= \sqrt{\frac{2}{\pi\hat{q}}}\left\langle \exp\left[-\frac{\left[\boldsymbol{\xi}_F^{\mathrm{T}}\mathbf{K}\mathbf{m}\right]^2}{2\hat{q}}\right]\right\rangle \\
&\equiv C\,.
\end{aligned} \tag{8.126}$$

The conjugated order parameter $\hat{q}$ is given by

$$\hat{q} = \alpha\int_0^1 du\frac{\Lambda^2(u)}{(1-C\Lambda(u))^2} + \mathbf{m}^{\mathrm{T}}\frac{\partial\mathbf{K}}{\partial C}\mathbf{m}, \tag{8.127}$$

where in the zero-temperature limit $q \to 1$.

The free energy at the zero-temperature limit is given by

$$-f = \frac{\alpha}{2}\int_0^1 du\frac{\Lambda(u)}{1-C\Lambda(u)} - \frac{C\hat{q}}{2} - \frac{1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{K}\mathbf{m} + \left\langle \frac{2a}{\sqrt{2\pi}}e^{-\frac{b^2}{2a^2}} + b\,\mathrm{erf}\left(\frac{b}{\sqrt{2}a}\right)\right\rangle, \tag{8.128}$$

where $a = \sqrt{\hat{q}}$ and $b = \hat{\mathbf{m}}^{\mathrm{T}}\boldsymbol{\xi}_F$.

#### 8.4.4.1 The Spin Glass Solution

In the spin glass solution of the SDE, $m_\mu = 0$ for all $\mu = 1, 2, \ldots, S$. Hence, we have

$$C = \sqrt{\frac{2}{\pi\hat{q}}}\,, \tag{8.129}$$

and

$$\hat{q} = \alpha \int_0^1 du \frac{\Lambda^2(u)}{(1 - C\Lambda(u))^2}. \tag{8.130}$$

We consider the simplest case of $\gamma = 0$ and $c = 1$. It immediately follows that

$$\hat{q} = \frac{\alpha}{(1 - C)^2}. \tag{8.131}$$

Therefore, $C = \left(1 + \sqrt{\frac{\pi\alpha}{2}}\right)^{-1}$ recovering previous results in the Hopfield model.

#### 8.4.4.2 The Retrieval Solution

The ferromagnetic phase have a single non-vanishing overlap, i.e., $m_\mu = m\delta_{\mu,1} \sim O(1)$. They are named retrieval states, captured by the following equations:

$$m = \left\langle \xi^1 \operatorname{erf}\left[ \frac{1}{\sqrt{2\hat{q}}} m \left[ \boldsymbol{\xi}_F^{\mathrm{T}} \mathbf{K} \right]_1 \right] \right\rangle, \tag{8.132a}$$

$$C = \sqrt{\frac{2}{\pi\hat{q}}} \left\langle \exp\left[ -\frac{\left[ m \left[ \boldsymbol{\xi}_F^{\mathrm{T}} \mathbf{K} \right]_1 \right]^2}{2\hat{q}} \right] \right\rangle, \tag{8.132b}$$

$$\hat{q} = \alpha \int_0^1 du \frac{\Lambda^2(u)}{[1 - C\Lambda(u)]^2} + \left[ \frac{\partial \mathbf{K}}{\partial C} \right]_{11} m^2. \tag{8.132c}$$

In the simplest case of $\gamma = 0$ and $c = 1$, we have $\mathbf{K} = \mathbb{I}$. The above equations thus reduce to

$$m = \operatorname{erf}\left( \frac{m}{\sqrt{2\hat{q}}} \right), \tag{8.133a}$$

$$C = \sqrt{\frac{2}{\pi\hat{q}}} e^{-\frac{m^2}{2\hat{q}}}, \tag{8.133b}$$

$$\hat{q} = \frac{\alpha}{(1 - C)^2}. \tag{8.133c}$$

This result gives the memory capacity of $\alpha_c \simeq 0.138$ beyond which $\mathbf{m} = 0$, which is exactly the memory capacity of the standard Hopfield network [3]. In the general case, we consider in this section, it is necessary to solve the general equation numerically.

Finally, we look at the phase diagram. As shown in Fig. 8.5a, we identify three phases. One is the retrieval phase where only one overlap component is of the order

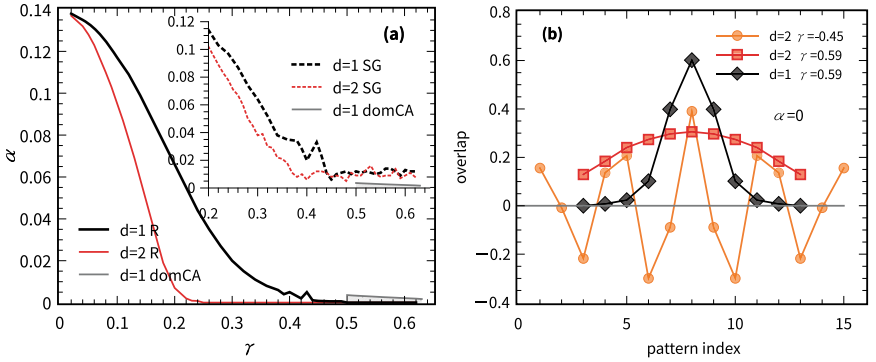



**Fig. 8.5** Phase diagram of the associative memory model in the $(\alpha, \gamma)$ plane given $c = 1$. **a** The phase boundary shown by the lines delimits the retrieval (R) phase from the region where the correlated-attractor (CA) and spin glass (SG) phases compete with each other (above the boundary). The boundary is the condition on which the retrieval phase loses its metastability from below. All shown transitions are of the discontinuous type. When $\alpha = 0$, the transition point is given by $\gamma_c = 0.5$ for $d = 1$, while $\gamma_c = 0.25$ for $d = 2$. The inset shows the boundary line above which the spin glass phase is dominant. Note that for $d = 1$, there exists a very narrow regime (indicated by the shadow) within which the correlated-attractor phase is dominant. **b** Overlap profiles obtained from the statistical mechanics theory. All overlap profiles are obtained by solving the saddle-point equations of the model when $\alpha = 0$ and $d = 2$ (or $d = 1$). All theoretical results are obtained by assuming that $S = 11$, except that for negative values of $\gamma$, we use $S = 15$. Note that the results are not sensitive to the value of $S$ (e.g., $S = 11$ or $S = 13$)

one, i.e., $m^\mu = m\delta_{\mu\nu}$, where $\nu$ indicates the stimulating pattern. Given the value of $\alpha$, increasing the value of $\gamma$ would finally make the retrieval phase lose its metastability, after which the correlated-attractor phase becomes metastable. The line separating these two phases is thus the first-order transition. The correlated-attractor phase is characterized by the stimulus-induced attractors being highly correlated with a finite number of patterns in the stored sequence. In other words, the value of the corresponding overlap decays with the distance between the patterns in the sequence and the one used as the stimulus. The numerical solutions of the saddle-point equations obtained by the replica theory (see the zero-temperature limit) reproduce the key features of the mean-field dynamics of the overlap [Fig. 8.5b], which corresponds to $\alpha = 0$ in our theory.

Our theory predicts that the value of $d$ can be used to expand the correlation span of the correlated-attractor, and moreover reshape significantly the phase diagram. When $\alpha = 0$, the threshold for the dominant retrieval phase is $\gamma_c = 0.5$ for $d = 1$, but $\gamma_c = 0.25$ for $d = 2$. In the presence of a finite $\alpha$, the retrieval phase loses its metastability at a smaller value of $\gamma$ for $d = 2$ than for $d = 1$ [Fig. 8.5a]. After that, the spin glass phase characterized by $m^\mu = 0$ $(\forall\mu)$ appears and competes with the correlated attractor phase, until the point where the spin glass phase becomes dominant (global minimum of the free energy), as shown in the inset of Fig. 8.5a. Remarkably, for $d = 1$, we identify a narrow regime for $\gamma > 0.5$ [the shadow in Fig. 8.5a], where the correlated-attractor phase becomes dominant. This regime shrinks gradually as

$\gamma$ increases. If noisy neural dynamics is allowed (e.g., at a non-zero temperature), the spin glass phase would be replaced by a paramagnetic phase at a continuous transition (see a detailed exploration in [17]). This transition line is also strongly affected by the Hebbian length.

As $\alpha$ gets close to the spin glass line [the inset of Fig. 8.5b], the peak value of the overlap in the correlated-attractor phase decreases, as expected from the significant memory interference at a relatively large memory load. At $\alpha = 0$, the correlation profile of the correlated attractor phase is more robust for $d = 2$ against increasing $\gamma$ than the case of $d = 1$. Further increasing $\gamma$ might leads to the result that the correlation is not localized any more, and the network loses the association ability about the stimuli.

In particular, our theoretical analysis also reproduces the unlearning effects observed in the mean-field dynamics [14]. Furthermore, a critical strength of $\gamma_c = -0.25$ for the oscillatory phase is predicted. $\gamma_c = -0.5$ for $d = 1$. When $\gamma < \gamma_c$, the unlearning effect of non-concurrent anti-Hebbian terms becomes more evident, preferring some particular patterns rather than their sign-reversed counterparts. In other words, the (spin reversal) symmetry in the Hamiltonian is broken, and the negative $\gamma$ selects particular patterns, which suggests that the energy landscape is reshaped and further the information storage is re-optimized [18–20]. This intriguing phenomenon thus establishes the connection between the Hebbian length, anti-Hebbian effect and memory function of unlearning.

## References

1. S.I. Amari, Biolog. cybern. **26**, 175 (1977)
2. J.J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982)
3. D.J. Amit, H. Gutfreund, H. Sompolinsky, Phys. Rev. Lett. **55**(14), 1530 (1985)
4. M. Mézard, G. Parisi, M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987)
5. D.J. Amit, H. Gutfreund, H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985)
6. D.J. Amit, H. Gutfreund, H. Sompolinsky, Ann. Phys. **173**(1), 30 (1987)
7. H. Huang, Phys. Rev. E **81**, 036104 (2010)
8. H. Huang, Phys. Rev. E **82**, 056111 (2010)
9. Y. Miyashita, Nature **335**, 817 (1988)
10. Y. Miyashita, H. Chang, Nature **331**, 68 (1988)
11. M. Griniasty, M.V. Tsodyks, D.J. Amit, Neural Comput. **5**(1), 1 (1993)
12. K.C. Bittner, A.D. Milstein, C. Grienberger, S. Romani, J.C. Magee, Science **357**(6355), 1033 (2017)
13. W. Gerstner, M. Lehmann, V. Liakoni, D. Corneil, J. Brea, Front. Neural Circuits **12**, 53 (2018)
14. Z. Jiang, J. Zhou, T. Hou, K.Y.M. Wong, H. Huang (2021). arXiv:2103.14317
15. L.F. Cugliandolo, M.V. Tsodyks, J. Phys. A: Math. General **27**(3), 741 (1994)
16. R.M. Gray, Found. Trends Commun. Inf. Theory **2**(3), 155 (2006)
17. J. Zhou, Z. Jiang, T. Hou, Z. Chen, K.Y.M. Wong, H. Huang (2021). arXiv: 2103.14324
18. A. Fachechi, E. Agliari, A. Barra, Neural Netw. **112**, 24 (2019)
19. V.S. Dotsenko, N.D. Yarunin, E.A. Dorotheyev, J. Phys. A **24**(10), 2419 (1991)
20. K. Nokura, J. Phys. A: Math. Gen. **31**(37), 7447 (1998)

# Chapter 9
# Replica Symmetry and Replica Symmetry Breaking


In this chapter, we introduce underlying physics behind the concept of replica symmetry, and replica symmetry breaking, which plays an important role in understanding the spin glass models of neural networks. Replica symmetry ansätz is considered as a first step of approximation to compute the quenched average of the free energy function. When the ansätz becomes unstable or yields unphysical results, the permutation symmetry of replica indexes must be broken, leading to a higher level of approximation—replica symmetry breaking.

## 9.1 Generalized Free Energy and Complexity of States

In previous chapters, replica symmetry is usually assumed as the first step for a statistical mechanical analysis of disordered systems (e.g., in the Hopfield model). The underlying physics is that a single giant pure state dominates the phase space of the model under investigation. In other words, the spin-spin correlation decays over their distance, satisfying the cluster decomposition (clustering) property [1], e.g., in a mean-field system of $N$ particles, the correlation magnitude is of the order $O(1/\sqrt{N})$ [1]. This usually occurs at a relatively high temperature, as shown in the dashed line of Fig. 9.1 As the temperature decreases, the giant state will split into many well-separated pure states, characterized by a free energy profile with many local minima separated by high barriers. Each minima corresponds to a fixed point of either TAP equation or belief propagation equation [2]. It contributes a statistical weight, i.e., $\frac{e^{-\beta F_\alpha}}{\sum_\alpha e^{-\beta F_\alpha}}$, where $F_\alpha$ indicates the free energy of the state with index $\alpha$, and $\beta$ is an inverse temperature.

To describe the decomposition of the Gibbs measure [3], we need to introduce an additional parameter characterizing the fluctuation of free energy levels, namely $y$, as follows:

$$e^{-y\Phi} = \sum_\alpha e^{-yF_\alpha} = \int \mathrm{d}f e^{N(-yf+\Sigma(f))}, \tag{9.1}$$



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_9

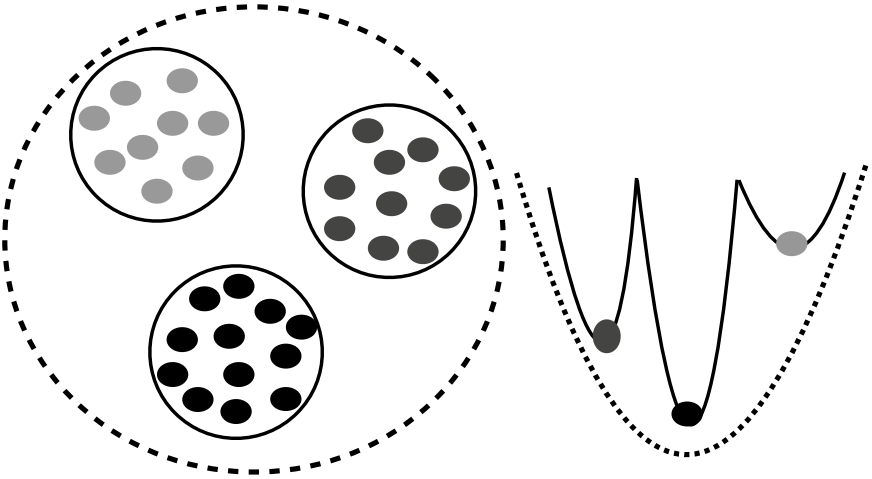


**Fig. 9.1** Schematic illustration of how a clustered organization of the phase space emerges. Here, we show only three clusters of configurations, within each of which the clustering property of a pure state holds. The right panel shows the corresponding free energy landscape

where $\Sigma(f)$ encodes the complexity of exponentially many states, an extension of the standard entropy in statistical mechanics. $\Phi$ denotes the replicated or generalized free energy, taking into account fluctuations across many local minima with free energy density $f_\alpha$ in the free energy landscape. This is the so-called one-step replica symmetry breaking (1RSB) scenario (later explained mathematically in detail in the last section of this chapter). Accordingly, we have $e^{-y\Phi} = \sum_\alpha Z_\alpha^m$, where the original partition function of the $\alpha$th state is weighted by a power $m$, which is thus called the Parisi RSB parameter or Parisi parameter [4]. It then follows that one can interpret $y$ as a product of $\beta m$, which we shall discuss in detail later.

In the thermodynamic limit, the intractable integral in Eq. (9.1) can be estimated by the Laplace approximation, resulting in

$$-y\phi = \max_f \{\Sigma(f) - yf\}, \tag{9.2}$$

$$y = \frac{\partial \Sigma(f)}{\partial f}. \tag{9.3}$$

$\phi$ denotes the replicated free energy density (i.e., per spin). By a Legendre transform, one obtains the following identities:

$$f = \frac{\partial(y\phi)}{\partial y}, \tag{9.4}$$

$$\Sigma = y(f - \phi) = y^2 \frac{\partial \phi}{\partial y}. \tag{9.5}$$

$\phi$ can be estimated by the cavity method at the 1RSB level. More precisely, by adding one variable node (e.g., spin) into the original factor graph of the model, and assuming a one-to-one correspondence among the pure states (at least with the lowest free energy) before and after the cavity operations (including also the operation of adding a function node, e.g., an interaction), we then have

$$\begin{aligned} e^{-y\phi_i^{\text{new}}} &= \sum_{\alpha} e^{-yF^{\alpha}-y\Delta F_i^{\alpha}} = e^{-y\phi^{\text{old}}} \sum_{\alpha} \omega(\alpha) e^{-y\Delta F_i^{\alpha}} \\ &= e^{-y\phi^{\text{old}}} \langle e^{-y\Delta F_i} \rangle, \end{aligned} \tag{9.6}$$

where the weight $\omega(\alpha) = \frac{e^{-yF^{\alpha}}}{\sum_a e^{-yF^{\alpha}}}$, and the angular bracket indicates an average over all equilibrium states.

Analogously, we have the contribution of adding a function node as follows:

$$e^{-y\phi_a^{\text{new}}} = e^{-y\phi^{\text{old}}} \langle e^{-y\Delta F_a} \rangle. \tag{9.7}$$

Therefore, the replicated free energy shift due to both cavity operations can be summarized as follows:

$$-y\Delta\phi_i = \ln\langle e^{-y\Delta F_i} \rangle, \tag{9.8}$$

$$-y\Delta\phi_a = \ln\langle e^{-y\Delta F_a} \rangle, \tag{9.9}$$

where $\Delta F_i$ and $\Delta F_a$ are the free energy shifts under the cavity operations, and can be estimated within each pure state, thereby having the same form with the RS theory. Finally, applying the Bethe approximation at the 1RSB level, the replicated free energy can be constructed by collecting two parts, given by

$$\phi = \sum_i \Delta\phi_i - \sum_a (|\partial a| - 1)\Delta\phi_a, \tag{9.10}$$

where $|\partial a|$ denotes the degree of the function node $a$ in the factor graph. The free energy density and the complexity can be derived based on Eqs. (9.4) and (9.5)

$$f = \frac{\langle \Delta F_i e^{-y\Delta F_i} \rangle}{\langle e^{-y\Delta F_i} \rangle} - \sum_a (|\partial a| - 1) \frac{\langle \Delta F_a e^{-y\Delta F_a} \rangle}{\langle e^{-y\Delta F_a} \rangle}, \tag{9.11}$$

$$\Sigma = y(f - \phi). \tag{9.12}$$

The mean-field spin glass models can be classified into two distinct categories. One is the SK model, where the low temperature phase can be described as an ultrametric hierarchy of states, or mathematically a full replica symmetry breaking (fRSB), explained in detail later. The transition to the spin glass phase is of a second order, accompanying a diverging correlation length. The other class is the p-spin ($p > 2$) models or discontinuous glass models [5]. The transition to the spin glass

phase is still second order (no latent heat) in the Ehrenfest sense, but the order parameter (e.g., Edwards–Anderson order parameter) jumps at the transition, being of the first-order characteristic. In spin glass theory, this transition is named the random first-order transition [6]. The 1RSB scheme is known to be correct for the $p$-spin spherical model [7, 8], where spin takes spherically-constrained continuous values, but for the general case of the $p$-spin Ising model, the spin glass phase may have a fRSB structure that occurs at a very low temperature [9]. Many complex systems, including structural glasses[1] and constraint satisfaction problems, fall in this category, sharing many interesting properties [4, 6, 10].

In a typical example of discontinuous spin glass models, there exists a maximal value of free energy such that $\Sigma(f_{\max})$ determines the number of metastable states, so-called threshold states trapping most local algorithms, e.g., simulated annealing [11]. At the other end, $\Sigma(f_{\rm gs}) = 0$ determines the lower-bound estimate of the ground state with the free energy $f_{\rm gs}$, corresponding to the maximum of $\Phi(y)$ at the 1RSB level. However, the 1RSB scheme becomes unstable for the free energy above $f_{\rm G}$, which is the Gardner energy where the fRSB scheme of a hierarchy of nested states should be assumed. As a classic example, the Ising $p$-spin glass undergoes a first discontinuous transition from a paramagnetic to a 1RSB phase at a relatively high temperature, and then a continuous transition to a fRSB phase as the temperature is lowered down to the Gardner temperature [5]. In addition, we have the following relationship $f_{\rm gs} \leq f_{\rm G} \leq f_{\max}$ [12].

## 9.2 Applications to Constraint Satisfaction Problems

At the 1RSB level, the cavity method can be classified into two cases, depending on different focuses on the probability measure of thermodynamics. We first introduce the energetic cavity method [13]. In this case, as mentioned above, we can write the 1RSB re-weighting parameter $y$ as a product $\beta m$. Then we obtain

$$-\beta m\phi(\beta, m) = \max_{s,\epsilon}\{\Sigma(s,\epsilon) + m(s - \beta\epsilon)\}, \tag{9.13}$$

where $s$ and $\epsilon$ denotes the entropy density and energy density, respectively. Taking the limit $\beta \to \infty$ and $m \to 0$ while keeping a finite value of $y$, we get

$$\phi_\epsilon(y) = \max_{\epsilon}\{\Sigma(\epsilon) - y\epsilon\}. \tag{9.14}$$

Note that the limit $\beta \to \infty$ is the zero temperature limit commonly took in an optimization problem to search for ground states of the model. $\Sigma(\epsilon)$ determines the number of clusters of configurations with the energy density $\epsilon$, and moreover its zero-value determines a SAT threshold, e.g., in random $K$-SAT problem (see Chap. 2).

[1] Many interacting particles move with a local random environment for each particle.

In a constraint satisfaction problem, each solution can be treated as an equilibrium configuration of a traditional statistical mechanics model. These solutions may be grouped into exponentially many clusters [14], and each cluster can be called a pure state, thus their statistics can be captured by the 1RSB scheme. In a SAT regime, where a solution satisfying all boolean constraints exists, or the ground sate energy remains zero, an optimal value of $y$ tends to be $\infty$. However, when the SAT threshold is crossed from below, the optimal $y$ takes immediately a finite value [15].

The energetic cavity method at the $y \to \infty$ (also $m = 0$) makes the Gibbs measure concentrate on the ground state with $\epsilon = 0$ (SAT configurations), leading to an efficient fully-distributed algorithm, namely survey propagation [13] for the random $K$-SAT problem. This algorithm goes beyond the standard belief propagation iteration that does not work when the constraint density (the number of boolean constraints or clauses per variable) is larger than some threshold (still below the SAT one). The salient feature is that, the cluster-to-cluster fluctuation is explicitly taken into account in this advanced algorithm, which tells us the exact probability of picking up a cluster randomly and finding a given variable frozen to one direction within that cluster [13]. The survey propagation can thus solve the NP-hard problem[2] in typical cases up to a threshold very close to the SAT threshold. This intriguing property inspires many following up works in other kinds of constraint satisfaction problems [4]. However, the algorithm does not work on the random $K$-XOR SAT problem (see the first chapter), due to the freezing effects in clusters of solutions [16]. In addition, the algorithm requires a high computational complexity in optimization problems where the ground state energy is non-zero.

If we focus our measure only on the SAT regime, i.e., $\epsilon = 0$, then we can shift our interest to the entropy part. This kind of cavity method is thus called the entropic one. It is easy to write first that

$$\Sigma(\epsilon = 0) = \max_s \Sigma(s, \epsilon = 0) = \Sigma(m = 0), \tag{9.15}$$

i.e., the energetic ($m = 0$) cavity method computes the complexity of the typical or most numerous clusters, and clusters are weighted equally independent of their sizes. This can be seen from the fact that $\frac{\partial \Sigma(s)}{\partial s} = -m$. Generalizing to the entropic case, we have

$$\phi(m) = \max_s \{\Sigma(s) + ms\}, \tag{9.16}$$

and by the Legendre transform

$$s = \frac{\partial \phi(m)}{\partial m}, \qquad \Sigma(s) = \phi(m) - ms. \tag{9.17}$$

[2] Whether the NP class is distinct or not from the P class that is solvable in polynomial time remains an open problem in mathematics.

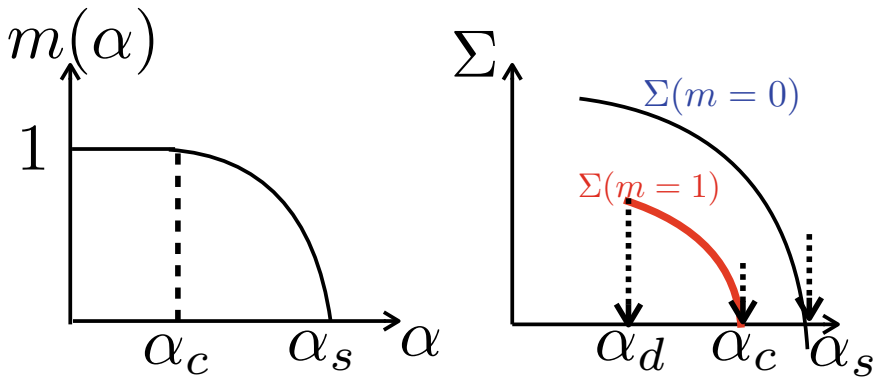



**Fig. 9.2** Schematic illustration of the Parisi parameter and the complexity of states. $\alpha$ denotes the constraint density (the number of constraints per degree of freedom) in a constraint satisfaction problem

The relationship $m = -\frac{\partial \Sigma(s)}{\partial s}$ can be easily derived under the saddle-point approximation. It is physically clear that a given value of $m$ selects the size of states (or clusters), like the temperature parameter selecting the configuration at the RS level. Note that, to detect if a 1RSB solution emerges in a model, a first test is to verify the appearance of a non-trivial solution of the 1RSB equation at $m = 1$ (Fig. 9.2). At the corresponding threshold, the point-to-set correlation function [14], an average of the correlation between a randomly chosen variable and a variable set at a distance $\ell$ from it, sets in discontinuously for a discontinuous transition or continuously for a continuous transition. This threshold is thus called the dynamical transition point ($\alpha_d$), many local algorithms (by local move—a few variables are changed at each step, like Monte Carlo algorithms) are affected by this transition, due to thermodynamically relevant (entropically dominant) clusters (at $m = 1$) prevail. Note that the local stability of the RS solution coincides exactly with the dynamical threshold for the continuous 1RSB transition. However, the local instability occurs after a discontinuous transition [14]. A non-trivial ergodicity breaking takes place at the dynamical transition, leading to impossible uniform sampling of solutions or equilibrium configurations after this transition.

Further, increasing the constraint density, the thermodynamic value of $m$ will start to decrease at the condensation threshold $\alpha_c$ (Fig. 9.2), where the Gibbs measure condensates on a few or sub-exponential (with the number of degrees of freedom) number of clusters, i.e., the equilibrium value of $m$ is determined by $\Sigma(m_{\rm eq}) = 0$. Depending on the specific problem, there may appear a freezing transition where the thermodynamically dominant clusters contain a finite fraction of variables frozen into the same specified direction [16]. The freezing transition at $\alpha_f$ forms an algorithmic barrier where a large-scale rearrangement of variables required for going from one cluster to another one. For example, in random $K$-XOR SAT, $\alpha_d = \alpha_f$, and all clusters have the same size, and for random 3-SAT problems, $\alpha_c = \alpha_d$.

In the 1RSB phase, cavity marginals fluctuate from one state (or cluster) to another one. In general, a 1RSB equation can be written into a compact form as

$$P(m_{i\to a}) = \frac{1}{\mathcal{Z}_{i\to a}} \prod_{b\in\partial i\setminus a} \int \mathrm{d}\hat{m}_{b\to i}\delta\left(m_{i\to a} - \mathcal{F}(\{\hat{m}_{b\to a}\})\right) Z^m_{i\to a}, \tag{9.18}$$

where the RS message $m_{i\to a}$ is now turned into a probability function, i.e., the survey of messages among states (or clusters), and $\mathcal{F}$ denotes the RS iteration, which holds within each state, and the cavity partition function $Z_{i\to a}$ is now weighted in a power $m$, acting as a statistical weight in a Monte Carlo sampling—$e^{-\beta m \Delta F_{i\to a}}$ where the free energy shift under the cavity operation $\Delta F_{i\to a}$ can be constructed from the RS theory. This re-weighting term discourages moving into states with high free energy [17, 18], in a similar way to a standard Monte Carlo sampling where a high energy state is highly undesired during the process of searching for low-energy configurations. The 1RSB iteration [Eq. (9.18)] can be derived from a variational principle on the 1RSB free energy function $\Phi(\beta, m)$ [19, 20], with respect to the functional order parameter (the probability measure over the messages) and the Parisi parameter $m$ or $y$. We lastly remark that if the distribution $P(m_{i\to a})$ does not peak on a few isolated values, then it cannot be parameterized by a few real numbers, thereby making it impossible to derive an efficient algorithm like survey propagation.

Now, let us analyze the special case of $m = 1$. This special Parisi parameter greatly simplifies the complex 1RSB equation, and make a numerical solution of the equation using population dynamics [14, 17, 19] much less time-demanding. Population dynamics is a special numerical techniques using a population of random variables (being updated) to represent a probability distribution, particularly suitable for solving the 1RSB equation that is a recursive probability function equation. In this special case, we have

$$\phi(\beta, m = 1) = \epsilon - \frac{\Sigma(s,\epsilon) + s}{\beta} = \epsilon - T s_{\mathrm{tot}}. \tag{9.19}$$

In the dynamical 1RSB regime, where the complexity $\Sigma(m = 1)$ is positive, the RS marginal probability and the free energy remains asymptotically exact. Moreover, the correct total entropy deviates from the RS one only in the condensation phase or phases after it. More precisely, $s_{\mathrm{tot}} = s^*(\Sigma(s^*) = 0) < s_{\mathrm{RS}}$ in this regime [4]. From this sense, the dynamical transition is not a genuine transition. When $\Sigma(m)$ vanishes, a genuine, or ideal glass transition occurs, namely the Kauzmann transition [21]. At this transition, the free energy has a discontinuity in its second derivative.

In some systems, there exists a frozen phase, e.g., in the random energy model, or the binary perceptron learning problem [22, 23]. Therefore, for the $\alpha$th state, $f_\alpha = \epsilon_\alpha$, we have

$$e^{-N\beta m\phi(\beta,m)} = \sum_\alpha e^{-N\beta m f_\alpha} = \sum_\alpha e^{-N\beta m \epsilon_\alpha} = e^{-N\beta m f_{\mathrm{RS}}(\beta m)}, \tag{9.20}$$

from which, we can define an inverse temperature $\beta_s$ where the entropy vanishes, $s_{\mathrm{RS}}(\beta_s) = 0$, i.e., when $\beta = \beta_s$, $m = \frac{\beta_s}{\beta} = 1$. As the temperature further decreases ($m$ decreases as well), the free energy is clamped to its zero-entropy value, like that occurs in the random energy model. In this kind of models, the zero-entropy condition is used to solve the entropy crisis, i.e., the free energy shows a maximum at a finite temperature [18, 24]. In addition, the RS instability usually takes place after the entropy crisis, there thus must exist a discontinuous transition before or at the zero-entropy point [25].

## 9.3 More Steps of Replica Symmetry Breaking

As we know, the RS solution may be incorrect if long-range correlations emerge in the system, e.g., the point-to-set correlation does not decay to zero [26]. This is also called the sufficient condition. A necessary condition for the correct RS solution is the non-divergence of the spin glass susceptibility $\chi_{\mathrm{SG}}$, defined as

$$\chi_{\mathrm{SG}} = \frac{1}{N} \sum_{i,j} \overline{(\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle)^2}, \tag{9.21}$$

where the angular brackets mean the thermal average and the overline means the disordered average over model parameters. When these conditions are not satisfied, high levels of approximation must be introduced, e.g., 1RSB, in other words, a small perturbation breaks the replica symmetry, in accord with the (de Aleida-Thouless) AT stability analysis within the replica scheme [27], i.e., via a perturbation analysis in the replica space around the symmetric order parameters. If the 1RSB solution is

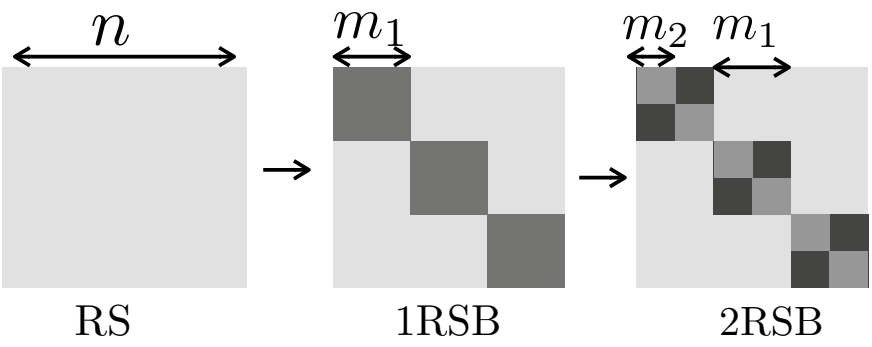



**Fig. 9.3** Schematic illustration of how the overlap matrix changes as more advanced approximations are introduced. $n$ denotes the number of replicas, while $m_i$ denotes the size of subblocks when considering a hierarchy of replica symmetry breaking

unstable against further perturbations, either in terms of small changes in distributions or in terms of messages, like at a Gardner temperature, more levels of replica symmetry breaking are required. In general, a 2-RSB theory involves an order parameter that is a distribution of distributions, and correspondingly, states can aggregate into different clusters (inter-state susceptibility diverges), or each state can further split into different states (intra-state susceptibility diverges) [25]. These susceptibilities depend on different ways of handling the overlarge over the states with their Boltzmann weights.

The overlap matrix in the replica theory can be interpreted in the matrix of the overlap between pure states $a$ and $b$ [28–31], defined by

$$Q_{ab} = \frac{1}{N}\sum_i \langle\sigma_i\rangle_a \langle\sigma_i\rangle_b, \tag{9.22}$$

where $\langle\bullet\rangle$ represents the thermal average within that state. In terms of the Parisi ansätz, the 1RSB corresponds to the $n$ replicas divided into $n/m_1$ identical clusters of size $m_1$. Within each cluster, the permutation symmetry of replicas still holds. Then we need two order parameters, $q_0$ and $q_1$ ($q_0 < q_1$) at the 1RSB level. In general, for a $k$-step RSB, we have $1 = m_{k+1} < m_k < \cdots < m_0 = n$, i.e., by adding one level, the diagonal block is further divided into $m_i/m_{i+1}$ subblocks, for each block $q_{i+1}$ is assigned (Fig. 9.3). In a mathematical form, we have

$$Q_{ab} = q_i, \quad \text{if} \quad \left\lceil \frac{a}{m_i} \right\rceil = \left\lceil \frac{b}{m_i} \right\rceil \quad \text{and} \quad \left\lceil \frac{a}{m_{i+1}} \right\rceil \neq \left\lceil \frac{b}{m_{i+1}} \right\rceil, \tag{9.23}$$

where $\lceil x \rceil$ takes the smallest integer not larger than $x$. $\{q_i\}$ then form a sequence of order parameters representing similarity of states in the hierarchical organization of the phase space. The permutation symmetry among replicas is clearly broken across different blocks. By an analytic continuation to $n \to 0$, as usually adopted in the last step of replica calculation, the relationship among $m_i$ becomes $0 = m_0 < m_1 < \cdots < m_{k+1} = 1$. An observable measure reflecting replica symmetry effects is defined by the realization-dependent distribution

$$P_J(q) = \sum_{a,b} \omega_a \omega_b \delta(q_{ab} - q), \tag{9.24}$$

where $J$ represents the model disorder, and $\omega_a$ denotes the $a$th state's statistical weight described as above. We remark that this distribution is not self-averaging, i.e., its profile depends on the specific realization of the model. It is thus unlike the free energy (being of the self-averaging property), as the system size increase, the fluctuation of the free energy value will be minimized, such that the free energy for a single instance roughly matches the typical value obtained by replica theory.

Taking another limit $k \to \infty$, one obtains the fRSB solution, where $q_i$ transforms to a continuous function $q(x)$, ranged as $q \in [0, q_{\max}]$. In the fRSB phase, pure states are organized according to an ultrametric structure. Note that the Edwards–Anderson

order parameter $q_{\rm EA} = \max_x q(x)$. The inverse function $x(q)$ is just the cumulative distribution

$$x(q) = \int_0^q dq' P(q'), \tag{9.25}$$

which gives the probability of observing an overlap less than or equal to $q$. Therefore, we have $0 \leq x \leq 1$. Note that $P(q) = \overline{P_J(q)} = \frac{dx(q)}{dq}$. In this way, the hierarchical clustering of replicas can be interpreted in a physical picture of pure states. RSB effects were investigated in the machine learning models of restricted Boltzmann machines [32]. Ultrametric structures in the state space were also revealed. The ultrametric property of overlaps among three states implies that two smallest overlaps are equal [1].

We finally remark that in the replica theory, the 1RSB free energy is always larger than the RS one, with increasing levels of RSB, the lower bound to the true free energy improves [33, 34]. A fRSB solution of the SK model, proposed by Giorgio Parisi, was proved to be rigorous in mathematics [35].

## References


1. M. Mézard, G. Parisi, M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987)
2. J. Ardelius, L. Zdeborova, Phys. Rev. E **78**, 040101 (2008)
3. R. Monasson, Phys. Rev. Lett. **75**(15), 2847 (1995)
4. M. Mézard, A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009)
5. E. Gardner, Nucl. Phys. **257**(6), 747 (1985)
6. M. Mezard, Phys. A-Stat. Mech. Appl. **306**, 25 (2002)
7. A. Crisanti, H.J. Sommers, Zeitschrift fur Physik B Condensed Matter **87**(3), 341 (1992)
8. A. Crisanti, H. Horner, H.J. Sommers, Zeitschrift fur Physik B Condensed Matter **92**(2), 257 (1993)
9. A. Montanari, F. Ricci-Tersenghi, Eur. Phys. J. B **33**(3), 339 (2003)
10. V. Lubchenko, Adv. Phys. **64**(3), 283 (2015)
11. S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science **220**(4598), 671 (1983)
12. A. Crisanti, L. Leuzzi, G. Parisi, T. Rizzo, Phys. Rev. Lett. **92**(12), 127203 (2004)
13. M. Mezard, G. Parisi, R. Zecchina, Science **297**(5582), 812 (2002)
14. A. Montanari, F. Ricci-Tersenghi, G. Semerjian, J. Stat. Mech.: Theory Exper. **2008**, 04004 (2008)
15. M. Mézard, R. Zecchina, Phys. Rev. E **66**(5), 056126 (2002)
16. G. Semerjian, J. Stat. Phys. **130**(2), 251 (2007)
17. H. Zhou, Phys. Rev. E **77**, 066102 (2008)
18. H. Huang, Commun. Theor. Phys. **63**(1), 115 (2015)
19. M. Mézard, G. Parisi, Eur. Phys. J. B **20**, 217 (2001)
20. F. Concetti (2019). arXiv:1911.00557
21. W. Kauzmann, Chem. Rev. **43**(2), 219 (1948)
22. W. Krauth, M. Mezard, J. Phys. **50**(20), 3057 (1989)
23. H. Huang, Y. Kabashima, Phys. Rev. E **90**, 052813 (2014)
24. O.C. Martin, M. Mezard, O. Rivoire, J. Stat. Mech.: Theory Exper. **2005**(9), 9006 (2005)
25. O. Rivoire, G. Biroli, O.C. Martin, M. Mézard, Eur. Phys. J. B - Conden. Matter Complex Syst. **37**, 55 (2004)

26. M. Mézard, A. Montanari, J. Stat. Phys. **124**(6), 1317 (2006)
27. J.R.L. de Almeida, D.J. Thouless, J. Phys. A **11**(5), 983 (1978)
28. G. Parisi, Phys. Lett. A **73**(3), 203 (1979)
29. G. Parisi, Phys. Rev. Lett. **43**(23), 1754 (1979)
30. G. Parisi, J. Phys. A **13**(3), 1101 (1980)
31. G. Parisi, J. Phys. A **13**(L115) (1980)
32. G.S. Hartnett, E. Parker, E. Geist, Phys. Rev. E **98**, 022116 (2018)
33. F. Guerra, Commun. Math. Phys. **233**(1), 1 (2003)
34. S. Franz, M. Leone, J. Stat. Phys. **111**, 535 (2003)
35. M. Talagrand, Ann. Math. **163**(1), 221 (2006)

# Chapter 10
# Statistical Mechanics of Restricted Boltzmann Machine


Energy-based model is an archetypal type of generative model, which can learn any distribution of data and generate new samples that follow the same distribution as the original one. In this chapter, two kinds of energy-based models are introduced—Boltzmann machine (BM) and restricted Boltzmann machine (RBM). The learning method of maximizing log-likelihoods is introduced and statistical mechanics analysis of restricted Boltzmann machines is performed. The free energy of RBMs is calculated based on the Bethe approximation. Then thermodynamic quantities related to learning, e.g., magnetizations as well as hidden-visible correlations are also derived, providing an alternative efficient way to train RBMs with continuous weights. In this chapter, we also introduce a powerful physics-inspired algorithm for training RBMs with discrete weights, which was previously thought to be out of reach until a very recent work (Huang in Phys. Rev. E 102:030301(R), 2020 [5]). Training RBMs plays an important role at the early stage of deep learning (Bengio et al. in Advances in Neural Information Processing Systems, pp. 153–160, 2007 [9]).

## 10.1 Boltzmann Machine

Boltzmann machine (BM) is an energy-based model, as shown in Fig. 10.1. It is an unsupervised learning network with the following energy:

$$E(\boldsymbol{\sigma}) = -\sum_i h_i \sigma_i - \sum_{i<j} w_{ij} \sigma_i \sigma_j, \tag{10.1}$$

where $\sigma_i = \pm 1$ is the state of node $i$, $h_i$ is the bias of node $i$, and $w_{ij}$ is the connection weight between node $i$ and node $j$. Configurations of $\boldsymbol{\sigma}$ obey the Boltzmann distribution

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{\sigma})}, \tag{10.2}$$



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_10

where $Z = \sum_{\boldsymbol{\sigma}} e^{-\beta E(\boldsymbol{\sigma})}$ is the partition function. To learn a given data set by a BM is known as the inverse Ising problem (see Chap. 3).

The Hopfield model can be considered as a special type of BM with $h_i = 0, \forall i$, and the coupling is constructed in the Hebbian rule. Generally speaking, the Hebbian rule is not enough to learn any distribution of data. Therefore, to learn the distribution of a given data set with $M$ configurations, $\{\boldsymbol{\sigma^1}, \boldsymbol{\sigma^2} \ldots, \boldsymbol{\sigma^M}\}$, the weights of a BM network are updated by maximizing the log-likelihood of the data

$$\begin{aligned} L(\boldsymbol{\theta}|\{\boldsymbol{\sigma}\}) &= \langle \log(p_{\boldsymbol{\theta}}(\boldsymbol{\sigma}))\rangle_{\text{data}} \\ &= -\langle E(\boldsymbol{\sigma}, \boldsymbol{\theta})\rangle_{\text{data}} - \log Z(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} h_i \langle \sigma_i \rangle_{\text{data}} + \sum_{i<j} w_{ij} \langle \sigma_i \sigma_j \rangle_{\text{data}} - \log Z(\boldsymbol{\theta}), \end{aligned} \tag{10.3}$$

where $\langle \ldots \rangle_{\text{data}}$ means the average carried out over the data, $\boldsymbol{\theta}$ denotes the parameters $\{\boldsymbol{W}, \boldsymbol{h}\}$, $p_{\boldsymbol{\theta}}(\boldsymbol{\sigma})$ is the distribution of $\boldsymbol{\sigma}$ with parameters $\boldsymbol{\theta}$, and $\beta$ is set to be 1 for convenience, as $\beta$ can be absorbed in both weights and biases. The gradient of $L(\boldsymbol{\theta}|\{\boldsymbol{\sigma}\})$ can be easily computed as follows:

$$\begin{aligned} \frac{\partial L}{\partial h_i} &= \langle \sigma_i \rangle_{\text{data}} - \langle \sigma_i \rangle_{\text{model}}; \\ \frac{\partial L}{\partial w_{ij}} &= \langle \sigma_i \sigma_j \rangle_{\text{data}} - \langle \sigma_i \sigma_j \rangle_{\text{model}}, \end{aligned} \tag{10.4}$$

where $\langle \ldots \rangle_{\text{model}}$ denotes the thermal average under the model measure. Network parameters can then be updated by gradient ascent

$$\begin{aligned} \Delta h_i &= \eta \frac{\partial L}{\partial h_i} = \eta(\langle \sigma_i \rangle_{\text{data}} - \langle \sigma_i \rangle_{\text{model}}); \\ \Delta w_{ij} &= \eta \frac{\partial L}{\partial w_{ij}} = \eta(\langle \sigma_i \sigma_j \rangle_{\text{data}} - \langle \sigma_i \sigma_j \rangle_{\text{model}}), \end{aligned} \tag{10.5}$$

where $\eta$ is the learning rate. The first terms of both equations in Eq. (10.5) are easy to compute by averaging over the data. But the model average terms are intractable to compute, as the computation of the partition function requires $O(2^N)$ time complexity. In practice, Monte Carlo method can give an approximate value of the model average terms. An alternative simple way is the mean field theory introduced in Chap. 3, which can be applied to calculate $\langle \sigma_i \rangle_{\text{model}}$, and then the two-point correlation can be obtained by using the linear-response theory: $\frac{\partial \langle \sigma_i \rangle_{\text{model}}}{\partial h_j} = \langle \sigma_i \sigma_j \rangle_{\text{model}} - \langle \sigma_i \rangle_{\text{model}} \langle \sigma_j \rangle_{\text{model}}$.

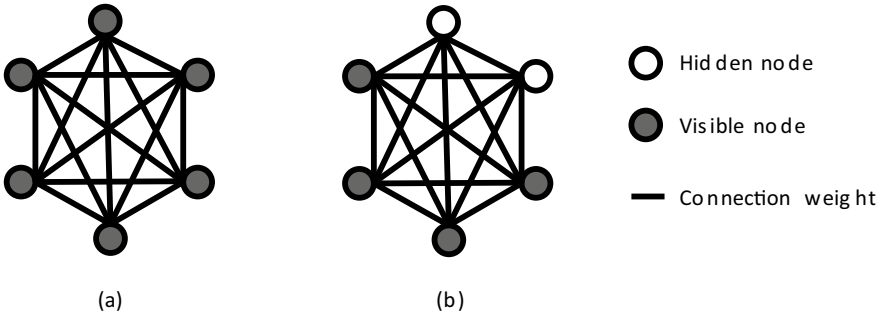



**Fig. 10.1** **a** Illustration of BM without hidden nodes. **b** Illustration of BM with hidden nodes

In the standard BM, hidden nodes can also be introduced by providing a latent encoding of the visible nodes (their number equals to the dimension of the data samples), as shown in Fig. 10.1b. However, learning the parameters involved in the hidden nodes is typically computationally demanding, as a Monte Carlo sampling of states of hidden nodes is required. This computational barrier motivates an alternative architecture, namely restricted Boltzmann machine to appear, which greatly reduces the computational cost by removing the connections among visible nodes (and hidden nodes) [1, 2].

## 10.2 Restricted Boltzmann Machine

The architecture of RBM is shown in Fig. 10.2. The hidden nodes play a role of an encoder of sensory inputs. The RBM has one hidden layer and one visible layer. Connections only exist between layers yet not within each layer. For the RBM network with $N$ visible nodes and $M$ hidden nodes, the energy function that the learning tries to minimize is given by:

$$E(\boldsymbol{\sigma}, \boldsymbol{s}) = -\sum_{i,a} \sigma_i w_{ia} s_a - \sum_i \phi_i \sigma_i - \sum_a h_a s_a, \tag{10.6}$$

where $\sigma_i$ is the state of visible node $i$ with bias $\phi_i$, $s_a$ is the state of hidden node $a$ with bias $h_a$ and $w_{ia}$ is the connection between them. The network state obeys the Boltzmann distribution

$$p(\boldsymbol{\sigma}, \boldsymbol{s}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{\sigma}, \boldsymbol{s})}, \tag{10.7}$$

where $Z = \sum_{\boldsymbol{\sigma}, \boldsymbol{s}} e^{-\beta E(\boldsymbol{\sigma}, \boldsymbol{s})}$. Here, we set $\beta = 1$, in that the temperature effect could be absorbed into the inferred couplings and biases. Nodes in the same layer are conditionally independent due to the absence of the lateral connections, and the conditional probability is specified by

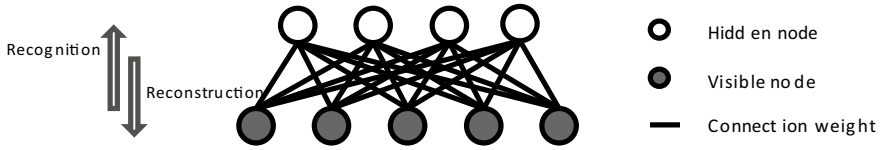



**Fig. 10.2** Schematic illustration of a RBM. The RBM has one hidden layer and one visible layer. Connections are only allowed between layers. The recognition process is defined as sampling hidden states given visible states. The reconstruction process is defined as sampling visible states given hidden states

$$
\begin{aligned}
p\left(\sigma_i|\boldsymbol{s}\right) &= \frac{\sum_{\{\sigma_j:j\neq i\}} p(\boldsymbol{\sigma},\boldsymbol{s})}{\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma},\boldsymbol{s})} \\
&= \frac{e^{\sigma_i(\phi_i+\sum_a w_{ia}s_a)}}{e^{\sigma_i(\phi_i+\sum_a w_{ia}s_a)}+e^{-\sigma_i(\phi_i+\sum_a w_{ia}s_a)}} \\
&= \frac{1}{1+e^{-2\sigma_i(\phi_i+\sum_a w_{ia}s_a)}}; \\
p\left(s_a|\boldsymbol{\sigma}\right) &= \frac{\sum_{\{s_b:b\neq a\}} p(\boldsymbol{\sigma},\boldsymbol{s})}{\sum_{\boldsymbol{s}} p(\boldsymbol{\sigma},\boldsymbol{s})} \\
&= \frac{e^{s_a(h_a+\sum_i w_{ia}\sigma_i)}}{e^{s_a(h_a+\sum_i w_{ia}\sigma_i)}+e^{-s_a(h_a+\sum_i w_{ia}\sigma_i)}} \\
&= \frac{1}{1+e^{-2s_a(h_a+\sum_i w_{ia}\sigma_i)}}.
\end{aligned}
\tag{10.8}
$$

Given states of visible nodes, the states of hidden nodes can be sampled easily, which is called the recognition process; and the converse process is called the reconstruction.

Similar to BM, the weights of RBM can be learned by maximizing the data log-likelihood. Given a data set, $\{\boldsymbol{\sigma^1}, \boldsymbol{\sigma^2}, \ldots, \boldsymbol{\sigma^M}\}$, the log-likelihood is formulated as

$$
\begin{aligned}
\mathcal{L}\left(\boldsymbol{\theta}|\{\boldsymbol{\sigma}\}\right) &= \left\langle \log\left(p_{\boldsymbol{\theta}}(\boldsymbol{\sigma})\right)\right\rangle_{\text{data}} \\
&= -\left\langle E\left(\boldsymbol{\sigma},\boldsymbol{\theta}\right)\right\rangle_{\text{data}} - \log Z\left(\{\boldsymbol{\theta}\}\right),
\end{aligned}
\tag{10.9}
$$

where $\boldsymbol{\theta}$ denotes the parameters $\{\boldsymbol{W}, \boldsymbol{\phi}, \boldsymbol{h}\}$ for convenience, and $p_{\boldsymbol{\theta}}(\boldsymbol{\sigma})$ is the distribution of $\boldsymbol{\sigma}$ with the parameters $\boldsymbol{\theta}$. The gradient of the parameters can be easily obtained as

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\{w_{ia},\phi_i,h_a\})}{\partial w_{ia}} &= \langle \sigma_i s_a\rangle_{\text{data}} - \langle \sigma_i s_a\rangle_{\text{model}}; \\
\frac{\partial \mathcal{L}(\{w_{ia},\phi_i,h_a\})}{\partial \phi_i} &= \langle \sigma_i\rangle_{\text{data}} - \langle \sigma_i\rangle_{\text{model}}; \\
\frac{\partial \mathcal{L}(\{w_{ia},\phi_i,h_a\})}{\partial h_a} &= \langle s_a\rangle_{\text{data}} - \langle s_a\rangle_{\text{model}}.
\end{aligned}
\tag{10.10}
$$

Nevertheless, the model average terms still require $O(2^{N+M})$ time complexity to compute. An efficient method is the well-known contrastive-divergence (CD) algorithm that performs an alternating Gibbs sampling [via Eq. (10.8)] that starts from the data samples [3]. For saving computation time, CD is usually truncated to a few Gibbs sampling steps, e.g., one step.

In next sections, we shall show that both the model average terms and the free energy of the system can be obtained analytically by performing the Bethe approximation, which allows us to understand the statistical mechanics of RBM [4]. For the sake of simplicity, we consider a random RBM with the property that all of $w_{ia}$ obey an i.i.d Gaussian distribution with zero mean and variance $\frac{g}{N}$, and biases for each layer also obey an i.i.d. Gaussian distribution with zero mean and variance $v$.

## 10.3 Free Energy Calculation

Using the conditional independence, we can derive the exact form of the marginal probability of visible nodes

$$\begin{aligned} p(\boldsymbol{\sigma}) &= \sum_{\boldsymbol{s}} p(\boldsymbol{\sigma}, \boldsymbol{s}) \\ &= \frac{1}{Z} \sum_{\boldsymbol{s}} e^{\sum_a (\sum_i \beta \sigma_i w_{ia} + \beta h_a) s_a + \sum_i \beta \sigma_i \phi_i} \\ &= \frac{1}{Z} e^{\sum_i \beta \sigma_i \phi_i} \sum_{\boldsymbol{s}} \prod_a e^{(\sum_i \beta \sigma_i w_{ia} + \beta h_a) s_a} \\ &= \frac{1}{Z} \prod_i e^{\beta \sigma_i \phi_i} \prod_a \sum_{s_a} e^{(\sum_i \beta \sigma_i w_{ia} + \beta h_a) s_a} \\ &= \frac{1}{Z} \prod_i e^{\beta \sigma_i \phi_i} \prod_a [2 \cosh(\beta \boldsymbol{w_a} \boldsymbol{\sigma} + \beta h_a)], \end{aligned} \tag{10.11}$$

where $\boldsymbol{w_a}$ denotes the weight vectors connecting to the hidden node $a$. A direct calculation of the partition function $Z$ requires a time complexity of $O(2^N)$, which becomes impossible as $N$ increases. Here, the Bethe approximation can be applied as a first-level approximation of the free energy. A factor graph can represent the current system after the marginalization [Eq. (10.11)]. As displayed in Fig. 10.3, the $i$th circle denotes the variable node $\sigma_i$, and the $a$th square denotes a Boltzmann factor $2\cosh(\beta \boldsymbol{w_a} \boldsymbol{\sigma} + \beta h_a)$, which corresponds to $f_a(\boldsymbol{x}_a)$ introduced in Chap. 3. $\beta$ is set to be 1 as explained above.

We then introduce a cavity probability $P_{i \to a}(\sigma_i)$, which denotes the probability of the state of the variable node $i$ in the absence of a factor node $a$, together with auxiliary quantity $\mu_{b \to i}$, summing the contribution of factor node $b$ when the variable node $i$ is frozen to the state $\sigma_i$. According to the belief propagation formula, we have

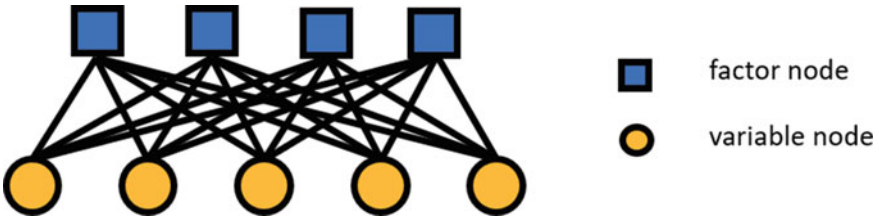



**Fig. 10.3** Factor graph representation of a RBM. Circles denote variable nodes and squares denote factor nodes

the following recursive equations:

$$P_{i\to a}(\sigma_i) = \frac{1}{Z_{i\to a}} e^{\phi_i \sigma_i} \prod_{b\in \partial i\backslash a} \mu_{b\to i}(\sigma_i); \tag{10.12a}$$

$$\mu_{b\to i}(\sigma_i) = \sum_{\{\sigma_j | j\in \partial b\backslash i\}} 2\cosh\left(\boldsymbol{w_b}\boldsymbol{\sigma} + h_b\right) \prod_{j\in\partial b\backslash i} P_{j\to b}\left(\sigma_j\right), \tag{10.12b}$$

where $Z_{i\to a} = e^{\phi_i} \prod_{b\in\partial i\backslash a} \mu_{b\to i}(+1) + e^{-\phi_i} \prod_{b\in\partial i\backslash a} \mu_{b\to i}(-1)$ is a normalization constant, $\partial i\backslash a$ represents the neighbors of variable node $i$ except factor node $a$, $\partial b\backslash i$ represents the neighbors of the factor node $b$ except the visible node $i$. Unfortunately, the sum in the second equation still needs $O(2^{N-1})$ time complexity, and resolving this difficulty requires further approximations.

Note that $\mu_{b\to i}$ estimates the mean of the Boltzmann factor $2\cosh\left(\boldsymbol{w_b}\boldsymbol{\sigma} + h_b\right)$ over the configuration $\{\sigma_j | j \in \partial b\backslash i\}$. Under the Bethe approximation, $\sigma_j$ around the function node ($b$ here) is approximately independent, provided that the cavity probability $P_{j\to a}(\sigma_j)$ is defined. As $\mathcal{U}_{b\to i} \equiv \sum_{j\in\partial b\backslash i} w_{jb}\sigma_j$ is the sum of $(N-1)$ variables, and the variables $\{\sigma_j\}$ are assumed to be nearly independent under the Bethe approximation, the central limit theorem (CLT) thus suggests that $\mathcal{U}_{b\to i}$ obeys a Gaussian distribution given a large value of $N$. The mean and variance of $\mathcal{U}_{b\to i}$ are, respectively, given by

$$\begin{aligned} G_{b\to i} &= \langle \mathcal{U}_{b\to i}\rangle_{\{\sigma_j | j\in\partial b\backslash i\}} = \sum_{j\in\partial b\backslash i} w_{jb} m_{j\to b}; \\ \Xi^2_{b\to i} &= \langle \mathcal{U}^2_{b\to i}\rangle_{\{\sigma_j | j\in\partial b\backslash i\}} - \langle \mathcal{U}_{b\to i}\rangle^2_{\{\sigma_j | j\in\partial b\backslash i\}} \\ &\simeq \sum_{j\in\partial b\backslash i} w^2_{jb}\left(1 - m^2_{j\to b}\right), \end{aligned} \tag{10.13}$$

where $m_{j\to b} \equiv \sum_{\sigma_j} \sigma_j P_{j\to b}\left(\sigma_j\right)$ is the cavity magnetization. According to the weakly correlated state assumption (the same spirit as the RS ansätz), a diagonal approximation is further applied to simplify the variance, i.e., only the sum of diagonal elements of the correlation matrix is calculated. Then $\mu_{b\to i}(\sigma_i)$ can be approx-

imately calculated as follows:

$$\begin{aligned}\mu_{b\to i}(\sigma_i) &= 2\int Dt\cosh(G_{b\to i}+\sqrt{\Xi_{b\to i}^2}t+h_b+w_{ib}\sigma_i)\\ &= 2e^{\frac{\Xi_{b\to i}^2}{2}}\cosh(G_{b\to i}+h_b+w_{ib}\sigma_i),\end{aligned} \tag{10.14}$$

where $Dt \equiv e^{-t^2/2}/\sqrt{2\pi}dt$. Inserting this result into the cavity probability $P_{i\to a}(\sigma_i)$, we obtain the cavity magnetization

$$\begin{aligned}m_{i\to a} &= \sum_{\sigma_j}\sigma_j P_{j\to b}(\sigma_j)\\ &= \frac{\sum_{\sigma_i}\sigma_i e^{\phi_i\sigma_i}\prod_{b\in\partial i\backslash a}\mu_{b\to i}(\sigma_i)}{\sum_{\sigma_i}e^{\phi_i\sigma_i}\prod_{b\in\partial i\backslash a}\mu_{b\to i}(\sigma_i)}\\ &= \tanh\left(\phi_i+\sum_{b\in\partial i\backslash a}u_{b\to i}\right);\\ u_{b\to i} &= \frac{1}{2}\ln\frac{\mu_{b\to i}(+1)}{\mu_{b\to i}(-1)} = \frac{1}{2}\ln\frac{\cosh(h_b+G_{b\to i}+w_{ib})}{\cosh(h_b+G_{b\to i}-w_{ib})},\end{aligned} \tag{10.15}$$

where $u_{b\to i}$ is the cavity bias (see Chap. 2). $m_{i\to a}$ represents the massage passing from variable node $i$ to factor node $a$, and $u_{b\to i}$ denotes the massage passing from factor node $b$ to variable node $i$. Iterating Eq. (10.15) can reach the fixed point. Then, the Bethe free energy can be calculated as follows:

$$\begin{aligned}F &= \sum_i F_i-(N-1)\sum_a F_a;\\ F_i &= -\ln Z_i = -\ln(e^{\phi_i}\prod_{b\in\partial i}\mu_{b\to i}(+1)+e^{-\phi_i}\prod_{b\in\partial i}\mu_{b\to i}(-1));\\ F_a &= -\ln Z_a = -\ln(2e^{\frac{\Xi_a^2}{2}}\cosh(G_a+h_a)),\end{aligned} \tag{10.16}$$

where $F_i$ and $F_a$ are local free energies of variable node $i$ and factor node $a$, respectively, $\Xi_a = \sum_{j\in\partial a}w_{ja}^2\left(1-m_{j\to a}^2\right)$, and $G_a = \sum_{j\in\partial a}w_{ja}m_{j\to a}$. The computation of $F_a$ is similar to that of $\mu_{a\to i}$. Here, we show an experiment result of the free energy computation via the Bethe approximation (Fig. 10.4).

## 10.4 Thermodynamic Quantities Related to Learning

To learn a RBM, computation of model averages in Eq. (10.10) is required, and the Bethe approximation can then be applied. After getting the fixed points of Eq. (10.15),

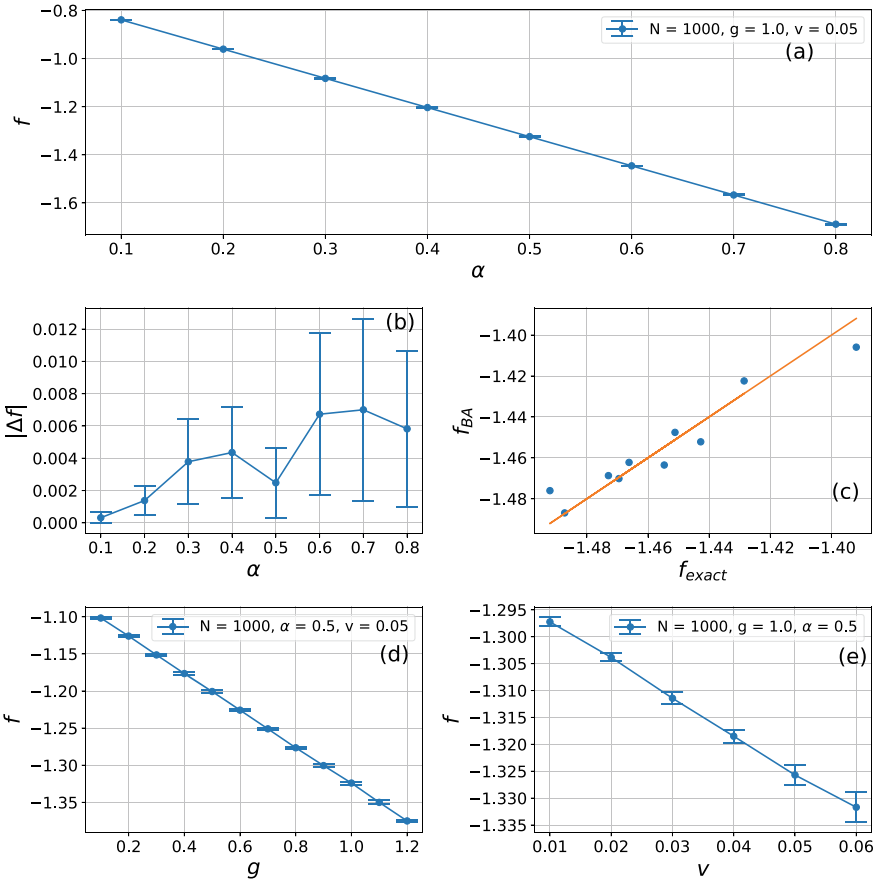



**Fig. 10.4** Free energy density ($f = F/N$) of random RBMs. The error bar is the standard deviation over ten trials. **a** $f$ versus $\alpha = M/N$ with $N = 1000$, $g = 1$, and $v = 0.05$. **b** The absolute difference between exact free energy (calculated by enumeration) and Bethe approximation of different $\alpha$ with N = 20, $g = 1$, and v = 0.05. **c** Comparison of exact free energy and Bethe approximation for $\alpha = 0.6$ in (**b**). The diagonal line denotes $f_{BA} = f_{exact}$. **d** $f$ versus $g$ with $N = 1000$, $v = 0.05$ and $\alpha = 0.5$. **e** $f$ versus $v$ with $N = 1000$, $g = 1$ and $\alpha = 0.5$

the magnetization of a visible node $i$, i.e., $m_i = \langle \sigma_i \rangle_{\text{model}}$, can be obtained as follows:

$$m_i = \tanh\left(\phi_i + \sum_{b \in \partial i} u_{b \to i}\right). \tag{10.17}$$

Then, the magnetization of a hidden node, i.e., $\hat{m}_a = \langle s_a \rangle_{\text{model}}$, can be calculated as follows:

$$\begin{aligned}\hat{m}_a &= \langle s_a \rangle_{\text{model}} = \langle \tanh\Big(\sum_i \sigma_i w_{ia} + h_a\Big)\rangle_{\boldsymbol{\sigma}} \\ &= \int Dx \tanh\left(\sqrt{\tilde{\Xi}_a^2}x + \tilde{G}_a\right),\end{aligned} \tag{10.18}$$

where $\tilde{G}_a = \sum_{i\in\partial a} w_{ia} m_i + h_a$, and $\tilde{\Xi}_a^2 \simeq \sum_{j\in\partial a} w_{ja}^2 \left(1 - m_j^2\right)$. Note that the sum in the mean and variance here involves in full magnetizations instead of cavity ones, which is different from those in Eq. (10.16).

To update the connection weights, correlations between hidden and visible nodes should be calculated

$$\mathbf{C}_{ai} = \langle s_a \sigma_i \rangle_{\text{model}} = \left\langle \tanh\Big(\sum_i \sigma_i w_{ia} + h_a\Big)\sigma_i \right\rangle_{\boldsymbol{\sigma}}, \tag{10.19}$$

where the average over $\boldsymbol{s}$ can be performed exactly due to the conditional independence. To calculate the intractable correlation matrix $\mathbf{C}$, we can first calculate the matrix-product $\mathbf{CW}$

$$[\mathbf{CW}]_{ab} = \left\langle \tanh\Big(\sum_i \sigma_i w_{ia} + h_a\Big) \sum_j \sigma_j w_{jb} \right\rangle. \tag{10.20}$$

We can then define $\mathcal{U}_a \equiv \sum_i w_{ia}\sigma_i$, and $\mathcal{U}_a$ that obeys a Gaussian distribution due to the CLT. The mean is given by $\sum_i w_{ia} m_i$. The covariance matrix of $\boldsymbol{\mathcal{U}} = (\mathcal{U}_1, \mathcal{U}_2, \ldots \mathcal{U}_a, \ldots)$ is defined as $\boldsymbol{\Delta}$, whose element $\Delta_{ab}$ is given by

$$\begin{aligned}\Delta_{ab} &= \langle \mathcal{U}_a \mathcal{U}_b \rangle - \langle \mathcal{U}_a \rangle \langle \mathcal{U}_b \rangle \\ &= \left\langle \sum_i w_{ia}\sigma_i \sum_j w_{jb}\sigma_j \right\rangle - \left\langle \sum_i w_{ia}\sigma_i \right\rangle \left\langle \sum_j w_{jb}\sigma_j \right\rangle \\ &= \sum_i w_{ia} \sum_j w_{jb} \left\langle \sigma_i \sigma_j \right\rangle - \sum_i w_{ia} \sum_j w_{jb} \left\langle \sigma_i \right\rangle \left\langle \sigma_j \right\rangle \\ &= \sum_i w_{ia} \sum_j w_{jb} \hat{C}_{ij},\end{aligned} \tag{10.21}$$

where $\hat{\mathbf{C}}$ is the covariance matrix of $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ and the entry $\hat{C}_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$. Hence the matrix $\boldsymbol{\Delta}$ can be expressed as

$$\boldsymbol{\Delta} = \mathbf{W}^{\mathrm{T}} \hat{\mathbf{C}} \mathbf{W}. \tag{10.22}$$

In particular, the diagonal element of $\boldsymbol{\Delta}$ is given by

$$\Delta_{aa} = \langle \mathcal{U}_a^2 \rangle - \langle \mathcal{U}_a \rangle^2 \simeq \sum_{i \in \partial a} w_{ia}^2 \left(1 - m_i^2\right), \tag{10.23}$$

which is the same as $\tilde{\Xi}_a$.

To proceed, we re-parametrize $\mathcal{U}_a = \sqrt{\Delta_{aa} - \Delta_{ab}} x + \sqrt{\Delta_{ab}} z + \langle \mathcal{U}_a \rangle$, and $\mathcal{U}_b = \sqrt{\Delta_{bb} - \Delta_{ab}} y + \sqrt{\Delta_{ab}} z + \langle \mathcal{U}_b \rangle$, which guarantees the covariance structure. In this parameterization, $x, y, z$ are independent random variables obeying the standard Gaussian distribution. Then the matrix-product $\mathbf{CW}$ can be calculated as

$$\begin{aligned}
\mathbf{CW} t_{ab} &= \langle \tanh(\mathcal{U}_a + h_a) \mathcal{U}_b \rangle \\
&= \langle \tanh(\mathcal{U}_a + h_a)(\mathcal{U}_b - \langle \mathcal{U}_b \rangle) \rangle + \langle \tanh(\mathcal{U}_a + h_a) \rangle \langle \mathcal{U}_b \rangle \\
&= \int Dx\, Dy\, Dz \tanh(\sqrt{\Delta_{aa} - \Delta_{ab}} x + \sqrt{\Delta_{ab}} z + \langle \mathcal{U}_a \rangle + h_a) \\
&\times (\sqrt{\Delta_{bb} - \Delta_{ab}} y + \sqrt{\Delta_{ab}} z) + \hat{m}_a \langle \mathcal{U}_b \rangle \\
&= \sqrt{\Delta_{ab}} \int Dx\, Dz \tanh(\sqrt{\Delta_{aa} - \Delta_{ab}} x + \sqrt{\Delta_{ab}} z + \tilde{G}_a) z + \left\langle \sum_i \sigma_i w_{ib} \right\rangle \hat{m}_a \\
&= \Delta_{ab} \int Dx\, Dz (1 - \tanh^2(\sqrt{\Delta_{aa} - \Delta_{ab}} x + \sqrt{\Delta_{ab}} z + \tilde{G}_a)) + \sum_i m_i w_{ib} \hat{m}_a,
\end{aligned} \tag{10.24}$$

where $Dx \equiv e^{-x^2/2}/\sqrt{2\pi} dx$, $Dy \equiv e^{-y^2/2}/\sqrt{2\pi} dy$, $Dz \equiv e^{-z^2/2}/\sqrt{2\pi} dz$, and the integral identity $\int Dz \tanh(z) z = \int Dz \tanh'(z)$ has been applied. Note that $\sqrt{\Delta_{aa} - \Delta_{ab}} x + \sqrt{\Delta_{ab}} z$ obeys a Gaussian distribution $\mathcal{N}(0, \Delta_{aa})$. The equation can be re-parameterized in a simpler form

$$[\mathbf{CW}]_{ab} = \Delta_{ab} \int Dt (1 - \tanh^2(\sqrt{\Delta_{aa}} t + \tilde{G}_a)) + \sum_i m_i w_{ib} \hat{m}_a, \tag{10.25}$$

where $Dt \equiv e^{-t^2/2}/\sqrt{2\pi} dt$. We can then rewrite the above equation to a matrix form for convenience. Here, we define a diagonal matrix $\mathbf{A}$ with diagonal elements $\mathbf{A}_{aa} = \int Dx (1 - \tanh^2(\sqrt{\Delta_{aa}} x + \tilde{G}_a))$, $\boldsymbol{m} = (m_1, m_2, \ldots, m_N)$ that is a matrix of dimension $1 \times N$, and $\hat{\boldsymbol{m}} = (\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_M)$ that is a matrix of dimension $1 \times M$. The matrix form of Eq. (10.25) is given by

$$\mathbf{CW} = \mathbf{A} \mathbf{W}^{\mathrm{T}} \hat{\mathbf{C}} \mathbf{W} + \hat{\boldsymbol{m}}^{\mathrm{T}} \boldsymbol{m} \mathbf{W}. \tag{10.26}$$

From the above equation, we can immediately obtain

$$\mathbf{C} = \mathbf{A} \mathbf{W}^{\mathrm{T}} \hat{C} + \hat{\boldsymbol{m}}^{\mathrm{T}} \boldsymbol{m}. \tag{10.27}$$

Each element of $\mathbf{C}$ can thus be read off

$$C_{ai} = A_{aa} \sum_j w_{ja} \hat{\mathbf{C}}_{ji} + \hat{m}_a m_i. \tag{10.28}$$

Considering the week correlation assumption, we apply the diagonal approximation that $\sum_j w_{ja}\hat{C}_{ij} \simeq w_{ia}\hat{C}_{ii} = w_{ia}(1-m_i^2)$. In sum, the model terms can now be obtained as follows:

$$\begin{aligned} m_i &= \tanh\left(\phi_i + \sum_b u_{b\to i}\right), \\ \hat{m}_a &= \int Dx \tanh\left(\sqrt{\tilde{\Xi}_a^2}x + \tilde{G}_a\right), \\ C_{ai} &= A_{aa}w_{ia}(1-m_i^2) + \hat{m}_a m_i, \\ A_{aa} &= \int Dx(1-\tanh^2(\sqrt{\tilde{\Xi}_a^2}x + \tilde{G}_a)), \end{aligned} \tag{10.29}$$

where $\tilde{G}_a = \sum_{i\in\partial a} w_{ia}m_i + h_a$, and $\tilde{\Xi}_a^2 \simeq \sum_{j\in\partial a} w_{ja}^2\left(1-m_j^2\right)$.

To evaluate the equilibrium properties of a random RBM model, BA requires $O(nMN)$ time complexity to compute the thermal average, where $n$ is the number of iteration steps before convergence and usually less then 100 (not around the transition point), while the CD-$k$ algorithm requires $O(kTMN)$ [4], where $T$ is the number of data samples, and $k$ is the length of the Gibbs sampling chain.

## 10.5 Stability Analysis

Bethe approximation requires that variance of weights and hidden-node density of the network, $\alpha = M/N$, should be small enough to ensure the visible nodes around cavity function nodes are statistically independent. With increasing weight-strength and hidden-node density, i.e., $g$ and $\alpha$ become larger, the Bethe approximation will be not self-consistent. It is thus necessary to perform the stability analysis to draw out the boundary within which the approximation works.

Considering a weak perturbation in the message from node $i$ to node $a$ as $\delta_{m_{i\to a}}$, we can write the actual passing message as $m_{i\to a} + \delta_{m_{i\to a}}$. Following the information flow in Fig. 10.5, we can write the recursive equation for the perturbation as follows [4]:

$$\delta_{m_{i\to a}} = \sum_{b\in\partial i\setminus a;\, j\in\partial b\setminus i} \frac{\partial m_{i\to a}}{\partial m_{j\to b}}\delta_{m_{j\to b}}. \tag{10.30}$$

To remove the sign dependence of the perturbation, we define the magnitude as the squared perturbation as

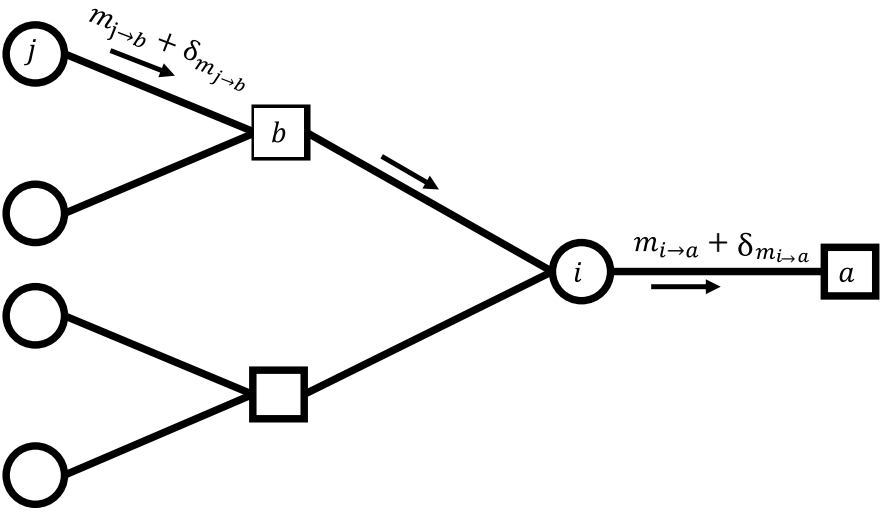



**Fig. 10.5** Propagation of message-perturbations through the factor graph of a RBM

$$\begin{aligned}\delta^2_{m_{i\to a}} &= \left(\sum_{b\in\partial i\backslash a;\, j\in\partial b\backslash i}\frac{\partial m_{i\to a}}{\partial m_{j\to b}}\delta_{m_{j\to b}}\right)^2 \\ &\simeq \sum_{b\in\partial i\backslash a}\sum_{j\in\partial b\backslash i}\left(\frac{\partial m_{i\to a}}{\partial m_{j\to b}}\right)^2(\delta_{m_{j\to b}})^2,\end{aligned} \tag{10.31}$$

where we ignore the correlation for different $\delta_{m_{j\to b}}$. Here, we use $\mathcal{V}_{i\to a}=\delta^2_{m_{i\to a}}$ to denote the strength. Using Eqs. (10.15), (10.31) can be simplified as

$$\mathcal{V}_{i\to a}=\frac{\left(1-m^2_{i\to a}\right)^2}{4}\sum_{b\in\partial i\backslash a}\mathcal{P}_{b\to i}\times\left[\tanh\left(\Gamma_{b\to i}\right)-\tanh\left(\Gamma_{b\to i}-2w_{bi}\right)\right]^2, \tag{10.32}$$

where $\Gamma_{b\to i}\equiv h_b+G_{b\to i}+w_{ib}$, and $\mathcal{P}_{b\to i}=\sum_{j\in\partial b\backslash i}w^2_{jb}\mathcal{V}_{j\to b}$. The total variance is defined by $S(t)=\sum_{(i,a)}\mathcal{V}_{i\to a}(t)$, where $t$ denotes the iteration step of Eq. (10.32)). To monitor the stability, we define $\lambda=S(t^c+1)/S(t^c)$, where $t^c$ denotes the time step when the iteration converges or reaches a prescribed maximal iteration number. If $\lambda>1$, the total variance will increase, thereby causing the instability of the message passing equation. When we perform the BA approximation, we should thus choose the suitable values of $g$ and $\alpha$ to ensure $\lambda\leq 1$. As Fig. 10.6 shows, the Bethe approximation becomes instable around $g=2.1$ with $N=1000$, $\alpha=0.5$ and $v=0.05$.

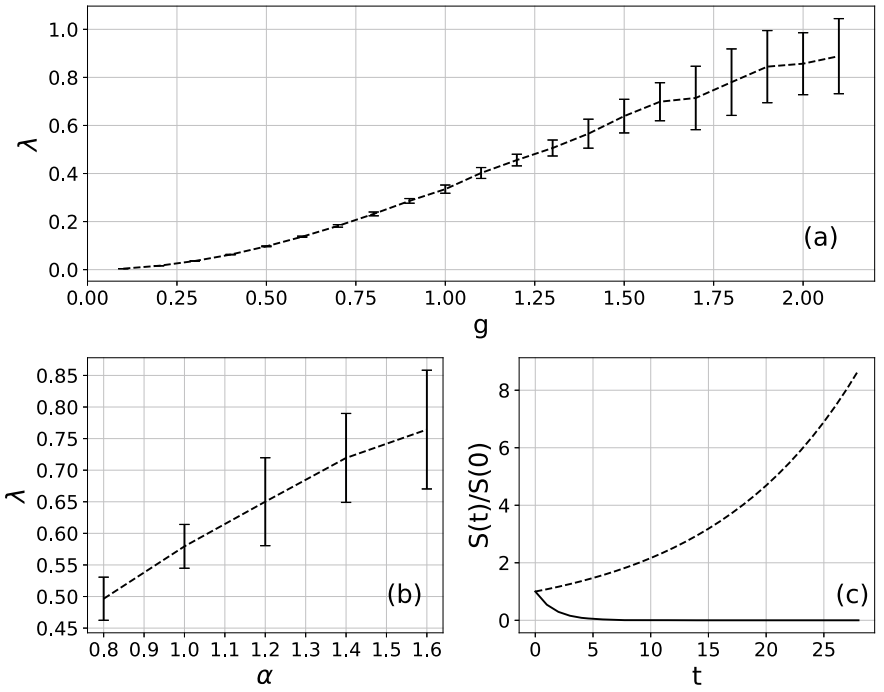



**Fig. 10.6** Stability analysis of random RBMs. The error bar is the standard deviation over 20 random realizations of the model. **a** Stability parameter $\lambda$ versus $g$ with $N = 1000$, $v = 0.05$ and $\alpha = 0.5$. $\lambda$ increases with $g$. When $g = 2.1$, $\lambda$ is near to the critical point ($\lambda = 1$). Some instances are unstable, and the others are stable as (**c**) shows. **b** $\lambda$ versus $\alpha$ with $N = 1000$, $v = 0.05$ and $g = 1$. $\lambda$ increases with $\alpha$. **c** Two instances of instability (dashed line) and stability (solid line) with $g = 2.1$, $N = 1000$, $\alpha = 0.5$ and $v = 0.05$

## 10.6 Variational Mean-Field Theory for Training Binary RBMs

In a restricted Boltzmann machine with continuous weights, maximizing the log-likelihood through computing gradient ascent can be applied in training process. However, in the binary RBM, the method fails in that the differentiation with respect to the binary weights is ill-defined. Here, we introduce a variational principle that maximizes the lower bound to the data log-likelihood, which results in a new algorithm combining message passing with gradient ascents [5].

### 10.6.1 RBMs with Binary Weights

Here, we consider a RBM with $N$ visible nodes and $P$ hidden nodes. The provided dataset consists of $M$ configurations, $\{\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2, \ldots, \boldsymbol{\sigma}^M\}$. Each configuration consists of a set of binary spin $\boldsymbol{\sigma} = \{\sigma_i = \pm 1\}_{i=1}^N$. The synaptic weights are denoted as $\boldsymbol{\xi}$, where $\xi_i^\mu = \pm 1$ denoting the synapse between visible node $i$ and hidden node $\mu$. The energy of such a RBM is thus given by $E(\boldsymbol{\sigma}, \boldsymbol{h}) = \sum_{i,\mu} \sigma_i \xi_i^\mu h_\mu$, where $h_\mu = \pm 1$ is the state of hidden node $\mu$. The Boltzmann distribution of visible nodes is then given by

$$P(\boldsymbol{\sigma}) = \frac{1}{Z(\boldsymbol{\xi})} \prod_\mu \cosh\left(\beta X_\mu\right), \tag{10.33}$$

where we denote a short-hand notation $X_\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{\xi}^\mu \cdot \boldsymbol{\sigma}$ , $\boldsymbol{\xi}^\mu$ is the vector of weights connecting to the $\mu$th hidden node, which is also called the receptive field of that hidden node, and the partition function reads $Z(\boldsymbol{\xi}) = \sum_{\boldsymbol{\sigma}} \prod_\mu \cosh\left(\beta X_\mu\right)$. The scaling factor $\frac{1}{\sqrt{N}}$ is added to $X_\mu$ to ensure that it is of the order $O(1)$. $\beta$ is the inverse temperature tuning the noise level of the input data [6].

In this model, we consider a weakly correlated data set which can be generated by sampling the planted model with a long Monte Carlo interval. This is also called the i.i.d data sample assumption widely used in deep learning community. Thus, the probability of a weakly correlated data set $\{\boldsymbol{\sigma}^a\}_{a=1}^M$ is modeled by

$$P\left(\left\{\boldsymbol{\sigma}^a\right\}_{a=1}^M | \boldsymbol{\xi}\right) = \prod_{a=1}^M P\left(\boldsymbol{\sigma}^a | \boldsymbol{\xi}\right) = \prod_{a=1}^M \frac{1}{Z(\boldsymbol{\xi})} \prod_\mu \cosh\left(\beta X_\mu^a\right), \tag{10.34}$$

where $X_\mu^a \equiv \frac{1}{\sqrt{N}} \boldsymbol{\xi}^\mu \cdot \boldsymbol{\sigma}^a$. According to the Bayes' rule, the posterior probability of synaptic weights given the raw data is

$$\begin{aligned} P\left(\boldsymbol{\xi} | \left\{\boldsymbol{\sigma}^a\right\}_{a=1}^M\right) &= \frac{P\left(\{\boldsymbol{\sigma}^a\}_{a=1}^M | \boldsymbol{\xi}\right) P(\boldsymbol{\xi})}{P\left(\{\boldsymbol{\sigma}^a\}_{a=1}^M\right)} \\ &= \frac{\prod_a P\left(\boldsymbol{\sigma}^a | \boldsymbol{\xi}\right) P(\boldsymbol{\xi})}{\sum_{\boldsymbol{\xi}} \prod_a P\left(\boldsymbol{\sigma}^a | \boldsymbol{\xi}\right) P(\boldsymbol{\xi})} \\ &= \frac{1}{\Omega} \exp\left(-M \ln Z(\boldsymbol{\xi}) + \sum_{a,\mu} \ln\cosh\left(\beta X_\mu^a\right)\right), \end{aligned} \tag{10.35}$$

where the partition function of the posterior probability is given by

$$\Omega = \sum_{\boldsymbol{\xi}} \exp\left(-M \ln Z(\boldsymbol{\xi}) + \sum_{a,\mu} \ln\cosh\left(\beta X_\mu^a\right)\right), \tag{10.36}$$

and $P(\boldsymbol{\xi})$ is assumed to be uniformly distributed, or no prior knowledge is assumed. Here, $\beta$ is a hyper-parameter during the training process. The nested partition function $Z(\boldsymbol{\xi})$ is intractable, let alone the posterior partition function $\Omega$. Computation of the nested partition function $\Omega$ with one or two hidden nodes will be introduced in next chapters. In the case of $P \geq 3$, analysis and calculation of $\Omega$ become extremely challenging. In the following, a training algorithm based on the variational principle will be introduced and moreover is applicable for any $P$.

### *10.6.2 Variational Principle*

Instead of analyzing the posterior, a variational principle tries to find an approximate distribution close to the exact learning posterior [7]. We define a variational distribution as $q_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ and the Kullback–Leibler(KL) divergence between $q_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ and $P\left(\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}_{a=1}^M\right)$ is used to measure how accurate the variational distribution is

$$\begin{aligned}\mathrm{KL}\left(q_\lambda(\boldsymbol{\xi})\|P(\boldsymbol{\xi}|\mathcal{D})\right) &= \mathbb{E}_q \ln\left(\frac{q_{\boldsymbol{\lambda}}(\boldsymbol{\xi})}{P(\boldsymbol{\xi}|\mathcal{D})}\right) = \mathbb{E}_q \ln q_{\boldsymbol{\lambda}}(\boldsymbol{\xi}) - \mathbb{E}_q \ln P(\boldsymbol{\xi}|\mathcal{D}) \\ &= \mathbb{E}_q \ln q_{\boldsymbol{\lambda}}(\boldsymbol{\xi}) - \mathbb{E}_q \ln P(\boldsymbol{\xi}) - \mathbb{E}_q \ln P(\mathcal{D}|\boldsymbol{\xi}) + \ln P(\mathcal{D}) \\ &= \mathrm{KL}\left(q_\lambda(\boldsymbol{\xi})\|P(\boldsymbol{\xi})\right) - \mathbb{E}_q \ln P(\mathcal{D}|\boldsymbol{\xi}) + \ln P(\mathcal{D}) \\ &= -\mathrm{LB}\left(q_{\boldsymbol{\lambda}}\right) + \ln P(\mathcal{D}),\end{aligned} \tag{10.37}$$

where $\mathcal{D}$ denotes $\{\boldsymbol{\sigma}^a\}_{a=1}^M$ for a short-hand notation, $\mathbb{E}_q$ denotes the average over distribution $q$, and $P(\boldsymbol{\xi}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\xi})P(\boldsymbol{\xi})}{P(\mathcal{D})}$. As the KL divergence is nonnegative, the lower bound to the data log-likelihood $P(\mathcal{D})$ is given as follows:

$$\mathrm{LB}\left(q_{\boldsymbol{\lambda}}\right) = \mathbb{E}_q \ln P(\mathcal{D}|\boldsymbol{\xi}) - \mathrm{KL}\left(q_\lambda(\boldsymbol{\xi})\|P(\boldsymbol{\xi})\right). \tag{10.38}$$

The learning process minimizing $\mathrm{KL}\left(q_\lambda(\boldsymbol{\xi})\|P(\boldsymbol{\xi}|\mathcal{D})\right)$ thus amounts to maximizing $\mathrm{LB}\left(q_{\boldsymbol{\lambda}}\right)$. To maximize the lower bound, the first term of expected log-likelihood should be increased, requiring $q_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ to explain the data. The second term is a regularization term pushing $q_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ to approach the prior.

To maximize the LB, we first parameterize $q_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ and $P(\boldsymbol{\xi})$. We assume that the synapses are independent in the prior

$$P(\boldsymbol{\xi}) = \prod_{i,\mu}\left[\frac{1+m_{i\mu}}{2}\delta_{\xi_i^\mu,+1} + \frac{1-m_{i\mu}}{2}\delta_{\xi_i^\mu,-1}\right] = \prod_{i,\mu}\frac{1+\xi_i^\mu m_{i\mu}}{2}, \tag{10.39}$$

where $m_{i\mu}$ corresponds to the mean of $\xi_i^\mu$, and $\delta_{x,y}$ denotes the Kronecker delta function. Similarly, the variational distribution is assumed to have the same form as prior yet with different parameters $\boldsymbol{\lambda}$,

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\xi}) = \prod_{i,\mu} \left[ \frac{1+\lambda_{i\mu}}{2} \delta_{\xi_i^\mu,+1} + \frac{1-\lambda_{i\mu}}{2} \delta_{\xi_i^\mu,-1} \right] = \prod_{i,\mu} \frac{1+\xi_i^\mu \lambda_{i\mu}}{2}. \tag{10.40}$$

Under these two assumed expressions, the KL divergence between $P(\boldsymbol{\xi})$ and $q_{\boldsymbol{\lambda}}$ can be calculated analytically. By substituting the explicit form of $P(\mathcal{D}|\boldsymbol{\xi})$ into the LB, we obtain

$$\mathrm{LB}\,(q_\lambda) = -\mathrm{KL}\,(q_\lambda(\boldsymbol{\xi}) \| P(\boldsymbol{\xi})) + \mathbb{E}_q \left[ \sum_{a,\mu} \ln\cosh\left(\beta X_\mu^a\right) - M \ln Z(\boldsymbol{\xi}) \right]. \tag{10.41}$$

Notice that $X_\mu^a \equiv \frac{1}{\sqrt{N}} \boldsymbol{\xi}^\mu \cdot \boldsymbol{\sigma}^a$. According to the central limit theorem (CLT), $X_\mu^a$ and $X_\mu$ obey a Gaussian distribution with the following mean and variance:

$$\begin{aligned}
G_\mu &= \langle X_\mu \rangle_q = \tfrac{1}{\sqrt{N}} \textstyle\sum_{i\in\partial\mu} \lambda_{i\mu}\sigma_i, \\
\Xi_\mu^2 &= \langle (X_\mu)^2 \rangle_q - (\langle X_\mu \rangle_q)^2 = \tfrac{1}{N} \textstyle\sum_{i\in\partial\mu} \left(1-\lambda_{i\mu}^2\right), \\
G_\mu^a &= \langle X_\mu^a \rangle_q = \tfrac{1}{\sqrt{N}} \textstyle\sum_{i\in\partial\mu} \lambda_{i\mu}\sigma_i^a, \\
\Xi_\mu^2 &= \langle (X_\mu^a)^2 \rangle_q - (\langle X_\mu^a \rangle_q)^2 = \tfrac{1}{N} \textstyle\sum_{i\in\partial\mu} \left(1-\lambda_{i\mu}^2\right).
\end{aligned} \tag{10.42}$$

Then the lower bound can be parameterized as

$$\begin{aligned}
\mathrm{LB}\,(q_\lambda) = &-\mathrm{KL}\,(q_\lambda(\boldsymbol{\xi}) \| P(\boldsymbol{\xi})) + \sum_{a,\mu} \int D\mathbf{z} \ln\cosh\left(\beta G_\mu^a + \beta \Xi_\mu z_\mu\right) \\
&- M \int D\mathbf{z} \ln \sum_{\boldsymbol{\sigma}} \prod_{\mu} \cosh\left(\beta G_\mu + \beta \Xi_\mu z_\mu\right),
\end{aligned} \tag{10.43}$$

where $D\mathbf{z} = \prod_{a,\mu} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z_\mu)^2}{2}} dz_\mu$. To train a binary RBM, gradients of the lower bound w.r.t the variational parameters must be computed. Next, we shall show the calculation of the lower bound and its gradients.

#### 10.6.2.1 Calculation of the Lower Bound

To compute the integral in Eq. (10.43), the Monte Carlo method can be applied. Therefore,

$$\begin{aligned}
\mathrm{LB}\,(q_\lambda) = &-\mathrm{KL}\,(q_\lambda(\boldsymbol{\xi}) \| P(\boldsymbol{\xi})) + \frac{1}{B_1} \sum_{a,\mu,s} \ln\cosh\left(\beta G_\mu^a + \beta \Xi_\mu z_\mu^s\right) \\
&- \frac{M}{B_2} \sum_{s} \ln \sum_{\boldsymbol{\sigma}} \prod_{\mu} \cosh\left(\beta G_\mu + \beta \Xi_\mu z_\mu^s\right),
\end{aligned} \tag{10.44}$$

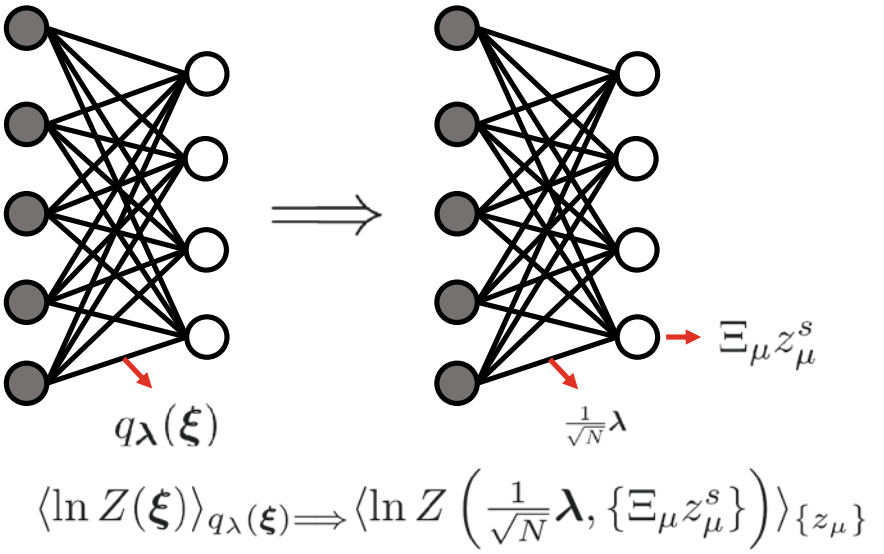



**Fig. 10.7** Calculation of the expected log-partition-function. The expected log-partition function over $q_{\boldsymbol{\lambda}}(\xi_i^\mu)$ is equivalent to the average of log-partition function over a dual RBM ensemble, whose synapses are specified by $\frac{1}{\sqrt{N}}\boldsymbol{\lambda}$, and hidden biases are specified by $\Xi_\mu z_\mu^s$

where $z_\mu^s$ is the $s$th standard Gaussian random number for the Monte Carlo integral, $B_1$ and $B_2$ is are the number of Monte Carlo samplers. In Eq. (10.44), the first two terms can be computed easily. The KL divergence can be directly computed under the parameterized form of $q_\lambda(\boldsymbol{\xi})$ and $P(\boldsymbol{\xi})$

$$
\begin{aligned}
-\mathrm{KL}\left(q_\lambda(\boldsymbol{\xi}) \| P(\boldsymbol{\xi})\right) &= \sum_{\boldsymbol{\xi}} q_{\boldsymbol{\lambda}}(\boldsymbol{\xi}) \ln \left( \prod_{i,\mu} \frac{1+\xi_i^\mu \lambda_{i\mu}}{1+\xi_i^\mu m_{i\mu}} \right) = \sum_{\boldsymbol{\xi}} q_{\boldsymbol{\lambda}}(\boldsymbol{\xi}) \sum_{i,\mu} \ln \left( \frac{1+\xi_i^\mu \lambda_{i\mu}}{1+\xi_i^\mu m_{i\mu}} \right) \\
&= \sum_{i,\mu} \sum_{\xi_i^\mu = \pm 1} q_{\boldsymbol{\lambda}}(\xi_i^\mu) \ln \left( \frac{1+\xi_i^\mu \lambda_{i\mu}}{1+\xi_{i\mu} m_{i\mu}} \right) \\
&= \sum_{x=\pm 1} \sum_{i,\mu} \left[ \mathcal{S}\left( \frac{1+\lambda_{i\mu} x}{2}, \frac{1+m_{i\mu} x}{2} \right) - \mathcal{S}\left( \frac{1+\lambda_{i\mu} x}{2}, \frac{1+\lambda_{i\mu} x}{2} \right) \right],
\end{aligned}
\tag{10.45}
$$

where $q_{\boldsymbol{\lambda}}(\xi_i^\mu) = \frac{1+\xi_i^\mu \lambda_{i\mu}}{2}$ is the variational distribution, and the entropy function $\mathcal{S}(z, y) \equiv z \ln y$.

The second term of the integral can be computed directly. However, the third term of Eq. (10.44) is intractable. But it has the same form as the partition function of a real-valued RBM with continuous weights, whose statistical mechanics properties have been already analyzed in previous sections of this chapter. Then it is necessary to consider a dual RBM ensemble, whose synapses are $\boldsymbol{\lambda}$ scaled by $\sqrt{N}$, and the bias of the $\mu$th hidden neuron is specified by $\Xi_\mu z_\mu^s$ (see Fig. 10.7).

For the dual RBM, the Bethe approximation can be applied to calculate $\ln Z$. As analyzed in the previous sections, the cavity magnetization $m_{i\to\nu}$ and cavity bias $u_{\mu\to i}$ are

$$\begin{aligned} m_{i\to\nu} &= \tanh\left(\sum_{\mu\in\partial i\setminus\nu} u_{\mu\to i}\right), \\ u_{\mu\to i} &= \tanh^{-1}\left(\tanh\left(\beta\chi_{\mu\to i}+\beta H_\mu\right)\tanh\left(\beta\lambda_{i\mu}/\sqrt{N}\right)\right), \end{aligned} \tag{10.46}$$

where $\chi_{\mu\to i}\equiv\frac{1}{\sqrt{N}}\sum_{j\in\partial\mu\setminus i}\lambda_{j\mu}m_{j\to\mu}$ denotes the message sent from a factor node $\mu$, $H_\mu=\Xi_\mu z_\mu^s$ represents the quenched-random hidden bias. The cavity magnetization $m_{i\to\nu}$ represents the message from $i$th visible node to $\nu$th hidden node, and the cavity bias $u_{\mu\to i}$ denotes the message from $\mu$th hidden node to $i$th visible node. After the BP equation converges, the Bethe log-partition function can be calculated as follows:

$$\ln Z=\sum_i F_i-(N-1)\sum_\mu F_\mu, \tag{10.47a}$$

$$\begin{aligned} F_i=&\sum_{\mu\in\partial i}\left[\beta^2\Lambda_{\mu\to i}^2/2+\ln\cosh\left(\beta\chi_{\mu\to i}+\beta H_\mu+\beta\lambda_{i\mu}/\sqrt{N}\right)\right] \\ &+\ln\left(1+\prod_{\mu\in\partial i}e^{-2u_{\mu\to i}}\right), \end{aligned} \tag{10.47b}$$

$$F_\mu=\beta^2\Lambda_\mu^2/2+\ln\cosh\left(\beta\chi_\mu+\beta H_\mu\right), \tag{10.47c}$$

where $\Lambda_{\mu\to i}^2\equiv\frac{1}{N}\sum_{j\in\partial\mu\setminus i}\lambda_{j\mu}^2\left(1-m_{j\to\mu}^2\right)$, $\Lambda_\mu^2\equiv\frac{1}{N}\sum_{j\in\partial\mu}\lambda_{j\mu}^2\left(1-m_{j\to\mu}^2\right)$, and $\chi_\mu\equiv\frac{1}{\sqrt{N}}\sum_{i\in\partial\mu}\lambda_{i\mu}m_{i\to\mu}$. Together with Eqs. (10.44), (10.45), (10.46), (10.47), the lower bound can be computed to measure the impacts of the approximations on the training.

#### 10.6.2.2 Calculation of Gradients

We first calculate the gradients of the first term in the LB [Eq. (10.45)]:

$$-\frac{\partial}{\partial\lambda_{i\mu}}\mathrm{KL}\left(q_\lambda(\xi)\|P(\xi)\right)=\sum_{x=\pm1}\frac{x}{2}\left(\ln\frac{1+xm_{i\mu}}{1+x\lambda_{i\mu}}-1\right). \tag{10.48}$$

If the variational distribution completely matches the prior, this term will vanish. The gradients of the second term of Eq. (10.43) is given by

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda_{i\mu}} \sum_{a,\mu} \int D\mathbf{z} \ln \cosh \left(\beta G_\mu^a + \beta \Xi_\mu z_\mu\right) \\
&= \sum_{a,\mu} \int D\mathbf{z} \left( \frac{\beta \sigma_i^a}{\sqrt{N}} \tanh(\beta G_\mu^a + \beta \Xi_\mu z_\mu) - \frac{\beta \lambda_{i\mu} z_\mu}{N \Xi_\mu} \tanh(\beta G_\mu^a + \beta \Xi_\mu z_\mu) \right) \\
&= \sum_{a,\mu} \int D\mathbf{z} \left( \frac{\beta \sigma_i^a}{\sqrt{N}} \tanh(\beta G_\mu^a + \beta \Xi_\mu z_\mu) - \frac{\beta^2 \lambda_{i\mu}}{N} (1 - \tanh^2(\beta G_\mu^a + \beta \Xi_\mu z_\mu)) \right) \\
&= \frac{\beta}{B_1 \sqrt{N}} \sum_{a,s} \sigma_i^a \tanh \left(\beta G_\mu^a + \beta \Xi_\mu z_\mu^s\right) - \frac{\beta^2 \lambda_{i\mu}}{N B_1} \sum_{a,s} \left[1 - \tanh^2 \left(\beta G_\mu^a + \beta \Xi_\mu z_\mu^s\right)\right].
\end{aligned}
\tag{10.49}
$$

where $\int Dz f(z) z = \int Dz f'(z)$ is applied.

Lastly, the gradient of the expected log-partition function is given by

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda_{i\mu}} \int D\mathbf{z} \ln \sum_{\boldsymbol{\sigma}} \prod_{\mu} \cosh \left(\beta G_\mu + \beta \Xi_\mu z_\mu\right) \\
&= \frac{\beta}{\sqrt{N} B_2} \sum_{s} \left\langle \sigma_i \tanh \left(\beta G_\mu + \beta \Xi_\mu z_\mu^s\right) \right\rangle \\
&- \frac{\beta \lambda_{i\mu}}{N B_2} \sum_{s} \left[ \frac{z_\mu^s}{\Xi_\mu} \left\langle \tanh \left(\beta G_\mu + \beta \Xi_\mu z_\mu^s\right) \right\rangle \right] \\
&= \frac{\beta}{\sqrt{N} B_2} \sum_{s} \left[ C_{i\mu} - \frac{\lambda_{i\mu} z_\mu^s}{\sqrt{N} \Xi_\mu} \hat{m}_\mu \right].
\end{aligned}
\tag{10.50}
$$

where $\langle \ldots \rangle$ denotes the thermal average on the dual RBM. $C_{i\mu}$ and $\hat{m}_\mu$ denote the correlation of visible and hidden nodes, and the magnetization of hidden nodes of the dual model, respectively. We now quote the results as derived in this chapter for estimating the equilibrium properties of the dual RBM, as shown below

$$
\begin{aligned}
m_i &= \tanh \left( \sum_{\mu \in \partial i} u_{\mu \to i} \right), \\
\hat{m}_\mu &= \int Dz \tanh \left( \beta \tilde{\chi}_\mu + \beta H_\mu + \beta \tilde{\Lambda}_\mu z \right), \\
C_{i\mu} &= \hat{m}_\mu m_i + \frac{\beta \lambda_{i\mu}}{\sqrt{N}} \left(1 - m_i^2\right) A_\mu, \\
A_\mu &= 1 - \int Dz \tanh^2 \left( \beta \tilde{\chi}_\mu + \beta H_\mu + \beta \tilde{\Lambda}_\mu z \right),
\end{aligned}
\tag{10.51}
$$

where $Dz \equiv e^{-z^2/2}/\sqrt{2\pi} dz$, $\tilde{\chi}_\mu \equiv \frac{1}{\sqrt{N}} \sum_{i \in \partial \mu} \lambda_{i\mu} m_i$, and $\tilde{\Lambda}_\mu^2 \equiv \frac{1}{N} \sum_{i \in \partial \mu} \lambda_{i\mu}^2 \left(1 - m_i^2\right)$.

To sum up, the final gradients of the lower bound w.r.t the variational parameters are given by

$$\Delta_{i\mu} = \sum_{x=\pm 1} \frac{x}{2} \left( \ln \frac{1 + x m_{i\mu}}{1 + x \lambda_{i\mu}} - 1 \right) + \frac{\beta}{B_1 \sqrt{N}} \sum_{a,s} \sigma_i^a \tanh \left( \beta G_\mu^a + \beta \Xi_\mu z_\mu^s \right) - \frac{\beta^2 \lambda_{i\mu}}{N B_1} \sum_{a,s} \left[ 1 - \tanh^2 \left( \beta G_\mu^a + \beta \Xi_\mu z_\mu^s \right) \right] - \frac{M\beta}{\sqrt{N} B_2} \sum_s \left[ C_{i\mu} - \frac{\lambda_{i\mu} z_\mu^s}{\sqrt{N} \Xi_\mu} \hat{m}_\mu \right]. \tag{10.52}$$

One can then train the network with the following learning rule:

$$\lambda_{i\mu}^{t+1} = \lambda_{i\mu}^t + \eta \Delta_{i\mu}. \tag{10.53}$$

$\boldsymbol{\xi}$ can be decoded as $\boldsymbol{\xi} = \text{sign}(\boldsymbol{\lambda})$, where sign() is the sign function. This decoding is also called the MPM (maximizing the marginal posterior) estimator in statistical inference [8].

### *10.6.3 Experiments*

We first carry out simulations of planted models, where $P$ hidden nodes are assumed. A ground-truth synapses, $\boldsymbol{\xi}^*$ is designed, and then a data set can be generated by Gibbs sampling. A long-time-interval sampling of two consecutive visible configurations is required to ensure the weak correlation of data samples in the synthetic dataset.

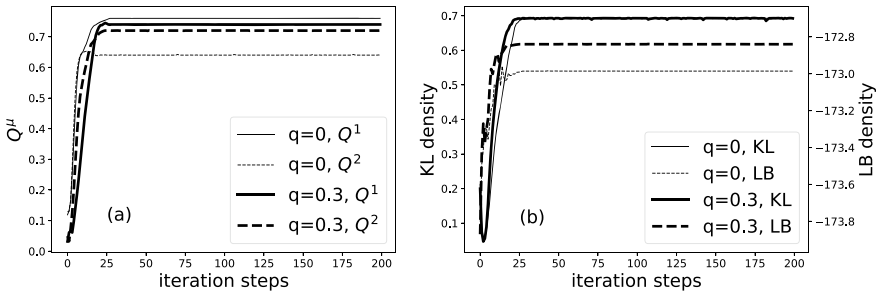



**Fig. 10.8** Training process of two-bit planted RBMs. The number of visible nodes $N = 100$, and the data density $\alpha = \frac{M}{N} = 5$. The x-axis denotes the training epoch. **a** The absolute overlap between decoded receptive field and ground-truth receptive field. $Q^\mu = \frac{1}{N} \sum_i \xi_i^\mu \xi_i^{\mu,*}$. $q = \frac{1}{N} \sum_i \xi_i^{1,*} \xi_i^{2,*}$ is the correlation of the ground-truth synapses. The algorithm works even when the ground-truth synapses are not independent. **b** KL divergence between $q_\lambda(\boldsymbol{\xi})$ and $P(\boldsymbol{\xi})$, and the approximate lower bound based on the cavity approximation. The variational distribution is pushed away from the assumed uniform (apparently incorrect) prior. The LB increases with training epochs, indicating the variational distribution is approaching the true posterior. The density means the corresponding physics value per model parameter

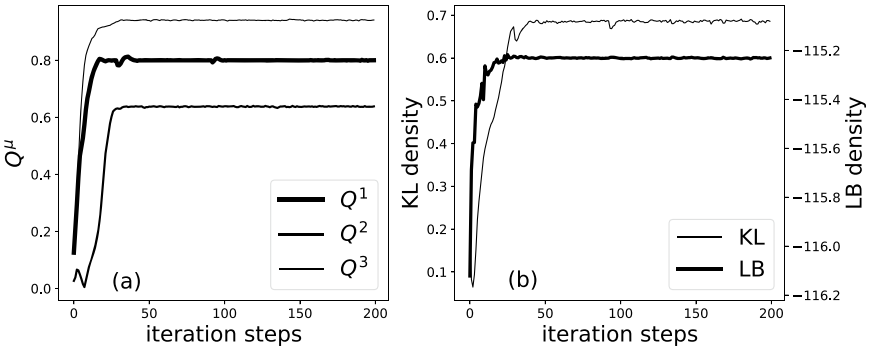



**Fig. 10.9** Training process of three-bit planted RBMs. The number of visible nodes $N = 100$, and the data density $\alpha = 5$. **a** The absolute overlap between decoded receptive field and ground-truth receptive field. **b** KL divergence between $q_\lambda(\boldsymbol{\xi})$ and $P(\boldsymbol{\xi})$, and the approximate lower bound

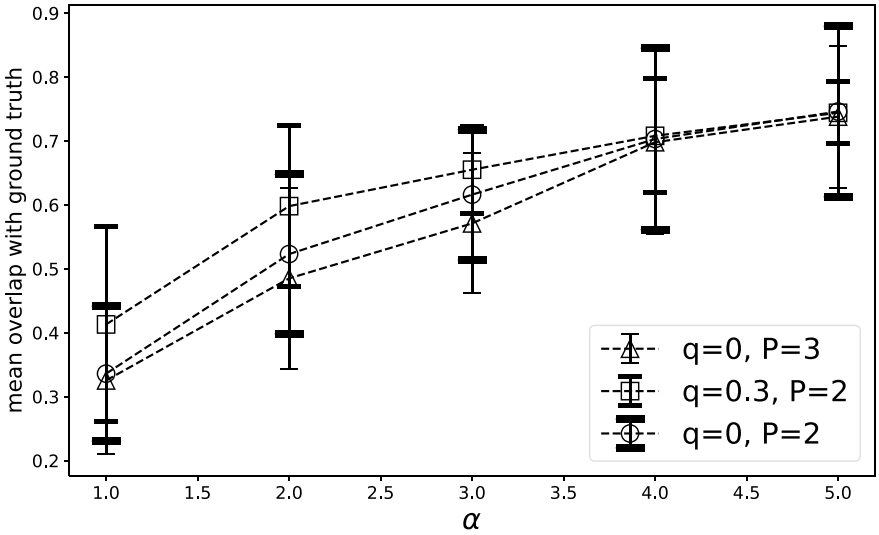



**Fig. 10.10** Average absolute overlap with different $\alpha$. $N = 100$. The error bar is the standard deviation over 10 trials terminated at the 100th epoch. The synapses with a correlation level $q = 0.3$ are relatively easier to learn with fewer examples, despite a finite-size rounding of the threshold

The overlap between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$, $Q^\mu = \frac{1}{N}\sum_i \xi_i^\mu \xi_i^{\mu,*}$ is considered as the measure of success in training (i.e., whether the ground truth can be recovered). Due to the reverse-symmetry of the model probability, we consider the absolute value of the overlap.

Due to the permutation symmetry-broken phenomenon, $P! = P(P-1)(P-2)\cdots 1$ situations should be considered. Figure 10.8 shows the results of a training process of two-bit RBMs ($P = 2$). The algorithm works even when the ground

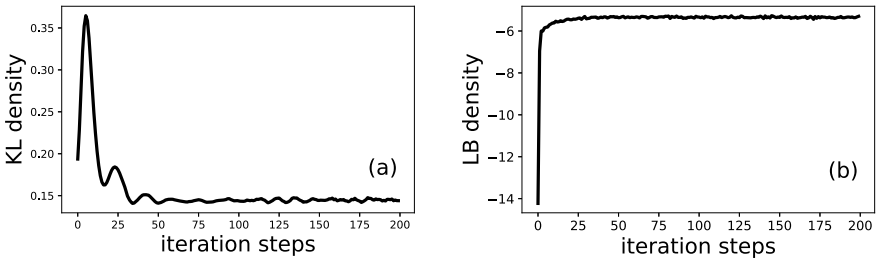



**Fig. 10.11** Training RBMs with binary synapses on MNIST. $N = 784$, $P = 100$ and $M = 2000$. **a** KL divergence between $q_\lambda(\boldsymbol{\xi})$ and $P(\boldsymbol{\xi})$. **b** The approximate lower bound

truth has correlations between two receptive fields. The approximate lower bound is increasing, implying that the variational distribution gets closer to the true posterior distribution during training. Training results on three-bit RBMs ($P = 3$) are also shown in Fig. 10.9. The learning effect of two-bit RBMs with different $\alpha$ is summarized in Fig. 10.10. Notice that the correlation reduces the necessary amount of data for learning as will be analytically derived in Chap. 12. Moreover, the method can be applied to structured data like MNIST, using RBMs with many hidden neurons, displayed in Fig. 10.11.

## References

1. P. Smolensky, Information processing in dynamical systems: foundations of harmony theory (MIT Press, Cambridge, 1986), pp. 194–281
2. Y. Freund, D. Haussler, Unsupervised learning of distributions on binary vectors using two layer networks. Technical Report, Santa Cruz, CA, USA (1994)
3. G. Hinton, Neural Comput. **14**, 1771 (2002)
4. H. Huang, T. Toyoizumi, Phys. Rev. E **91**, 050101 (2015)
5. H. Huang, Phys. Rev. E **102**, 030301(R) (2020)
6. H. Huang, J. Stat. Mech.: Theory Exper. **2017**(5), 053302 (2017)
7. D.M. Blei, A. Kucukelbir, J.D. McAuliffe, J. Amer. stat. Assoc. **112**(518), 859 (2017)
8. H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, Oxford, 2001)
9. Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, in *Advances in Neural Information Processing Systems* (2007), pp. 153–160

# Chapter 11
# Simplest Model of Unsupervised Learning with Binary Synapses


Learning features hidden in unlabeled data is called unsupervised learning. Unsupervised feature learning has been thought of as a fundamental learning process found in brains of humans and non-human animals. In standard machine learning algorithms, a large number of samples are needed to uncover hidden features. However, biological brains only require a few samples to learn the features. It is thus important to understand how the number of samples affects the learning process. In this chapter, we propose a simplest unsupervised learning model to provide statistical physics insights about inner workings of neural networks (Huang and Toyoizumi in Phys. Rev. E 94:062310, 2016 [1]; Huang in J. Stat. Mech.: Theory Exper. 2017(5):053302, 2017 [2]).

## 11.1 Model Setting

Our simplest model of unsupervised feature learning is built upon the restricted Boltzmann Machine (RBM), which has been analyzed in the previous chapter as a statistical mechanics model where random couplings and fields are considered. As mentioned in the previous chapter (see also Fig. 11.1), a RBM consists of two layers of neurons, including a visible layer receiving the input data and a hidden layer building an internal representation of the input. Neurons of the RBM are fully connected across layers but with no lateral connections within each layer. The symmetric synapses between visible and hidden neurons are considered as features that the network tries to learn from the training examples.

It is impossible to analytically study the commonly-used gradient-descent method in feature learning process, like the CD algorithm, due to nested complexity. In this chapter, we simplify the problem and study feature extraction within a Bayesian learning framework.

We first define a teacher–student setting for an analytic study. Finite samples are generated by a simple RBM (Fig. 11.2), where only one hidden node is considered. $\boldsymbol{\sigma}$ and $h$ are defined as the visible configuration and the state of hidden node,



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_11

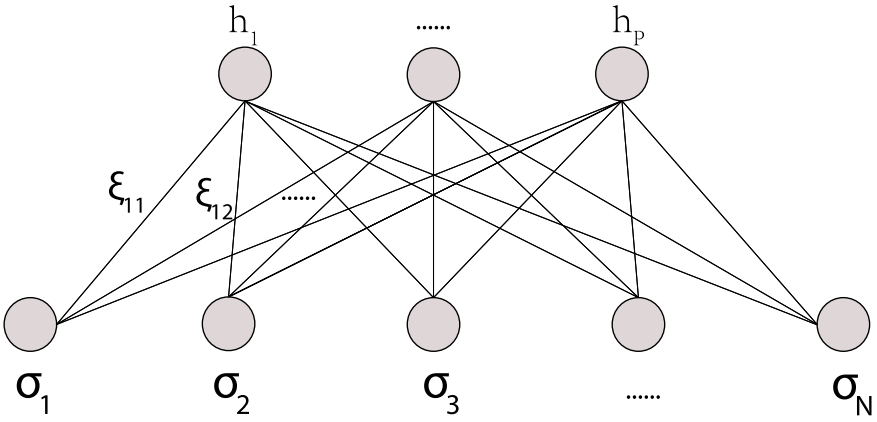



**Fig. 11.1** Schematic illustration of a general restricted Boltzmann Machine. $(\sigma_1, \sigma_2, \ldots, \sigma_N)$ is a sequence of input data, $(h_1, h_2, \ldots, h_P)$ are the hidden representations and $(\xi_{11}, \ldots, \xi_{PN})$ encodes learned features

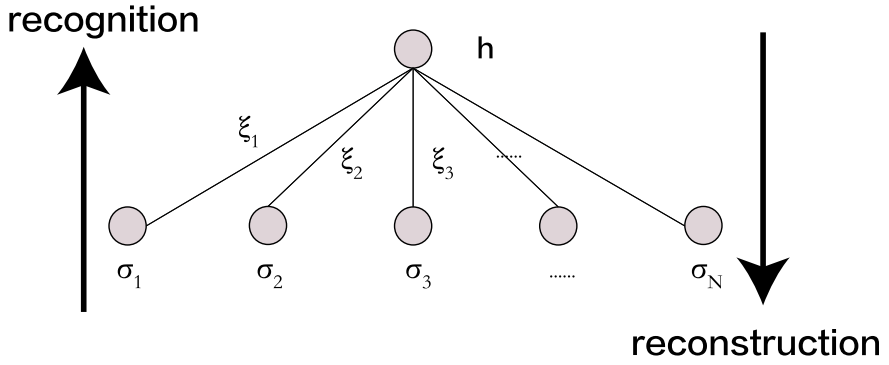



**Fig. 11.2** Schematic illustration of a simple RBM with only one hidden node. $(\sigma_1, \sigma_2, \ldots, \sigma_N)$ is a sequence of input data, $h$ is the state of the hidden node, and $(\xi_1, \ldots, \xi_N)$ encodes true features. The directions of reconstruction and recognition are illustrated by two arrows

respectively. Both the components of $\boldsymbol{\sigma}$ and $h$ take binary values $(\pm 1)$. Meanwhile, we assume that the components of true feature vector $\boldsymbol{\xi}$ generating the data samples takes only two values, i.e., $+1$ or $-1$, with equal probabilities. For simplicity, we consider the case of neurons without any external biases (fields). Under these conditions, independent samples are generated according to the joint distribution $P(\boldsymbol{\sigma}, h) \propto e^{-\beta \frac{E(\boldsymbol{\sigma}, h)}{\sqrt{N}}}$, where $E(\boldsymbol{\sigma}, h) = -\sum_i h \xi_i \sigma_i$. A rescaled factor by the model size $\sqrt{N}$ is considered for a statistical physics analysis. $\beta$ denotes an inverse temperature.

Given the true feature vector $\boldsymbol{\xi}$, the distribution of $\boldsymbol{\sigma}$ can be obtained by the marginalization of the hidden node's state $h$ on the joint distribution $P(\boldsymbol{\sigma}, h)$ as

$$P(\boldsymbol{\sigma}|\boldsymbol{\xi}) = \frac{\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}\right)}{\sum_{\boldsymbol{\sigma}}\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}\right)}, \tag{11.1}$$

where $\bullet^{\mathrm{T}}$ denotes the transpose operation, and the normalization can be obtained exactly as

$$\begin{aligned}\sum_{\boldsymbol{\sigma}}\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}\right) &= \sum_{\boldsymbol{\sigma}}\frac{\exp\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}\right)+\exp\left(-\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}\right)}{2} \\ &= \frac{1}{2}\left[\prod_i\sum_{\sigma_i}\exp\left(\frac{\beta}{\sqrt{N}}\xi_i\sigma_i\right)+\prod_i\sum_{\sigma_i}\exp\left(-\frac{\beta}{\sqrt{N}}\xi_i\sigma_i\right)\right] \\ &= \left[2\cosh\frac{\beta}{\sqrt{N}}\right]^N.\end{aligned} \tag{11.2}$$

Independent samples of the model can be generated by Gibbs sampling through the above conditional probability [Eq. (11.1)].

The feature vector $\boldsymbol{\xi}$ can also be learned through Bayesian learning framework, given $M$ independent samples $\{\boldsymbol{\sigma}^a\}_{a=1}^M$. Hence, the posterior distribution of the feature vector can be obtained by the Bayesian rule as

$$\begin{aligned}P(\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}) &= \frac{P(\boldsymbol{\xi},\{\boldsymbol{\sigma}^a\})}{P(\{\boldsymbol{\sigma}^a\})} \\ &= \frac{P\left(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi}\right)\times P\left(\boldsymbol{\xi}\right)}{\sum_{\boldsymbol{\xi}}P\left(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi}\right)\times P\left(\boldsymbol{\xi}\right)} \\ &= \frac{1}{Z}\prod_a\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^a\right),\end{aligned} \tag{11.3}$$

where $Z$ is the partition function of this learning model, and $a$ goes from $a = 1$ to $M$. For simplicity, we consider the prior probability for the feature vector as a uniform one. The inverse temperature $\beta$ tunes the noise level of the provided data. From Eq. (11.3), a large $\beta$ implies that the feature in the data is strong and can be revealed by a few samples, while a small $\beta$ requires a large number of samples. In the process of inferring the feature vector, each sample serves as a constraint, which makes the model non-trivial. The parameter $\alpha$ is defined as the data density as $\alpha = \frac{M}{N}$. In the following sections, we omit the conditional dependence of $P(\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\})$ on $\{\boldsymbol{\sigma}^a\}$; and the dependence is clear.

## 11.2 Derivation of sMP and AMP Equations

In the framework of the Bayesian learning, the main purpose is to maximize the posterior distribution $P(\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\})$, in order to to get the correct inference on the true

feature vector $\boldsymbol{\xi}$ in a probabilistic way, with the form as $\hat{\xi}_i = \arg\max_{\xi_i} P_i(\xi_i)$. In order to measure the efficiency of the inference quantitatively, we define an overlap between the inferred feature vector and the true feature vector as $q = \langle \frac{1}{N} \sum_i \xi_i^{\text{true}} \langle \xi_i \rangle \rangle$. The inner average is a thermal average, while the outer average is taken over many different true feature vectors (also called quenched-disorder average). If $q = 0$, the examples for learning do not give any useful information about the true feature vectors. On the other hand, $q = 1$ implies that the feature vectors hidden in the examples are perfectly inferred. Because of the interactions among an extensive number of data samples, how to calculate $P_i(\xi_i)$ is highly non-trivial. In this section, we shall achieve this goal by the message passing or cavity approximation.

First, we make a weak correlation assumption (also named Bethe approximation) in the factor graph. By defining a cavity probability distribution $P_{i\to a}(\xi_i)$ denoting the probability distribution of $\xi_i$ in the absence of the sample constraint $a$, we can arrive at the following belief propagation equations:

$$P_{i\to a}(\xi_i) = \frac{1}{Z_{i\to a}} \prod_{b\in\partial i\backslash a} \mu_{b\to i}(\xi_i), \tag{11.4a}$$

$$\mu_{b\to i}(\xi_i) = \sum_{\{\xi_j | j\in\partial b\backslash i\}} \cosh\Big(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{\sigma}^b\Big) \prod_{j\in\partial b\backslash i} P_{j\to b}(\xi_i), \tag{11.4b}$$

where $\partial i\backslash a$ denotes the constraints connecting to $i$ except the constraint $a$. The auxiliary variable $\mu_{b\to i}(\xi_i)$ indicates the contribution from the constraint $b$ to the node $i$, and can be understood in physics as an average of the Boltzmann factor over the joint distribution of $\{\xi_j | j \in \partial b\backslash i\}$. In the large $N$ limit, we can apply the central limit theorem, and calculate the auxiliary variable $\mu_{b\to i}(\xi_i)$ with a Gaussian integral. More precisely, we define $G_{b\to i} = \frac{1}{\sqrt{N}} \sum_{j\in\partial b\backslash i} \sigma_j^b m_{j\to b}$ as the average of $\frac{1}{\sqrt{N}} \sum_{j\in\partial b\backslash i} \xi_j \sigma_j^b$. Under this assumption, we can obtain the simplified message passing equation (sMP) as

$$m_{i\to a} = \tanh\Big(\sum_{b\in\partial i\backslash a} u_{b\to i}\Big), \tag{11.5a}$$

$$u_{b\to i} = \tanh^{-1}\left(\tanh\left(\beta G_{b\to i}\right) \tanh\left(\frac{\beta\sigma_i^b}{\sqrt{N}}\right)\right), \tag{11.5b}$$

where $m_{i\to a} = \sum_{\xi_i} \xi_i P_{i\to a}(\xi_i)$ denotes the cavity magnetization interpreted as the message passing from feature $i$ to data constraint $a$, and $u_{b\to i}$ can be interpreted as the massage passing from data constraint $a$ to feature $i$. If the Bethe approximation is self-consistent, the sMP would converge to a stationary point $\{m_{i\to a}, u_{a\to i}\}$ after a certain number of iterations. By calculating the marginal probability as $P_i(\xi_i) = \frac{1+m_i\xi_i}{2}$, where $m_i = \tanh\left(\sum_{b\in\partial i} u_{b\to i}\right)$, we may finally extract useful information about the true feature vector from the given data. The sMP equation [Eq. (11.5)] therefore offers a practical way to compute the marginal probability distribution $P_i(\xi_i)$. Nevertheless, this method is still computationally expensive with the time complexity and memory

of the order $O(MN)$, which motivates the derivation of approximate message passing (AMP) equations [3, 4] in the remaining part of this section.

In the large-$N$ limit, a Taylor expansion of Eq. (11.5b) w.r.t $\beta\sigma_i^b/\sqrt{N}$ can be carried out. We use the fact that $\tanh x \approx x$ and $\tanh^{-1}(x) \approx x$ when $x$ is close to zero, and obtain

$$u_{b\to i} \approx \beta\sigma_i^b/\sqrt{N}\tanh\left(\beta G_{b\to i}\right). \tag{11.6}$$

In addition, $m_{i\to a}$ can be rewritten as

$$m_{i\to a} = \tanh\left(\tanh^{-1}\left(m_i\right) - u_{a\to i}\right), \tag{11.7}$$

Notice again that $\tanh(x+\epsilon) \approx \tanh x + \tanh'(x)\epsilon = \tanh x + (1-\tanh^2 x)\epsilon$, where $\epsilon$ is a small number. Applying this expansion together with Eq. (11.6) to Eq. (11.7), we can obtain

$$m_{i\to a} \simeq m_i - \left(1-m_i^2\right)\frac{\beta\sigma_i^a}{\sqrt{N}}\tanh\beta G_{a\to i}. \tag{11.8}$$

In addition, $G_b = \frac{1}{\sqrt{N}}\sum_{j\in\partial b}\sigma_j^b m_{j\to b}$ can be rewritten as

$$G_{b\to i} = G_b - \frac{1}{\sqrt{N}}\sigma_i^b m_{i\to b}. \tag{11.9}$$

Our goal is now to eliminate all the subscripts—$a\to i$ and $i\to a$, thereby reducing the total computation cost through saving the computer memory. In other words, we need to find a set of equations involving only the site indexes, i.e., $G_a$ and $m_i$. By definition

$$G_a = \frac{1}{\sqrt{N}}\sum_{i\in\partial a}\sigma_i^a m_{i\to a}. \tag{11.10}$$

Applying Eq. (11.8) to Eq. (11.10), we can get

$$G_a = \frac{1}{\sqrt{N}}\sum_{i\in\partial a}(\sigma_i^a m_i) - \frac{1}{N}\sum_{i\in\partial a}\beta(1-m_i^2)\tanh\left(\beta G_{b\to i}\right), \tag{11.11}$$

where we have used that $(\sigma_i^a)^2 = 1$.

Because $G_{a\to i} \simeq G_a$, Eq. (11.11) can be simplified as follows:

$$G_a = \frac{1}{\sqrt{N}}\sum_{i\in\partial a}\sigma_i^a m_i - \beta(1-Q)\tanh\beta G_a, \tag{11.12}$$

where $Q \equiv \frac{1}{N}\sum_i m_i^2$. We then define the local field $H_i = \sum_{b\in\partial i}\frac{\sigma_i^b}{\sqrt{N}}\tanh\beta G_{b\to i}$ where $G_{b\to i}$ is given in Eq. (11.9). Then we do a Taylor expansion w.r.t $\frac{\sigma_i^b}{\sqrt{N}}m_{i\to b}$ and approximate $m_{i\to b}$ by $m_i$. Finally, a closed-form of $H_i$ and $m_i$ is obtained as

$$H_i \simeq \sum_{b\in\partial i} \frac{\sigma_i^b}{\sqrt{N}} \tanh \beta G_b - \frac{\beta m_i}{N} \sum_{b\in\partial i} \left(1 - \tanh^2 \beta G_b\right), \tag{11.13}$$

$$m_i = \tanh(\beta H_i) \simeq \tanh\left(\sum_{b\in\partial i} \frac{\beta\sigma_i^b}{\sqrt{N}} \tanh \beta G_b - \frac{\beta^2 m_i}{N} \sum_{b\in\partial i} \left(1 - \tanh^2 \beta G_b\right)\right). \tag{11.14}$$

To sum up, the AMP equation for the unsupervised learning is given by an iterative form

$$G_a = \frac{1}{\sqrt{N}} \sum_{i\in\partial a} \sigma_i^a m_i - \beta(1-Q) \tanh \beta G_a, \tag{11.15a}$$

$$m_i = \tanh\left(\sum_{b\in\partial i} \frac{\beta\sigma_i^b}{\sqrt{N}} \tanh \beta G_b - \frac{\beta^2 m_i}{N} \sum_{b\in\partial i} \left(1 - \tanh^2 \beta G_b\right)\right). \tag{11.15b}$$

The correct iteration order to implement the AMP equation follows the order of the above theoretical derivation, as summarized by

$$G_a^{t-1} = \frac{1}{\sqrt{N}} \sum_{i\in\partial a} \sigma_i^a m_i^{t-1} - \beta\left(1 - Q^{t-1}\right) \tanh \beta G_a^{t-2}, \tag{11.16a}$$

$$m_i^t \simeq \tanh\left(\sum_{b\in\partial i} \frac{\beta\sigma_i^b}{\sqrt{N}} \tanh \beta G_b^{t-1} - \frac{\beta^2 m_i^{t-1}}{N} \sum_{b\in\partial i} \left(1 - \tanh^2 \beta G_b^{t-1}\right)\right), \tag{11.16b}$$

where $t$ denotes the update temporal order. Another remarkable feature of the AMP equation is that the essential physics is closely related to the TAP equation in mean-field models, with the extra advantage of requiring a much lower memory demand compared with sMP. The AMP equation is also helpful for a theoretical analysis of the typical properties of the model [3].

## 11.3 Replica Computation

The basic idea of replica computation is to compute the disorder average of an integer power of $Z$, instead of the disorder average of $\ln Z$. Therefore, the free energy function can be obtained as

$$-\beta f = \lim_{N\to\infty} \frac{\langle \ln Z\rangle}{N} = \lim_{n\to 0, N\to\infty} \frac{\ln\langle Z^n\rangle}{nN}, \tag{11.17}$$

where $N$ represents the number of neurons in the visible layer.

### 11.3.1 Explicit form of $\langle Z^n \rangle$

$Z$ is the partition function of the learning posterior $P(\boldsymbol{\xi} \mid \{\boldsymbol{\sigma}^a\})$, which can be written as

$$Z = \sum_{\boldsymbol{\xi}} \prod_a \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{\sigma}^a\right). \tag{11.18}$$

$Z^n$ is actually the product of the partition functions of different replicated systems

$$Z^n = \sum_{\{\boldsymbol{\xi}^\gamma\}} \prod_{a,\gamma} \cosh\left(\frac{\beta}{\sqrt{N}} (\boldsymbol{\xi}^\gamma)^{\mathrm{T}} \boldsymbol{\sigma}^a\right). \tag{11.19}$$

Let us define $\langle \cdot \rangle$ as the disorder average about the true feature vector $\boldsymbol{\xi}^{\text{true}}$, or simply $\boldsymbol{\xi}^*$, as well as the generated data samples given the true feature vector. Therefore, we can obtain

$$\left\langle Z^n \right\rangle = \frac{1}{2^N} \sum_{\{\boldsymbol{\sigma}^a\}, \boldsymbol{\xi}^*} P(\{\boldsymbol{\sigma}^a\} \mid \boldsymbol{\xi}^*) \sum_{\{\boldsymbol{\xi}^\gamma\}} \prod_{a,\gamma} \cosh\left(\frac{\beta}{\sqrt{N}} (\boldsymbol{\xi}^\gamma)^{\mathrm{T}} \boldsymbol{\sigma}^a\right), \tag{11.20}$$

where the joint distribution $P(\{\boldsymbol{\sigma}^a\}, \boldsymbol{\xi}^*)$ is used for the disorder average. Because we have $M$ independent samples $\{\boldsymbol{\sigma}^a\}_{a=1}^M$,

$$\begin{aligned} P(\{\boldsymbol{\sigma}^a\} \mid \boldsymbol{\xi}^*) &= \prod_a P(\boldsymbol{\sigma}^a \mid \boldsymbol{\xi}^*) = \prod_a \frac{\cosh(\frac{\beta}{\sqrt{N}} (\boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\sigma}^a)}{\sum_{\boldsymbol{\sigma}} \cosh(\frac{\beta}{\sqrt{N}} (\boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\sigma}^a)} \\ &= \prod_a \frac{\cosh(\frac{\beta}{\sqrt{N}} (\boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\sigma}^a)}{(2 \cosh \frac{\beta}{\sqrt{N}})^N} \\ &\simeq \frac{1}{2^{NM}} \prod_a e^{-\frac{\beta^2}{2}} \cosh\left(\beta \frac{(\boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\sigma}^a}{\sqrt{N}}\right), \end{aligned} \tag{11.21}$$

where we have used $\ln \cosh(x) \simeq \frac{x^2}{2}$ when $x$ is a small quantity. Substituting Eq. (11.21) to Eq. (11.20), we can obtain the form of $\langle Z^n \rangle$ as follows:

$$\langle Z^n \rangle = \frac{1}{2^N} \frac{1}{2^{NM}} \sum_{\{\boldsymbol{\sigma}^a\}, \boldsymbol{\xi}^*} \prod_a e^{-\frac{\beta^2}{2}} \cosh\left(\frac{\beta}{\sqrt{N}} (\boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\sigma}^a\right) \sum_{\{\boldsymbol{\xi}^\gamma\}} \prod_{a,\gamma} \cosh\left(\frac{\beta}{\sqrt{N}} (\boldsymbol{\xi}^\gamma)^{\mathrm{T}} \boldsymbol{\sigma}^a\right). \tag{11.22}$$

### 11.3.2 Estimation of $\langle Z^n \rangle$ Under Replica Symmetry Ansätz

To proceed, we first define $u^a = \frac{(\boldsymbol{\xi}^*)^{\mathrm{T}}\boldsymbol{\sigma}^a}{\sqrt{N}}$, and $v^{a\gamma} = \frac{(\boldsymbol{\xi}^\gamma)^{\mathrm{T}}\boldsymbol{\sigma}^a}{\sqrt{N}}$. Both $u^a$ and $v^{a\gamma}$ are random variables subject to the covariance structure: $\langle u \rangle = 0$, $\langle u^2 \rangle = 1$, $\langle v^\gamma \rangle = 0$, $\langle (v^\gamma)^2 \rangle = 1$, $\langle u v^\gamma \rangle = q^\gamma$, $\langle v^\gamma v^{\gamma'} \rangle = r^{\gamma\gamma'}$, where we have dropped off the index $a$ because of the independence of data samples, and we also define the overlap between the true feature vector and the estimated one as $q^\gamma = \frac{1}{N}\sum_i \xi_i^\gamma \xi_i^*$, and the overlap between two estimated feature vectors as $r^{\gamma\gamma'} = \frac{1}{N}\sum_i \xi_i^\gamma \xi_i^{\gamma'}$. After introducing the definition of $q^\gamma$ and $r^{\gamma\gamma'}$ by delta functions, we can obtain

$$\begin{aligned}\langle Z^n \rangle =& \frac{1}{2^N}\frac{1}{2^{NM}} \sum_{\{\boldsymbol{\sigma}^a\},\boldsymbol{\xi}^*,\{\boldsymbol{\xi}^\gamma\}} \int \frac{dq^\gamma d\hat{q}^\gamma}{2\pi} e^{-\mathrm{i}\sum_\gamma \hat{q}^\gamma (q^\gamma - \frac{1}{N}\sum_i \xi_i^\gamma \xi_i^*)} \\ &\times \int \frac{dr^{\gamma\gamma'} d\hat{r}^{\gamma\gamma'}}{2\pi} e^{-\mathrm{i}\sum_{\gamma<\gamma'} \hat{r}^{\gamma\gamma'}(r^{\gamma\gamma'} - \frac{1}{N}\sum_i \xi_i^\gamma \xi_i^{\gamma'})} \\ &\times \prod_a e^{-\frac{\beta^2}{2}} \cosh\left(\frac{\beta}{\sqrt{N}}(\boldsymbol{\xi}^*)^{\mathrm{T}}\boldsymbol{\sigma}^a\right) \prod_{a,\gamma} \cosh\left(\frac{\beta}{\sqrt{N}}(\boldsymbol{\xi}^\gamma)^{\mathrm{T}}\boldsymbol{\sigma}^a\right).\end{aligned} \tag{11.23}$$

Under the replica symmetry assumption (similar to that used in the analysis of Hopfield model), $q^\gamma = q$ and $r^{\gamma\gamma'} = r$, we have

$$-\mathrm{i}\sum_\gamma \hat{q}^\gamma q^\gamma = -\mathrm{i}nq\hat{q}, \tag{11.24}$$

$$-\mathrm{i}\sum_{\gamma<\gamma'} \hat{r}^{\gamma\gamma'} r^{\gamma\gamma'} = -\mathrm{i}\frac{n(n-1)}{2} r\hat{r}. \tag{11.25}$$

For the sake of a concise physics representation, we define the entropy term $G_S$ and energy term $G_E$. We first derive the entropy term.

$$
\begin{aligned}
(G_S)^N &= \sum_{\xi^*,\{\xi^\gamma\}} e^{\hat{q}\sum_\gamma \sum_i \xi_i^\gamma \xi_i^* + \sum_{\gamma<\gamma'} \sum_i \hat{r}\xi_i^\gamma \xi_i^{\gamma'}} \\
&= \sum_{\xi^*,\{\xi^\gamma\}} e^{\hat{q}\sum_\gamma \sum_i \xi_i^\gamma \xi_i^* + \sum_i \frac{\hat{r}}{2}(\sum_{\gamma,\gamma'} \xi_i^\gamma \xi_i^{\gamma'} - n)} \\
&= \sum_{\xi^*,\{\xi^\gamma\}} e^{\hat{q}\sum_\gamma \sum_i \xi_i^\gamma \xi_i^* + \sum_i [\frac{\hat{r}}{2}(\sum_\gamma \xi_i^\gamma)^2 - \frac{\hat{r}}{2}n]} \\
&= \sum_{\xi^*,\{\xi^\gamma\}} \prod_i e^{\hat{q}\sum_\gamma \xi_i^\gamma \xi_i^* + \frac{\hat{r}}{2}(\sum_\gamma \xi_i^\gamma)^2 - \frac{\hat{r}}{2}n} \\
&= \prod_i \sum_{\xi^*,\{\xi^\gamma\}} e^{\hat{q}\sum_\gamma \xi^\gamma \xi^* - \frac{\hat{r}}{2}n} \int Dz e^{\sqrt{\hat{r}}\sum_\gamma \xi^\gamma z} \\
&= \prod_i \int Dz e^{-\frac{\hat{r}}{2}n} \sum_{\xi^*,\{\xi^\gamma\}} \prod_\gamma e^{\hat{q}\xi^\gamma \xi^* + \sqrt{\hat{r}}\xi^\gamma z} \\
&= \prod_i \int Dz e^{-\frac{\hat{r}}{2}n} \sum_{\xi^*} \prod_\gamma 2\cosh(\hat{q}\xi^* + \sqrt{\hat{r}}z) \\
&= \prod_i \int Dz e^{-\frac{\hat{r}}{2}n} \sum_{\xi^*} (2\cosh(\hat{q}\xi^* + \sqrt{\hat{r}}z))^n \\
&= \left[\int Dz e^{-\frac{\hat{r}}{2}n} 2(2\cosh(\hat{q} + \sqrt{\hat{r}}z))^n\right]^N \\
&= 2^N \left[\int Dz e^{-\frac{\hat{r}}{2}n} (2\cosh(\hat{q} + \sqrt{\hat{r}}z))^n\right]^N,
\end{aligned}
\tag{11.26}
$$

where $Dz \equiv \frac{e^{z^2/2}\, dz}{\sqrt{2\pi}}$, and the prefactor $2^N$ cancels with the $\frac{1}{2^N}$ term in Eq. (11.23). Note that in Eq. (11.26), we remove the dependence on the site index $i$ after the order exchange of the summation and product. We also apply the Hubbard–Stratonovich transformation to derive the fifth equality. We finally arrive at

$$
G_S = \int Dz e^{-\frac{\hat{r}}{2}n} (2\cosh(\hat{q} + \sqrt{\hat{r}}z))^n. \tag{11.27}
$$

To compute the energy term $G_E^M$, we should first parametrize $u$ and $v$ with mutually-independent standard Gaussian random variables $(t, x^\gamma, y)$ as follows:

$$
u = t, \tag{11.28}
$$

$$
v^\gamma = qt + \sqrt{1-r}x^\gamma + \sqrt{r-q^2}y. \tag{11.29}
$$

It is easy to check that the above parameterization gives the same covariance structure required in our problem (see the first paragraph in the this subsection). $G_E^M$ is thus given by

$$G_E^M = \left\langle \prod_a e^{-\frac{\beta^2}{2}} \cosh(\beta u) \prod_{a,\gamma} \cosh(\beta v^\gamma) \right\rangle$$
$$= \prod_a \left[ e^{-\frac{\beta^2}{2}} \int Dt \int Dy \int Dx^\gamma \cosh(\beta t) \prod_\gamma \cosh(\beta q t + \beta\sqrt{1-r}x^\gamma + \beta\sqrt{r-q^2}y) \right]. \tag{11.30}$$

According to the identity $\int Dz \cosh(az + c) = e^{\frac{a^2}{2}} \cosh c$, we have

$$G_E^M = \left[ e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t)(e^{\frac{\beta^2}{2}(1-r)} \cosh(\beta q t + \beta\sqrt{r-q^2}y))^n \right]^M, \tag{11.31}$$

where $Dt \equiv \frac{-e^{t^2/2}\,dt}{\sqrt{2\pi}}$, and $Dy \equiv \frac{-e^{y^2/2}\,dy}{\sqrt{2\pi}}$. We obtain the energy term $G_E$ as

$$G_E = e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t)(e^{\frac{\beta^2}{2}(1-r)} \cosh(\beta q t + \beta\sqrt{r-q^2}y))^n. \tag{11.32}$$

Taken together, we can estimate the disorder-averaged integer power $\langle Z^n \rangle$ as

$$\langle Z^n \rangle = \int \frac{dq d\hat{q}}{2\pi \mathrm{i}/N} \int \frac{dr d\hat{r}}{2\pi \mathrm{i}/N} \times \exp\left[ -Nnq\hat{q} - Nr\hat{r}\frac{n(n-1)}{2} - Nn\frac{\hat{r}}{2} \right.$$
$$\left. + N \ln \int Dz (2\cosh(\hat{q} + \sqrt{\hat{r}}z))^n \right] \times \exp\left[ \alpha N \ln \left\{ e^{\beta^2(1-r)n/2} \right.\right. \tag{11.33}$$
$$\left.\left. \times e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t)(\cosh(\beta q t + \beta\sqrt{r-q^2}y))^n \right\} \right].$$

where $\alpha = M/N$, and i is absorbed into $\hat{q}$. Using the saddle-point method (in the large-$N$ limit), the disorder average $\langle Z^n \rangle$ can be approximated by the argument in the exponential function, or so-called action in physics.

### 11.3.3 Derivation of Free Energy and Saddle-Point Equations

Under the saddle-point approximation in the large $N$ limit, the free energy function is given below

$$-\beta f_{\mathrm{RS}} = \lim_{n\to 0, N\to\infty} \frac{\ln\langle Z^n \rangle}{nN} = \lim_{n\to 0} \left[ -q\hat{q} - r\hat{r}\frac{n-1}{2} + \frac{\ln G_S}{n} + \alpha\frac{\ln G_E}{n} \right], \tag{11.34}$$

where

$$G_S = \int Dz e^{-\frac{\hat{r}}{2}n}(2\cosh(\hat{q}+\sqrt{\hat{r}}z))^n, \tag{11.35}$$

$$G_E = e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t)(e^{\frac{\beta^2}{2}(1-r)}\cosh(\beta q t + \beta\sqrt{r-q^2}y))^n. \tag{11.36}$$

To proceed, we have to perform two limits: $\lim_{n\to\infty}\frac{\ln G_S}{n}$ and $\lim_{n\to\infty}\frac{\ln G_E}{n}$. First

$$\begin{aligned}\lim_{n\to 0}\frac{\ln G_S}{n} &= \lim_{n\to 0}\frac{-\frac{\hat{r}}{2}n + \ln\int Dz(2\cosh(\hat{q}+\sqrt{\hat{r}}z))^n}{n} \\ &= -\frac{\hat{r}}{2} + \int Dz \ln(2\cosh(\hat{q}+\sqrt{\hat{r}}z)),\end{aligned} \tag{11.37}$$

and second,

$$\begin{aligned}\lim_{n\to 0}\frac{\ln G_E}{n} &= \lim_{n\to 0}\frac{\beta^2\frac{1-r}{2}n + \ln\left[e^{-\frac{\beta^2}{2}}\int Dt\int Dy\cosh(\beta t)(\cosh(\beta q t+\beta\sqrt{r-q^2}y))^n\right]}{n} \\ &= \beta^2\frac{1-r}{2} + \frac{\int Dt\int Dy\cosh(\beta t)\ln(\cosh\beta(qt+\sqrt{r-q^2}y))}{\int Dt\int Dy\cosh(\beta t)} \\ &= \beta^2\frac{1-r}{2} + e^{-\frac{\beta^2}{2}}\int Dt\int Dy\cosh(\beta t)\ln(\cosh\beta(qt+\sqrt{r-q^2}y)).\end{aligned} \tag{11.38}$$

Finally, we obtain the following free energy function:

$$\begin{aligned}-\beta f_{\mathrm{RS}} = &-q\hat{q} + \frac{\hat{r}(r-1)}{2} + \frac{\alpha\beta^2}{2}(1-r) + \int Dz\ln(2\cosh(\hat{q}+\sqrt{\hat{r}}z)) \\ &+ \alpha e^{-\frac{\beta^2}{2}}\int Dt\int Dy\cosh(\beta t)\ln(\cosh\beta(qt+\sqrt{r-q^2}y)).\end{aligned} \tag{11.39}$$

In Eq. (11.33), we use the saddle-point value to approximate the integral in the large-N limit. Therefore, the order parameters $\{q,\hat{q},r,\hat{r}\}$ must be the values making the free energy a lowest value. The saddle-point equations for the order parameters $\{q,\hat{q},r,\hat{r}\}$ can be derived from the stationary condition—$\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial\hat{q}}=0$, $\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial\hat{r}}=0$, $\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial q}=0$, and $\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial r}=0$, as are precisely given by

$$\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial \hat{q}} = -q + \int Dz \tanh(\hat{q} + \sqrt{\hat{r}}z), \tag{11.40a}$$

$$\begin{aligned}\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial \hat{r}} &= \frac{r-1}{2} + \int Dz \tanh(\hat{q} + \sqrt{\hat{r}}z)\frac{z}{2\sqrt{\hat{r}}} \\ &= \frac{r-1}{2} + \frac{1}{2\sqrt{\hat{r}}}\left[\int Dz(1 - \tanh^2(\hat{q} + \sqrt{\hat{r}}z))\sqrt{\hat{r}}\right] \\ &= \frac{r}{2} + \frac{1}{2}\int Dz(-\tanh^2(\hat{q} + \sqrt{\hat{r}}z)),\end{aligned} \tag{11.40b}$$

$$\begin{aligned}\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial q} &= -\hat{q} + \alpha e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t) \tanh(\beta q t + \beta\sqrt{r - q^2}y) \\ &\quad \times \left[\beta t + \beta \frac{y}{2\sqrt{r-q^2}} \times (-2q)\right] \\ &= -\hat{q} + \alpha e^{-\frac{\beta^2}{2}} \int Dt \int Dy \Bigg\{\Bigg[\beta^2 \sinh(\beta t) \tanh(\beta q t + \beta\sqrt{r - q^2}y) \\ &\quad + \cosh(\beta t)\beta(1 - \tanh^2(\beta q t + \beta\sqrt{r-q^2}))\beta q\Bigg] \\ &\quad + \cosh(\beta t)\left[\beta \frac{-2q}{2\sqrt{r-q^2}}(1 - \tanh^2(\beta q t + \beta\sqrt{r-q^2}y))\beta\sqrt{r-q^2}\right]\Bigg\} \\ &= -\hat{q} + \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt \int Dy \sinh(\beta t) \tanh(\beta q t + \beta\sqrt{r-q^2}y),\end{aligned} \tag{11.40c}$$

$$\begin{aligned}\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial r} &= \frac{\hat{r}}{2} - \frac{\alpha\beta^2}{2} + \alpha e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t) \tanh(\beta q t + \beta\sqrt{r-q^2}y) \\ &\quad \times \beta y \frac{1}{2\sqrt{r-q^2}} \\ &= \frac{\hat{r}}{2} - \frac{\alpha\beta^2}{2} + \alpha e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t)(1 - \tanh^2(\beta q t + \beta\sqrt{r-q^2}y)) \\ &\quad \times \beta \frac{1}{2\sqrt{r-q^2}} \times \beta\sqrt{r-q^2} \\ &= \frac{\hat{r}}{2} - \frac{\alpha\beta^2}{2} + \frac{1}{2}\alpha e^{-\frac{\beta^2}{2}} \times e^{\frac{\beta^2}{2}}\beta^2 - \frac{1}{2}\alpha\beta^2 e^{-\frac{\beta^2}{2}} \\ &\quad \times \int Dt \int Dy \cosh(\beta t) \tanh^2(\beta q t + \beta\sqrt{r-q^2}y) \\ &= \frac{\hat{r}}{2} - \frac{1}{2}\alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t) \tanh^2(\beta q t + \beta\sqrt{r-q^2}y).\end{aligned} \tag{11.40d}$$

Finally, the saddle-point equations are expressed as

$$q = \int Dz \tanh(\hat{q} + \sqrt{\hat{r}}z), \tag{11.41a}$$

$$r = \int Dz \tanh^2(\hat{q} + \sqrt{\hat{r}}z), \tag{11.41b}$$

$$\hat{q} = \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt \int Dy \sinh(\beta t) \tanh(\beta q t + \beta\sqrt{r - q^2}y), \tag{11.41c}$$

$$\hat{r} = \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t) \tanh^2(\beta q t + \beta\sqrt{r - q^2}y). \tag{11.41d}$$

## 11.4 Phase Transitions

For a statistical mechanics analysis of the system in the thermodynamic limit, we first define the cavity field

$$H_{i\to a} = \frac{1}{\sqrt{N}} \sum_{b\in\partial i\setminus a} \sigma_i^b \tanh \beta G_{b\to i}, \tag{11.42}$$

where $G_{b\to i}$ has the form as

$$G_{b\to i} = \frac{1}{\sqrt{N}} \sum_{j\in\partial b\setminus i} \sigma_j^b m_{j\to b}. \tag{11.43}$$

Under the replica symmetric assumption, $H_{i\to a}$ follows a Gaussian distribution with mean zero and variance $\alpha\hat{Q}$ in the large-N limit, which can be described as follows:

$$\langle H_{i\to a}\rangle = 0, \tag{11.44a}$$

$$\langle (H_{i\to a})^2\rangle = \frac{M}{N}\langle \tanh^2 \beta G_{b\to i}\rangle = \alpha\hat{Q}, \tag{11.44b}$$

where $\hat{Q} \equiv \langle \tanh^2 \beta G_{b\to i}\rangle$, and $\frac{M}{N}$ denotes the data density. Note that the average refers to the disorder average. The Gaussian assumption can be checked with a comparison with numerical simulations. As for $G_{b\to i}$, we can also obtain a similar structure as follows:

$$\langle G_{b\to i}\rangle = 0, \tag{11.45a}$$

$$\langle (G_{b\to i})^2\rangle = \frac{1}{N}\sum_j m_{j\to b}^2 = Q. \tag{11.45b}$$

$Q$ and $\hat{Q}$ can thus be written in a compact form

$$\hat{Q} = \int Dz \tanh^2 \beta \sqrt{Q} z, \tag{11.46a}$$

$$Q = \int Dz \tanh^2 (\beta \sqrt{\alpha \hat{Q}} z), \tag{11.46b}$$

where $Dz = \frac{dz e^{-z^2/2}}{\sqrt{2\pi}}$. It is easy to check that $Q = 0$ is a solution of Eq. (11.46). $Q = 0$ implies that $m_{j \to b} = 0$, the information flow characterized by passing messages contains no information anyway; the whole system is thus in a disordered/symmetric state. Next, we use a linear stability analysis method to measure on which condition the stability of this trivial solution is not guaranteed.

When $\alpha$ is small, only one solution of $Q = 0$ exists for the above mean-field equations. However, at some critical point $\alpha_c$, there appears continuously a non-trivial solution of $Q \neq 0$, which signals the fixed point of sMP or AMP starts to contain information about the underlying true feature vector. We then assume $Q$ is of the order of $\epsilon$, a very small quantity. Notice that $\tanh x \approx x$ when the argument $x$ approaches zero, and thus $\hat{Q} \approx \beta^2 \epsilon$. Therefore, we can expand the following equation around $\hat{Q} = 0$.

$$Q = \int Dz \tanh^2 (\beta \sqrt{\alpha \hat{Q}} z) = \int Dz \beta^2 \alpha \hat{Q} z^2 = \beta^2 \alpha \beta^2 \epsilon \int Dz z^2. \tag{11.47}$$

Using $\int Dz z^2 = 1$, we arrive at a very simple form of $Q$.

$$Q = \beta^4 \alpha \epsilon. \tag{11.48}$$

Putting back the iteration step $t$, we then have

$$\epsilon^{t+1} = \alpha \beta^4 \epsilon^t. \tag{11.49}$$

This result shows that when $\alpha \beta^4 < 1$, the $Q = 0$ solution is stable. The transition point is thus set to $\alpha_c = \beta^{-4}$. As shown in Fig. 11.3, a critical-slowing-down phenomenon is observed, suggesting a continuous phase transition at $\alpha_c$, where the trivial solution $Q = 0$ loses its stability.

Note that the statistical analysis of the sMP equation does not give the correct value of $Q$ after the transition, probably due to the invalid Gaussian field assumption in the case of biased messages immediately after the transition. Hence, a deeper analysis from replica computation is needed.

According to the Nishimori condition, $q = r$, implying that the true feature vector follows the same posterior distribution in the optimal Bayesian inference (the algorithm can have access to the true temperature). This can also be verified through the numerical solution of the saddle-point equations [Eq. (11.41)].

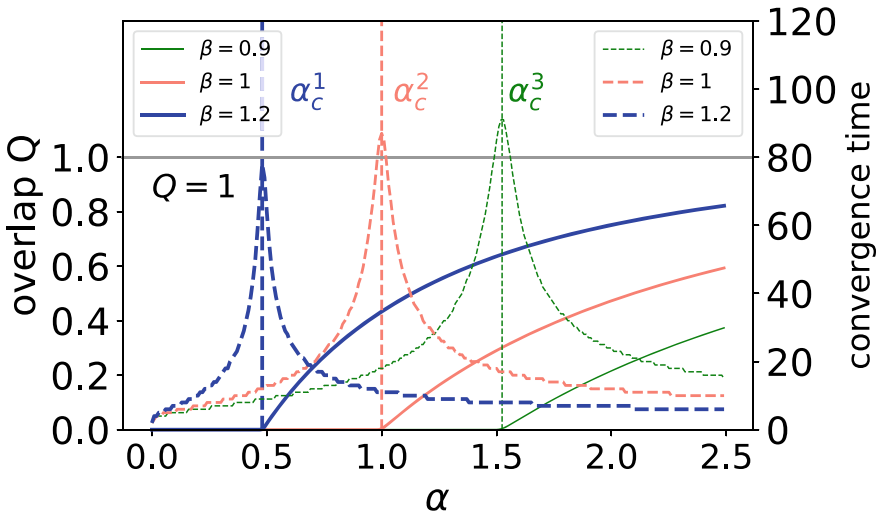



**Fig. 11.3** An illustration of critical point $\alpha_c$. The simulation is made based on three different temperatures: $\beta = 1.2$, $\beta = 1$ and $\beta = 0.9$. The fixed point of Q and convergence time of sMP are also shown. The dashed lines show convergence time measured by iteration steps

Assuming $q$ and $r$ are both small values close to zero, we obtain

$$\hat{q} = \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt \int Dy \sinh(\beta t)\beta q t = \alpha\beta^4 q, \tag{11.50a}$$

$$\begin{aligned}\hat{r} &= \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t)[\beta q t + \beta\sqrt{r - q^2}y]^2 \\ &= \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt \int Dy \cosh(\beta t)[(\beta q t)^2 + (\beta\sqrt{r - q^2}y)^2] \\ &= \alpha\beta^4(q^2\beta^2 + r),\end{aligned} \tag{11.50b}$$

$$q = \hat{q} = \alpha\beta^4 q, \tag{11.50c}$$

$$r = \hat{q}^2 + \hat{r} = (\alpha\beta^4 q)^2 + \alpha\beta^4 q^2\beta^2 + \alpha\beta^4 r \simeq \alpha\beta^4 r. \tag{11.50d}$$

Note that Eqs. (11.50c) and (11.50d) imply the same critical point $\alpha_c = \frac{1}{\beta^4}$ for both $q$ and $r$, above which $q = 0$ is not a stable solution any more. This transition is continuous, and is called spontaneous symmetry breaking transition as well, since our model has original symmetry with respect to changing the sign of feature components (e.g., $\boldsymbol{\xi} \to -\boldsymbol{\xi}$). In addition, this theoretical prediction coincides well with the numerical results of sMP iterations on single instances of learning (Fig. 11.4).

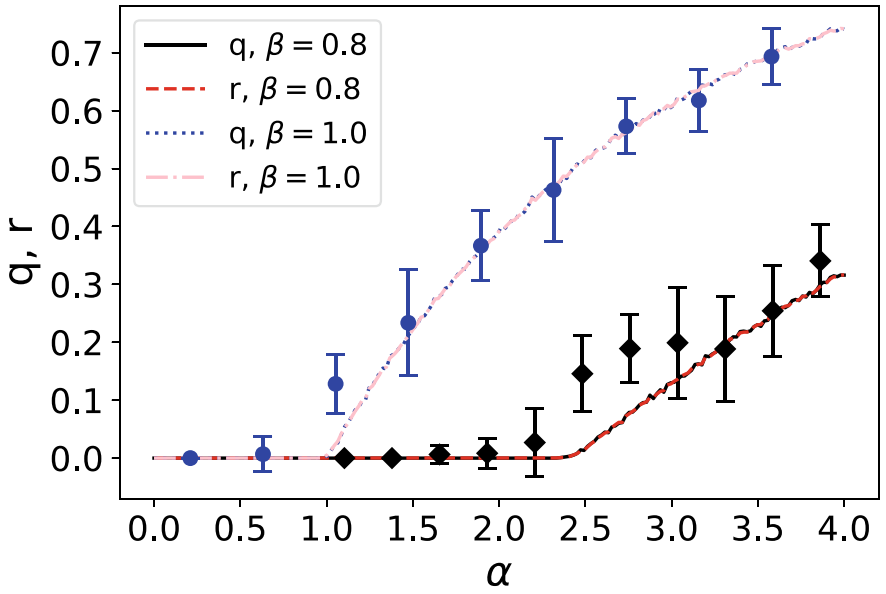



**Fig. 11.4** Order parameters versus $\alpha$. The theoretical predictions of replica computation are indicated by lines, while the numerical simulations of $q$ on single instances are indicated by symbols (solid circles denote the case of $\beta = 1.0$, and solid diamonds denote the case of $\beta = 0.8$)

## 11.5 Measuring the Temperature of Dataset

We already know that the critical number of data samples to trigger unsupervised learning is clearly determined by the inverse temperature $\beta$, a measure of noise level in the raw inputs. Is it possible to infer the true temperatures used to generate the data itself? If we can learn the temperature parameter, we can know the typical properties of phase transitions intrinsic in the system. Here, we will apply the Bayesian rule to infer the true temperature parameters, which is further consistent with the fact that the data itself reflect how noisy a data sample is.

The posterior probability of $\beta$ given the data $\{\boldsymbol{\sigma}^a\}_{a=1}^M$ is given by [2]

$$P(\beta|\{\sigma^a\}) = \sum_{\boldsymbol{\xi}} P(\beta, \boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}) = \frac{\sum_{\boldsymbol{\xi}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi}, \beta) P_0(\boldsymbol{\xi}, \beta)}{\int d\beta \sum_{\boldsymbol{\xi}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi}, \beta) P_0(\boldsymbol{\xi}, \beta)}, \tag{11.51}$$

where we have used the uniform prior probability $P_0$ for the hyper-parameters. We then apply the property of the generative model

$$P(\{\sigma^a\}|\beta, \boldsymbol{\xi}) = \prod_a P(\boldsymbol{\sigma}^a|\beta, \boldsymbol{\xi}) = \prod_a \frac{\cosh(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^a)}{\sum_{\boldsymbol{\sigma}} \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^a\right)}, \tag{11.52}$$

We can then rewrite the posterior probability as follows:

$$P(\beta|\{\boldsymbol{\sigma}^a\}) = \frac{1}{Z(\{\boldsymbol{\sigma}^a\})} \sum_{\boldsymbol{\xi}} \frac{1}{\tilde{Z}^M} \prod_a \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{\sigma}^a\right), \tag{11.53}$$

where $\tilde{Z} = \left[2\cosh\left(\frac{\beta}{\sqrt{N}}\right)\right]^N$, and $Z(\{\boldsymbol{\sigma}^a\}) = \int d\beta \sum_{\boldsymbol{\xi}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi}, \beta)$. We can then write an explicit form of the posterior as follows:

$$P(\beta|\{\boldsymbol{\sigma}^a\}) = \frac{1}{Z(\{\boldsymbol{\sigma}^a\})} \sum_{\boldsymbol{\xi}} e^{-MN \ln\left(2\cosh\frac{\beta}{\sqrt{N}}\right)} \prod_a \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{\sigma}^a\right), \tag{11.54}$$

where in the large-$N$ limit, we approximate the factor $e^{-MN \ln(2\cosh\frac{\beta}{\sqrt{N}})}$ by $e^{-M\frac{\beta^2}{2}}$. With this approximation, we can get the final form

$$P(\beta|\{\boldsymbol{\sigma}^a\}) \propto \frac{1}{Z(\{\boldsymbol{\sigma}^a\})} e^{-M\frac{\beta^2}{2}} \sum_{\boldsymbol{\xi}} \prod_a \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{\sigma}^a\right) \propto e^{-M\frac{\beta^2}{2}} Z(\beta, \{\boldsymbol{\sigma}^a\}), \tag{11.55}$$

where $Z(\{\boldsymbol{\sigma}^a\})$ does not depend on $\beta$, and $Z(\beta, \{\boldsymbol{\sigma}^a\})$ is the very partition function of our original unsupervised learning model.

If we want to get the true temperature parameters of this system, we can try to maximize the probability of $\beta$ given the data $\{\boldsymbol{\sigma}^a\}_{a=1}^M$, which is denoted by $\beta = \mathrm{argmax}_\beta P(\beta|\{\boldsymbol{\sigma}^a\})$. Applying the method of Maximum Likelihood Estimation (MLE), we obtain the self-consistent equation that $\beta$ must satisfy

$$\frac{\partial P(\beta|\{\boldsymbol{\sigma}^a\})}{\partial \beta} = 0, \tag{11.56a}$$

$$\frac{\partial Z(\beta, \{\boldsymbol{\sigma}^a\})}{\partial \beta} = ZM\beta, \tag{11.56b}$$

$$\frac{1}{N} \frac{\partial \ln Z(\beta, \{\boldsymbol{\sigma}^a\})}{\partial \beta} = \frac{M}{N}\beta = \alpha\beta. \tag{11.56c}$$

In statistic physics, Eq. (11.56c) is defined as the negative energy density $(-\epsilon)$. Therefore, $\beta = -\frac{\epsilon}{\alpha}$, which is considered as the Nishimori condition, i.e., in the optimal Bayesian setting, the internal energy of the disorder model is analytic, like the case in the $p$-spin model. When $N$ is not very large, the equation determining $\beta$ is given by $\beta = \sqrt{N}\tanh^{-1}(-\frac{\epsilon}{\alpha\sqrt{N}})$. Under the Bethe approximation, the energy per neuron $\epsilon$ can be computed by $N\epsilon = -\sum_i \Delta\epsilon_i + (N-1)\sum_a \Delta\epsilon_a$, where $\Delta\epsilon_i$ and $\Delta\epsilon_a$ are given, respectively, by

$$\Delta\epsilon_i = \frac{\sum_{a\in\partial i}\mathcal{H}_{a\to i}(+1) + \left(\prod_{a\in\partial i}\mathcal{G}_{a\to i}\right)\sum_{a\in\partial i}\mathcal{H}_{a\to i}(-1)}{\beta + \beta\prod_{a\in\partial i}\mathcal{G}_{a\to i}}, \tag{11.57a}$$

$$\Delta\epsilon_a = \beta\Xi_a^2 + G_a\tanh(\beta G_a), \tag{11.57b}$$

where $\Xi_a^2 = \frac{1}{N}\sum_{i\in\partial a}(1 - m_{i\to a}^2)$, and

$$\mathcal{H}_{a\to i}(\xi_i) = \beta^2\Xi_{a\to i}^2 + (\beta G_{a\to i} + \beta\sigma_i^a\xi_i/\sqrt{N})\tanh(\beta G_{a\to i} + \beta\sigma_i^a\xi_i/\sqrt{N}), \tag{11.58}$$

and $\mathcal{G}_{a\to i} = \mathrm{e}^{-2u_{a\to i}}$. These quantities can be easily derived from the cavity approximation of the free energy function of the model.

We apply the expectation-maximization (EM) procedure [5], to implement the update of the hyper-parameter $\beta$—$\beta(t) = -\frac{\epsilon(t)}{\alpha}$, where $t$ denotes the iteration step. In this algorithm, the message updates are called E-step, and the temperature update is called M-step. First, we start from some initial value of $\beta_0$. One can iteratively update the value of $\beta$ until convergence within some precision. After one updating of the temperature, the messages in the sMP equation are also updated. To avoid numerical instability, the damping technique is recommended, i.e., $\beta(t) = \eta\beta(t) + (1-\eta)\beta(t-1)$, where $t$ denotes the iteration step and $\eta \in [0, 1]$ is a damping factor. Results of this algorithm are shown in Fig. 11.5.

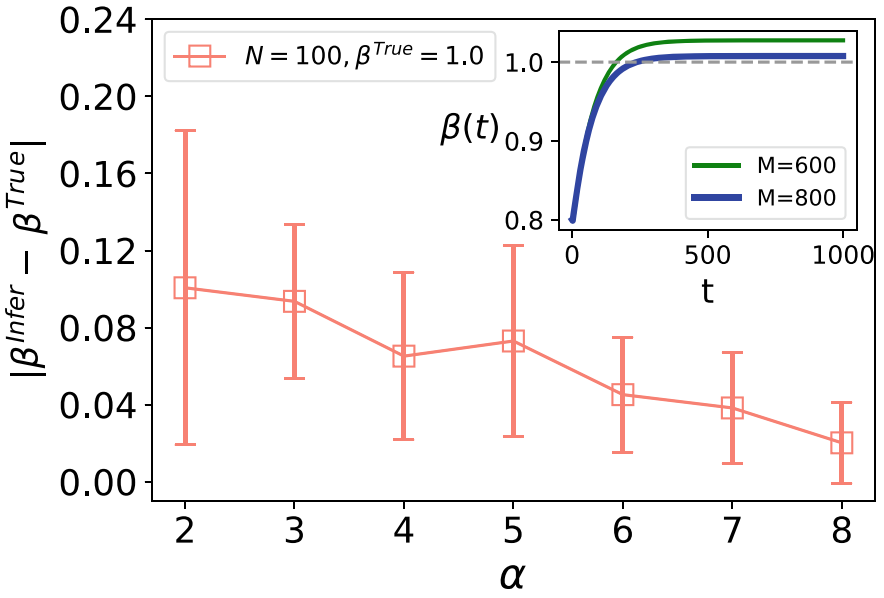



**Fig. 11.5** The inference performance of the hyper-parameter $\beta$. Deviation of inferred $\beta$ from the true value decreases with the data size. In simulations, we consider 20 instances of size N = 100, and use $\eta = 0.02$ and initial value of $\beta_0 = 0.8$. Two representative trajectories of $\beta(t)$ are shown in the inset

## References

1. H. Huang, T. Toyoizumi, Phys. Rev. E **94**, 062310 (2016)
2. H. Huang, J. Stat. Mech.: Theory Exper. **2017**(5), 053302 (2017)
3. Y. Kabashima, J. Phys. A **36**(43), 11111 (2003)
4. D. Donoho, A. Maleki, A. Montanari, Proc. Natl. Acad. Sci. U.S.A. **106**, 18914 (2009)
5. A.P. Dempster, N.M. Laird, D.B. Rubin, J. R. Stat. Soc. Ser. B **39**, 1 (1977)

# Chapter 12
# Inherent-Symmetry Breaking in Unsupervised Learning


In this chapter, we introduced a toy model of unsupervised learning, which exhibits inherent reverse-spin symmetry and permutation symmetry of any two hidden neurons. These symmetries can be broken by the increasing amount of data, reflecting the nature of unsupervised learning in its simplest setting (Hou et al. in J. Phys. A: Math. Theor. 52(41):414001, 2019 [1]; Hou and Huang in Phys. Rev. Lett. 124:248302, 2020 [2]).

## 12.1 Model Setting

We use the two-bit binary RBM (Fig. 12.1), which has two hidden neurons to learn embedded features in input data samples. This is a simple model to learn the internal representation from the raw unlabeled data, which we call unsupervised learning. Each data sample is specified by a binary configuration $\boldsymbol{\sigma} = \{\sigma_i = \pm 1\}_{i=1}^{N}$, where $N$ is the input dimensionality. A collection of $M$ samples is denoted as $\{\boldsymbol{\sigma}^a\}_{a=1}^{M}$. Synaptic values connecting visible and hidden neurons are characterized by $\boldsymbol{\xi}$, where each component takes a binary value ($\pm 1$) as well. Because of two hidden neurons, $\boldsymbol{\xi} = (\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$ where the superscript indicates the hidden neuron's index, are also called receptive fields (RFs) of the first and second hidden neurons, respectively. The joint distribution of hidden unit and input data in this RBM model, given the two receptive fields, is thus described by the Boltzmann distribution as

$$P(\boldsymbol{\sigma}, h_1, h_2 | \boldsymbol{\xi}^1, \boldsymbol{\xi}^2) = \frac{1}{Z(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)} e^{\frac{\beta}{\sqrt{N}}(\boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma} h_1 + \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma} h_2)}, \tag{12.1}$$

where $h_i$ is the $i$th hidden neural activity, $X = \frac{1}{\sqrt{N}} \boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma}$ and $Y = \frac{1}{\sqrt{N}} \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}$. Hereafter, $\boldsymbol{ab}$ denotes the inner product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The scaling factor $\frac{1}{\sqrt{N}}$ ensures that the argument of the hyperbolic cosine function is of the order of unity. $\beta$ represents the inverse-temperature, and $Z(\boldsymbol{\xi})$ is the partition function depending



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_12

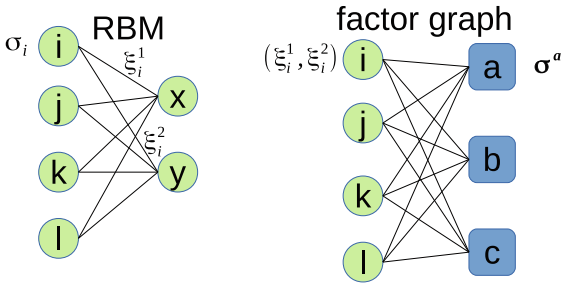



**Fig. 12.1** A schematic illustration of the two-bit RBM model. $N = 4$ in this example (say, $i$, $j$, $k$ and $l$). (Left panel) The original model with only two hidden neurons (say, $x$ and $y$). (Right panel) The corresponding factor graph where the data node is represented by a square, and the paired-synapses (feature vector) is indicated by a circle. In this example, $M = 3$ (say, $a$, $b$ and $c$). The circle is an augmented version of single synapse considered in the one-bit RBM [3]. The plot is taken from Ref. [1]

on the feature $\boldsymbol{\xi}$. $\boldsymbol{\sigma}$ can be arbitrary one of the $M$ samples. The marginal distribution of input data $\boldsymbol{\sigma}$ will be obtained when the two hidden neurons' activities ($\pm 1$) have been marginalized out

$$P(\boldsymbol{\sigma}) = \sum_{h_1,h_2} P(\boldsymbol{\sigma}, h_1, h_2|\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) = \frac{1}{Z(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)} \cosh \beta X \cosh \beta Y, \tag{12.2}$$

where the dependence of $P(\boldsymbol{\sigma})$ on the hidden feature $\boldsymbol{\xi}$ is omitted.

When the embedded feature is randomly generated, the inverse-temperature $\beta$ tunes the noise level of generated data samples from the feature $\boldsymbol{\xi}$. Clearly, the data distribution is invariant with respect to (w.r.t) the exchange of the hidden neurons, which is called the permutation symmetry (PS), i.e., the distribution $P(\boldsymbol{\sigma}|\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) = P(\boldsymbol{\sigma}|\boldsymbol{\xi}^2, \boldsymbol{\xi}^1)$. The required number of hidden neurons to produce this symmetry is at least two. Therefore, this setting defines a minimal model to study the permutation symmetry in unsupervised learning.

In this model, the embedded feature follows the distribution $P(\boldsymbol{\xi}) = P(\boldsymbol{\xi}^1)P(\boldsymbol{\xi}^2|\boldsymbol{\xi}^1)$ in which $P(\boldsymbol{\xi}^1) = \prod_{i=1}^{N} \left[\frac{1}{2}\delta(\xi_i^1 - 1) + \frac{1}{2}\delta(\xi_i^1 + 1)\right]$ and

$$P(\boldsymbol{\xi}^2|\boldsymbol{\xi}^1) = \prod_{i=1}^{N} \left[p_{\mathrm{d}}\delta(\xi_i^2 = -\xi_i^1) + (1 - p_{\mathrm{d}})\delta(\xi_i^2 = \xi_i^1)\right], \tag{12.3}$$

where $p_{\mathrm{d}}$ specifies the fraction of components taking different values in the two feature maps associated with the two hidden neurons.

First, we consider the case of no prior knowledge about $\boldsymbol{\xi}$. Given the $M$ data samples, one gets the posterior probability of the embedded feature according to the Bayes' rule

$$
\begin{aligned}
P(\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}_{a=1}^M) &= \frac{\prod_a P(\boldsymbol{\sigma}^a|\boldsymbol{\xi})}{\sum_{\boldsymbol{\xi}} \prod_a P(\boldsymbol{\sigma}^a|\boldsymbol{\xi})} \\
&= \frac{1}{\Omega} \prod_a \frac{1}{Z(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)} \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}^a\right),
\end{aligned}
\tag{12.4}
$$

where $\Omega$ is the partition function of the minimal model. In addition, we use the same temperature as that used to generate data. Because we do not use the true prior $\prod_i P_i(\xi_i^1, \xi_i^2|p_{\mathrm{d}})$, the current setting does not require the value of $p_{\mathrm{d}}$, and is, therefore, not the Bayes optimal setting which corresponds to Nishimori condition in physics.

One challenging issue to compute the posterior probability is the nested partition function $Z(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$. Fortunately, this partition function can be further simplified in the large-$N$ limit. More precisely

$$
\begin{aligned}
Z(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) &= \sum_{\boldsymbol{\sigma}} \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma}\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}\right) \\
&= \frac{1}{2} \sum_{\boldsymbol{\sigma}} \left[\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma} + \frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}\right) + \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma} - \frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}\right)\right] \\
&= \frac{1}{2} \left[\prod_i 2\cosh\left(\frac{\beta}{\sqrt{N}}(\xi_i^1 + \xi_i^2)\right) + \prod_i 2\cosh\left(\frac{\beta}{\sqrt{N}}(\xi_i^1 - \xi_i^2)\right)\right] \\
&= \frac{1}{2} \left[\prod_i e^{\ln 2 + \frac{\beta^2}{2N}(\xi_i^1 + \xi_i^2)^2} + \prod_i e^{\ln 2 + \frac{\beta^2}{2N}(\xi_i^1 - \xi_i^2)^2}\right] \\
&= \frac{1}{2} \left[\prod_i e^{\ln 2 + \frac{\beta^2}{N} + \frac{\beta^2}{N}\xi_i^1\xi_i^2} + \prod_i e^{\ln 2 + \frac{\beta^2}{N} - \frac{\beta^2}{N}\xi_i^1\xi_i^2}\right] \\
&= \frac{1}{2} \left[e^{N\ln 2 + \beta^2 + \frac{\beta^2}{N}\sum_i \xi_i^1\xi_i^2} + e^{N\ln 2 + \beta^2 - \frac{\beta^2}{N}\sum_i \xi_i^1\xi_i^2}\right] \\
&\simeq 2^N e^{\beta^2} \cosh(\beta^2 Q),
\end{aligned}
\tag{12.5}
$$

where we have used $\ln\cosh(x) \simeq \frac{x^2}{2}$ for small $x$ to arrive at the final equality, and defined $Q \equiv \frac{1}{N}\sum_i \xi_i^1\xi_i^2$, which is the very overlap between the two feature maps. Finally, we move all the irrelevant constants into the partition function $\Omega$, the posterior probability can thus be rewritten into the following form:

$$
P(\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}_{a=1}^M) = \frac{1}{\Omega} \prod_a \frac{1}{\cosh(\beta^2 Q)} \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}^a\right),
\tag{12.6}
$$

which constructs the Boltzmann distribution of the minimal model. We are interested in the case of $M = \alpha N$, where $\alpha$ specifies the data (constraint) density.

### 12.1.1 Cavity Approximation

Our goal is to compute the maximum of the posterior marginals (MPM) estimator $(\hat{\xi}_i^1, \hat{\xi}_i^2) = \arg\max_{\xi_i^1,\xi_i^2} P_i(\xi_i^1, \xi_i^2)$. Hence, the task is to compute marginal probabilities, i.e., $P_i(\xi_i^1, \xi_i^2)$, which is, in general, intractable due to the interaction among data constraints (the product over $a$ in Eq. (12.6)). However, we can represent the problem in a graphical model, where data constraints and paired-synapses are treated, respectively, as factor (data) nodes and variable nodes. Then, the computation of the marginal probability can be achieved by running a message passing iteration among factor and variable nodes. We further assume that the paired-synapses on the graphical model are weakly correlated, which is called the Bethe approximation in physics.

We first define a cavity probability $P_{i\to a}(\xi_i^1, \xi_i^2)$ with the data node $a$ removed. Under the weak correlation assumption, $P_{i\to a}(\xi_i^1, \xi_i^2)$ obeys a self-consistent equation

$$P_{i\to a}(\xi_i^1, \xi_i^2) = \frac{1}{Z_{i\to a}} \prod_{b\in\partial i\backslash a} \mu_{b\to i}(\xi_i^1, \xi_i^2), \tag{12.7a}$$

$$\begin{aligned}\mu_{b\to i}(\xi_i^1, \xi_i^2) = &\sum_{\boldsymbol{\xi}\backslash\xi_i^1,\xi_i^2} \frac{1}{\cosh\left(\beta^2 Q_c + \frac{\beta^2}{N}\xi_i^1\xi_i^2\right)} \cosh\left(\beta X_b + \frac{\beta}{\sqrt{N}}\xi_i^1\sigma_i^b\right) \\ &\times \cosh\left(\beta Y_b + \frac{\beta}{\sqrt{N}}\xi_i^2\sigma_i^b\right) \prod_{j\in\partial b\backslash i} P_{j\to b}(\xi_j^1, \xi_j^2),\end{aligned} \tag{12.7b}$$

where $Z_{i\to a}$ is a normalization constant, $\partial i\backslash a$ denotes neighbors of the feature node $i$ except the data node $a$, $\partial b\backslash i$ denotes neighbors of the data node $b$ except the feature node $i$ and the auxiliary quantity $\mu_{b\to i}(\xi_i^1, \xi_i^2)$ denotes the contribution from data node $b$ to feature node $i$ given the value of $(\xi_i^1, \xi_i^2)$. Products in Eq. (12.7) are due to the weak correlation assumption. In addition, $X_b \equiv \frac{1}{\sqrt{N}}\sum_{j\neq i}\xi_j^1\sigma_j^b$, $Y_b \equiv \frac{1}{\sqrt{N}}\sum_{j\neq i}\xi_j^2\sigma_j^b$, and the cavity version of $Q$ is defined as $Q_c \equiv \frac{1}{N}\sum_{j\neq i}\xi_j^1\xi_j^2$.

However, the above self-consistent equation is still intractable due to the summation in $\mu_{b\to i}$. Nevertheless, $X_b$ and $Y_b$ are approximately correlated Gaussian random variables due to the central limit theorem. As a result, the intractable summation can be replaced by an integral which is easy to work out in this model. Hence, we just need to compute the following means, variances and covariances among these random variables

$$G_{b\to i}^1 = \frac{1}{\sqrt{N}}\sum_{j\neq i}\sigma_j^b m_{j\to b}^1, \tag{12.8a}$$

$$G_{b\to i}^2 = \frac{1}{\sqrt{N}}\sum_{j\neq i}\sigma_j^b m_{j\to b}^2, \tag{12.8b}$$

$$\Gamma^1_{b\to i} = \frac{1}{N}\sum_{j\neq i}\left(1-(m^1_{j\to b})^2\right), \tag{12.8c}$$

$$\Gamma^2_{b\to i} = \frac{1}{N}\sum_{j\neq i}\left(1-(m^2_{j\to b})^2\right), \tag{12.8d}$$

$$\Xi_{b\to i} = \frac{1}{N}\sum_{j\neq i}\left(q_{j\to b}-m^1_{j\to b}m^2_{j\to b}\right), \tag{12.8e}$$

where $G$ and $\Gamma$ denote the mean and variance of the Gaussian random variables, respectively, and the last quantity denotes the covariance between $X_b$ and $Y_b$. As a result, we can express the first and second statistics of $X_b$ and $Y_b$ as follows:

$$\begin{aligned}
&\langle X_b\rangle = G^1_{b\to i}, \quad \langle Y_b\rangle = G^2_{b\to i},\\
&\langle X_b^2\rangle - \langle X_b\rangle^2 = \Gamma^1_{b\to i}, \quad \langle Y_b^2\rangle - \langle Y_b\rangle^2 = \Gamma^2_{b\to i},\\
&\langle X_bY_b\rangle - \langle X_b\rangle\langle Y_b\rangle = \Xi_{b\to i}.
\end{aligned} \tag{12.9}$$

By the reparametrization trick, we express $X_b$ and $Y_b$ by two standard Gaussian variables $x$, $y$ as

$$\begin{aligned}
X_b &= G^1_{b\to i} + \sqrt{\Gamma^1_{b\to i}}x,\\
Y_b &= G^2_{b\to i} + \sqrt{\Gamma^2_{b\to i}}(\psi x + \sqrt{1-\psi^2}y),\\
\psi &= \frac{\Xi_{b\to i}}{\sqrt{\Gamma^1_{b\to i}\Gamma^2_{b\to i}}}.
\end{aligned} \tag{12.10}$$

To compute Eq. (12.7b) under the joint Gaussian distribution $P(X_b, Y_b)$, we use the following analytic integral:

$$\begin{aligned}
I &= \iint Dx\,Dy\cosh(Ax+D)\cosh(Bx+Cy+E)\\
&= \frac{1}{2}e^{\frac{C^2}{2}}\left[e^{\frac{1}{2}(A+B)^2}\cosh(D+E) + e^{\frac{1}{2}(A-B)^2}\cosh(D-E)\right].
\end{aligned} \tag{12.11}$$

The cavity distribution $P_{j\to b}(\xi^1_j, \xi^2_j)$ can be parameterized as $P_{j\to b}(\xi^1_j, \xi^2_j) = \frac{1+m^1_{j\to b}\xi^1_j + m^2_{j\to b}\xi^2_j + q_{j\to b}\xi^1_j\xi^2_j}{4}$. The cavity magnetization is thus defined as $m^{1,2}_{j\to b} = \sum_{\xi^1_j,\xi^2_j}\xi^{1,2}_j P_{j\to b}(\xi^1_j,\xi^2_j)$, and the cavity correlation is defined as $q_{j\to b} = \sum_{\xi^1_j,\xi^2_j}\xi^1_j\xi^2_j P_{j\to b}(\xi^1_j,\xi^2_j)$. Finally, using the above parameters of the correlated Gaussian distribution, we rewrite $\mu_{b\to i}(\xi^1_i, \xi^2_i)$ as

$$\begin{aligned}
\mu_{b\to i}(\xi^1_i,\xi^2_i) = \frac{1}{\cosh\left(\beta^2 Q_{b\to i} + \frac{\beta^2}{N}\xi^1_i\xi^2_i\right)}\iint Dx\,Dy\cosh\left(\beta\sqrt{\Gamma^1_{b\to i}}x + \beta G^1_{b\to i} + \frac{\beta}{\sqrt{N}}\xi^1_i\sigma^b_i\right)\\
\times\cosh\left(\beta\sqrt{\Gamma^2_{b\to i}}(\psi x+\sqrt{1-\psi^2}y) + \beta G^2_{b\to i} + \frac{\beta}{\sqrt{N}}\xi^2_i\sigma^b_i\right),
\end{aligned} \tag{12.12}$$

where $Dx \equiv \frac{e^{-x^2/2}dx}{\sqrt{2\pi}}$, $\psi = \frac{\Xi_{b\to i}}{\sqrt{\Gamma^1_{b\to i}\Gamma^2_{b\to i}}}$, and $Q_{b\to i} = \frac{1}{N}\sum_{j\neq i} q_{j\to b}$ (coming from $Q_c$) replaced by its cavity mean. The above integral representation of $\mu_{b\to i}(\xi_i^1, \xi_i^2)$ can be analytically worked out. Then, $u_{b\to i} \stackrel{\text{def}}{=} \ln \mu_{b\to i}$ can be expressed as follows:

$$\begin{aligned} u_{b\to i}(\xi_i^1, \xi_i^2) = {} & \frac{\beta^2\Gamma^2_{b\to i}(1-\psi^2)}{2} - \ln\left(2\cosh\left(\beta^2 Q_{b\to i} + \frac{\beta^2\xi_i^1\xi_i^2}{N}\right)\right) + \frac{\beta^2}{2}\left(\sqrt{\Gamma^1_{b\to i}}\right. \\ & \left. + \sqrt{\Gamma^2_{b\to i}}\psi\right)^2 + \ln\cosh\left(\beta G^1_{b\to i} + \beta G^2_{b\to i} + \frac{\beta}{\sqrt{N}}\sigma_i^b(\xi_i^1 + \xi_i^2)\right) \\ & + \ln\left[1 + e^{-2\beta^2\sqrt{\Gamma^1_{b\to i}\Gamma^2_{b\to i}}\psi}\frac{\cosh\left(\beta G^1_{b\to i} - \beta G^2_{b\to i} + \frac{\beta}{\sqrt{N}}\sigma_i^b(\xi_i^1 - \xi_i^2)\right)}{\cosh\left(\beta G^1_{b\to i} + \beta G^2_{b\to i} + \frac{\beta}{\sqrt{N}}\sigma_i^b(\xi_i^1 + \xi_i^2)\right)}\right], \end{aligned} \tag{12.13}$$

where the integral identity in Eq. (12.11) has been used.

To close the iteration equation, we need to compute the cavity magnetization and correlation as follows:

$$m^1_{j\to a} = \frac{\prod_{b\in\partial i\setminus} \mu^{++}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{+-}_{b\to i} - \prod_{b\in\partial i\setminus a} \mu^{-+}_{b\to i} - \prod_{b\in\partial i\setminus a} \mu^{--}_{b\to i}}{\prod_{b\in\partial i\setminus a} \mu^{++}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{+-}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{-+}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{--}_{b\to i}}, \tag{12.14a}$$

$$m^2_{j\to a} = \frac{\prod_{b\in\partial i\setminus a} \mu^{++}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{-+}_{b\to i} - \prod_{b\in\partial i\setminus a} \mu^{+-}_{b\to i} - \prod_{b\in\partial i\setminus a} \mu^{--}_{b\to i}}{\prod_{b\in\partial i\setminus a} \mu^{++}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{+-}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{-+}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{--}_{b\to i}}, \tag{12.14b}$$

$$q_{j\to a} = \frac{\prod_{b\in\partial i\setminus a} \mu^{++}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{--}_{b\to i} - \prod_{b\in\partial i\setminus a} \mu^{-+}_{b\to i} - \prod_{b\in\partial i\setminus a} \mu^{+-}_{b\to i}}{\prod_{b\in\partial i\setminus a} \mu^{++}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{+-}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{-+}_{b\to i} + \prod_{b\in\partial i\setminus a} \mu^{--}_{b\to i}}, \tag{12.14c}$$

where $\mu^{\pm,\pm}_{b\to i} \equiv \mu_{b\to i}(\xi_i^1 = \pm 1, \xi_i^2 = \pm 1)$. We define $u_{b\to i}(\xi_i^1, \xi_i^2) \equiv \ln \mu_{b\to i}(\xi_i^1, \xi_i^2)$ before for the purpose to recast $m^1_{j\to b}$, $m^2_{j\to b}$, and $q_{j\to b}$ in a compact form

$$m^1_{i\to a} = \frac{\sum_{\xi^1=\pm1,\xi^2=\pm1} \xi^1 e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi^1,\xi^2)}}{\sum_{\xi^1=\pm1,\xi^2=\pm1} e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi^1,\xi^2)}}, \tag{12.15a}$$

$$m^2_{i\to a} = \frac{\sum_{\xi^1=\pm1,\xi^2=\pm1} \xi^2 e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi^1,\xi^2)}}{\sum_{\xi^1=\pm1,\xi^2=\pm1} e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi^1,\xi^2)}}, \tag{12.15b}$$

$$q_{i\to a} = \frac{\sum_{\xi^1=\pm1,\xi^2=\pm1} \xi^1\xi^2 e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi^1,\xi^2)}}{\sum_{\xi^1=\pm1,\xi^2=\pm1} e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi^1,\xi^2)}}. \tag{12.15c}$$

$m_{i\to a}^{1,2}$ can be interpreted as the message passing from feature node $i$ to data node $a$ ($q_{i\to a}$ is also similarly interpreted), while $u_{b\to i}$ can be interpreted as the message passing from data node $b$ to feature node $i$.

Suppose the weak correlation assumption is self-consistent, starting from randomly initialized messages, the learning equations will converge to a fixed point corresponding to a thermodynamically dominant minimum of the Bethe free energy function, which is given by $-\beta f_{\text{Bethe}} = \frac{1}{N}\sum_i \Delta f_i - \frac{N-1}{N}\sum_a \Delta f_a$, where $\Delta f_i = \ln Z_i$ and $\Delta f_a = \ln Z_a$. According to the cavity approximation, the free energy contributions of variable node and data node are derived, respectively, by

$$
\begin{aligned}
Z_i &= \sum_{\xi_i^1,\xi_i^2} \prod_{b\in\partial i\setminus a} \mu_{b\to i}(\xi_i^1,\xi_i^2),\\
Z_a &= \sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2} \frac{1}{\cosh \beta^2 Q_a} \cosh\left(\frac{\beta}{\sqrt{N}}\sum_j \xi_j^1\sigma_j^a\right)\cosh\left(\frac{\beta}{\sqrt{N}}\sum_j \xi_j^2\sigma_j^a\right)\prod_{j\in\partial a} P_{j\to a}(\xi_j^1,\xi_j^2),
\end{aligned}
\tag{12.16}
$$

where $Q_a = \frac{1}{N}\sum_{j\in\partial a} q_{j\to a}$. We then denote $X_a = \frac{1}{\sqrt{N}}\sum_j \xi_j^1\sigma_j^a$ and $Y_a = \frac{1}{\sqrt{N}}\sum_j \xi_j^2\sigma_j^a$. A full (non-cavity) version of relevant quantities to parameterize the above weighted-sums can be defined as

$$
\begin{aligned}
G_a^1 &= \frac{1}{\sqrt{N}}\sum_j \sigma_j^a m_{j\to a}^1,\\
G_a^2 &= \frac{1}{\sqrt{N}}\sum_j \sigma_j^a m_{j\to a}^1,\\
\Gamma_a^1 &= \frac{1}{N}\sum_j (1-(m_{j\to a}^1)^2),\\
\Gamma_a^2 &= \frac{1}{N}\sum_j (1-(m_{j\to a}^2)^2),\\
\Xi_a &= \frac{1}{N}\sum_j (q_{j\to a} - m_{j\to a}^1 m_{j\to a}^2).
\end{aligned}
\tag{12.17}
$$

Thus, $X_a$ and $Y_a$ can be parameterized by standard Gaussian variables $x$ and $y$ as

$$
\begin{aligned}
X_a &= G_a^1 + \sqrt{\Gamma_a^1}x,\\
Y_a &= G_a^2 + \sqrt{\Gamma_a^2}(\psi x + \sqrt{1-\psi^2}y),\\
\psi &= \frac{\Xi_a}{\sqrt{\Gamma_a^1\Gamma_a^2}}.
\end{aligned}
\tag{12.18}
$$

Hence, $Z_a$ can be worked out, leading to

$$\Delta f_a = \frac{\beta^2\Gamma_a^2(1-\tilde{\psi}^2)}{2} - \ln\left(2\cosh(\beta^2 Q_a)\right) + \frac{\beta^2}{2}\left(\sqrt{\Gamma_a^1} + \sqrt{\Gamma_a^2}\tilde{\psi}\right)^2 + \ln\cosh\left(\beta G_a^1 + \beta G_a^2\right) + \ln\left[1 + e^{-2\beta^2\Xi_a}\frac{\cosh\left(\beta G_a^1 - \beta G_a^2\right)}{\cosh\left(\beta G_a^1 + \beta G_a^2\right)}\right], \tag{12.19}$$

where $\tilde{\psi} = \frac{\Xi_a}{\sqrt{\Gamma_a^1\Gamma_a^2}}$. We can also get the marginal probability $P_i(\xi_i^1, \xi_i^2)$, which is defined as $P_i(\xi_i^1, \xi_i^2) = \frac{1+m_i^1\xi_i^1+m_i^2\xi_i^2+q_i\xi_i^1\xi_i^2}{4}$, where $m_i^1$ and $m_i^2$ are the magnetizations of $\xi_i^1$, and $\xi_i^2$, respectively. The marginal posterior probability is given by

$$P_i(\xi_i^1, \xi_i^2) = \frac{1}{Z_i}\prod_{b\in\partial i}\mu_{b\to i}(\xi_i^1, \xi_i^2), \tag{12.20}$$

and $m_j^1$, $m_j^2$, and $q_j$ are given, respectively, by

$$m_j^1 = \frac{\sum_{\xi^1=\pm1,\xi^2=\pm1}\xi^1 e^{\sum_{b\in\partial j}u_{b\to j}(\xi^1,\xi^2)}}{\sum_{\xi^1=\pm1,\xi^2=\pm1}e^{\sum_{b\in\partial j}u_{b\to j}(\xi^1,\xi^2)}}, \tag{12.21a}$$

$$m_j^2 = \frac{\sum_{\xi^1=\pm1,\xi^2=\pm1}\xi^2 e^{\sum_{b\in\partial j}u_{b\to j}(\xi^1,\xi^2)}}{\sum_{\xi^1=\pm1,\xi^2=\pm1}e^{\sum_{b\in\partial j}u_{b\to j}(\xi^1,\xi^2)}}, \tag{12.21b}$$

$$q_j = \frac{\sum_{\xi^1=\pm1,\xi^2=\pm1}\xi^1\xi^2 e^{\sum_{b\in\partial j}u_{b\to j}(\xi^1,\xi^2)}}{\sum_{\xi^1=\pm1,\xi^2=\pm1}e^{\sum_{b\in\partial j}u_{b\to j}(\xi^1,\xi^2)}}. \tag{12.21c}$$

If we consider the prior information $P_0(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$, the posteriori probability $P(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2|\{\boldsymbol{\sigma}^a\}_{a=1}^M)$ is given by

$$P(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2|\{\sigma\}_{a=1}^M) = \frac{\prod_a P(\boldsymbol{\sigma}^a|\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)\prod_{i=1}^N P_0(\xi_i^1, \xi_i^2)}{\sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2}\prod_a P(\boldsymbol{\sigma}^a|\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)\prod_{i=1}^N P_0(\xi_i^1, \xi_i^2)} = \frac{1}{\Omega}\prod_a\frac{1}{\cosh(\beta^2 Q)}\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^1\boldsymbol{\sigma}^a\right)\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^2\boldsymbol{\sigma}^a\right)\prod_{i=1}^N P_0(\xi_i^1, \xi_i^2). \tag{12.22}$$

We have assumed that the prior is factorized over $i$. The self-consistent equations for the cavity distribution $P_{i\to a}(\xi_i^1, \xi_i^2)$ and the auxiliary quantity $\mu_{b\to i}(\xi_i^1, \xi_i^2)$ read as follows:

$$P_{i\to a}(\xi_i^1, \xi_i^2) = \frac{1}{Z_{i\to a}}P_0(\xi_i^1, \xi_i^2)\prod_{b\in\partial i\backslash a}\mu_{b\to i}(\xi_i^1, \xi_i^2), \tag{12.23a}$$

$$\mu_{b\to i}(\xi_i^1, \xi_i^2) = \sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2\backslash\xi_i^1,\xi_i^2}\frac{1}{\cosh\left(\beta^2 Q_c + \frac{\beta^2}{N}\xi_i^1\xi_i^2\right)}\cosh\left(\beta X_b + \frac{\beta}{\sqrt{N}}\xi_i^1\sigma_i^b\right) \times\cosh\left(\beta Y_b + \frac{\beta}{\sqrt{N}}\xi_i^2\sigma_i^b\right)\prod_{j\in\partial b\backslash i}P_{j\to b}(\xi_j^1, \xi_j^2). \tag{12.23b}$$

The cavity magnetization $m^1_{i\to a}, m^2_{i\to a}$ and correlation $q_{i\to a}$ are computed as follows:

$$
\begin{aligned}
m^1_{i\to a} &= \frac{\sum_{\xi_i^1,\xi_i^2} \xi_i^1 e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi_i^1,\xi_i^2)} \times P_0(\xi_i^1,\xi_i^2)}{\sum_{\xi_i^1,\xi_i^2} e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi_i^1,\xi_i^2)} \times P_0(\xi_i^1,\xi_i^2)}, \\
m^2_{i\to a} &= \frac{\sum_{\xi_i^1,\xi_i^2} \xi_i^2 e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi_i^1,\xi_i^2)} \times P_0(\xi_i^1,\xi_i^2)}{\sum_{\xi_i^1,\xi_i^2} e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi_i^1,\xi_i^2)} \times P_0(\xi_i^1,\xi_i^2)}, \\
q_{i\to a} &= \frac{\sum_{\xi_i^1,\xi_i^2} \xi_i^1\xi_i^2 e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi_i^1,\xi_i^2)} \times P_0(\xi_i^1,\xi_i^2)}{\sum_{\xi_i^1,\xi_i^2} e^{\sum_{b\in\partial i\setminus a} u_{b\to i}(\xi_i^1,\xi_i^2)} \times P_0(\xi_i^1,\xi_i^2)}.
\end{aligned}
\tag{12.24}
$$

The free energy shifts can be obtained in the form of $\Delta f_i$ and $\Delta f_a$ given, respectively, by

$$
\Delta f_i = \ln \sum_{\xi_i^1,\xi_i^2} P_0(\xi_i^1,\xi_i^2) \prod_{b\in\partial i} \mu_{b\to i}(\xi_i^1,\xi_i^2), \tag{12.25a}
$$

$$
\begin{aligned}
\Delta f_a = {} & \frac{\beta^2\Gamma_a^2(1-\tilde{\psi}^2)}{2} - \ln\left(2\cosh(\beta^2 Q_a)\right) + \frac{\beta^2}{2}\left(\sqrt{\Gamma_a^1} + \sqrt{\Gamma_a^2}\tilde{\psi}\right)^2 \\
& + \ln\cosh\left(\beta G_a^1 + \beta G_a^2\right) + \ln\left[1 + e^{-2\beta^2\Xi_a}\frac{\cosh\left(\beta G_a^1 - \beta G_a^2\right)}{\cosh\left(\beta G_a^1 + \beta G_a^2\right)}\right].
\end{aligned}
\tag{12.25b}
$$

The above belief propagation equations for either prior-free or prior cases provide us the practical algorithms for the unsupervised learning problem at hand. An easy implementation is carried out on a teacher–student setting. Note that, teacher here does not provide labels of data, unlike the supervised learning. Instead, the teacher setting means that the raw data is generated from a teacher (or ground truth) architecture with specified feature vectors $(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$. Then the student uses the above belief propagation to infer which feature vectors underlie the data, given that only the temperature is known or both the temperature and the correlation prior are known.

### *12.1.2 Replica Computation*

As we show in the previous section, the partition function of the two-bit RBM model is defined as

$$
\Omega = \sum_{\{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2\}} \prod_{a=1}^{M} \frac{\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^1\boldsymbol{\sigma}^a\right)\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^2\boldsymbol{\sigma}^a\right)}{2^N e^{\beta^2}\cosh\left(\beta^2 q\right)}. \tag{12.26}
$$

In order to have an analytical argument about the typical performance and the critical point where the spontaneous symmetry breaking (SSB) phase transition appears, we

have to calculate the free energy in the thermodynamic limit by the replica method. Instead of calculating the disorder average of $\ln \Omega$, the replica method calculates the disorder average of the $n$th moment of $\Omega$, i.e., $\langle \Omega^n \rangle$, where $n$ is the replica number, which means copying $n$ replicas of the original system. The disorder average $\langle \bullet \rangle$ is taken over all possible sampling data and the random realization of the true feature vectors. Using the replica trick, i.e., $\ln x = \lim_{n\to 0} \frac{\partial}{\partial n} x^n$, the free energy density can be obtained as

$$\beta f = \lim_{n\to 0} \lim_{N\to\infty} \frac{\langle \Omega^n \rangle}{nN}. \tag{12.27}$$

Given the two true feature vectors, the data distribution generated by these feature vectors are

$$P(\{\sigma^a\}) = \prod_{a=1}^{M} \frac{\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,true} \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,true} \boldsymbol{\sigma}^a\right)}{Z(\boldsymbol{\xi}^{1,true}, \boldsymbol{\xi}^{2,true})}, \tag{12.28}$$

where the nested partition function

$$\begin{aligned} Z(\boldsymbol{\xi}^{1,true}, \boldsymbol{\xi}^{2,true}) &= \sum_{\boldsymbol{\sigma}} \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,true} \boldsymbol{\sigma}\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,true} \boldsymbol{\sigma}\right) \\ &= 2^N e^{\beta^2} \cosh(\beta^2 q), \end{aligned} \tag{12.29}$$

where $q$ is defined as the overlap between the true feature vectors: $q = \frac{1}{N} \boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{2,true}$.

Next, we show how to compute $\langle \Omega^n \rangle$, which is defined as

$$\begin{aligned} \langle \Omega^n \rangle = &\sum_{\{\boldsymbol{\xi}^{true}, \boldsymbol{\sigma}^a\}} \prod_{i=1}^{N} [P(\xi_i^{1,true}, \xi_i^{2,true})] \prod_{a=1}^{M} \frac{\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,true} \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,true} \boldsymbol{\sigma}^a\right)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \\ &\times \sum_{\{\boldsymbol{\xi}^{1,\gamma}, \boldsymbol{\xi}^{2,\gamma}\}} \prod_{a,\gamma} \frac{\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,\gamma} \boldsymbol{\sigma}^a\right)}{\cosh(\beta^2 R^\gamma)}, \end{aligned} \tag{12.30}$$

where $\gamma$ indicates the replica index, $\boldsymbol{\xi}^{true} = \{\boldsymbol{\xi}^{1,true}, \boldsymbol{\xi}^{2,true}\}$, and $R^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma}$.

To further calculate $\langle \Omega^n \rangle$, we have to introduce the order parameters as follows:

$$T_1^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{1,\gamma}, \qquad T_2^\gamma = \frac{1}{N} \boldsymbol{\xi}^{2,true} \boldsymbol{\xi}^{2,\gamma}, \tag{12.31a}$$

$$\tau_1^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{2,\gamma}, \qquad \tau_2^\gamma = \frac{1}{N} \boldsymbol{\xi}^{2,true} \boldsymbol{\xi}^{1,\gamma}, \tag{12.31b}$$

$$q_1^{\gamma,\gamma'} = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{1,\gamma'}, \qquad q_2^{\gamma,\gamma'} = \frac{1}{N} \boldsymbol{\xi}^{2,\gamma} \boldsymbol{\xi}^{2,\gamma'}, \tag{12.31c}$$

$$R^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma}, \qquad r^{\gamma,\gamma'} = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma'}. \tag{12.31d}$$

Note that these order parameters construct a complete set to describe the problem at hand, although the necessary number of order parameters may be reduced due to symmetry. These order parameters capture the emergent behavior of our model. $T_1$ and $T_2$ characterize the overlap between prediction and ground truth. $q_1$ and $q_2$ characterize the self-overlap (Edwards–Anderson order parameter in physics). $\tau_1$ and $\tau_2$ characterize the permutation-type overlap. $R$ and $r$ characterize the students' guess on the correlation level of the planted receptive fields.

By using these order parameters, the disorder average $\langle \Omega^n \rangle$ can be expressed as

$$
\begin{aligned}
\langle \Omega^n \rangle = & \sum_{\{\boldsymbol{\sigma}^a, \boldsymbol{\xi}^{true}\}} \prod_{i=1}^{N} P(\xi_i^{1,true}, \xi_i^{2,true}) \sum_{\{\boldsymbol{\xi}^{1,\gamma}, \boldsymbol{\xi}^{2,\gamma}\}} \int \prod_{\gamma=1}^{n} dR^\gamma \delta(\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma} - NR^\gamma) \\
& \times \int \prod_{\gamma=1}^{n} dT_1^\gamma \delta(\boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{1,\gamma} - NT_1^\gamma) \int \prod_{\gamma=1}^{n} dT_2^\gamma \delta(\boldsymbol{\xi}^{2,true} \boldsymbol{\xi}^{2,\gamma} - NT_2^\gamma) \\
& \times \int \prod_{\gamma=1}^{n} d\tau_1^\gamma \delta(\boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{2,\gamma} - N\tau_1^\gamma) \int \prod_{\gamma=1}^{n} d\tau_2^\gamma \delta(\boldsymbol{\xi}^{2,true} \boldsymbol{\xi}^{1,\gamma} - N\tau_2^\gamma) \\
& \times \int \prod_{\gamma<\gamma'} dq_1^{\gamma,\gamma'} \delta(\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{1,\gamma'} - Nq_1^{\gamma,\gamma'}) \int \prod_{\gamma<\gamma'} dq_2^{\gamma,\gamma'} \delta(\boldsymbol{\xi}^{2,\gamma} \boldsymbol{\xi}^{2,\gamma'} - Nq_2^{\gamma,\gamma'}) \\
& \times \int \prod_{\gamma<\gamma'} dr^{\gamma,\gamma'} \delta(\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma'} - Nr^{\gamma,\gamma'}) \\
& \times \prod_{a=1}^{M} \left\{ \frac{\cosh(\beta X_a^0) \cosh(\beta Y_a^0)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \prod_{\gamma=1}^{n} \frac{\cosh(\beta X_a^\gamma) \cosh(\beta Y_a^\gamma)}{\cosh(\beta^2 R^\gamma)} \right\} \\
= & \sum_{\{\boldsymbol{\sigma}^a, \boldsymbol{\xi}^{true}\}} \prod_{i=1}^{N} P(\xi_i^{1,true}, \xi_i^{2,true}) \sum_{\{\boldsymbol{\xi}^{1,\gamma}, \boldsymbol{\xi}^{2,\gamma}\}} \int \prod_{\gamma=1}^{n} \left( \frac{dR^\gamma d\hat{R}^\gamma}{2\pi} \right) \int \prod_{\gamma=1}^{n} \left( \frac{dT_1^\gamma d\hat{T_1}^\gamma}{2\pi} \right) \\
& \times \int \prod_{\gamma=1}^{n} \left( \frac{dT_2^\gamma d\hat{T_2}^\gamma}{2\pi} \right) \int \prod_{\gamma=1}^{n} \left( \frac{d\tau_1^\gamma d\hat{\tau}_1^\gamma}{2\pi} \right) \int \prod_{\gamma=1}^{n} \left( \frac{d\tau_2^\gamma d\hat{\tau}_2^\gamma}{2\pi} \right) \\
& \times \int \prod_{\gamma<\gamma'} \left( \frac{dq_1^{\gamma,\gamma'} d\hat{q}_1^{\gamma,\gamma'}}{2\pi} \right) \int \prod_{\gamma<\gamma'} \left( \frac{dq_2^{\gamma,\gamma'} d\hat{q}_2^{\gamma,\gamma'} dr^{\gamma,\gamma'} d\hat{r}^{\gamma,\gamma'}}{4\pi^2} \right) \\
& \times \exp\left( \sum_{\gamma=1}^{n} \mathrm{i}\hat{R}^\gamma (\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma} - NR^\gamma) + \sum_{\gamma=1}^{n} \mathrm{i}\hat{T_1}^\gamma (\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{1,true} - NT_1^\gamma) + \sum_{\gamma=1}^{n} \mathrm{i}\hat{T_2}^\gamma (\boldsymbol{\xi}^{2,\gamma} \boldsymbol{\xi}^{2,true} - NT_2^\gamma) \right) \\
& \times \exp\left( \sum_{\gamma=1}^{n} \mathrm{i}\hat{\tau_1}^\gamma (\boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{2,\gamma} - N\tau_1^\gamma) + \sum_{\gamma=1}^{n} \mathrm{i}\hat{\tau_2}^\gamma (\boldsymbol{\xi}^{2,true} \boldsymbol{\xi}^{1,\gamma} - N\tau_2^\gamma) + \sum_{\gamma<\gamma'} \mathrm{i}\hat{q_1}^{\gamma,\gamma'} (\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{1,\gamma'} - Nq_1^{\gamma,\gamma'}) \right) \\
& \times \exp\left( \sum_{\gamma<\gamma'} \mathrm{i}\hat{q_2}^{\gamma,\gamma'} (\boldsymbol{\xi}^{2,\gamma} \boldsymbol{\xi}^{2,\gamma'} - Nq_2^{\gamma,\gamma'}) + \sum_{\gamma<\gamma'} \mathrm{i}\hat{r}^{\gamma,\gamma'} (\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma'} - Nr^{\gamma,\gamma'}) \right) \\
& \times \prod_{a=1}^{M} \left\{ \frac{\cosh(\beta X_a^0) \cosh(\beta Y_a^0)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \prod_{\gamma=1}^{n} \frac{\cosh(\beta X_a^\gamma) \cosh(\beta Y_a^\gamma)}{\cosh(\beta^2 R^\gamma)} \right\},
\end{aligned}
\tag{12.32}
$$

where we have defined $X_a^0 = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^{1,true} \sigma_i^a$, $Y_a^0 = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^{2,true} \sigma_i^a$, and $X_a^\gamma = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^{1,\gamma} \sigma_i^a$, $Y_a^\gamma = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^{2,\gamma} \sigma_i^a$. To get the second equality, we have used the integral representation of the delta function $\delta(x) = \int \frac{d\hat{x}}{2\pi} e^{\mathrm{i}\hat{x}x}$. Hence, we have to introduce the conjugate order parameters

$$(\hat{T}_1{}^{\gamma}, \hat{T}_2{}^{\gamma}, \hat{\tau}_1{}^{\gamma}, \hat{\tau}_2{}^{\gamma}, \hat{q}_1{}^{\gamma,\gamma'}, \hat{q}_2{}^{\gamma,\gamma'}, \hat{R}^{\gamma}, \hat{r}^{\gamma,\gamma'}) \tag{12.33}$$

corresponding to the non-conjugated (physical) order parameters

$$(T_1^{\gamma}, T_2^{\gamma}, \tau_1^{\gamma}, \tau_2^{\gamma}, q_1^{\gamma,\gamma'}, q_2^{\gamma,\gamma'}, R^{\gamma}, r^{\gamma,\gamma'}). \tag{12.34}$$

To further compute an explicit form of the free energy, we assume a simple ansatz, i.e., all order parameters do not depend on their specific replica indexes, which is called the replica-symmetry assumption. To be more precise, we assume

$$R^{\gamma} = R, \qquad \mathrm{i}\hat{R}^{\gamma} = \hat{R}, \tag{12.35a}$$
$$T_1^{\gamma} = T_1, \qquad \mathrm{i}\hat{T}_1{}^{\gamma} = \hat{T}_1, \tag{12.35b}$$
$$T_2^{\gamma} = T_2, \qquad \mathrm{i}\hat{T}_2{}^{\gamma} = \hat{T}_2, \tag{12.35c}$$
$$\tau_1^{\gamma} = \tau_1, \qquad \mathrm{i}\hat{\tau}_1^{\gamma} = \hat{\tau}_1, \tag{12.35d}$$
$$\tau_2^{\gamma} = \tau_2, \qquad \mathrm{i}\hat{\tau}_2^{\gamma} = \hat{\tau}_2, \tag{12.35e}$$

for any $\gamma$. We also assume that

$$q_1^{\gamma,\gamma'} = q_1, \qquad \mathrm{i}\hat{q}_1{}^{\gamma,\gamma'} = \hat{q}_1, \tag{12.36a}$$
$$q_2^{\gamma,\gamma'} = q_2, \qquad \mathrm{i}\hat{q}_2{}^{\gamma,\gamma'} = \hat{q}_2, \tag{12.36b}$$
$$r^{\gamma,\gamma'} = r, \qquad \mathrm{i}\hat{r}^{\gamma,\gamma'} = \hat{r}, \tag{12.36c}$$

for any $\gamma$ and $\gamma'$. Then we can express $\langle \Omega^n \rangle$ as

$$\langle \Omega^n \rangle = \int dO d\hat{O} e^{N\mathcal{A}(O,\hat{O},\alpha,\beta,n)}. \tag{12.37}$$

In the thermodynamics limit, $\langle \Omega^n \rangle$ can be approximated as $e^{N\mathcal{A}(O^*,\hat{O}^*,\alpha,\beta,n)}$ (namely the saddle-point method), where $O^*$ and $\hat{O}^*$ represent all non-conjugated order parameters and conjugated order parameters evaluated at the maximal value of the action, respectively. The expression for the action $\mathcal{A}(O, \hat{O}, \alpha, \beta, n)$ (we omit $*$ hereafter) can be written by

$$\begin{aligned}\mathcal{A} = &-nR\hat{R} - nT_1\hat{T}_1 - nT_2\hat{T}_2 - n\tau_1\hat{\tau}_1 - n\tau_2\hat{\tau}_2 - \frac{n(n-1)}{2}q_1\hat{q}_1 \\ &- \frac{n(n-1)}{2}q_2\hat{q}_2 - \frac{n(n-1)}{2}r\hat{r} + G_S + \alpha G_E,\end{aligned} \tag{12.38}$$

where $G_S$ is the entropy term, and $G_E$ is the energy term.

To derive the entropy term $G_S$, we use the following identities:

$$\sum_{\gamma<\gamma'} \xi^{1,\gamma}\xi^{1,\gamma'} = \frac{1}{2}\left(\sum_{\gamma}\xi^{1,\gamma}\right)^2 - \frac{1}{2}\sum_{\gamma}(\xi^{1,\gamma})^2, \tag{12.39a}$$

$$\sum_{\gamma<\gamma'} \xi^{2,\gamma}\xi^{2,\gamma'} = \frac{1}{2}\left(\sum_{\gamma}\xi^{2,\gamma}\right)^2 - \frac{1}{2}\sum_{\gamma}(\xi^{2,\gamma})^2, \tag{12.39b}$$

$$\begin{aligned}\sum_{\gamma<\gamma'} \xi^{1,\gamma}\xi^{2,\gamma'} &= \frac{1}{2}\sum_{\gamma,\gamma'}\xi^{1,\gamma}\xi^{2,\gamma'} - \frac{1}{2}\sum_{\gamma}\xi^{1,\gamma}\xi^{2,\gamma} \\ &= \frac{1}{4}\left(\sum_{\gamma}\xi^{1,\gamma} + \sum_{\gamma'}\xi^{2,\gamma'}\right)^2 - \frac{1}{4}\left(\sum_{\gamma}\xi^{1,\gamma}\right)^2 - \frac{1}{4}\left(\sum_{\gamma'}\xi^{2,\gamma'}\right)^2 - \frac{1}{2}\sum_{\gamma}\xi^{1,\gamma}\xi^{2,\gamma}.\end{aligned} \tag{12.39c}$$

The above non-linear terms can be reduced to linear terms in the exponential functions of Eq. (12.37) by the Hubbard–Stratonovich transformation $\int Dt e^{bt} = e^{\frac{1}{2}b^2}$. Then, we obtain $G_S$ as

$$\begin{aligned}G_S = \ln\Bigg[\sum_{\{\xi^{1,\gamma},\xi^{2,\gamma}\}} &\exp\left(\hat{R}\sum_{\gamma=1}^{n}\xi^{1,\gamma}\xi^{2,\gamma} + \hat{T}_1\sum_{\gamma=1}^{n}\xi^{1,\gamma}\xi^{1,true} + \hat{T}_2\sum_{\gamma=1}^{n}\xi^{2,\gamma}\xi^{2,true}\right. \\ &\left.+\hat{\tau}_1\sum_{\gamma=1}^{n}\xi^{1,true}\xi^{2,\gamma}\right) \times \exp\left(\hat{\tau}_2\sum_{\gamma=1}^{n}\xi^{1,\gamma}\xi^{2,true} + \hat{q}_1\sum_{\gamma<\gamma'}\xi^{1,\gamma}\xi^{1,\gamma'}\right. \\ &\left.+\hat{q}_2\sum_{\gamma<\gamma'}\xi^{2,\gamma}\xi^{2,\gamma'} + \hat{r}\sum_{\gamma<\gamma'}\xi^{1,\gamma}\xi^{2,\gamma'}\right)\Bigg]_{\xi^{1,true},\xi^{2,true}} \\ = \ln\Bigg[\sum_{\{\xi^{1,\gamma},\xi^{2,\gamma}\}} &\exp\left(\frac{\hat{q}_1 - \frac{\hat{r}}{2}}{2}\left(\sum_{\gamma}\xi^{1,\gamma}\right)^2 + \frac{\hat{q}_2 - \frac{\hat{r}}{2}}{2}\left(\sum_{\gamma}\xi^{2,\gamma}\right)^2 + \hat{T}_1\sum_{\gamma}\xi^{1,\gamma}\xi^{1,true}\right) \\ &\times \exp\left(\frac{\hat{r}}{4}\left(\sum_{\gamma}\xi^{1,\gamma} + \sum_{\gamma'}\xi^{2,\gamma'}\right)^2 + \hat{T}_2\sum_{\gamma}\xi^{2,\gamma}\xi^{2,true} + (\hat{R} - \frac{\hat{r}}{2})\sum_{\gamma}\xi^{1,\gamma}\xi^{2,\gamma}\right) \\ &\times \exp\left(\hat{\tau}_1\sum_{\gamma}\xi^{1,true}\xi^{2,\gamma} + \hat{\tau}_2\sum_{\gamma}\xi^{2,true}\xi^{1,\gamma} - \frac{n}{2}\hat{q}_1 - \frac{n}{2}\hat{q}_2\right)\Bigg]_{\xi^{1,true},\xi^{2,true}} \\ = \ln\Bigg[\sum_{\{\xi^{1,\gamma},\xi^{2,\gamma}\}} &\int D\mathbf{z}\exp\left(\sum_{\gamma}\sqrt{\hat{q}_1 - \frac{\hat{r}}{2}}\xi^{1,\gamma}z_1 + \sum_{\gamma}\sqrt{\hat{q}_2 - \frac{\hat{r}}{2}}\xi^{2,\gamma}z_2\right. \\ &\left.+\sqrt{\frac{\hat{r}}{2}}z_3\left(\sum_{\gamma}\xi^{1,\gamma} + \sum_{\gamma'}\xi^{2,\gamma'}\right)\right) \\ &\times \exp\left(\hat{T}_1\sum_{\gamma}\xi^{1,true}\xi^{1,\gamma} + \hat{T}_2\sum_{\gamma}\xi^{2,\gamma}\xi^{2,true} + \hat{\tau}_1\sum_{\gamma}\xi^{1,true}\xi^{2,\gamma}\right) \\ &\times \exp\left(\hat{\tau}_2\sum_{\gamma}\xi^{2,true}\xi^{1,\gamma} + (\hat{R} - \frac{\hat{r}}{2})\sum_{\gamma}\xi^{1,\gamma}\xi^{2,\gamma} - \frac{n}{2}\hat{q}_1 - \frac{n}{2}\hat{q}_2\right)\Bigg]_{\xi^{1,true},\xi^{2,true}}.\end{aligned} \tag{12.40}$$

Finally, we can express the entropy term $G_S$ in a compact form as

$$G_S = \ln \left[ \int D\mathbf{z} \left( \sum_{\xi^1,\xi^2} e^{b_1\xi^1 + b_2\xi^2 + b_3\xi^1\xi^2} \right)^n \right]_{\xi^{1,true},\xi^{2,true}} - \frac{n}{2}\hat{q}_1 - \frac{n}{2}\hat{q}_2, \quad (12.41)$$

where we have defined $D\mathbf{z} = Dz_1 Dz_2 Dz_3$, and the auxiliary variables $b_1$, $b_2$, and $b_3$ as

$$b_1 = \sqrt{\hat{q}_1 - \frac{\hat{r}}{2}} z_1 + \sqrt{\frac{\hat{r}}{2}} z_3 + \hat{T}_1 \xi^{1,true} + \hat{\tau}_2 \xi^{2,true}, \quad (12.42a)$$

$$b_2 = \sqrt{\hat{q}_2 - \frac{\hat{r}}{2}} z_2 + \sqrt{\frac{\hat{r}}{2}} z_3 + \hat{T}_2 \xi^{2,true} + \hat{\tau}_1 \xi^{1,true}, \quad (12.42b)$$

$$b_3 = \hat{R} - \frac{\hat{r}}{2}. \quad (12.42c)$$

We remark that in the expression of $G_S$, the inner summation over $\xi^1, \xi^2$ can be thought as a two-spin interaction partition function, which is defined as $Z_{\text{eff}} = \sum_{\xi^1,\xi^2} e^{b_1\xi^1 + b_2\xi^2 + b_3\xi^1\xi^2} = 2e^{b_3}\cosh(b_1 + b_2) + 2e^{-b_3}\cosh(b_1 - b_2)$. $[\bullet]_{\xi^{1,true},\xi^{2,true}}$ means an average w.r.t $P(\boldsymbol{\xi}^{1,true}, \boldsymbol{\xi}^{2,true})$. This simplification is due to the introduction of replicas, i.e., the original spin interaction decouples, being transformed into the overlap matrix.

Next, we turn to compute the energy term $G_E$. The expression of $G_E$ is given by

$$G_E = \ln \left\langle \frac{\cosh(\beta X^0)\cosh(\beta Y^0)}{\cosh(\beta^2 q)} \prod_{\gamma=1}^{n} \frac{\cosh(\beta X^\gamma)\cosh(\beta Y^\gamma)}{\cosh(\beta^2 R^\gamma)} \right\rangle, \quad (12.43)$$

where $\langle \bullet \rangle$ defines the disorder average. $X^0$, $Y^0$, $X^\gamma$, $Y^\gamma$ are correlated Gaussian random variables, which are the same as before but the data index $a$ has been dropped off. They have zero mean and unit variance. Their covariances are determined by the aforementioned order parameters as follows:

$$\langle X^0 Y^0 \rangle = q, \qquad \langle X^0 X^\gamma \rangle = T_1, \qquad \langle X^0 Y^\gamma \rangle = \tau_1, \quad (12.44a)$$

$$\langle X^\gamma X^{\gamma'} \rangle = q_1, \qquad \langle Y^\gamma Y^{\gamma'} \rangle = q_2, \qquad \langle X^\gamma Y^\gamma \rangle = R, \quad (12.44b)$$

$$\langle Y^0 Y^\gamma \rangle = T_2, \qquad \langle Y^0 X^\gamma \rangle = \tau_2, \qquad \langle X^\gamma Y^{\gamma'} \rangle = r. \quad (12.44c)$$

The random variables $X^0$, $Y^0$, $X^\gamma$, $Y^\gamma$ can thus be parameterized by six standard Gaussian variables of zero mean and unit variance $(t_0, x_0, u, u', y_\gamma, \omega_\gamma)$ as follows:

$$X^0 = t_0, \quad (12.45a)$$

$$Y^0 = q t_0 + \sqrt{1 - q^2} x_0, \quad (12.45b)$$

$$X^{\gamma} = T_1 t_0 + \frac{\tau_2 - T_1 q}{\sqrt{1-q^2}} x_0 + Bu + \sqrt{1-q_1}\omega_{\gamma}, \tag{12.45c}$$

$$Y^{\gamma} = \tau_1 t_0 + \frac{T_2 - \tau_1 q}{\sqrt{1-q^2}} x_0 + \frac{r-A}{B} u + \frac{R-r}{\sqrt{1-q_1}}\omega_{\gamma} + Ku' + \sqrt{1 - q_2 - \frac{(R-r)^2}{1-q_1}}\, y_{\gamma}, \tag{12.45d}$$

where $A = T_1\tau_1 + \frac{(\tau_2 - T_1 q)(T_2 - \tau_1 q)}{1-q^2}$, $B = \sqrt{q_1 - (T_1)^2 - \frac{(\tau_2 - T_1 q)^2}{1-q^2}}$, and $K = \sqrt{q_2 - (\tau_1)^2 - \frac{(T_2 - \tau_1 q)^2}{1-q^2} - (\frac{r-A}{B})^2}$. One can easily verify that the above parameterization satisfies their covariance structures. Therefore, the $G_E$ term can be calculated by a standard Gaussian integration given by

$$G_E = \ln\left[\int Dt_0 Dx_0 Du Du' \frac{\cosh(\beta t^0)\cosh\beta(qt_0 + \sqrt{1-q^2}x_0)}{\cosh(\beta^2 q)} \times \left(\int D\omega Dy \frac{1}{\cosh(\beta^2 R)} \cosh\beta\left(T_1 t_0 + \frac{\tau_2 - T_1 q}{\sqrt{1-q^2}} x_0 + Bu + \sqrt{1-q_1}\omega\right) \times \cosh\beta\left(\tau_1 t_0 + \frac{T_2 - \tau_1 q}{\sqrt{1-q^2}} x_0 + \frac{r-A}{B} u + + \frac{R-r}{\sqrt{1-q_1}}\omega + Ku' + Cy\right)\right)^n\right], \tag{12.46}$$

where $C \equiv \sqrt{1 - q_2 - \frac{(R-r)^2}{1-q_1}}$.

To proceed, we first define the auxiliary quantities as

$$\Lambda_+ = (T_1 + \tau_1)t_0 + \frac{(T_2 + \tau_2) - q(T_1 + \tau_1)}{\sqrt{1-q^2}} x_0 + \left(B + \frac{r-A}{B}\right) u + Ku', \tag{12.47a}$$

$$\Lambda_- = (T_1 - \tau_1)t_0 + \frac{(\tau_2 - T_2) - q(T_1 - \tau_1)}{\sqrt{1-q^2}} x_0 + \left(B - \frac{r-A}{B}\right) u - Ku'. \tag{12.47b}$$

Then we compute the integral inside the power $n$, which is defined by $I$ whose result is given by

$$
\begin{aligned}
I \equiv & \int D\omega Dy \Bigg[ \cosh\beta\Bigg(\tau_1 t_0 + \frac{T_2 - \tau_1 q}{\sqrt{1-q^2}} x_0 + \frac{r-A}{B} u + \frac{R-r}{\sqrt{1-q_1}}\omega \\
& + K u' + \sqrt{1 - q_2 - \frac{(R-r)^2}{1-q_1}}\, y\Bigg) \times \cosh\beta(T_1 t_0 + \frac{\tau_2 - T_1 q}{\sqrt{1-q^2}} x_0 + Bu + \sqrt{1-q_1}\omega)\Bigg] \\
= & \frac{1}{4}\int D\omega Dy \Bigg[ e^{\beta\{\Lambda_+ + (\sqrt{1-q_1} + \frac{R-r}{\sqrt{1-q_1}})\omega + \sqrt{1-q_2-\frac{(R-r)^2}{1-q_1}}\, y\}} + e^{-\beta\{\Lambda_+ + (\sqrt{1-q_1} + \frac{R-r}{\sqrt{1-q_1}})\omega + \sqrt{1-q_2-\frac{(R-r)^2}{1-q_1}}\, y\}} \\
& + e^{\beta\{\Lambda_- + (\sqrt{1-q_1} - \frac{R-r}{\sqrt{1-q_1}})\omega - \sqrt{1-q_2-\frac{(R-r)^2}{1-q_1}}\, y\}} + e^{-\beta\{\Lambda_- + (\sqrt{1-q_1} - \frac{R-r}{\sqrt{1-q_1}})\omega - \sqrt{1-q_2-\frac{(R-r)^2}{1-q_1}}\, y\}} \Bigg] \\
= & \frac{1}{2} e^{\beta^2(1 - \frac{q_1+q_2}{2})} \Big[ e^{\beta^2(R-r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta\Lambda_-) \Big].
\end{aligned}
\tag{12.48}
$$

For simplicity, we also define the following auxiliary quantities $Z_E, G_c^-, G_s^+, G_s^-$:

$$
\begin{aligned}
Z_E &= e^{\beta^2(R-r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta\Lambda_-), \\
G_c^- &= \frac{e^{\beta^2(R-r)} \cosh(\beta\Lambda_+) - e^{-\beta^2(R-r)} \cosh(\beta\Lambda_-)}{e^{\beta^2(R-r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta\Lambda_-)}, \\
G_s^+ &= \frac{e^{\beta^2(R-r)} \sinh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \sinh(\beta\Lambda_-)}{e^{\beta^2(R-r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta\Lambda_-)}, \\
G_s^- &= \frac{e^{\beta^2(R-r)} \sinh(\beta\Lambda_+) - e^{-\beta^2(R-r)} \sinh(\beta\Lambda_-)}{e^{\beta^2(R-r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta\Lambda_-)}.
\end{aligned}
\tag{12.49}
$$

Following the replica trick, we can get

$$
\lim_{n\to 0} \frac{G_E}{n} = \frac{\int Dt_0 Dx_0 Du Du' \frac{\cosh\beta t_0 \cosh\beta(q t_0 + \sqrt{1-q^2} x_0)}{\cosh\beta^2 q} \ln\left[\frac{I}{\cosh\beta^2 R}\right]}{\int Dt_0 Dx_0 Du Du' \frac{\cosh(\beta t_0)\cosh\beta(q t_0 + \sqrt{1-q^2} x_0)}{\cosh(\beta^2 q)}},
\tag{12.50}
$$

where the integral in the denominator can be exactly computed with the result [see also Eq. (12.11)] given by

$$
\begin{aligned}
& \int Dt_0 Dx_0 Du Du' \cosh(\beta t_0) \cosh\beta(q t_0 + \sqrt{1-q^2} x_0) \\
& = \frac{1}{2}\left( e^{\frac{\beta^2}{2}(1-q)^2 + \frac{\beta^2}{2}(1-q^2)} + e^{\frac{\beta^2}{2}(1+q)^2 + \frac{\beta^2}{2}(1-q^2)} \right) = e^{\beta^2} \cosh\beta^2 q.
\end{aligned}
\tag{12.51}
$$

Finally, by collecting all the above relevant terms, we have the following estimation of $\langle \Omega^n \rangle$ given by:

$$\begin{aligned}\langle\Omega^n\rangle = &\int dOd\hat{O}\exp\left(-NnR\hat{R} - NnT_1\hat{T}_1 - NnT_2\hat{T}_2 - Nn\tau_2\hat{\tau}_2 - \frac{N}{2}n(n-1)q_1\hat{q}_1\right)\\ &\times\exp\left(-\frac{N}{2}n(n-1)q_2\hat{q}_2 - \frac{N}{2}n(n-1)r\hat{r} - \frac{nN}{2}\hat{q}_1 - \frac{nN}{2}\hat{q}_2 + N\ln\left[\int D\mathbf{z}Z_{\text{eff}}^n\right]_{\xi^{1,true},\xi^{2,true}}\right.\\ &\left.+\alpha N\ln\left\{\int D\mathbf{t}\frac{\cosh(\beta t_0)\cosh\beta(qt_0+\sqrt{1-q^2}x_0)}{\cosh(\beta^2 q)}\left[\frac{I}{\cosh(\beta^2 R)}\right]^n\right\}\right),\end{aligned} \tag{12.52}$$

where in shorthand $D\mathbf{t} = Dt_0 Dx_0 Du Du'$. By computing $\lim_{n\to 0}\frac{\ln\langle\Omega^n\rangle}{n}$ and using Eq. (12.50), we get the expression $F_\beta = -\beta f_{\text{RS}}$ as

$$\begin{aligned}F_\beta = &-R\hat{R} - T_1\hat{T}_1 - T_2\hat{T}_2 - \tau_1\hat{\tau}_1 - \tau_2\hat{\tau}_2 + \frac{\hat{q}_1}{2}(q_1-1) + \frac{\hat{q}_2}{2}(q_2-1)\\ &+\frac{r\hat{r}}{2} + \int D\mathbf{z}\left[\ln Z_{\text{eff}}\right]_{\xi^{1,true},\xi^{2,true}} - \alpha\ln\left(2\cosh(\beta^2 R)\right) + \alpha\beta^2\left(1 - \frac{q_1+q_2}{2}\right)\\ &+\frac{\alpha e^{-\beta^2}}{\cosh(\beta^2 q)}\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)\ln Z_E.\end{aligned} \tag{12.53}$$

Note that we have used $\lim_{n\to 0}\frac{\ln\left[\int D\mathbf{z}Z_{\text{eff}}^n\right]_{\xi^{1,true},\xi^{2,true}}}{n} = \int D\mathbf{z}\left[\ln Z_{\text{eff}}\right]_{\xi^{1,true},\xi^{2,true}}$ to arrive at the final expression.

By the saddle-point analysis, these non-conjugated order parameters $O$ should obey the following stationary conditions:

$$\frac{\partial F_\beta}{\partial R} = 0, \quad \frac{\partial F_\beta}{\partial r} = 0, \quad \frac{\partial F_\beta}{\partial q_1} = 0, \quad \frac{\partial F_\beta}{\partial q_2} = 0, \tag{12.54a}$$

$$\frac{\partial F_\beta}{\partial T_1} = 0, \quad \frac{\partial F_\beta}{\partial T_2} = 0, \quad \frac{\partial F_\beta}{\partial \tau_1} = 0, \quad \frac{\partial F_\beta}{\partial \tau_2} = 0. \tag{12.54b}$$

Similarly, for conjugated order parameters $\hat{O}$, the following stationary conditions should be satisfied:

$$\frac{\partial F_\beta}{\partial \hat{R}} = 0, \quad \frac{\partial F_\beta}{\partial \hat{r}} = 0, \quad \frac{\partial F_\beta}{\partial \hat{q}_1} = 0, \quad \frac{\partial F_\beta}{\partial \hat{q}_2} = 0, \tag{12.55a}$$

$$\frac{\partial F_\beta}{\partial \hat{T}_1} = 0, \quad \frac{\partial F_\beta}{\partial \hat{T}_2} = 0, \quad \frac{\partial F_\beta}{\partial \hat{\tau}_1} = 0, \quad \frac{\partial F_\beta}{\partial \hat{\tau}_2} = 0. \tag{12.55b}$$

We first evaluate the self-consistent equations those non-conjugated order parameters obey. For $R$, we have the following equation:

$$\frac{\partial F_\beta}{\partial \hat{R}} = -R + \left[\int D\mathbf{z}\frac{\partial \ln Z_{\text{eff}}}{\partial \hat{R}}\right]_{\xi^{1,true},\xi^{2,true}} = 0. \tag{12.56}$$

Thus the saddle-point equation of $R$ is given by

$$R = [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \tag{12.57}$$

where the thermal average $\langle \bullet \rangle$ is computed under the partition function $Z_{\text{eff}}$ (a two-spin interaction partition function), and the outer average indicates the disorder average over Gaussian random variables $\mathbf{z}$ and the distribution $P(\xi^{1,true}, \xi^{2,true})$.

Similarly, for the order parameter $T_1$, we have the following equation:

$$\frac{\partial F_\beta}{\partial \hat{T}_1} = -T_1 + \int D\mathbf{z} \left[ \frac{1}{Z_{\text{eff}}} \frac{\partial Z_{\text{eff}}}{\partial \hat{T}_1} \right]_{\xi^{1,true}, \xi^{2,true}} = 0. \tag{12.58}$$

Noting that $\frac{\partial Z_{\text{eff}}}{\partial \hat{T}_1} = \sum_{\xi^1, \xi^2} \xi^{1,true} \xi^1 e^{b_1 \xi^1 + b_2 \xi^2 + b_3 \xi^1 \xi^2}$, we get the final expression of $T_1$ as

$$T_1 = [\langle \xi^1 \rangle \xi^{1,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}. \tag{12.59}$$

The expressions of $T_2$, $\tau_1$ and $\tau_2$ can be derived in the same way as follows:

$$T_2 = [\langle \xi^2 \rangle \xi^{2,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \tag{12.60}$$

and

$$\tau_1 = [\langle \xi^2 \rangle \xi^{1,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \tag{12.61}$$

and

$$\tau_2 = [\langle \xi^1 \rangle \xi^{2,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}. \tag{12.62}$$

Next, we turn to the saddle-point equation of $q_1$, i.e.,

$$\frac{\partial F_\beta}{\partial \hat{q}_1} = \frac{1}{2}(q_1 - 1) + \int D\mathbf{z} \left[ \frac{1}{Z_{\text{eff}}} \frac{\partial Z_{\text{eff}}}{\partial \hat{q}_1} \right]_{\xi^{1,true}, \xi^{2,true}} = 0. \tag{12.63}$$

Noticing that $\frac{\partial Z_{\text{eff}}}{\partial \hat{q}_1} = \frac{1}{2}(\hat{q}_1 - \frac{\hat{r}}{2})^{-\frac{1}{2}} \sum_{\xi^1, \xi^2} \xi^1 z_1 e^{b_1 \xi^1 + b_2 \xi^2 + b_3 \xi^1 \xi^2}$, we get the expression of $q_1$ as

$$q_1 - 1 + \left( \hat{q}_1 - \frac{\hat{r}}{2} \right)^{-\frac{1}{2}} [\langle \xi^1 \rangle z_1]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}} = 0. \tag{12.64}$$

To proceed, we use the following identity:

$$\int Dz f(z) z = \int Dz f'(z), \tag{12.65}$$

where $f(z)$ is any differentiable function of $z$. Thus, we have the following equality:

$$[\langle \xi^1 \rangle z_1]_{\mathbf{z}} = \left[ \frac{\partial}{\partial z_1} \left( \frac{\sum_{\xi^1,\xi^2} \xi^1 e^{b_1\xi^1 + b_2\xi^2 + b_3\xi^1\xi^2}}{Z_{\text{eff}}} \right) \right]_{\mathbf{z}} = \sqrt{\hat{q}_1 - \frac{\hat{r}}{2}} [1 - \langle \xi^1 \rangle^2]_{\mathbf{z}}. \tag{12.66}$$

Finally, the expression of $q_1$ is given by

$$q_1 = [\langle \xi^1 \rangle^2]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}. \tag{12.67}$$

Similarly, $q_2$ should obey the following equation, which is given by:

$$q_2 = [\langle \xi^2 \rangle^2]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}. \tag{12.68}$$

Following the same line of computation, we get the following stationary condition for $\hat{r}$ as:

$$\frac{r}{2} + \int D\mathbf{z} \Big[ \frac{\partial}{\partial \hat{r}} \ln Z_{\text{eff}} \Big]_{\xi^{1,true},\xi^{2,true}} = 0. \tag{12.69}$$

Note that

$$\begin{aligned} \frac{\partial}{\partial \hat{r}} \ln Z_{\text{eff}} = &-\frac{1}{4} \left( \hat{q}_1 - \frac{\hat{r}}{2} \right)^{-\frac{1}{2}} \langle \xi^1 \rangle z_1 + \frac{1}{4} \left( \frac{\hat{r}}{2} \right)^{-\frac{1}{2}} \langle \xi^1 \rangle z_3 \\ &- \frac{1}{4} \left( \hat{q}_2 - \frac{\hat{r}}{2} \right)^{-\frac{1}{2}} \langle \xi^2 \rangle z_2 + \frac{1}{4} \left( \frac{\hat{r}}{2} \right)^{-\frac{1}{2}} \langle \xi^2 \rangle z_3 - \frac{1}{2} \langle \xi^1 \xi^2 \rangle. \end{aligned} \tag{12.70}$$

By applying Eq. (12.65), we can obtain the following three identities:

$$\begin{aligned} [\langle \xi^2 \rangle z_2]_{\mathbf{z}} &= \sqrt{\hat{q}_2 - \frac{\hat{r}}{2}} \big(1 - [\langle \xi^2 \rangle^2]_{\mathbf{z}}\big), \\ [\langle \xi^1 \rangle z_3]_{\mathbf{z}} &= \sqrt{\frac{\hat{r}}{2}} \big(1 - [\langle \xi^1 \rangle^2]_{\mathbf{z}} + [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}} - [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z}}\big), \\ [\langle \xi^2 \rangle z_3]_{\mathbf{z}} &= \sqrt{\frac{\hat{r}}{2}} \big(1 - [\langle \xi^2 \rangle^2]_{\mathbf{z}} + [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}} - [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z}}\big). \end{aligned} \tag{12.71}$$

Using the above three identities together with Eq. (12.66), we get the expression of the saddle-point equation for $r$ as follows:

$$r = [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}. \tag{12.72}$$

Given the result that $Z_{\text{eff}} = 2e^{b_3} \cosh(b_1 + b_2) + 2e^{-b_3} \cosh(b_1 - b_2)$, the thermal average like $\langle \xi^1 \rangle$, $\langle \xi^2 \rangle$, and $\langle \xi^1 \xi^2 \rangle$ can be easily calculated as follows:

$$\begin{aligned}\langle \xi^1 \xi^2 \rangle_{Z_{\text{eff}}} &= \frac{\partial}{\partial b_3} \ln Z_{\text{eff}} \\ &= \frac{e^{b_3}\cosh(b_1+b_2) - e^{-b_3}\cosh(b_1-b_2)}{e^{b_3}\cosh(b_1+b_2) + e^{-b_3}\cosh(b_1-b_2)} \\ &= \frac{e^{b_3}(\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) - e^{-b_3}(\cosh b_1 \cosh b_2 - \sinh b_1 \sinh b_2)}{e^{b_3}(\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) + e^{-b_3}(\cosh b_1 \cosh b_2 - \sinh b_1 \sin b_2)}, \\ &= \frac{\sinh b_3 \cosh b_1 \cosh b_2 + \cosh b_3 \sinh b_1 \sinh b_2}{\cosh b_3 \cosh b_1 \cosh b_2 + \sinh b_3 \sinh b_1 \sinh b_2} \\ &= \frac{\tanh b_3 + \tanh b_1 \tanh b_2}{1 + \tanh b_1 \tanh b_2 \tanh b_3},\end{aligned} \tag{12.73}$$

and

$$\begin{aligned}\langle \xi^1 \rangle_{Z_{\text{eff}}} &= \frac{\partial}{\partial b_1} \ln Z_{\text{eff}} \\ &= \frac{e^{b_3}\sinh(b_1+b_2) + e^{-b_3}\sinh(b_1-b_2)}{e^{b_3}\cosh(b_1+b_2) + e^{-b_3}\cosh(b_1-b_2)} \\ &= \frac{e^{b_3}(\sinh b_1 \cosh b_2 + \cosh b_1 \sinh b_2) + e^{-b_3}(\sinh b_1 \cosh b_2 - \cosh b_1 \sinh b_2)}{e^{b_3}(\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) + e^{-b_3}(\cosh b_1 \cosh b_2 - \sinh b_1 \sinh b_2)} \\ &= \frac{\cosh b_3 \sinh b_1 \cosh b_2 + \sinh b_3 \cosh b_1 \sinh b_2}{\cosh b_3 \cosh b_1 \cosh b_2 + \sinh b_3 \sinh b_1 \sinh b_2} \\ &= \frac{\tanh b_1 + \tanh b_2 \tanh b_3}{1 + \tanh b_1 \tanh b_2 \tanh b_3},\end{aligned} \tag{12.74}$$

and finally

$$\begin{aligned}\langle \xi^2 \rangle_{Z_{\text{eff}}} &= \frac{\partial}{\partial b_2} \ln Z_{\text{eff}} \\ &= \frac{e^{b_3}\sinh(b_1+b_2) - e^{-b_3}\sinh(b_1-b_2)}{e^{b_3}\cosh(b_1+b_2) + e^{-b_3}\cosh(b_1-b_2)} \\ &= \frac{e^{b_3}(\sinh b_1 \cosh b_2 + \cosh b_1 \sinh b_2) - e^{-b_3}(\sinh b_1 \cosh b_2 - \cosh b_1 \sinh b_2)}{e^{b_3}(\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) + e^{-b_3}(\cosh b_1 \cosh b_2 - \sinh b_1 \sinh b_2)} \\ &= \frac{\cosh b_2 \sinh b_1 \sinh b_3 + \sinh b_2 \cosh b_1 \cosh b_3}{\cosh b_3 \cosh b_1 \cosh b_2 + \sinh b_3 \sinh b_1 \sinh b_2} \\ &= \frac{\tanh b_2 + \tanh b_1 \tanh b_3}{1 + \tanh b_1 \tanh b_2 \tanh b_3}.\end{aligned} \tag{12.75}$$

In case of $\hat{r} < 0$, we can re-parameterize $b_1$ and $b_2$ as

$$b_1 = \sqrt{\hat{q}_1} z_1 + \hat{T}_1 \xi^{1,true} + \hat{\tau}_2 \xi^{2,true}, \tag{12.76a}$$

$$b_2 = \sqrt{\hat{q}_2}\left(\psi z_1 + \sqrt{1-\psi^2} z_2\right) + \hat{T}_2 \xi^{2,true} + \hat{\tau}_1 \xi^{1,true}, \tag{12.76b}$$

$$\psi = \frac{\hat{r}}{2\sqrt{\hat{q}_1 \hat{q}_2}}. \tag{12.76c}$$

We remark that this re-parameterization does not change the final results of multidimensional Gaussian integrations in the saddle-point equations.

To sum up, the saddle-point equations for non-conjugated order parameters are given by

$$T_1 = [\xi^{1,true}\langle\xi^1\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.77a}$$
$$T_2 = [\xi^{2,true}\langle\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.77b}$$
$$q_1 = [\langle\xi^1\rangle^2]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.77c}$$
$$q_2 = [\langle\xi^2\rangle^2]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.77d}$$
$$\tau_1 = [\xi^{1,true}\langle\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.77e}$$
$$\tau_2 = [\xi^{2,true}\langle\xi^1\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.77f}$$
$$R = [\langle\xi^1\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.77g}$$
$$r = [\langle\xi^1\rangle\langle\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}. \tag{12.77h}$$

Next, we derive the saddle-point equations for those conjugated order parameters. For $\hat{R}$, we obtain the saddle-point equation as

$$\frac{\partial F_\beta}{\partial R} = -\hat{R} - \alpha\beta^2\tanh(\beta^2 R) + \frac{\alpha e^{-\beta^2}}{\cosh(\beta^2 q)}\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)\frac{\partial}{\partial R}\ln Z_E = 0, \tag{12.78}$$

where $\frac{\partial}{\partial R}\ln Z_E = \beta^2\frac{e^{\beta^2(R-r)}\cosh(\beta\Lambda_+)-e^{-\beta^2(R-r)}\cosh(\beta\Lambda_-)}{e^{\beta^2(R-r)}\cosh(\beta\Lambda_+)+e^{-\beta^2(R-r)}\cosh(\beta\Lambda_-)} = \beta^2 G_c^-$. Therefore, the saddle-point equation of $\hat{R}$ is given by

$$\hat{R} = \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)}\int D\mathbf{t}[\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)G_c^- - \alpha\beta^2\tanh(\beta^2 R). \tag{12.79}$$

For convenience, we define the measure $\langle\bullet\rangle$ as $\frac{e^{-\beta^2}}{\cosh(\beta^2 q)}\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)\bullet$. As a result,

$$\hat{R} = \alpha\beta^2\langle G_c^-\rangle - \alpha\beta^2\tanh(\beta^2 R). \tag{12.80}$$

For $\hat{T}_1$, we have the following condition:

$$\frac{\partial F_\beta}{\partial T_1} = -\hat{T}_1 + \frac{\alpha e^{-\beta^2}}{\cosh(\beta^2 q)}\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)\frac{\partial}{\partial T_1}\ln Z_E = 0. \tag{12.81}$$

To proceed, we first get the derivation of $\Lambda_+$ and $\Lambda_-$ w.r.t $T_1$ as follows:

$$\frac{\partial\Lambda_+}{\partial T_1} = t_0 - \frac{q}{\sqrt{1-q^2}}x_0 + \frac{\partial}{\partial T_1}\left(B+\frac{r-A}{B}\right)u + \frac{\partial K}{\partial T_1}u', \tag{12.82a}$$

$$\frac{\partial \Lambda_-}{\partial T_1} = t_0 - \frac{q}{\sqrt{1-q^2}} x_0 + \frac{\partial}{\partial T_1}\left(B - \frac{r-A}{B}\right) u - \frac{\partial K}{\partial T_1} u'. \tag{12.82b}$$

Then, the derivation of $\ln Z_E$ w.r.t $T_1$ can be simplified into the form as

$$\frac{\partial \ln Z_E}{\partial T_1} = \beta \left[ G_s^+ t_0 - \frac{q}{\sqrt{1-q^2}} G_s^+ x_0 + \frac{\partial B}{\partial T_1} G_s^+ u + \frac{\partial}{\partial T_1}\left(\frac{r-A}{B}\right) G_s^- u + \frac{\partial K}{\partial T_1} G_s^- u' \right]. \tag{12.83}$$

To further simplify the result, we need to evaluate the following integral formulas. The first one is derived by applying Eq. (12.65) as

$$\begin{aligned}
&\int D\mathbf{t} \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) G_s^+ t_0 \\
&= \int D\mathbf{t} \frac{\partial}{\partial t_0} \left( \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) G_s^+ \right) \\
&= \beta \int D\mathbf{t} \left[ \sinh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) + q \cosh \beta t_0 \sinh \beta (q t_0 + \sqrt{1-q^2} x_0) \right] G_s^+ \\
&+ \beta \int D\mathbf{t} \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) \left[ T_1 + \tau_1 G_c^- - T_1 (G_s^+)^2 - \tau_1 G_s^+ G_s^- \right].
\end{aligned} \tag{12.84}$$

The second one is derived as

$$\begin{aligned}
&\int D\mathbf{t} \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) G_s^+ x_0 \\
&= \int D\mathbf{t} \frac{\partial}{\partial x_0} \left( \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) G_s^+ \right) \\
&= \beta \sqrt{1-q^2} \int D\mathbf{t} \cosh \beta t_0 \sinh \beta (q t_0 + \sqrt{1-q^2} x_0) G_s^+ \\
&+ \frac{\beta}{\sqrt{1-q^2}} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) \\
&\times \left[ (\tau_2 - q T_1) + (T_2 - q\tau_1) G_c^- - (\tau_2 - q T_1)(G_s^+)^2 - (T_2 - q\tau_1) G_s^+ G_s^- \right].
\end{aligned} \tag{12.85}$$

The third one is derived as

$$\begin{aligned}
&\int D\mathbf{t} \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) G_s^+ u \\
&= \int D\mathbf{t} \frac{\partial}{\partial u} \left( \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) G_s^+ \right) \\
&= \beta \int D\mathbf{t} \cosh \beta t_0 \cosh \beta (q t_0 + \sqrt{1-q^2} x_0) \left[ B + \frac{r-A}{B} G_c^- - B (G_s^+)^2 - \frac{r-A}{B} G_s^+ G_s^- \right].
\end{aligned} \tag{12.86}$$

The fourth one is derived as

$$\begin{aligned}
&\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)G_s^- u\\
&=\int D\mathbf{t}\frac{\partial}{\partial u}\Big(\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)G_s^-\Big)\\
&=\beta\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)\Big[BG_c^-+\frac{r-A}{B}-\frac{r-A}{B}(G_s^-)^2-BG_s^+G_s^-\Big].
\end{aligned}\tag{12.87}$$

The last one is given by

$$\begin{aligned}
&\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)G_s^- u'\\
&=\int D\mathbf{t}\frac{\partial}{\partial u'}\Big(\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)G_s^-\Big)\\
&=\beta K\int D\mathbf{t}\cosh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)\Big[1-(G_s^-)^2\Big].
\end{aligned}\tag{12.88}$$

Through a bit lengthy algebraic manipulations, we get

$$\hat{T}_1=\frac{\alpha\beta^2e^{-\beta^2}}{\cosh(\beta^2q)}\int D\mathbf{t}\sinh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)G_s^+.\tag{12.89}$$

We thus define another measure $\langle\langle\bullet\rangle\rangle=\frac{e^{-\beta^2}}{\cosh(\beta^2q)}\int D\mathbf{t}\sinh\beta t_0\cosh\beta(qt_0+\sqrt{1-q^2}x_0)\bullet$, and it then follows that

$$\hat{T}_1=\alpha\beta^2\langle\langle G_s^+\rangle\rangle.\tag{12.90}$$

Similarly, we can obtain the saddle-point equation of $\hat{\tau}_1$ as

$$\hat{\tau}_1=\alpha\beta^2\langle\langle G_s^-\rangle\rangle.\tag{12.91}$$

Next, we turn to the saddle-point equations for $\hat{T}_2$ and $\hat{\tau}_2$. We first get the derivation of $\Lambda_+$ and $\Lambda_-$ w.r.t $T_2$ as

$$\frac{\partial\Lambda_+}{\partial T_2}=\frac{x_0}{\sqrt{1-q^2}}-\frac{1}{B}\frac{\partial A}{\partial T_2}u+\frac{\partial K}{\partial T_2}u',\tag{12.92a}$$

$$\frac{\partial\Lambda_-}{\partial T_2}=-\frac{x_0}{\sqrt{1-q^2}}+\frac{1}{B}\frac{\partial A}{\partial T_2}u-\frac{\partial K}{\partial T_2}u'.\tag{12.92b}$$

Based on the above equations, we get the derivative of $\ln Z_E$ w.r.t $T_2$ given by

$$\frac{\partial\ln Z_E}{\partial T_2}=\beta\left[\frac{x_0}{\sqrt{1-q^2}}G_s^--\frac{1}{B}\frac{\partial A}{\partial T_2}G_s^-u+\frac{\partial K}{\partial T_2}G_s^-u'\right].\tag{12.93}$$

Then we have

$$\hat{T}_2 = \frac{\alpha\beta e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh\beta t_0 \cosh\beta(qt_0 + \sqrt{1-q^2}x_0) \times \left[ \frac{x_0}{\sqrt{1-q^2}} G_s^- - \frac{1}{B}\frac{\partial A}{\partial T_2} G_s^- u + \frac{\partial K}{\partial T_2} G_s^- u' \right]. \tag{12.94}$$

For a further simplification, we need to derive the following integral identity:

$$\begin{aligned} &\int D\mathbf{t} \cosh\beta t_0 \cosh\beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- x_0 \\ &= \int D\mathbf{t} \frac{\partial}{\partial x_0} \left( \cosh\beta t_0 \cosh\beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- \right) \\ &= \beta\sqrt{1-q^2} \int D\mathbf{t} \cosh\beta t_0 \sinh\beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- \\ &+ \frac{\beta}{\sqrt{1-q^2}} \int D\mathbf{t} \cosh\beta t_0 \cosh\beta(qt_0 + \sqrt{1-q^2}x_0) \\ &\times \left[ (\tau_2 - qT_1) G_c^- - (\tau_2 - qT_1) G_s^+ G_s^- - (T_2 - q\tau_1)(G_s^-)^2 + (T_2 - q\tau_1) \right]. \end{aligned} \tag{12.95}$$

Using Eq. (12.95) together with Eqs. (12.87) and (12.88), we finally arrive at the saddle-point equation of $\hat{T}_2$

$$\hat{T}_2 = \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh\beta t_0 \sinh\beta(qt_0 + \sqrt{1-q^2}x_0) G_s^-. \tag{12.96}$$

We thus define the third measure $\langle\langle\langle \bullet \rangle\rangle\rangle = \frac{e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh\beta t_0 \sinh\beta(qt_0 + \sqrt{1-q^2}x_0) \bullet$. We then write the saddle-point equation in a compact form as

$$\hat{T}_2 = \alpha\beta^2 \langle\langle\langle G_s^- \rangle\rangle\rangle. \tag{12.97}$$

Similarly, we obtain the saddle-point equation for $\hat{\tau}_2$ as

$$\hat{\tau}_2 = \alpha\beta^2 \langle\langle\langle G_s^+ \rangle\rangle\rangle. \tag{12.98}$$

Then we turn to the saddle-point equations of $\hat{q}_1$ and $\hat{q}_2$. From $\frac{\partial F_\beta}{\partial q_1} = 0$, we get

$$\frac{1}{2}\hat{q}_1 - \frac{\alpha\beta^2}{2} + \frac{\alpha\beta e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh\beta t_0 \cosh\beta(qt_0 + \sqrt{1-q^2}x_0) \frac{\partial \ln Z_E}{\partial q_1} = 0. \tag{12.99}$$

The derivation of $\ln Z_E$ w.r.t $q_1$ is given by

$$\frac{\partial \ln Z_E}{\partial q_1} = \frac{\partial B}{\partial q_1} G_s^+ u + \frac{\partial}{\partial q_1}\left(\frac{r-A}{B}\right) G_s^- u + \frac{\partial K}{\partial q_1} G_s^- u'. \tag{12.100}$$

Most terms in the above equation cancel each other, leading to

$$\begin{aligned} \hat{q}_1 &= \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh(\beta t_0) \cosh\beta(qt_0 + \sqrt{1-q^2}x_0)(G_s^+)^2 \\ &= \alpha\beta^2 \langle (G_s^+)^2 \rangle. \end{aligned} \tag{12.101}$$

Similarly, we can derive the saddle-point equation for $\hat{q}_2$ as

$$\begin{aligned} \hat{q}_2 &= \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh(\beta t_0) \cosh\beta(qt_0 + \sqrt{1-q^2}x_0)(G_s^-)^2 \\ &= \alpha\beta^2 \langle (G_s^-)^2 \rangle. \end{aligned} \tag{12.102}$$

Lastly, we derive the saddle-point equation for $\hat{r}$ as

$$\frac{\hat{r}}{2} + \frac{\alpha e^{-\beta^2}}{\cosh\beta^2 q} \int D\mathbf{t} \cosh\beta t_0 \cosh\beta(qt_0 + \sqrt{1-q^2}x_0) \frac{\partial \ln Z_E}{\partial r} = 0. \tag{12.103}$$

Noting that $\frac{\partial \ln Z_E}{\partial r} = -\beta^2 G_c^- + \beta\left(\frac{1}{B} G_s^- u + \frac{\partial K}{\partial r} G_s^- u'\right)$, we get the saddle-point equation of $\hat{r}$ as

$$\hat{r} = 2\alpha\beta^2 \langle G_s^+ G_s^- \rangle. \tag{12.104}$$

To sum up, the saddle-point equations of our minimal model are listed as follows:

$$\hat{T}_1 = \alpha\beta^2 \langle \langle G_s^+ \rangle \rangle, \tag{12.105a}$$

$$\hat{T}_2 = \alpha\beta^2 \langle \langle \langle G_s^- \rangle \rangle \rangle, \tag{12.105b}$$

$$\hat{\tau}_1 = \alpha\beta^2 \langle \langle G_s^- \rangle \rangle, \tag{12.105c}$$

$$\hat{\tau}_2 = \alpha\beta^2 \langle \langle \langle G_s^+ \rangle \rangle \rangle, \tag{12.105d}$$

$$\hat{q}_1 = \alpha\beta^2 \langle (G_s^+)^2 \rangle, \tag{12.105e}$$

$$\hat{q}_2 = \alpha\beta^2 \langle (G_s^-)^2 \rangle, \tag{12.105f}$$

$$\hat{r} = 2\alpha\beta^2 \langle G_s^+ G_s^- \rangle, \tag{12.105g}$$

$$\hat{R} = \alpha\beta^2 \langle G_c^- \rangle - \alpha\beta^2 \tanh(\beta^2 R). \tag{12.105h}$$

In the case of $q = 0$ (correlation-free scenario), the saddle-point equation of the correlation-prior-free minimal model has the solution: $q_1 = q_2 = T_1 = T_2$ and other order parameters vanish. Thus, we can simplify $\Lambda_+$ and $\Lambda_-$ as follows:

$$\Lambda_+ = T_1 t_0 + T_2 x_0 + \sqrt{q_1 - (T_1)^2}u + \sqrt{q_2 - (T_2)^2}u', \tag{12.106a}$$

$$\Lambda_- = T_1 t_0 - T_2 x_0 + \sqrt{q_1 - (T_1)^2}u - \sqrt{q_2 - (T_2)^2}u'. \tag{12.106b}$$

We then define $\chi_1 = T_1 t_0 + \sqrt{q_1 - (T_1)^2}u$, and $\chi_2 = T_2 x_0 + \sqrt{q_2 - (T_2)^2}u'$. The saddle-point equation of $\hat{T}_1$ is given by

$$\begin{aligned}
\hat{T}_1 &= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \sinh\beta t_0 \cosh\beta x_0 \left[\frac{\sinh\beta\Lambda_+ + \sinh\beta\Lambda_-}{\cosh\beta\Lambda_+ + \cosh\beta\Lambda_-}\right] \\
&= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \sinh\beta t_0 \cosh\beta x_0 \left[\frac{\sinh\beta\chi_1 \cosh\beta\chi_2}{\cosh\beta\chi_1 \cosh\beta\chi_2}\right] \\
&= \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt_0 Du \sinh\beta t_0 \tanh\beta(T_1 t_0 + \sqrt{q_1 - (T_1)^2}u),
\end{aligned} \tag{12.107}$$

where we have used the identity $\int Dx_0 \cosh(\beta x_0) = e^{\beta^2/2}$. In an analogous way, one can prove that $\hat{T}_1 = \hat{T}_2$. As for $\hat{q}_1$, we will have

$$\begin{aligned}
\hat{q}_1 &= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \cosh\beta t_0 \cosh\beta x_0 \left[\frac{\sinh\beta\Lambda_+ + \sinh\beta\Lambda_-}{\cosh\beta\lambda_+ + \cosh\beta\Lambda_-}\right]^2 \\
&= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \cosh\beta t_0 \cosh\beta x_0 \left[\frac{\sinh\beta\chi_1 \cosh\beta\chi_2}{\cosh\beta\chi_1 \cosh\beta\chi_2}\right]^2 \\
&= \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt_0 Du \cosh\beta t_0 \tanh^2\beta(T_1 t_0 + \sqrt{q_1 - (T_1)^2}u).
\end{aligned} \tag{12.108}$$

Similarly, one can prove that $\hat{q}_1 = \hat{q}_2$.

It is also straightforward to prove that $\hat{\tau}_1 = 0$, $\hat{\tau}_2 = 0$ and $\hat{R} = 0$, $\hat{r} = 0$, then we can compute $b_1$, $b_2$ and $b_3$ as

$$b_1 = \hat{T}_1 \xi^{1,true} + \sqrt{\hat{q}_1} z_1, \tag{12.109a}$$

$$b_2 = \hat{T}_2 \xi^{2,true} + \sqrt{\hat{q}_2} z_2, \tag{12.109b}$$

$$b_3 = 0. \tag{12.109c}$$

Therefore, $\int D\mathbf{z}[\ln Z_{\text{eff}}]_{\xi^{1,true},\xi^{2,true}}$ can be simplified as $2\int Dz \ln 2\cosh(\hat{T}_1 + \sqrt{\hat{q}_1}z)$. In addition, $T_1$ becomes

$$\begin{aligned}
T_1 &= \left[\int Dz_1 Dz_2 Dz_3 \xi^{1,true} \tanh\left(\hat{T}_1 \xi^{1,true} + \sqrt{q_1} z_1\right)\right]_{\xi^{1,true},\xi^{2,true}} \\
&= \int Dz_1 \frac{1}{2}\left[\tanh\left(\hat{T}_1 + \sqrt{\hat{q}_1} z_1\right) - \tanh\left(-\hat{T}_1 + \sqrt{q}_1 z_1\right)\right] \\
&= \int Dz_1 \frac{1}{2}\left[\tanh(\hat{T}_1 + \sqrt{\hat{q}_1} z_1) - \tanh\left(-\hat{T}_1 - \sqrt{q}_1 z_1\right)\right] \\
&= \int Dz_1 \tanh\left(\hat{T}_1 + \sqrt{q_1} z_1\right).
\end{aligned} \tag{12.110}$$

One can easily prove that $T_1 = T_2$. Similarly, for the order parameter $q_2$, we can also get

$$\begin{aligned} q_2 &= \left[ \int Dz_2 \tanh^2 (\hat{T}_2 \xi^{1,true} + \sqrt{\hat{q}_2} z_1) \right]_{\xi^{1,true}, \xi^{2,true}} \\ &= \frac{1}{2} \int Dz_2 \left[ \tanh^2 (\hat{T}_2 + \sqrt{\hat{q}_2} z_2) + \tanh^2 (-\hat{T}_2 + \sqrt{\hat{q}_2} z_2) \right] \\ &= \int Dz_2 \tanh^2 (\hat{T}_2 + \sqrt{\hat{q}_2} z_2). \end{aligned} \tag{12.111}$$

On can similarly show that $q_1 = q_2$, and moreover $R = r = \tau_1 = \tau_2 = 0$. To sum up, we recover the saddle-point equations of one-bit RBM.

Next, we show the $q = 0$ version of the free energy function. It is easy to show that $Z_{\mathrm{E}} = \cosh \beta(\chi_1 + \chi_2) + \cosh \beta(\chi_1 - \chi_2) = 2 \cosh \beta \chi_1 \cosh \beta \chi_2$. Therefore, we have the following integral

$$\begin{aligned} &\alpha e^{-\beta^2} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta x_0 \ln Z_{\mathrm{E}} = \alpha e^{-\beta^2} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta x_0 \ln(2 \cosh \beta \chi_1 \cosh \beta \chi_2) \\ &= \alpha \ln 2 + 2\alpha e^{-\frac{\beta^2}{2}} \int Du Dt_0 \cosh \beta t_0 \ln \cosh \beta (T_1 t_0 + \sqrt{q_1 - (T_1)^2} u). \end{aligned} \tag{12.112}$$

Collecting all the relevant terms, one shows that the free energy of our minimal model with $q = 0$ is merely two times as large as that of one-bit RBM (see Chap. 11), which can also be intuitively understood by the argument that the partition function factorizes as $\Omega = \Omega^2_{\mathrm{one-bit-RBM}}$. Therefore, we draw the conclusion that the critical data size for spontaneous symmetry breaking does not change even if an additional hidden node is added. This conclusion is expected to hold in the case of more hidden nodes following the principle of the partition function's factorization.

Next, we turn to the two-bit RBM model with the prior knowledge about the embedded correlation level. For the replica analysis, we need to evaluate the disorder average of an integer power of the partition function $\langle \Omega^n \rangle$, where $\langle \bullet \rangle$ is the disorder average over the true features distribution $P_0(\boldsymbol{\xi}^{1,true}, \boldsymbol{\xi}^{2,true})$ and the corresponding data distribution $P(\{\boldsymbol{\sigma}^a\}_{a=1}^M | \boldsymbol{\xi}^{1,true}, \boldsymbol{\xi}^{2,true})$ as:

$$\begin{aligned} \langle \Omega^n \rangle = &\sum_{\{\boldsymbol{\xi}^{true}, \boldsymbol{\sigma}^a\}} \prod_{i=1}^{N} [P_0(\xi_i^{1,true}, \xi_i^{2,true})] \prod_{a=1}^{M} \frac{\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,true} \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,true} \boldsymbol{\sigma}^a\right)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \\ &\times \sum_{\{\boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma}\}} \prod_{a,\gamma} \frac{\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,\gamma} \boldsymbol{\sigma}^a\right)}{\cosh(\beta^2 R^\gamma)} \prod_{i,\gamma} P_0(\xi_i^{1,\gamma}, \xi_i^{2,\gamma}). \end{aligned} \tag{12.113}$$

Under the RS assumption, $\langle \Omega^n \rangle$ can be expressed as $\langle \Omega^n \rangle = \int d\mathcal{O}\hat{\mathcal{O}} e^{N\mathcal{A}(\mathcal{O}, \hat{\mathcal{O}}, n, \beta, \beta)}$, where $\mathcal{A} = G_0 + G_S + \alpha G_E$. The term $G_0$ reads

$$G_0 = -nR\hat{R} - nT_1\hat{T}_1 - nT_2\hat{T}_2 - n\tau_1\hat{\tau}_1 - n\tau_2\hat{\tau}_2 + \frac{n(n-1)}{2}\hat{q}_1 q_1 + \frac{n(n-1)}{2}\hat{q}_2 q_2 + \frac{n}{2}\hat{r}r. \tag{12.114}$$

The entropic term $G_S$ reads

$$G_S = \ln\left[\sum_{\{\xi^{1,\gamma},\xi^{2,\gamma}\}} \exp\left(\hat{R}\sum_{\gamma=1}^{n}\xi^{1,\gamma}\xi^{2,\gamma} + \hat{T}_1\sum_{\gamma=1}^{n}\xi^{1,\gamma}\xi^{1,\text{true}} + \hat{T}_2\sum_{\gamma=1}^{n}\xi^{2,\gamma}\xi^{2,\text{true}} + \hat{\tau}_1\sum_{\gamma=1}^{n}\xi^{1,\text{true}}\xi^{2,\gamma}\right) \times \exp\left(\hat{\tau}_2\sum_{\gamma=1}^{n}\xi^{1,\gamma}\xi^{2,\text{true}} + \sum_{\gamma<\gamma'}\left(\hat{q}_1\xi^{1,\gamma}\xi^{1,\gamma'} + \hat{q}_2\xi^{2,\gamma}\xi^{2,\gamma'} + \hat{r}\xi^{1,\gamma}\xi^{2,\gamma'}\right) + \sum_{\gamma=1}^{n}\ln P_0(\xi^{1,\gamma},\xi^{2,\gamma})\right)\right]_{\xi^{1,\text{true}},\xi^{2,\text{true}}}. \tag{12.115}$$

In an analogous way to the prior-free (not Bayes optimal) case, we can express the entropy term $G_S$ in a compact form as

$$G_S = \ln\left[\int D\mathbf{z}\left(\sum_{\xi^1,\xi^2} e^{b_1\xi^1 + b_2\xi^2 + b_3\xi^1\xi^2 + \ln P_0(\xi^1,\xi^2)}\right)^n\right]_{\xi^{1,true},\xi^{2,true}} - \frac{n}{2}\hat{q}_1 - \frac{n}{2}\hat{q}_2, \tag{12.116}$$

where we have defined $D\mathbf{z} = Dz_1 Dz_2 Dz_3$, random variables $z_1$, $z_2$, $z_3$ are standard Gaussian variables, $[\bullet]$ is the disorder average under the true features distribution $P_0(\xi^{1,true}, \xi^{2,true})$, and the auxiliary variables $b_1$, $b_2$, and $b_3$ are given, respectively, by

$$b_1 = \sqrt{\hat{q}_1 - \frac{\hat{r}}{2}}z_1 + \sqrt{\frac{\hat{r}}{2}}z_3 + \hat{T}_1\xi^{1,true} + \hat{\tau}_2\xi^{2,true}, \tag{12.117a}$$

$$b_2 = \sqrt{\hat{q}_2 - \frac{\hat{r}}{2}}z_2 + \sqrt{\frac{\hat{r}}{2}}z_3 + \hat{T}_2\xi^{2,true} + \hat{\tau}_1\xi^{1,true}, \tag{12.117b}$$

$$b_3 = \hat{R} - \frac{\hat{r}}{2}. \tag{12.117c}$$

In particular, we obtain an effective partition function $Z_{\text{eff}}$ as

$$\begin{aligned} Z_{\text{eff}} &= \sum_{\xi^1,\xi^2} e^{b_1\xi^1 + b_2\xi^2 + b_3\xi^1\xi^2 + \ln P_0(\xi^1,\xi^2)} \\ &= \frac{1+q}{2}e^{b_3}\cosh(b_1+b_2) + \frac{1-q}{2}e^{-b_3}\cosh(b_1-b_2). \end{aligned} \tag{12.118}$$

The saddle-point equations for non-conjugated order parameters are given by:

$$T_1 = [\xi^{1,true}\langle\xi^1\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.119a}$$
$$T_2 = [\xi^{2,true}\langle\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.119b}$$
$$q_1 = [\langle\xi^1\rangle^2]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.119c}$$
$$q_2 = [\langle\xi^2\rangle^2]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.119d}$$
$$\tau_1 = [\xi^{1,true}\langle\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.119e}$$
$$\tau_2 = [\xi^{2,true}\langle\xi^1\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.119f}$$
$$R = [\langle\xi^1\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}, \tag{12.119g}$$
$$r = [\langle\xi^1\rangle\langle\xi^2\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}}. \tag{12.119h}$$

where $\langle\bullet\rangle$ is the average under the distribution $P(\xi^1,\xi^2) = \frac{1}{Z_{\text{eff}}}e^{b_1\xi^1+b_2\xi^2+b_3\xi^1\xi^2+\ln P_0(\xi^1,\xi^2)}$. For $\langle\xi^1\rangle_{Z_{\text{eff}}}$, we can get

$$\begin{aligned}
\langle\xi^1\rangle_{Z_{\text{eff}}} &= \frac{\partial}{\partial b_1}\ln Z_{\text{eff}} \\
&= \frac{(1+q)e^{b_3}\sinh(b_1+b_2)+(1-q)e^{-b_3}\sinh(b_1-b_2)}{(1+q)e^{b_3}\cosh(b_1+b_2)+(1-q)e^{-b_3}\cosh(b_1-b_2)} \\
&= \frac{(\cosh b_3\sinh b_1\cosh b_2+\sinh b_3\cosh b_1\sinh b_2)+q(\cosh b_2\sinh b_1\sinh b_3+\sinh b_2\cosh b_1\cosh b_3)}{\cosh b_1\cosh b_2\cosh b_3+\sinh b_1\sinh b_2\sinh b_3+q(\cosh b_1\cosh b_2\sinh b_3+\sinh b_1\sinh b_2\cosh b_3)} \\
&= \frac{\tanh b_1+\tanh b_2\tanh b_3+q\tanh b_2+q\tanh b_1\tanh b_3}{1+\tanh b_1\tanh b_2\tanh b_3+q\tanh b_3+q\tanh b_1\tanh b_2}.
\end{aligned} \tag{12.120}$$

Similarly, for $\langle\xi^2\rangle_{Z_{\text{eff}}}$ and $\langle\xi^2\rangle_{Z_{\text{eff}}}$, we have

$$\begin{aligned}
\langle\xi^2\rangle_{Z_{\text{eff}}} &= \frac{\partial}{\partial b_2}\ln Z_{\text{eff}} \\
&= \frac{\tanh b_2+\tanh b_1\tanh b_3+q\tanh b_1+q\tanh b_2\tanh b_3}{1+\tanh b_1\tanh b_2\tanh b_3+q\tanh b_3+q\tanh b_1\tanh b_2},
\end{aligned} \tag{12.121}$$

and

$$\begin{aligned}
\langle\xi^1\xi^2\rangle_{Z_{\text{eff}}} &= \frac{\partial}{\partial b_3}\ln Z_{\text{eff}} \\
&= \frac{\tanh b_3+\tanh b_1\tanh b_2+q\tanh b_1\tanh b_2\tanh b_3+q}{1+\tanh b_1\tanh b_2\tanh b_3+q\tanh b_3+q\tanh b_1\tanh b_2}.
\end{aligned} \tag{12.122}$$

The saddle-point equations for conjugated order parameters are same with the prior-free case

$$\hat{T}_1 = \alpha\beta^2\langle\langle G_s^+\rangle\rangle, \tag{12.123a}$$
$$\hat{T}_2 = \alpha\beta^2\langle\langle\langle G_s^-\rangle\rangle\rangle, \tag{12.123b}$$
$$\hat{q}_1 = \alpha\beta^2\langle(G_s^+)^2\rangle, \tag{12.123c}$$
$$\hat{q}_2 = \alpha\beta^2\langle(G_s^-)^2\rangle, \tag{12.123d}$$

$$\hat{\tau}_1 = \alpha\beta^2\langle\langle G_s^-\rangle\rangle, \tag{12.123e}$$
$$\hat{\tau}_2 = \alpha\beta^2\langle\langle\langle G_s^+\rangle\rangle\rangle, \tag{12.123f}$$
$$\hat{R} = \alpha\beta^2\langle G_c^-\rangle - \alpha\beta^2\tanh(\beta^2 R), \tag{12.123g}$$
$$\hat{r} = 2\alpha\beta^2\langle G_s^+ G_s^-\rangle. \tag{12.123h}$$

### 12.1.3 Stability Analysis

It is reasonable that near a continuous transition point, all order parameters are very small (a trivial state) such that we can expand them to leading order. We first analyze the prior-free unsupervised learning. According to Eq. (12.77), when the critical point is approached from below, $\langle\xi^1\rangle \simeq \tanh b_1 \simeq b_1$. Analogously, $\langle\xi^2\rangle \simeq b_2$, and $\langle\xi^1\xi^2\rangle \simeq b_3$. We thus have the following results in this limit:

$$T_1 = [\xi^{1,true}\langle\xi^1\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}} = \hat{T}_1 + q\hat{\tau}_2, \tag{12.124}$$
$$\tau_2 = [\xi^{2,true}\langle\xi^1\rangle]_{\mathbf{z},\xi^{1,true},\xi^{2,true}} = \hat{\tau}_2 + q\hat{T}_1. \tag{12.125}$$

Similarly, in the limit of vanishing order parameters, we have the following approximation:

$$\begin{aligned} G_s^+ &= \frac{e^{\beta^2(R-r)}\sinh(\beta\Lambda_+) + e^{-\beta^2(R-r)}\sinh(\beta\Lambda_-)}{e^{\beta^2(R-r)}\cosh(\beta\Lambda_+) + e^{-\beta^2(R-r)}\cosh(\beta\Lambda_-)} \\ &= \frac{\beta}{2}(\Lambda_+ + \Lambda_-). \end{aligned} \tag{12.126}$$

Substituting this approximation into the saddle-point equations of $\hat{T}_1$ and $\hat{\tau}_2$, we get the approximate results of $\hat{T}_1$ and $\hat{\tau}_2$ as

$$\begin{aligned} \hat{T}_1 &= \alpha\beta^2\langle\langle G_s^+\rangle\rangle \simeq \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)}\int D\mathbf{t}\sinh\beta t_0\cosh\beta(qt_0 + \sqrt{1-q^2}x_0)\frac{\beta}{2}[\Lambda_+ + \Lambda_-] \\ &= \alpha\beta^4[T_1 + \tanh(\beta^2 q)\tau_2], \\ \hat{\tau}_2 &= \alpha\beta^2\langle\langle\langle G_s^+\rangle\rangle\rangle \simeq \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)}\int D\mathbf{t}\cosh\beta t_0\sinh\beta(qt_0 + \sqrt{1-q^2}x_0)\frac{\beta}{2}[\Lambda_+ + \Lambda_-] \\ &= \alpha\beta^4[\tau_2 + \tanh(\beta^2 q)T_1]. \end{aligned} \tag{12.127}$$

We recast the equations for all these four order parameters in a matrix form as

$$\begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix}\begin{pmatrix} \hat{T}_1 \\ \hat{\tau}_2 \end{pmatrix}, \tag{12.128}$$

$$\begin{pmatrix} \hat{T}_1 \\ \hat{\tau}_2 \end{pmatrix} = \alpha\beta^4 \begin{pmatrix} 1 & \tanh(\beta^2 q) \\ \tanh(\beta^2 q) & 1 \end{pmatrix} \begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix}. \tag{12.129}$$

From the Eqs. (12.128) and (12.129), $T_1$ and $\tau_2$ can be worked out as

$$\begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix} = \alpha\beta^4 \begin{pmatrix} 1 + q\tanh(\beta^2 q) & q + \tanh(\beta^2 q) \\ q + \tanh(\beta^2 q) & 1 + q\tanh(\beta^2 q) \end{pmatrix} \begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix} = \mathcal{M} \begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix}, \tag{12.130}$$

where the matrix $\mathcal{M}$ is the so-called stability matrix, whose largest eigenvalue determines the critical value of the learning data size $\alpha_c$. In detail, the stability matrix has two eigenvalues:

$$\lambda_+ = \alpha\beta^4 \left(1 + q\tanh(\beta^2 q) + |q + \tanh(\beta^2 q)|\right), \tag{12.131}$$

$$\lambda_- = \alpha\beta^4 \left(1 + q\tanh(\beta^2 q) - |q + \tanh(\beta^2 q)|\right). \tag{12.132}$$

The $\alpha_c$ can be read off from $\lambda_+ = 1$, i.e.

$$\alpha_c = \frac{\beta^{-4}}{1 + q\tanh(\beta^2 q) + |q + \tanh(\beta^2 q)|}. \tag{12.133}$$

A physics understanding of why the smaller eigenvalue could not be used to determine the threshold $\alpha_c$ can be carried out, in the sense that the result is in contradiction with the expectation that learning should be easier given noise-free data.

Next, we analyze two interesting limits of the critical threshold equation [Eq. (12.133)]. As the first limit, $|q| \to 1$, $\alpha_c \to \frac{1}{4}\beta^{-4}$ provided that $\beta$ is relatively large such that $\tanh\beta^2 \simeq 1$. The second limit is that $|q| \to 0$, i.e., $q$ takes a small value but not zero, suggesting a weak correlation among feature maps. Based on the order of magnitude of $q$, we have the following analytic result given a relatively large $\beta$:

$$\lim_{\beta\to\infty} \alpha_c\beta^4 = \begin{cases} 1 & \text{if } |q| \ll \beta^{-2}, \\ \frac{1}{1+|\tanh q_0|} & \text{if } q = q_0\beta^{-2} \text{ or } |q| \sim \beta^{-2}, \\ \frac{1}{2(1+|q|)} & \text{if } |q| \gg \beta^{-2}. \end{cases} \tag{12.134}$$

Note that $\infty$ means any large value of $\beta$ making $\tanh\beta \simeq 1$, rather than a definite value of infinity. Equation (12.134) shows that once the two feature maps are weakly correlated, the minimal learning data size for a transition can be further (or even significantly) reduced compared to the correlation-free case, especially in the case that $q$ is not very small but still larger than the order of magnitude set by $\beta^{-2}$. We show this result in Fig. 12.2.

We thus deduce a *significant hypothesis* for the triggering of concept formation that a bit large (compared with $\beta^{-2}$) yet still small value of the correlation level is highly favored for unsupervised learning from a dataset of smaller size (compared with the correlation-free case). Regularization techniques such as locally enforcing feature orthogonality [4] has been introduced to deep learning. Weakly-correlated recep-

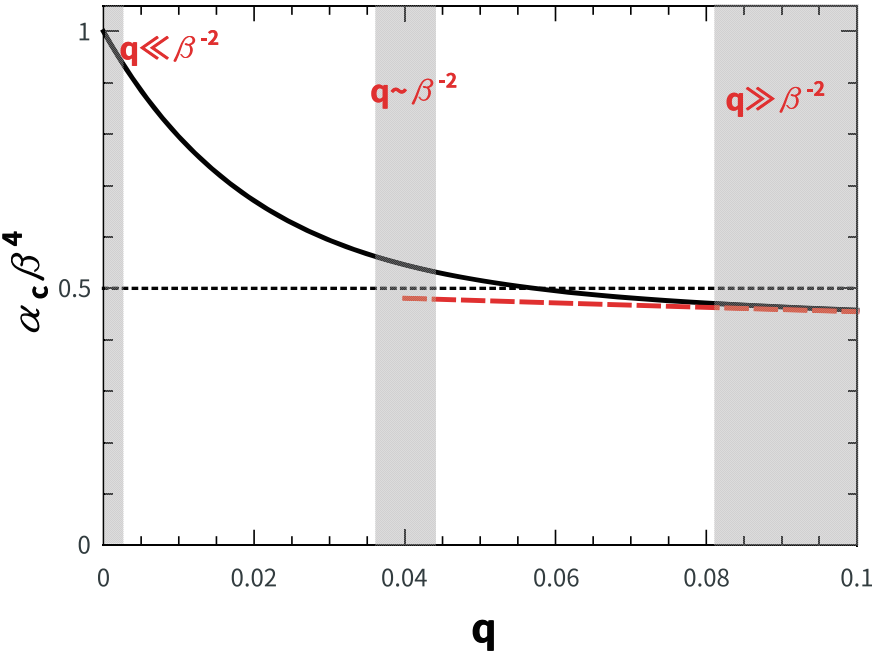



**Fig. 12.2** The critical value of data size (Eq. (12.133)) as a function of the correlation level $q$. The weak-feature-correlation limit at different orders of magnitude compared with $\beta^{-2}$ is considered. $\beta = 5$ for this example. The dashed line shows the third case of Eq. (12.134). This plot is adapted from Ref. [1]

tive fields are also favored from the perspective of neural computation, because the redundancy among synaptic weights is reduced and thus different feature detectors inside the network can encode efficiently stimuli features rather than capturing only noise in the data. A similar decorrelation in hidden activities was recently theoretically analyzed in feedforward neural networks [5]. We anticipate in specific machine learning tasks, and even in neuroscience experiments the relationship among the minimal data size for learning, the correlation level of synapses (or receptive fields) and the noise level in stimuli can be jointly established. Therefore, from the Bayesian learning perspective, the non-orthogonal-feature case yields a much lower threshold for the phase transition toward the concept formation, in comparison with the correlation-free case [3, 6, 7].

In the optimal Bayes inference case, when $\alpha$ approaches the SSB threshold from below, all order parameters get close to zero, except for $R$ which is always equal to $q$ due to the prior information. It is straightforward to show that $\hat{R}$ is also zero below the SSB threshold. Therefore, $b_1$, $b_2$ and $b_3$ are all small quantities. Then we can expand our order parameters to leading order. Note that $\langle \xi^1 \rangle \simeq b_1 + q b_2$, and $\langle \xi^2 \rangle \simeq b_2 + q b_1$. It then follows that

$$T_1 = [\xi^{1,\text{true}}\langle\xi^1\rangle] \simeq \hat{T}_1 + q\hat{\tau}_2 + q\hat{\tau}_1 + q^2\hat{T}_2, \tag{12.135a}$$

$$T_2 = [\xi^{2,\text{true}}\langle\xi^2\rangle] \simeq \hat{T}_2 + q\hat{\tau}_2 + q\hat{\tau}_1 + q^2\hat{T}_1, \tag{12.135b}$$

$$\tau_1 = [\xi^{1,\text{true}}\langle\xi^2\rangle] \simeq \hat{\tau}_1 + q\hat{T}_1 + q\hat{T}_2 + q^2\hat{\tau}_2, \tag{12.135c}$$

$$\tau_2 = [\xi^{2,\text{true}}\langle\xi^1\rangle] \simeq \hat{\tau}_2 + q\hat{T}_1 + q\hat{T}_2 + q^2\hat{\tau}_1. \tag{12.135d}$$

Because $R = q$, by defining $W(q) = \frac{e^{\beta^2 q}}{2\cosh(\beta^2 q)}$, one arrives at the approximation $G_s^{\pm} \simeq \beta W(q)(\Lambda_+ \mp \Lambda_-) \pm \beta\Lambda_-$. To proceed, it is worth noticing that

$$\langle\langle\Lambda_+\rangle\rangle = \beta[T_1 + \tau_1 + \tau_2\tanh(\beta^2 q) + T_2\tanh(\beta^2 q)], \tag{12.136a}$$

$$\langle\langle\Lambda_-\rangle\rangle = \beta[T_1 - \tau_1 + \tau_2\tanh(\beta^2 q) - T_2\tanh(\beta^2 q)], \tag{12.136b}$$

$$\langle\langle\langle\Lambda_+\rangle\rangle\rangle = \beta[T_2 + \tau_2 + \tau_1\tanh(\beta^2 q) + T_1\tanh(\beta^2 q)], \tag{12.136c}$$

$$\langle\langle\langle\Lambda_-\rangle\rangle\rangle = \beta[\tau_2 - T_2 + T_1\tanh(\beta^2 q) - \tau_1\tanh(\beta^2 q)]. \tag{12.136d}$$

Based on the above approximations, it is easy to derive the following approximate values of the relevant conjugated quantities

$$\hat{T}_1 \simeq \alpha\beta^4[T_1 + \Upsilon\tau_1 + \tau_2\tanh(\beta^2 q) + \Upsilon T_2\tanh(\beta^2 q)], \tag{12.137a}$$

$$\hat{T}_2 \simeq \alpha\beta^4[T_2 + \Upsilon\tau_2 + \tau_1\tanh(\beta^2 q) + \Upsilon T_1\tanh(\beta^2 q)], \tag{12.137b}$$

$$\hat{\tau}_1 \simeq \alpha\beta^4[\tau_1 + \Upsilon T_1 + T_2\tanh(\beta^2 q) + \Upsilon\tau_2\tanh(\beta^2 q)], \tag{12.137c}$$

$$\hat{\tau}_2 \simeq \alpha\beta^4[\tau_2 + \Upsilon T_2 + T_1\tanh(\beta^2 q) + \Upsilon\tau_1\tanh(\beta^2 q)], \tag{12.137d}$$

where $\Upsilon \equiv 2W(q) - 1$.

The above approximations of $(T_1, T_2, \tau_1, \tau_2)$ and $(\hat{T}_1, \hat{T}_2, \hat{\tau}_1, \hat{\tau}_2)$ can be easily recasted into a compact matrix form as follows:

$$\begin{pmatrix} T_1 \\ T_2 \\ \tau_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} 1 & q^2 & q & q \\ q^2 & 1 & q & q \\ q & q & 1 & q^2 \\ q & q & q^2 & 1 \end{pmatrix} \begin{pmatrix} \hat{T}_1 \\ \hat{T}_2 \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{pmatrix}, \tag{12.138}$$

and

$$\begin{pmatrix} \hat{T}_1 \\ \hat{T}_2 \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{pmatrix} = \alpha\beta^4 \begin{pmatrix} 1 & \Upsilon\tanh(\beta^2 q) & \Upsilon & \tanh(\beta^2 q) \\ \Upsilon\tanh(\beta^2 q) & 1 & \tanh(\beta^2 q) & \Upsilon \\ \Upsilon & \tanh(\beta^2 q) & 1 & \Upsilon\tanh(\beta^2 q) \\ \tanh(\beta^2 q) & \Upsilon & \Upsilon\tanh(\beta^2 q) & 1 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ \tau_1 \\ \tau_2 \end{pmatrix}. \tag{12.139}$$

A linear stability analysis implies that the stability matrix $\mathcal{M}$ can be organized in this case as a block matrix of the form $\mathcal{M} = \begin{pmatrix} A & B \\ B & A \end{pmatrix}$, where the matrices $A$ and $B$ are derived from Eqs. (12.138) and (12.139), and given, respectively, by

$$A = \alpha\beta^4 \begin{pmatrix} (1+q\tanh(\beta^2 q))(1+q\Upsilon) & (\tanh(\beta^2 q)+q)(q+\Upsilon) \\ (\tanh(\beta^2 q)+q)(\Upsilon+q) & (1+q\tanh(\beta^2 q))(1+q\Upsilon) \end{pmatrix}, \tag{12.140a}$$

$$B = \alpha\beta^4 \begin{pmatrix} (\Upsilon+q)(1+q\tanh(\beta^2 q)) & (\Upsilon q+1)(q+\tanh(\beta^2 q)) \\ (\Upsilon q+1)(q+\tanh(\beta^2 q)) & (\Upsilon+q)(1+q\tanh(\beta^2 q)) \end{pmatrix}. \tag{12.140b}$$

According to the determinant identity for a block matrix, $|\mathcal{M}-\lambda I| = |A+B-\lambda I|\,|A-B-\lambda I|$, the eigenvalues of the stability matrix can be determined by the following two equations:

$$\begin{vmatrix} \alpha\beta^4(1+q)(1+q\tanh(\beta^2 q))(1+\Upsilon)-\lambda & \alpha\beta^4(1+q)(\Upsilon+1)(q+\tanh(\beta^2 q)) \\ \alpha\beta^4(1+q)(\Upsilon+1)(q+\tanh(\beta^2 q)) & \alpha\beta^4(1+q)(1+q\tanh(\beta^2 q))(1+\Upsilon)-\lambda \end{vmatrix} = 0, \tag{12.141}$$

and

$$\begin{vmatrix} \alpha\beta^4(1-q)(1+q\tanh(\beta^2 q))(1-\Upsilon)-\lambda & \alpha\beta^4(1-q)(\Upsilon-1)(q+\tanh(\beta^2 q)) \\ \alpha\beta^4(1-q)(\Upsilon-1)(q+\tanh(\beta^2 q)) & \alpha\beta^4(1-q)(1+q\tanh(\beta^2 q))(1-\Upsilon)-\lambda \end{vmatrix} = 0. \tag{12.142}$$

Using the mathematical identity $\max(1-q, 1+q) = 1+|q|$, and $\max(1-\Upsilon, 1+\Upsilon) = 1+|\Upsilon|$, we conclude that the maximal value of all eigenvalues is given by $\lambda_{\max} = \alpha\beta^4(1+|q|)(1+|\Upsilon|)(1+q\tanh(\beta^2 q)+|q+\tanh(\beta^2 q)|)$. The critical data density for the SSB phase is thus given by

$$\alpha_c = \frac{\beta^{-4}}{(1+|q|)(1+|\Upsilon|)(1+q\tanh(\beta^2 q)+|q+\tanh(\beta^2 q)|)}. \tag{12.143}$$

This SSB critical data density is compared with that of the prior-free case in Fig. 12.3. We see that the prior knowledge about $q$ significantly reshapes the critical data density surface for the SSB phase, which provides deep insights about roles of prior information.

## 12.2 Phase Diagram

In this section, we provide a detailed explanation of phase transitions caused by increasing data size for the model with prior. The difference from the prior-free scenario is also highlighted. Interestingly, when $\alpha$ is small, trivial (null values) order parameters except $R$ are a stable solution of Eq. (12.119), thereby suggesting a random guess (RG) phase. As expected, $R$ captures the prior information, thus being equal to $q$ irrespective of $\alpha$. In this phase, $\langle \xi^1 \rangle = \langle \xi^2 \rangle = 0$, the weight thus takes $\pm 1$ with equal probabilities, implying that the data does not provide any useful information to bias the weight's direction during learning. The underlying physics is that the posterior [Eq. (12.22)] is invariant under the reverse operation $\boldsymbol{\xi} \to -\boldsymbol{\xi}$, and this symmetry is unbroken in the RG phase.

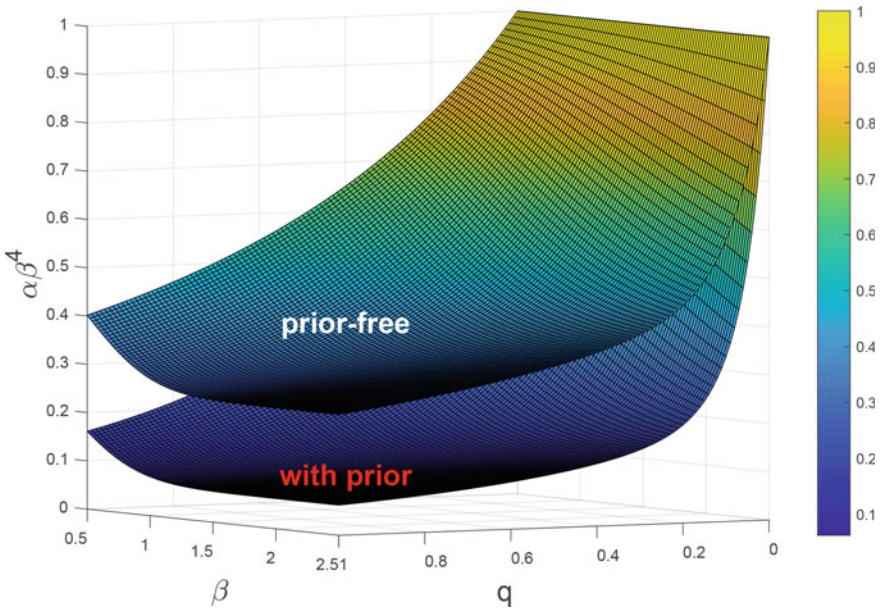



**Fig. 12.3** Comparison of SSB critical data densities in models with/without prior knowledge. This plot is adapted from Ref. [2]

Surprisingly, as more data is supplied, the RG phase would lose its stability at a critical data density. By a linear stability analysis as shown above, this threshold can be analytically obtained as

$$\alpha_c = \frac{\Lambda(\beta, q)}{(1 + |q|)(1 + |\tanh(\beta^2 q)|)}, \tag{12.144}$$

where $\Lambda(\beta, q) = \frac{\beta^{-4}}{1+q\tanh(\beta^2 q)+|q+\tanh(\beta^2 q)|}$ denotes the learning threshold for the prior-free scenario [1]. In the correlation-free case ($q = 0$, more than one hidden nodes allowed), the known threshold $\alpha_c = \beta^{-4}$ is recovered [3, 6]. Compared to the prior-free scenario, the prior knowledge contributes to a further reduction of the threshold (∼60% of the prior-free one for $q = 0.3$ and $\beta = 1$). Most interestingly, in the weak correlation limit, where $q \sim \beta^{-2}$ with a proportional constant $q_0$ in the presence of less noisy data (large $\beta$), $\alpha_c\beta^4 = \frac{1}{(1+|\tanh q_0|)^2}$, which demonstrates that the learning threshold can be decreased to only 32% of the correlation-free case for $q_0 = 1$. This demonstrates the same benefit of the weak correlation between synapses as shown in the prior-free scenario.

When $\alpha > \alpha_c$, the RG phase is replaced by the symmetry-broken phase, where $\langle \xi^1 \rangle = \langle \xi^2 \rangle \neq 0$. Note that the inherent-reverse-symmetry is spontaneously broken. We thus call the second phase a spontaneous symmetry breaking (SSB) phase. The SSB implies a non-zero solution of $q_1 = q_2 = T_1 = T_2 = \tau_1 = \tau_2 = r$. As a reason-

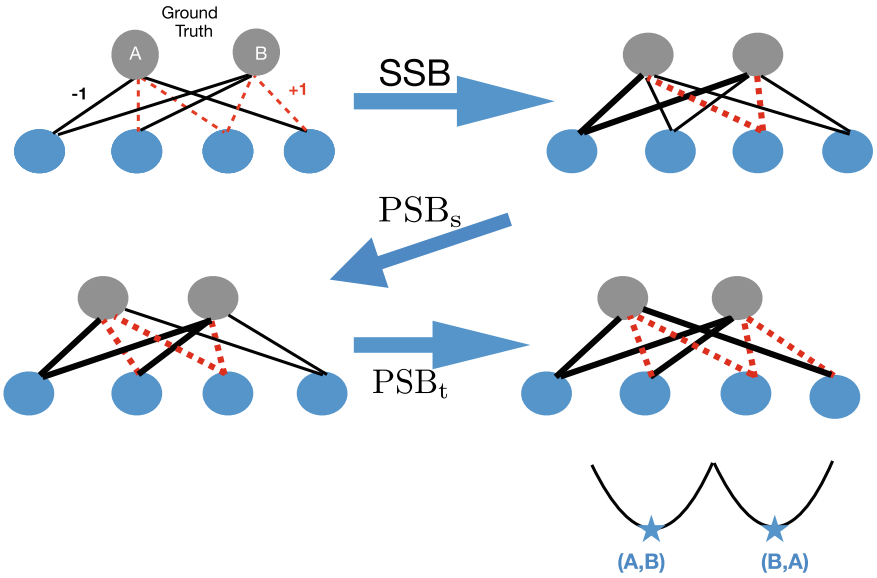
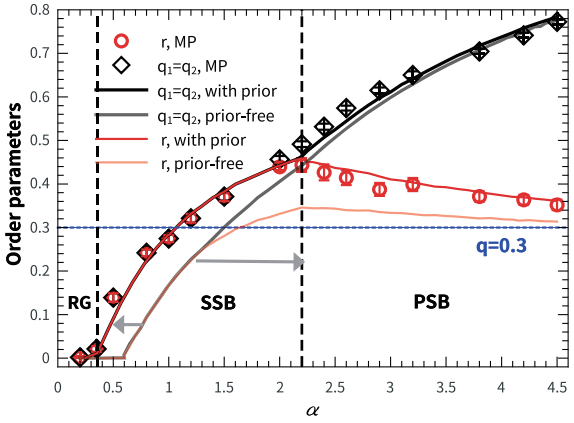


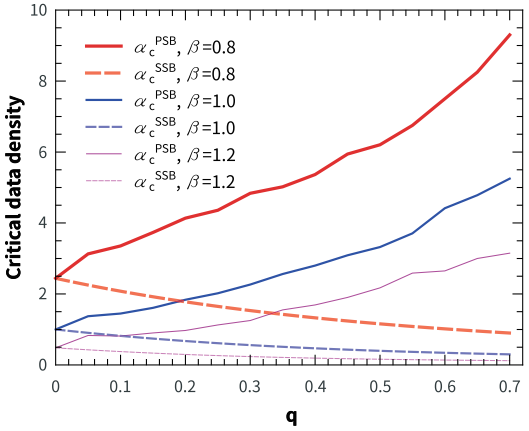

**Fig. 12.4** A schematic illustration of various kinds of inherent symmetry breaking in unsupervised learning. As the data density $\alpha$ increases, a first continuous transition related to the reverse symmetry breaking occurs, where the student machine starts to infer the common parts of ground-truth receptive fields. This type of transition is named spontaneous symmetry breaking (SSB), as encountered in a standard Ising model. As $\alpha$ further increases, the student starts to infer the distinct part of the ground truth. This is called the first type of permutation symmetry breaking (PSB), i.e., the student starts to realized that its two receptive fields are not the same. We name this transition as PSB$_s$, where the subscript means student. As the data density further increases, the student starts to be capable of distinguishing the intrinsic order of two hidden nodes in the teacher's (or ground truth) architecture. We call this transition as PSB$_t$, where the subscript means teacher. Only after this transition, the free energy has two equally important valleys (for an arbitrary number $P$ of hidden neurons, there are reasonably $P!$ valleys). These two valleys corresponds to two possible orders of $(A, B)$ or $(B, A)$ for the ground truth, which is also the inherent permutation symmetry in the model to generate the data of the identical Gibbs-Boltzmann distribution

able interpretation, the student infers only the common part of the two planted RFs in this new phase (see Fig. 12.4, see also a numerical simulation proof in the previous work [1]). Thus the PS still holds for the student's hidden neurons. Moreover, $\boldsymbol{\xi}^{1,\text{true}}$ and $\boldsymbol{\xi}^{2,\text{true}}$ have the PS property as well, providing a physics explanation of the solution we obtained. The SSB phase is thus permutation symmetric and stable until a turnover of the order parameter $r$ is reached [Fig. 12.5a].

At the turnover, the PS is also spontaneously broken, thereby leading to a permutation symmetry breaking (PSB) phase. The third phase is characterized by two fixed points: (i) $q_1 = q_2 = T_1 = T_2$, and $\tau_1 = \tau_2 = r$; (ii) $q_1 = q_2 = \tau_1 = \tau_2$, and $T_1 = T_2 = r$. We remark that these two fixed points share the same free energy, representing two possible choices of ground truth— $(\boldsymbol{\xi}^{1,\text{true}}, \boldsymbol{\xi}^{2,\text{true}})$ or $(\boldsymbol{\xi}^{2,\text{true}}, \boldsymbol{\xi}^{1,\text{true}})$, resembling the well-known free energy picture of ferromagnetic Ising model. In fact,




(a)




(b)

**Fig. 12.5** Phase diagram of unsupervised learning with priors. **a** Order parameters versus data densities with $(\beta, q)$=(1.0, 0.3). Lines are replica results compared with symbols obtained from the message passing (MP) procedure (instances of $N = 200$). Results of the prior-free unsupervised learning are also plotted for comparison. The arrows indicate the role of priors in shifting the phase transition points. **b** Critical data densities for SSB and PSB are obtained from replica analysis and plotted for increasing values of $\beta$. These plots are adapted from Ref. [2]

the PSB phase has two subtypes: the first one is a $\text{PSB}_\text{s}$ phase where the permutation symmetry between $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ is broken on the student's side, i.e., $\langle \xi^1 \rangle$ can point conversely to $\langle \xi^2 \rangle$ but with the same magnitude, thereby $q_1 = q_2 \neq r$, and the second one is a $\text{PSB}_\text{t}$ phase where the PSB occurs on the teacher's side, i.e., $\boldsymbol{\xi}^{1,\text{true}}$ and $\boldsymbol{\xi}^{2,\text{true}}$ cannot be freely permuted, thereby $T_{1,2} \neq \tau_{2,1}$ (see Fig. 12.4). Interestingly, the self-overlap deviates from $r$ at the turnover, thereby merging $\text{PSB}_\text{s}$ phase and $\text{PSB}_\text{t}$ phase into a single PSB phase, rather than separating these two subtypes as in the prior-free scenario (Fig. 12.5a). With the help of prior knowledge, the student is able to distinguish two planted RFs ($\text{PSB}_\text{t}$, recognizing the exact order) *at the same time* when starting to infer different components of the student's RFs ($\text{PSB}_\text{s}$). This process is also pictorially shown in Fig. 12.4. Furthermore, the prior does not change the $\text{PSB}_\text{t}$ transition point of the prior-free case, in that knowing $q$ does not help to accelerate the recognition of two choices of ground truth. The only effect is that the

knowledge of $q$ does elevate the overlap values before the turnover, resulting in a larger value of $r$ in the post-turnover regime compared to the prior-free case. After the turnover, the overlap equal to $\min(T_1, \tau_1)$ or $\min(T_2, \tau_2)$ has the same value with $r$, since $(\boldsymbol{\xi}^{1,\text{true}}, \boldsymbol{\xi}^{2,\text{true}})$ follows the same posterior as $(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$, as can be deduced from the Nishimori condition of the Bayes optimal learning. As expected, $r$ finally tends to $q$ at a finite but large value of $\alpha$ [Fig. 12.5a], suggesting that the unsupervised learning is completed.

We conclude that with/without the prior knowledge, the data stream drives the SSB and PSB phase transitions of continuous type [1, 2]. Thresholds of the transitions for the prior case are summarized in Fig. 12.5b. This conclusion can be verified by numerical simulations on single instances of the model by applying the corresponding message-passing-based learning algorithm (Fig. 12.5a, technical details have been given in the previous sections). We finally remark that in a general RBM with more than two hidden neurons, the message passing does not apply, or even we cannot have a closed-form for the equation. However, a variational mean-field theory, working at the model ensemble level, can be used to treat arbitrary many hidden neurons, as we already introduce in detail in Chap. 10.

## 12.3 Hyper-Parameters Inference

In this section, we show how to infer the hyper-parameters of our unsupervised learning model. We first write the posterior probability of the hyper-parameters $\beta$ and $q$ as [2]

$$P(\beta, q|\mathcal{D}) = \sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2} P(\beta, q, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2|\mathcal{D}) = \sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2} \frac{P(\mathcal{D}|\beta, q, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2) P_0(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2|q)}{\int\int d\beta dq \sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2} P(\mathcal{D}|\beta, q, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2) P_0(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2|q)}, \tag{12.145}$$

where $\mathcal{D}$ indicates the dataset, we have used the Bayes' rule, and we assume that $P_0(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \beta, q) = P_0(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2|q)\tilde{P}_0(\beta, q)$ where $\tilde{P}_0(\beta, q)$ is a constant or equivalently we have no prior knowledge about the true values of the hyper-parameters. Therefore, we have

$$P(\beta, q|\mathcal{D}) \propto \sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2} \prod_{a=1}^{M} P(\boldsymbol{\sigma}^a|\beta, q, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2) \prod_{i=1}^{N} P_0(\xi_i^1, \xi_i^2|q). \tag{12.146}$$

Note that the data distribution can be expressed as

$$P(\boldsymbol{\sigma}^a|\beta, q, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2) = \frac{\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}^a\right)}{2^N e^{\beta^2} \cosh(\beta^2 Q)}. \tag{12.147}$$

The posterior probability of the hyper-parameters can be finally simplified as $P(\beta, q|\mathcal{D}) \propto e^{-\beta^2 M}\Omega$, where $\Omega$ is exactly the partition function of the posterior $P(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2|\mathcal{D})$. This partition function can be written explicitly as follows:

$$\Omega(\beta, q) = \sum_{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2} \prod_{a=1}^{M} \frac{\cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}^a\right)}{\cosh(\beta^2 Q)} \prod_{i=1}^{N} P_0(\xi_i^1, \xi_i^2|q). \tag{12.148}$$

Searching for consistent hyper-parameters $(\beta, q)$ compatible with the supplied dataset is equivalent to maximizing the posterior $P(\beta, q|\mathcal{D})$. Following this principle, we first derive the temperature equation as

$$\frac{\partial \ln P(\beta, q|\mathcal{D})}{\partial \beta} = -2M\beta + \frac{\partial}{\partial \beta} \ln \Omega(\beta, q). \tag{12.149}$$

Note that in statistical physics, the energy function is given by $N\epsilon = -\frac{\partial \ln \Omega}{\partial \beta}$, where $\epsilon(\beta, q)$ denotes the energy density (per degree of freedom). We thus conclude that $\beta$ should obey the following temperature equation

$$\beta = -\frac{\epsilon(\beta, q)}{2\alpha}. \tag{12.150}$$

Note that when the true prior is taken into account, the energy density of the model is analytic with the result $\epsilon = -2\alpha\beta$ independent of $q$. This is exactly what the Nishmori model shows (see Chap. 6). Note that rare model of spin glass can have analytic energy in general.

Given the dataset and an initial guess of $\beta$, the aforementioned message passing scheme with prior knowledge can be used to estimate the energy density of the system as $N\epsilon = -\sum_i \Delta\epsilon_i + (N-1)\sum_a \Delta\epsilon_a$ based on the Bethe approximation. The energy contribution of one synapse-pair reads

$$\Delta\epsilon_i = \frac{\sum_{\xi_i^1, \xi_i^2} \sum_{b \in \partial i} \frac{\partial u_{b\to i}(\xi_i^1, \xi_i^2)}{\partial \beta} e^{\sum_{b\in\partial i} u_{b\to i}(\xi_i^1, \xi_i^2) + \ln P_0(\xi_i^1, \xi_i^2)}}{\sum_{\xi_i^1, \xi_i^2} e^{\sum_{b\in\partial i} u_{b\to i}(\xi_i^1, \xi_i^2) + \ln P_0(\xi_i^1, \xi_i^2)}}, \tag{12.151}$$

where $\frac{\partial u_{b\to i}(\xi_i^1, \xi_i^2)}{\partial \beta}$ reads as follows,

$$\begin{aligned}
\beta \frac{\partial u_{b\to i}(\xi_i^1, \xi_i^2)}{\partial \beta} &= \beta^2[\Gamma_{b\to i}^1 + \Gamma_{b\to i}^1 + 2\Xi_{b\to i}] - 2\beta^2 \left(Q_{b\to i} + \frac{\xi_i^1 \xi_i^2}{N}\right) \\
&\times \tanh\left(\beta^2 Q_{b\to i} + \frac{\beta^2}{N}\xi_i^1 \xi_i^2\right) + Y_{b\to i} \tanh Y_{b\to i} \\
&+ \frac{\Delta_{b\to i}}{1 + \Delta_{b\to i}} \left(-4\beta^2 \Xi_{b\to i} + X_{b\to i} \tanh X_{b\to i} - Y_{b\to i} \tanh Y_{b\to i}\right),
\end{aligned} \tag{12.152}$$

where $X_{b\to i} \equiv \beta G^1_{b\to i} - \beta G^2_{b\to i} + \frac{\beta}{\sqrt{N}}\sigma_i^b(\xi_i^1 - \xi_i^2)$, $Y_{b\to i} \equiv \beta G^1_{b\to i} + \beta G^2_{b\to i} + \frac{\beta}{\sqrt{N}}\sigma_i^b(\xi_i^1 + \xi_i^2)$, and $\Delta_{b\to i} \equiv e^{-2\beta^2\Xi_{b\to i}}\frac{\cosh X_{b\to i}}{\cosh Y_{b\to i}}$. The energy contribution of one data sample is given by

$$\begin{aligned}\beta\Delta\epsilon_a &= \beta^2(\Gamma_a^1 + \Gamma_a^2 + 2\Xi_a) - 2\beta^2 Q_a \tanh(\beta^2 Q_a) + Y_a \tanh Y_a \\ &+ \frac{\Delta_a}{1+\Delta_a}\left(-4\beta^2\Xi_a + X_a \tanh X_a - Y_a \tanh Y_a\right),\end{aligned} \tag{12.153}$$

where $X_a \equiv \beta G_a^1 - \beta G_a^2$, $Y_a \equiv \beta G_a^1 + \beta G_a^2$, and $\Delta_a = e^{-2\beta^2\Xi_a}\frac{\cosh X_a}{\cosh Y_a}$.

Next, we derive the correlation equation. Note that $P_0(\xi_i^1, \xi_i^2) = \frac{e^{J_0\xi_i^1\xi_i^2}}{4\cosh J_0}$, where $J_0 = \tanh^{-1} q$. This prior contributes an extra coupling term in the effective Hamiltonian in the replica computation. We then have

$$\frac{\partial P(\beta, q|\mathcal{D})}{\partial q} = e^{-M\beta^2}\frac{\partial\Omega}{\partial q} = 0, \tag{12.154}$$

which requires that $\frac{\partial\Omega}{\partial q} = 0$. It then follows that

$$\begin{aligned}\frac{\partial\Omega}{\partial q} &= \Omega\sum_{\boldsymbol{\xi}^1,\boldsymbol{\xi}^2} P(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2|\mathcal{D})\sum_i(\xi_i^1\xi_i^2 - \tanh J_0)\frac{\partial J_0}{\partial q} \\ &= \Omega\left(\sum_i \langle\xi_i^1\xi_i^2\rangle_{P(\boldsymbol{\xi}^1,\boldsymbol{\xi}^2|\mathcal{D})} - N\tanh J_0\right)\frac{\partial J_0}{\partial q} = 0.\end{aligned} \tag{12.155}$$

To satisfy Eq. (12.155), the following correlation equation must be solved:

$$q = \frac{1}{N}\sum_i q_i, \tag{12.156}$$

where $q_i$ can be computed in a single instance by iterating the message passing scheme. More precisely

$$q_i = \frac{\sum_{\xi_i^1,\xi_i^2}\xi_i^1\xi_i^2 e^{\sum_{b\in\partial i}u_{b\to i}(\xi_i^1,\xi_i^2)}P_0(\xi_i^1,\xi_i^2)}{\sum_{\xi_i^1,\xi_i^2}e^{\sum_{b\in\partial i}u_{b\to i}(\xi_i^1,\xi_i^2)}P_0(\xi_i^1,\xi_i^2)}. \tag{12.157}$$

In addition, the negative log-likelihood of the hyper-parameter posterior per neuron can also be estimated as $\frac{\mathcal{L}}{N} = C - \frac{\ln\Omega}{N} + \alpha\beta^2$, where $C$ is an irrelevant constant, and the second term can be approximated by $\beta f_{\text{Bethe}}$. This measure is helpful to characterize the quality of the inference performance, as was analyzed in our work [2]. Given only the data samples, the inference of hyper-parameters $(\beta,q)$ underlying the data can be achieved by iteratively imposing the Nishimori constraint to reach a consistent value of $(\beta, q)$ to explain the data. In statistics, this iterative scheme

is called the expectation-maximization algorithm [8], where the message update to compute $(\epsilon, \{q_i\})$ is called an expectation-step, while the hyper-parameter update is called a maximization-step. The hyper-parameter space, especially when the amount of data samples is not sufficient, is not guaranteed to be convex, instead being highly non-convex in general. A high relative inference error with a large fluctuation in a data-deficient regime would be observed.

## References

1. T. Hou, K.Y.M. Wong, H. Huang, J. Phys. A: Math. Theor. **52**(41), 414001 (2019)
2. T. Hou, H. Huang, Phys. Rev. Lett. **124**, 248302 (2020)
3. H. Huang, T. Toyoizumi, Phys. Rev. E **94**, 062310 (2016)
4. P. Rodriguez, J. Gonzalez, G. Cucurull, J.M. Gonfaus, X. Roca, in *ICLR 2017* (2016). arXiv:1611.01967
5. H. Huang, Phys. Rev. E **98**, 062313 (2018)
6. H. Huang, J. Stat. Mech.: Theory Exper. **2017**(5), 053302 (2017)
7. A. Barra, G. Genovese, P. Sollich, D. Tantari, Phys. Rev. E **96**, 042156 (2017)
8. A.P. Dempster, N.M. Laird, D.B. Rubin, J. R. Stat. Soc. Ser. B **39**, 1 (1977)

# Chapter 13
# Mean-Field Theory of Ising Perceptron


Learning problem asks one to find a group of synapses to store $P$ patterns in a network with $N$ neurons. For a feedforward structure, it can also be seen as a classification problem of $P$ patterns with specified labels. For this purpose, we can design a simplest architecture with only one output unit but with binary synapses connecting input nodes to the output. Although this binary Perceptron is not a practical setting for complex learning (e.g., non-linear-separable datasets), the toy model received a substantial research interest especially in statistical physics community. In particular, many important theoretical insights are gained from studies of this model. In this chapter, we bring some important progresses in recent years about the theoretical studies of the Ising/binary Perceptron (Braunstein and Zecchina in Phys. Rev. Lett. 96:030201, 2006 [1]; Huang and Kabashima in Phys. Rev. E 90:052813, 2014 [2]; Baldassi et al. in Phys. Rev. Lett. 115(12): 128101, 2015 [3]).

## 13.1 Ising Perceptron model

Perceptron models [4] were first studied by physicists in 1980s. Continuous weights were first analyzed as a statistical mechanics problem. From an information storage perspective, the capacity, denoted as the maximal ratio ($\alpha_c$) between the number of random patterns classified correctly by the machine and the number of input neurons, was claimed to be $\alpha_c = 2$ [5]; later, this setting was generalized to the perceptron with binary ($\pm 1$) synapses (also called Ising-type), and the capacity decreases below one (the upper-bound from an information-theoretical perspective) [6–8]. The spherical perceptron with continuous weights has the continuous space of solutions below the capacity, and thus training is easy. However, the binary perceptron has isolated equilibrium solutions [2], and the training in the worst cases belongs to the NP-complete class [9, 10].



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_13

The Ising perceptron is a simple and abstract model of a biological neuronal network (e.g., cerebellar Purkinje cells). The output of the Ising perceptron is the sign of an weighted summation of its input (see Fig. 13.1), given by

$$y^{\mu} = \operatorname{sgn}\left(\sum_i \xi_i^{\mu} J_i\right), \tag{13.1}$$

where $\xi_i^{\mu}$ is the $i$th component of the $\mu$th pattern, $J_i$ is the $i$th synapse and $\operatorname{sgn}(\cdot)$ is the sign function. Note that the patterns are randomly selected with equal probabilities for their entries, i.e., $P(\xi_i^{\mu} = \pm 1) = 1/2$. The corresponding label is also random and independent of the input signals. In the case of labels generated from a teacher perceptron, the learning problem turns out to be a generalization problem [11]. We will not analyze this case, because methods introduced in this chapter can be easily adapted to the generalization case, in which there emerges interesting first-order transitions for learning [12, 13]. If the output $y^{\mu}$ is equal to a prescribed label $\sigma^{\mu}$, we say the perceptron successfully classifies $\boldsymbol{\xi}^{\mu}$. The energy cost of the network is then given by

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} \Theta\left(-\frac{\sigma^{\mu}}{\sqrt{N}} \sum_i J_i \xi_i^{\mu}\right). \tag{13.2}$$

$E$ ranges from 0 to $P$, taking 0 for a complete storage, and $P$ for a complete failure of learning. Thus our goal is to find an optimal $\mathbf{J}$ to minimize $E$. The learning problem is thus formulated as an optimization problem in the space of neural interactions. The joint distribution of $\mathbf{J}$ can be formulated in the following Boltzmann–Gibbs form

$$P(\mathbf{J}) = \frac{1}{Z} \exp(-\beta E(\mathbf{J})), \tag{13.3}$$

where $\beta$ is an inverse-temperature characterizing the learning noise. In the zero-temperature limit, the distribution can be written as

$$P(\mathbf{J}) = \frac{1}{Z} \prod_{\mu} \Theta\left(\frac{\sigma^{\mu}}{\sqrt{N}} \sum_i J_i \xi_i^{\mu}\right), \tag{13.4}$$

where $\Theta$ is the Heaviside step function. $\ln Z$ thus counts the number of solutions (valid $\mathbf{J}$) to the learning problem (a definition of an entropy $S$). In other words, Eq. (13.4) indicates a uniform sampling of the solution space composed of all valid $\mathbf{J}$.

One can expect that the number of solutions will decrease with the increase of the number of patterns $P$, because it is more and more difficult to satisfy more and more constraints of patterns. We are interested in large values of $P$ and $N$, but keeping a finite pattern density $\alpha \equiv P/N$. How $S$ changes with $\alpha$ is of theoretical interests, which determines the maximal density (capacity) of the network that can be compared with numerical experiments.

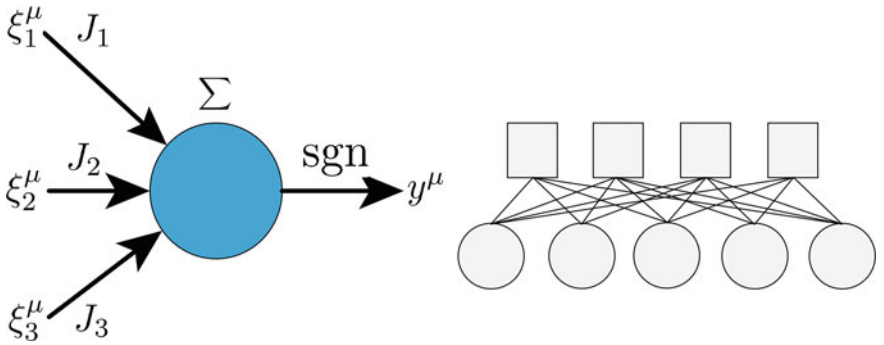



**Fig. 13.1** Left: A binary perceptron with three synapses. Right: the factor graph of the binary perceptron. Circles (called variable nodes) represent synapses, and squares (called function nodes) represent patterns to be learned

## 13.2 Message-Passing-Based Learning

To calculate $S$, one has to compute the partition function $Z$, but direct calculation is unrealistic when $N$ is large. Message passing algorithm, which is an application of Bethe approximation in statistical physics, can provide a reasonable approximation about the partition function with a much less computation, as we already see in Chap. 2. Under this approximation, the joint distribution is factorized as the product of pattern $\mu$ (except for a site-dependent factor for normalization). Therefore, the message passing equation (see the factor graph in Fig. 13.1) is given by [1]

$$\begin{aligned} P_{i\to a}(J_i) &= \frac{1}{Z_{i\to a}} \prod_{b\in\partial i\backslash a} \hat{P}_{b\to i}(J_i), \\ \hat{P}_{b\to i}(J_i) &= \sum_{\{J_j|j\in\partial b\backslash i\}} \Theta\left(\frac{\sigma^b}{\sqrt{N}} \sum_j J_j \xi_j^b\right) \times \prod_{j\in\partial b\backslash i} P_{j\to b}(J_j), \end{aligned} \tag{13.5}$$

where $Z_{i\to a} = \prod_{b\neq a} \hat{P}_{b\to i}(+1) + \prod_{b\neq a} \hat{P}_{b\to i}(-1)$.

The second equation of Eq. (13.5) needs to sum up all $2^{N-1}$ configurations, which is practically impossible. Notice that this summation is exactly the average of the factor term under the cavity distributions, then $\hat{P}_{b\to i}$ can be written as

$$\hat{P}_{b\to i}(J_i) = \sum_{\mathbf{J}_{\backslash i}} f(\mathbf{J}_{\backslash i}) P(\mathbf{J}_{\backslash i}). \tag{13.6}$$

Since $\mathbf{J}_{\backslash i}$ take discrete values, we cannot directly replace the summation by an integral. If we could find an auxiliary variable $w(\mathbf{J}_{\backslash i})$ which is a function of $\mathbf{J}_{\backslash i}$ and takes continual values in the large $N$ limit, the average can be replaced by

$$\hat{P}_{b\to i}(J_i) \approx \int dw P(w) g(w), \tag{13.7}$$

where $g(w(\mathbf{J}_{\setminus i})) \equiv f(\mathbf{J}_{\setminus i})$. Naturally, we define $w_{b\to i} \equiv \frac{1}{\sqrt{N}} \sum_{j\in\partial b\setminus i} J_j \xi_j^b$. Without loss of generality, we set $\sigma^b \equiv 1$ for any input patterns in the remaining part of this chapter, since our learning setting is invariant under the transformation $\xi_i^b \to \sigma^b \xi_i^b$. Then the exact form of Eq. (13.7) is given by

$$\hat{P}_{b\to i}(J_i) \approx \int dw_{b\to i} P(w_{b\to i}) \Theta\left(w_{b\to i} + \frac{1}{\sqrt{N}} J_i \xi_i^b\right). \tag{13.8}$$

Due to the central limit theorem (CLT), $w_{b\to i} \sim \mathcal{N}(\hat{w}_{b\to i}, \hat{\sigma}_{b\to i})$, where

$$\hat{w}_{b\to i} = \langle w_{b\to i}\rangle = \frac{1}{\sqrt{N}} \sum_{j\neq i} m_{j\to b} \xi_j^b, \tag{13.9a}$$

$$\hat{\sigma}_{b\to i} = \langle w_{b\to i}^2\rangle - \langle w_{b\to i}\rangle^2 = \frac{1}{N} \sum_{j\neq i} (1 - m_{j\to b}^2). \tag{13.9b}$$

It then follows that $P(w_{b\to i}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{b\to i}}} \exp\left[\frac{-(w_{b\to i}-\hat{w}_{b\to i})^2}{2\hat{\sigma}_{b\to i}}\right]$. Notice that the step function equals zero when its argument is less than zero. Then we obtain

$$\hat{P}_{b\to i}(J_i) = \int_{-\frac{1}{\sqrt{N}} J_i \xi_i^b}^{\infty} P(w_{b\to i}) dw_{b\to i} = H\left(-\frac{\frac{1}{\sqrt{N}} J_i \xi_i^b + \hat{w}_{b\to i}}{\sqrt{\hat{\sigma}_{b\to i}}}\right). \tag{13.10}$$

The function $H(x) = \int_x^\infty \frac{e^{-\frac{x^2}{2}} dz}{\sqrt{2\pi}}$, which is related to the error function $H(x) = \frac{1-\mathrm{erf}(x/\sqrt{2})}{2}$.

For a further simplification, we apply the definition of cavity magnetization $m_{i\to a}$,

$$\begin{aligned} m_{i\to a} &= P_{i\to a}(+1) - P_{i\to a}(-1) \\ &= \frac{\prod_{b\neq a} \hat{P}_{b\to i}(+1) - \prod_{b\neq a} \hat{P}_{b\to i}(-1)}{\prod_{b\neq a} \hat{P}_{b\to i}(+1) + \prod_{b\neq a} \hat{P}_{b\to i}(-1)} \\ &= \frac{\exp(\sum_{b\neq a} \ln \hat{P}_{b\to i}(+1)) - \exp(\sum_{b\neq a} \ln \hat{P}_{b\to i}(-1))}{\exp(\sum_{b\neq a} \ln \hat{P}_{b\to i}(+1)) + \exp(\sum_{b\neq a} \ln \hat{P}_{b\to i}(-1))} \\ &= \tanh\left(\sum_{b\neq a} \frac{1}{2} \ln \frac{\hat{P}_{b\to i}(+1)}{\hat{P}_{b\to i}(-1)}\right). \end{aligned} \tag{13.11}$$

Finally, the message passing equations are summarized as follows:

$$m_{i\to a} = \tanh\left(\sum_{b\neq a} u_{b\to i}\right), \tag{13.12a}$$

$$u_{b\to i} = \frac{1}{2}\left[\ln H\left(-\frac{\frac{1}{\sqrt{N}}\xi_i^b + \hat{w}_{b\to i}}{\sqrt{\hat{\sigma}_{b\to i}}}\right) - \ln H\left(-\frac{-\frac{1}{\sqrt{N}}\xi_i^b + \hat{w}_{b\to i}}{\sqrt{\hat{\sigma}_{b\to i}}}\right)\right], \tag{13.12b}$$

where $\hat{w}_{b\to i} = \frac{1}{\sqrt{N}}\sum_{j\neq i} m_{j\to b}\xi_j^b$ and $\hat{\sigma}_{b\to i} = \frac{1}{\sqrt{N}}\sum_{j\neq i}(1 - m_{j\to b}^2)$.

Meanwhile, the Bethe free energy can be calculated as (see also explanations in Chap. 2)

$$\beta F = \sum_i \beta F_i - (N-1)\sum_a \beta F_a, \tag{13.13}$$

$$\beta F_i = -\ln Z_i = -\ln\left[\prod_b \hat{P}_{b\to i}(+1) + \prod_b \hat{P}_{b\to i}(-1)\right], \tag{13.14}$$

$$\beta F_a = -\ln Z_a = -\ln H\left(\frac{-\hat{w}_b}{\sqrt{\hat{\sigma}_b}}\right), \tag{13.15}$$

where $\hat{w}_b = \frac{1}{\sqrt{N}}\sum_j m_{j\to b}\xi_j^b$, and $\hat{\sigma}_b = \frac{1}{\sqrt{N}}\sum_j(1 - m_{j\to b}^2)$. The entropy is exactly the value of $-\beta F$ when the energy is zero. Hence, we have

$$s = \frac{S}{N} = -\frac{\beta F}{N}. \tag{13.16}$$

Figure 13.2 shows how the entropy changes with the pattern density. Two points can help us to examine whether the entropy is correct: (i) When $\alpha$ is zero, each synapse can take arbitrary values, suggesting that the total configuration is $2^N$, and thus the entropy should be $\ln 2$; (ii) The shape of entropy as a function of $\alpha$ must be concave. A monotonic decrease of the entropy profile is confirmed. The capacity above which the entropy becomes negative is estimated to be about 0.833, which will be exactly computed by the replica theory in the next section.

## 13.3 Replica Analysis

Since MP only gives the approximation of the capacity when specific patterns are given, it is hard to find the precise capacity (the $N$-independent one), due to the fluctuation caused by selections of $\boldsymbol{\xi}$. Instead, we turn to the replica method to calculate the precise free energy averaged over all possible realizations of random patterns. The replica trick is given by

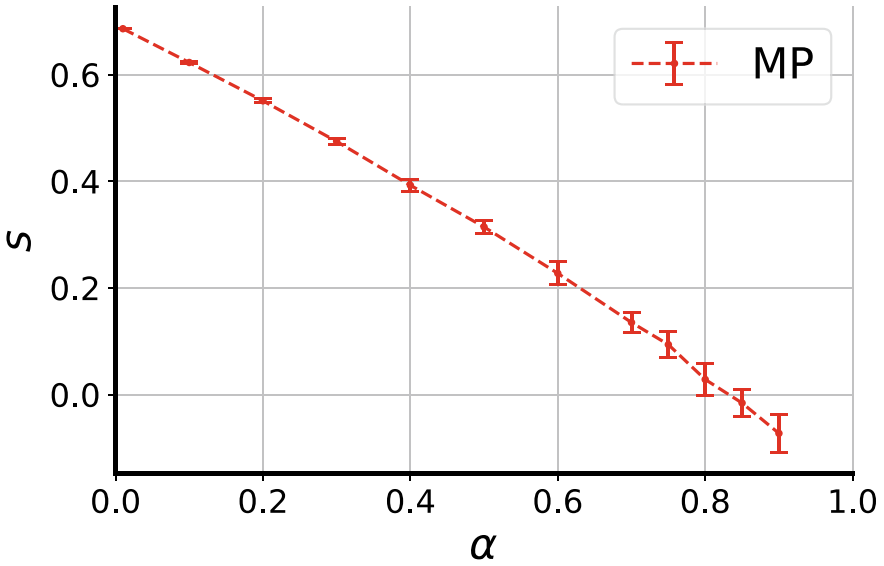



**Fig. 13.2** Entropy versus pattern density. Results are estimated by the message passing algorithm (MP) running on single instances of $N = 1000$. 20 random realizations of the model are considered

$$\langle \ln Z \rangle = \lim_{n \to 0} \frac{\ln \langle Z^n \rangle}{n}, \tag{13.17}$$

where $\langle \cdots \rangle$ indicates the quenched-disorder average over different random patterns. Here, the disorder average of a logarithm can be transformed into computing the average of an integer power of the replicated partition function $\langle Z^n \rangle$. Replica refers to the process we copy the original system for $n$ times. Correlations among synapses, which precludes an analytic study, will be transformed into correlations among replicas, which is amenable for further assumptions. In other words, synapses are decoupled, and instead an overlap among replicas of the original system has to be introduced.

Therefore, $\langle Z^n \rangle$ is given by

$$\langle Z^n \rangle = \left\langle \sum_{\{\mathbf{J}^a\}} \prod_{a=1}^{n} \prod_{\mu=1}^{P} \Theta\left( \frac{1}{\sqrt{N}} \mathbf{J}^a \boldsymbol{\xi}^\mu \right) \right\rangle = \sum_{\{\mathbf{J}^a\}} \left\langle \prod_{a=1}^{n} \prod_{\mu=1}^{p} \Theta\left( \frac{1}{\sqrt{N}} \mathbf{J}^a \boldsymbol{\xi}^\mu \right) \right\rangle. \tag{13.18}$$

To proceed, we first define the weighted sum as

$$u_\mu^a = \frac{1}{\sqrt{N}} \mathbf{J}^a \boldsymbol{\xi}^\mu. \tag{13.19}$$

The covariance structure for the sum is given by

$$\langle u_\mu^a \rangle = 0, \tag{13.20}$$

$$\langle u_\mu^a u_\nu^b \rangle = \delta_{\mu\nu} q^{ab}, \tag{13.21}$$

where $\delta_{\mu\nu}$ is a Kronecker delta function, and $q^{ab} = \frac{1}{N}\sum_i J_i^a J_i^b$ being the desired overlap (order parameter) due to the replica operation. Then, we introduce $q$ by the delta function

$$\langle Z^n \rangle = \sum_{\{\mathbf{J}^a\}} \int \prod_{a<b} dq^{ab} \delta \left( \sum_i J_i^a J_i^b - N q^{ab} \right) \left\langle \prod_{\mu=1}^{P} \prod_{a=1}^{n} \Theta(u_\mu^a) \right\rangle_{\{u_\mu^a\}}. \tag{13.22}$$

Using the Fourier representation, $\delta(x-a) = \int d\hat{q}/2\pi \exp\left[\mathrm{i}(x-a)\hat{q}\right]$, we recast Eq. (13.22) into the following form

$$\begin{aligned} \langle Z^n \rangle = & \sum_{\{\mathbf{J}^a\}} \int \left( \prod_{a<b} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi} \right) \exp\left( N(-\mathrm{i} \sum_{a<b} q^{ab}\hat{q}^{ab}) + \mathrm{i} \sum_{a<b} \hat{q}^{ab} J^a J^b \right) \\ & \times \left\langle \prod_{a}^{n} \prod_{\mu=1}^{P} \Theta(u_\mu^a) \right\rangle_{\{u_\mu^a\}}, \end{aligned} \tag{13.23}$$

where $J^a J^b$ is a vector inner product.

### 13.3.1 Replica Symmetry

To get physics results, we have to make an assumption about the form of the overlap matrix. Here, we use the RS ansätz, i.e., the overlap entries do not depend on specific replica index, or permutation symmetry holds for the matrix. Specifically,

$$q^{ab} = \delta_{ab} + (1 - \delta_{ab}) q. \tag{13.24}$$

Under this first-level approximation, we can first simplify terms involving $q$:

$$\sum_{a<b} q^{ab}\hat{q}^{ab} = \frac{n(n-1)}{2} q\hat{q}, \tag{13.25a}$$

$$\begin{aligned} \sum_{a<b} \hat{q}^{ab} J^a J^b &= \frac{\hat{q}}{2} \left( \sum_{a,b} J^a J^b - \sum_a J^a J^a \right) \\ &= \frac{\hat{q}}{2} \left( \sum_{a,b,i} J_i^a J_i^b - nN \right) = \frac{\hat{q}}{2} \left( \sum_i \Big( \sum_a J_i^a \Big)^2 - nN \right). \end{aligned} \tag{13.25b}$$

By making the variable transformation, $\hat{q} \to \mathrm{i}\hat{q}$, we have

$$\langle Z^n \rangle = \int \prod_{a<b} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi \mathrm{i}} \underbrace{\exp\left(-\frac{Nn(n-1)}{2} q\hat{q}\right) \overbrace{\sum_{\{\mathbf{J}^a\}} \exp\left(\frac{\hat{q}}{2}\left(\sum_i \left(\sum_a J_i^a\right)^2 - nN\right)\right)}^{\mathrm{A}}}_{\text{Entropy term}}$$
$$\times \underbrace{\left\langle \prod_{a=1}^{n} \prod_{\mu=1}^{P} \Theta(u_\mu^a) \right\rangle_{\{u_\mu^a\}}}_{\text{Energy term}} . \tag{13.26}$$

By applying the Gaussian integral identity: $\int Dz e^{bz} = e^{\frac{b^2}{2}}$, we compute the part $A$ as follows:

$$\begin{aligned}
A &\equiv \sum_{\{\mathbf{J}^a\}} \exp\left(\frac{\hat{q}}{2}\left(\sum_i \left(\sum_a J_i^a\right)^2 - nN\right)\right) \\
&= \exp(-nN\hat{q}/2) \sum_{\{\mathbf{J}^a\}} \prod_i \exp\left(\frac{\hat{q}}{2}\left(\sum_a J_i^a\right)^2\right) \\
&= \exp(-nN\hat{q}/2) \prod_i \sum_{\{J_i^a\}} \exp\left(\frac{\hat{q}}{2}\left(\sum_a J_i^a\right)^2\right) \\
&= \exp(-nN\hat{q}/2) \prod_i \sum_{\{J_i^a\}} \int Dz \exp\left(\sqrt{\hat{q}} \sum_a J_i^a z\right) \\
&= \exp(-nN\hat{q}/2) \prod_i \int Dz \sum_{\{J_i^a\}} \prod_a \exp(\sqrt{\hat{q}} J_i^a z) \\
&= \exp(-nN\hat{q}/2) \prod_i \int Dz \prod_a \sum_{J_i^a} \exp(\sqrt{\hat{q}} J_i^a z) \\
&= \exp(-nN\hat{q}/2) \prod_i \int Dz \prod_a 2\cosh(\sqrt{\hat{q}} z) \\
&= \exp(-nN\hat{q}/2) \prod_i \int Dz (2\cosh(\sqrt{\hat{q}} z))^n \\
&= \left\{\exp(-n\hat{q}/2) \int Dz (2\cosh(\sqrt{\hat{q}} z))^n\right\}^N .
\end{aligned} \tag{13.27}$$

Then, we start to compute the energy term. Notice that $\{u_\mu^a\}$ are independent for different patterns. It then follows that

$$\begin{aligned}\text{Energy term} &\equiv \left\langle \prod_{a=1}^{n}\prod_{\mu=1}^{P}\Theta(u_\mu^a)\right\rangle_{\{u_\mu^a\}} \\ &= \prod_{\mu=1}^{P}\left\langle \prod_{a=1}^{n}\Theta(u_\mu^a)\right\rangle_{\{u_\mu^a\}} \\ &= \left[\left\langle \prod_{a=1}^{n}\Theta(u^a)\right\rangle_{\{u^a\}}\right]^P .\end{aligned} \tag{13.28}$$

Under the RS ansätz, the mean and covariance of $u^a$ is given by $\langle u^a\rangle = 0$; $\langle u^a u^b\rangle = \delta_{ab} + (1-\delta_{ab})q^{ab}$. According to the CLT, $u$ obeys a multivariate Gaussian distribution subject to their covariance structure constraints. Let $u^a = Ax^a + Bz$, where $x^a$ and $z$ are mutually independent standard Gaussian random variables. The variance is then given by $\langle u^a u^b\rangle = A^2\langle x^a x^b\rangle + B^2\langle z^2\rangle = \delta_{ab} + (1-\delta_{ab})q^{ab}$. To satisfy the covariance constraint, we have $B^2 + A^2 = 1$, $B^2 = q$, resulting in $u^a = \sqrt{1-q}x^a + \sqrt{q}z$. Then, the energy term can be written in the form of a probability distribution integral

$$\begin{aligned}\text{Energy term} &= \left[\int Dz \prod_{a=1}^{n}\int Dx^a \Theta(\sqrt{1-q}x^a + \sqrt{q}z)\right]^P \\ &= \left\{\left[\int Dz H\left(-\sqrt{\frac{q}{1-q}}z\right)\right]^n\right\}^P .\end{aligned} \tag{13.29}$$

Taken together, we have the final result of $\langle Z^n\rangle$

$$\begin{aligned}\langle Z^n\rangle = \int \frac{dqd\hat{q}}{2\pi \mathrm{i}} \exp\left[-N\frac{n(n-1)}{2}\hat{q}q - \frac{Nn}{2}\hat{q} + N\ln\left(\int Dz[2\cosh(\sqrt{\hat{q}}z)]^n\right)\right] \\ \times \exp\left[N\alpha \ln\int Dz\left(H(-\sqrt{\frac{q}{1-q}}z)\right)^n\right],\end{aligned} \tag{13.30}$$

where $\alpha = \frac{P}{N}$. We define $F(q,\hat{q},n) = -\frac{n(n-1)}{2}\hat{q}q - \frac{n}{2}\hat{q} + \ln(\int Dz[2\cosh(\sqrt{\hat{q}}z)]^n) + \alpha\ln\int Dz[H(-\sqrt{\frac{q}{1-q}}z)]^n$. Then, we get the free energy of the perceptron under the replica symmetry ansätz as follows:

$$-\beta f_{\mathrm{RS}} = \lim_{n\to 0}\frac{\ln\langle Z^n\rangle}{Nn} = \lim_{n\to 0}\frac{\ln\int\frac{dqd\hat{q}}{2\pi\mathrm{i}}e^{NF}}{Nn}. \tag{13.31}$$

To get around a high-dimensional integral, we apply the Laplace approximation

$$
\begin{aligned}
-\beta f_{\mathrm{RS}} &= \lim_{n\to 0;N\to\infty} \frac{\ln\langle Z^n\rangle}{Nn} \simeq \lim_{n\to 0} \frac{F_{max}}{n} \\
&= \frac{1}{2}\hat{q}q - \frac{1}{2}\hat{q} + \lim_{n\to 0} \frac{\ln\left(\int Dz[2\cosh(\sqrt{\hat{q}}z)]^n\right)}{n} + \lim_{n\to 0} \frac{\alpha \ln \int Dz[H(-\sqrt{\frac{q}{1-q}}z)]^n}{n}.
\end{aligned}
\tag{13.32}
$$

Using L'Hospital's rule for computing limits, we have

$$
-\beta f_{\mathrm{RS}} = \frac{1}{2}\hat{q}q - \frac{1}{2}\hat{q} + \int Dz \ln[2\cosh(\sqrt{\hat{q}}z)] + \alpha \int Dz \ln[H(-\sqrt{\frac{q}{1-q}}z)].
\tag{13.33}
$$

Using $\int DzF(z)z = \int DzF'(z)$, $\tanh'(x) = 1 - \tanh^2(x)$, and $H''(y) = -yH'(y)$, we then derive the saddle-point equations as follows:

$$
\begin{aligned}
\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial\hat{q}} &= \frac{1}{2}(q-1) + \int Dz \frac{z}{2\sqrt{\hat{q}}} \tanh(\sqrt{\hat{q}}z) \\
&= \frac{1}{2}q - \frac{1}{2}\int Dz \tanh^2(\sqrt{\hat{q}}z) \\
&= 0, \\
\frac{\partial(-\beta f_{\mathrm{RS}})}{\partial q} &= \frac{1}{2}\hat{q} - \frac{\alpha}{2\sqrt{q(1-q)^3}} \int Dz \frac{zH'(-\sqrt{\frac{q}{1-q}}z)}{H(-\sqrt{\frac{q}{1-q}}z)} \\
&= \frac{1}{2}\hat{q} - \frac{\alpha}{2\sqrt{q(1-q)^3}} \int Dz \left(\frac{H'(-\sqrt{\frac{q}{1-q}}z)}{H(-\sqrt{\frac{q}{1-q}}z)}\right)' \\
&= \frac{1}{2}\hat{q} - \frac{\alpha}{2(1-q)} \int Dz \left(\frac{H'(-\sqrt{\frac{q}{1-q}}z)}{H(-\sqrt{\frac{q}{1-q}}z)}\right)^2 \\
&= 0,
\end{aligned}
\tag{13.34}
$$

which leads to the final saddle-point equations of the Perceptron model,

$$
\begin{aligned}
q &= \int Dz \tanh^2(\sqrt{\hat{q}}z), \\
\hat{q} &= \frac{\alpha}{1-q} \int Dz \left(\frac{H'(-\sqrt{\frac{q}{1-q}}z)}{H(-\sqrt{\frac{q}{1-q}}z)}\right)^2.
\end{aligned}
\tag{13.35}
$$

As shown in Fig. 13.3, the saddle-point equation solution is not physical anymore once $\alpha > \alpha_c \simeq 0.833$ [8], because a negative entropy is impossible for a system of

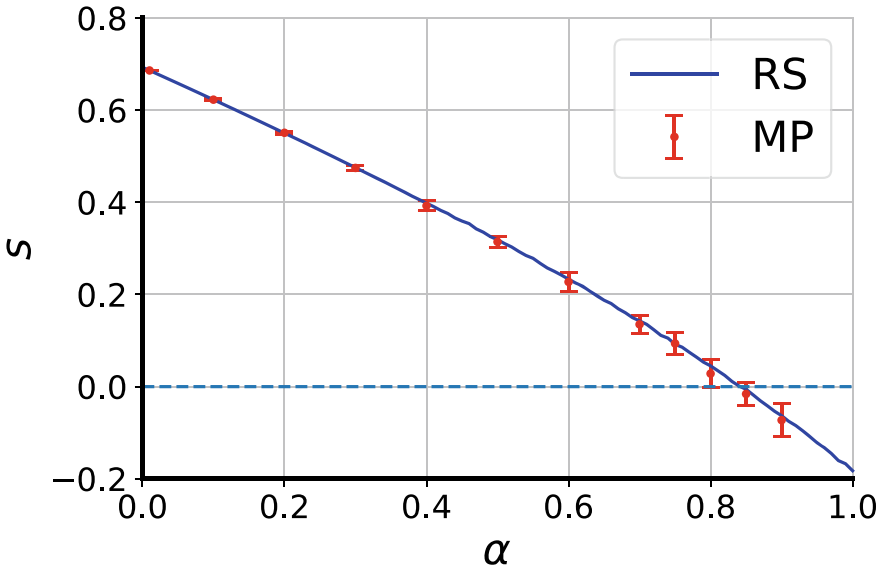



**Fig. 13.3** Entropy versus pattern density. Results are computed by replica symmetry theory, compared with the results obtained on single instances of the learning problem by running MP (indicated by the symbol). 20 random instances of network size $N = 1000$ are considered

discrete state-variables. One can further check the AT stability condition,[1] showing that $\alpha_{\mathrm{AT}} \simeq 1.015$ [6, 14]. Therefore, the RS solution is still stable in the negative-entropy regime. To gain deeper insights, we have to consider the replica symmetry breaking effect in the next subsection.

### *13.3.2 Replica Symmetry Breaking*

We consider the one-step replica symmetry breaking (1RSB) ansätz, i.e., the form of the overlap matrix $Q$ is assumed to have the following shape

$$Q = \begin{pmatrix} 1 & q_1 & q_0 & q_0 & q_0 & q_0 \\ q_1 & 1 & q_0 & q_0 & q_0 & q_0 \\ q_0 & q_0 & 1 & q_1 & q_0 & q_0 \\ q_0 & q_0 & q_1 & 1 & q_0 & q_0 \\ q_0 & q_0 & q_0 & q_0 & 1 & q_1 \\ q_0 & q_0 & q_0 & q_0 & q_1 & 1 \end{pmatrix},$$

[1] It is interesting to show that the microscopic instability condition around the fixed point of the MP algorithm is identical to the instability for breaking the RS in equilibrium, which is left as an exercise for interested readers.

where we assume $n = 6,\ m = 2$ for an example, and the matrix is divided into $n/m \times n/m$ small blocks in general, and $m$ is the width of each small block. Diagonal blocks have elements $q_0$ and off-diagonal ones have elements $q_1$. All diagonal elements take 1 by definition. The physical meaning of $q_1$ is the average overlap between replicas in the same state, and $q_0$ is the average overlap of two replicas from different states. Consequently, we have $q_0 < q_1$. Under the 1RSB ansätz, we have

$$\langle Z^n \rangle = \int \prod_{a<b} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi} \underbrace{\overbrace{\exp\left(N(-\sum_{a<b} q^{ab}\hat{q}^{ab})\right)}^{\text{A}} \overbrace{\sum_{\{\mathbf{J}^a\}} \exp\left(\sum_{a<b} \hat{q}^{ab} J^a J^b\right)}^{\text{B}}}_{\text{Entropy term}} \times \underbrace{\left\langle \prod_a^n \prod_{\mu=1}^P \Theta(u_\mu^a) \right\rangle_{\{u_\mu^a\}}}_{\text{Energy term}}. \tag{13.36}$$

The part A is then computed as

$$\sum_{a<b} q^{ab}\hat{q}^{ab} = \frac{n}{m}\frac{m(m-1)}{2} q_1 \hat{q}_1 + \frac{n(n-m)}{2} q_0 \hat{q}_0. \tag{13.37}$$

The summation over $a < b$ inside the part B is then calculated as

$$\begin{aligned}
&\sum_{a<b} \hat{q}^{ab} J^a J^b = \sum_i \sum_{a<b} \hat{q}^{ab} J_i^a J_i^b \\
&= \sum_i \sum_{a<b} \hat{q}_0 J_i^a J_i^b + \sum_i \sum_{a<b} (\hat{q}^{ab} - \hat{q}_0) J_i^a J_i^b \\
&= \sum_i \sum_{a<b} \hat{q}_0 J_i^a J_i^b + \sum_i \sum_c^{n/m} \sum_{a,b\in c: a<b} (\hat{q}_1 - \hat{q}_0) J_i^a J_i^b \\
&= \frac{\hat{q}_0}{2} \sum_i \left( \left( \sum_{a=1}^n J_i^a \right)^2 - n \right) + \frac{\hat{q}_1 - \hat{q}_0}{2} \sum_i \left( \sum_c^{n/m} \left( \sum_{a\in c} J_i^a \right)^2 - n \right).
\end{aligned} \tag{13.38}$$

Then, the part B can be explicitly calculated out as follows:

$$
\begin{aligned}
&\sum_{\{\mathbf{J}^a\}} \exp\left(\sum_{a<b} \hat{q}^{ab} J^a J^b\right) = \sum_{\{\mathbf{J}^a\}} \prod_i \exp\left(\frac{\hat{q}_0}{2}\left(\sum_{a=1}^n J_i^a\right)^2 + \frac{\hat{q}_1 - \hat{q}_0}{2} \sum_c^{n/m} \left(\sum_{a\in c} J_i^a\right)^2 - \frac{\hat{q}_1}{2} n\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \sum_{\{\mathbf{J}^a\}} \prod_i \exp\left(\frac{\hat{q}_0}{2}\left(\sum_{a=1}^n J_i^a\right)^2 + \frac{\hat{q}_1 - \hat{q}_0}{2} \sum_c^{n/m} \left(\sum_{a\in c} J_i^a\right)^2\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \sum_{\{J_i^a\}} \exp\left(\frac{\hat{q}_0}{2}\left(\sum_{a=1}^n J_i^a\right)^2 + \frac{\hat{q}_1 - \hat{q}_0}{2} \sum_c^{n/m} \left(\sum_{a\in c} J_i^a\right)^2\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \sum_{\{J_i^a\}} \exp\left(\frac{\hat{q}_0}{2}\left(\sum_{a=1}^n J_i^a\right)^2\right) \prod_c^{n/m} \exp\left(\frac{\hat{q}_1 - \hat{q}_0}{2} \left(\sum_{a\in c} J_i^a\right)^2\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \sum_{\{J_i^a\}} \int Dz_0 \exp\left(\sqrt{\hat{q}_0} z_0 \sum_{a=1}^n J_i^a\right) \prod_c^{n/m} \int Dz_1 \exp\left(\sqrt{\hat{q}_1 - \hat{q}_0} \sum_{a\in c} J_i^a z_1\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \sum_{\{J_i^a\}} \int Dz_0 \exp\left(\sqrt{\hat{q}_0} \left(\sum_c^{n/m} \sum_{a\in c} J_i^a\right) z_0\right) \prod_c^{n/m} \int Dz_1 \exp\left(\sqrt{\hat{q}_1 - \hat{q}_0} \sum_{a\in c} J_i^a z_1\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \sum_{\{J_i^a\}} \int Dz_0 \prod_c^{n/m} \int Dz_1 \prod_{a\in c} \exp\left(\sqrt{\hat{q}_0} J_i^a z_0 + \sqrt{\hat{q}_1 - \hat{q}_0} J_i^a z_1\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \int Dz_0 \prod_c^{n/m} \int Dz_1 \sum_{\{J_i^a\}} \prod_{a\in c} \exp\left(\sqrt{\hat{q}_0} J_i^a z_0 + \sqrt{\hat{q}_1 - \hat{q}_0} J_i^a z_1\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \int Dz_0 \prod_c^{n/m} \int Dz_1 \prod_{a\in c} \sum_{J_i^a} \exp\left(\sqrt{\hat{q}_0} J_i^a z_0 + \sqrt{\hat{q}_1 - \hat{q}_0} J_i^a z_1\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \int Dz_0 \prod_c^{n/m} \int Dz_1 \prod_{a\in c} 2\cosh\left(\sqrt{\hat{q}_0} z_0 + \sqrt{\hat{q}_1 - \hat{q}_0} z_1\right) \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \prod_i \int Dz_0 \prod_c^{n/m} \int Dz_1 \left(2\cosh\left(\sqrt{\hat{q}_0} z_0 + \sqrt{\hat{q}_1 - \hat{q}_0} z_1\right)\right)^m \\
&= \exp\left(-\frac{nN\hat{q}_1}{2}\right) \left\{\int Dz_0 \left[\int Dz_1 \left(2\cosh\left(\sqrt{\hat{q}_0} z_0 + \sqrt{\hat{q}_1 - \hat{q}_0} z_1\right)\right)^m\right]^{n/m}\right\}^N .
\end{aligned}
\tag{13.39}
$$

Next, we are going to compute the energy term

$$
\begin{aligned}
&\left\langle \prod_{a=1}^n \prod_{\mu=1}^P \Theta(u_\mu^a) \right\rangle_{\{u_\mu^a\}} \\
&= \prod_{\mu=1}^P \left\langle \prod_{a=1}^n \Theta(u_\mu^a) \right\rangle_{\{u_\mu^a\}} \\
&= \prod_{\mu=1}^P \left\langle \prod_{a=1}^n \Theta(u^a) \right\rangle_{\{u^a\}} .
\end{aligned}
\tag{13.40}
$$

The mean and variance of $u^a$ are specified by

$$\begin{aligned}&\langle u^a\rangle = 0,\\ &\langle u^a u^b\rangle = \begin{cases}1, & a=b;\\ q_1, & |b-a|<m \text{ and } a\neq b;\\ q_0, & \text{otherwise.}\end{cases}\end{aligned} \tag{13.41}$$

To obey this hierarchical statistical structure, we suppose $u^a = AX + BY + CZ$, $X, Y, Z$ are independent standard Gaussian random variables. We then have to solve the following constraint equations

$$\begin{cases}A^2+B^2+C^2=1, & \forall a=b;\\ B^2+C^2=q_1, & |b-a|<m \text{ and } a\neq b;\\ C^2=q_0.\end{cases} \tag{13.42}$$

Therefore, $X$ gets a superscript $a$, and $Y$ gets a superscript $c$, where c is the index of the small block that $a$ belongs to. More precisely, $u^a = \sqrt{1-q_1}X^a + \sqrt{q_1-q_0}Y^c + \sqrt{q_0}Z$. Thus, the energy term can be written in the form of probability distribution integrals as follows:

$$\begin{aligned}&\prod_{\mu=1}^{P}\left\langle\prod_{a=1}^{n}\Theta(u^a)\right\rangle_{\{u^a\}}\\ &=\prod_{\mu=1}^{P}\int DZ\prod_{c}^{n/m}\int DY^c\prod_{a\in c}\int DX^a\Theta(\sqrt{1-q_1}X^a+\sqrt{q_1-q_0}Y^c+\sqrt{q_0}Z)\\ &=\prod_{\mu=1}^{P}\int DZ\prod_{c}^{n/m}\int DY^c\prod_{a\in c}H\left(-\frac{\sqrt{q_1-q_0}Y^c+\sqrt{q_0}Z}{\sqrt{1-q_1}}\right)\\ &=\prod_{\mu=1}^{P}\int DZ\prod_{c}^{n/m}\int DY^c\left[H\left(-\frac{\sqrt{q_1-q_0}Y^c+\sqrt{q_0}Z}{\sqrt{1-q_1}}\right)\right]^m\\ &=\prod_{\mu=1}^{P}\int DZ\left\{\int DY\left[H\left(-\frac{\sqrt{q_1-q_0}Y+\sqrt{q_0}Z}{\sqrt{1-q_1}}\right)\right]^m\right\}^{n/m}\\ &=\left\{\int DZ\left\{\int DY\left[H\left(-\frac{\sqrt{q_1-q_0}Y+\sqrt{q_0}Z}{\sqrt{1-q_1}}\right)\right]^m\right\}^{n/m}\right\}^{P}.\end{aligned} \tag{13.43}$$

By applying the Laplace approximation, we finally arrive at

$$\frac{\ln\langle Z^n\rangle}{N} = -\frac{n(m-1)}{2}q_1\hat{q}_1 - \frac{n(n-m)}{2}q_0\hat{q}_0 - \frac{n\hat{q}_1}{2} + \ln\left(\int Dz_0\left[\int Dz_1\left(2\cosh\left(\sqrt{\hat{q}_0}z_0 + \sqrt{\hat{q}_1-\hat{q}_0}z_1\right)\right)^m\right]^{n/m}\right) + \alpha\ln\left(\int DZ\left\{\int DY\left[H\left(-\frac{\sqrt{q_1-q_0}Y+\sqrt{q_0}Z}{\sqrt{1-q_1}}\right)\right]^m\right\}^{n/m}\right). \tag{13.44}$$

Taking the limit: $n \to 0$, we get the 1RSB free energy

$$-\beta f_{\text{1RSB}} = \lim_{n\to 0}\frac{\ln\langle Z^n\rangle}{Nn} = \frac{1-m}{2}q_1\hat{q}_1 + \frac{m}{2}q_0\hat{q}_0 - \frac{\hat{q}_1}{2} + \frac{1}{m}\int Dz_0\ln\left\{\int Dz_1\left[2\cosh(\sqrt{\hat{q}_0}z_0 + \sqrt{\hat{q}_1-\hat{q}_0}z_1)\right]^m\right\} + \frac{\alpha}{m}\int DZ\ln\left\{\int DY\left[H\left(-\frac{\sqrt{q_1-q_0}Y+\sqrt{q_0}Z}{\sqrt{1-q_1}}\right)\right]^m\right\}, \tag{13.45}$$

where $m \in [0, 1]$ due to $n \to 0$. In fact, $m$ is called the Parisi parameter for the 1RSB analysis [15]. Saddle-point equations are derived by requiring that

$$\frac{\partial[-\beta f]}{\partial q_0} = \frac{\partial[-\beta f]}{\partial \hat{q}_0} = \frac{\partial[-\beta f]}{\partial q_1} = \frac{\partial[-\beta f]}{\partial \hat{q}_1} = \frac{\partial[-\beta f]}{\partial m} = 0, \tag{13.46}$$

where $f$ indicates the 1RSB approximate value of the true free energy.

The transition from RS to RSB takes place at the zero-entropy line: $S_{\text{RS}}(\alpha, T) = 0$, where we introduce the temperature parameter [see Eq. (13.3)]. This is also called the frozen-RSB solution, widely existing in a broad class of constraint satisfaction problems [16]. The transition is of the first-order type, in the sense that $q_1 = 1$ becomes a RSB solution at the transition point [8], which also suggests that $\hat{q}_1 \to \infty$. This solution implies further that

$$F_{\text{1RSB}}(q_0, \hat{q}_0, 1, \infty, \beta, m) = \frac{1}{m}F_{\text{RS}}(q_0, m^2\hat{q}_0, \beta m), \tag{13.47}$$

where $F \equiv -\beta f$, $m = \beta_c/\beta$ and $\beta_c$ are determined by the zero entropy condition. Moreover, the stationary requirement of the 1RSB free energy w.r.t $m$ reduces to the zero-entropy condition. The free energy is equal to the RS one at $\beta_c$, independent of the temperature when $T < T_c$, like that in the random energy model. Then, the distribution of the order parameter $q$ is specified by [8]

$$P(q) = m\delta(q - q_0) + (1 - m)\delta(q - 1), \tag{13.48}$$

where $m$ is now interpreted as the probability (see also Chap. 9).

## 13.4 Further Theory Development

To find a solution for the Ising perceptron is typically very hard [10, 17, 18]; whereas, a reinforced message passing algorithm was proposed [1], and is able to solve the binary perceptron problem up to a pattern density $\alpha \sim 0.7$. These two facts seem to conflict with each other. This puzzle was first explained by Huang and Kabashima, who adapted the Franz–Parisi framework, originally proposed to study spherical spin glass models [19], to the neural network learning problems. In this work, they demonstrated the origin of the computation hardness of the Ising perceptron problem, by a theory-grounded picture about the weight space, i.e., isolated solutions emerge in the entire finite $\alpha$ regime, and the typical distance separating any two solutions grows rapidly with $\alpha$ [2].

The basic idea is to first choose an equilibrium configuration $\mathbf{J}$ at a temperature $T'$, then constrain its overlap with another equilibrium configuration $\mathbf{w}$ at a different temperature $T$, which results in a constrained free energy [19]

$$F(T, T', x) = \left\langle \frac{1}{Z(T')} \sum_{\mathbf{J}} e^{-\beta' E(\mathbf{J})} \ln \sum_{\mathbf{w}} e^{-\beta E(\mathbf{w}) + x \mathbf{J}\cdot\mathbf{w}} \right\rangle, \tag{13.49}$$

after taking the quenched-disorder average (over the pattern distribution $\boldsymbol{\xi}$, denoted by the angular bracket) and the average over the distribution of $\mathbf{J}$, which is $e^{-\beta' E(\mathbf{J})}/Z(T')$. $Z(T')$ is the partition function for the original Boltzmann measure, and $\beta$ (or $\beta'$) denotes the inverse temperature.

A ground-state focus leads to the following replica representation of the framework

$$F(x) = \lim_{\substack{n\to 0 \\ m\to 0}} \frac{\partial}{\partial m} \left\langle \sum_{\{\mathbf{J}^a, \mathbf{w}^\gamma\}} \prod_{\mu} \left[ \prod_{a,\gamma} \Theta(u_a^\mu)\Theta(v_\gamma^\mu) \right] e^{x \sum_{\gamma,i} J_i^1 w_i^\gamma} \right\rangle, \tag{13.50}$$

where $u_a^\mu \equiv \sum_i J_i^a \xi_i^\mu/\sqrt{N}$ and $v_\gamma^\mu \equiv \sum_i w_i^\gamma \xi_i^\mu/\sqrt{N}$. Detailed replica calculation is given in the original paper [2]. The Franz–Parisi potential $\mathcal{V}(p)$ is obtained through a Legendre transform of $f(x) = \lim_{N\to\infty} F(x)/N$, i.e., $\mathcal{V}(p) = f(x) - xp$ and $\frac{\mathrm{d}f(x)}{\mathrm{d}x} = p$. $\mathcal{V}(p)$ has the meaning of the entropy characterizing the growth rate of the number of solutions ($e^{N\mathcal{V}(p)}$) lying apart at a normalized distance $(1-p)/2$ (Hamming distance divided by $N$) from the fixed equilibrium solution.

At the point $p \to 1$ ($\epsilon \equiv 1 - p \to 0$), we have $\frac{\mathrm{d}\mathcal{V}(p)}{\mathrm{d}p} = \alpha C_p \epsilon^{-1/2} + (\ln \epsilon)/2 + C$ [2] where $C$ is a finite constant and $C_p$ is a positive constant. The first term dominates the divergent behavior in the limit $\epsilon \to 0$. This means that, for any finite $\alpha > 0$, the entropy curve has a negative infinite slope ($\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}d} = -2\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}p}$) at $p = 1$, supporting the existence of the convex part in the entropy curve, thereby confirming the point-like clusters in the Ising perceptron problem.

On the other hand, the isolated solution is not accessible by the reinforced message passing algorithm. This heuristic strategy, working by progressively enhancing or

weakening the local field ($h_i = \sum_b u_{b\to i}$) each synapse feels by an increasing probability as a function of iteration steps, may search for sub-dominant dense regions of the weight space. This hypothesis was proposed as a large-deviation analysis [3, 20]

$$\mathcal{F}(d, y) = -\frac{1}{Ny} \ln \left( \sum_{\tilde{\mathbf{w}}} \mathbb{I}_{\boldsymbol{\xi}}(\tilde{\mathbf{w}}) \mathcal{N}(\tilde{\mathbf{w}}, d)^y \right), \tag{13.51}$$

where $\mathbb{I}$ constrains $\tilde{\mathbf{w}}$ to be a solution, $\mathcal{N}(\tilde{\mathbf{w}}, d) = \sum_{\boldsymbol{w}} \mathbb{I}_{\boldsymbol{\xi}}(\boldsymbol{w}) \delta(\mathbf{w}\cdot\tilde{\mathbf{w}}, N(1-2d))$. Then, the local entropy $\mathcal{S}_L(d, y) = \frac{1}{N}\langle \ln \mathcal{N}(\tilde{\mathbf{w}}, d)\rangle_{\boldsymbol{\xi},\tilde{\mathbf{w}}}$ can be obtained by the above generating function as $\mathcal{S}_L = \partial_y(y\mathcal{F}(d, y))$. In this new measure, individual solutions are favored provided that they are surrounded by a large number of other solutions. These solutions are not necessary to be equilibrium (e.g., isolated ones). The sub-dominant dense region is then characterized by a nonzero $\mathcal{S}_L(d, y) > 0$ around $d = 0$. In fact, the previous work considering the solution-pair entropy landscape falls into the case of $y = 1$ [21], while the frozen picture falls into the case of $y \to 0$ [2].

This large-deviation analysis inspires new entropy-driven algorithms [20, 22], suggesting that heuristic learning algorithms are biased towards a flat region in the high-dimensional weight landscape. In particular, these flat regions have better generalization performances compared to those narrow regions. However, how to measure the flatness of the weight space is still under heated debate [23, 24].

Although this chapter is restricted to the single-layer Perceptron model, the same statistical mechanics techniques can be applied to multi-layer models with specific topology of the architectures. Interested readers can go through several papers related to multi-layered toy models [25–27]. In the next chapter, we shall explore the statistical mechanics analysis of an arbitrary topology of multi-layered networks.

## References


1. A. Braunstein, R. Zecchina, Phys. Rev. Lett. **96**, 030201 (2006)
2. H. Huang, Y. Kabashima, Phys. Rev. E **90**, 052813 (2014)
3. C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, Phys. Rev. Lett. **115**(12), 128101 (2015)
4. F. Rosenblatt, Psychol. Rev. **65**, 386 (1958)
5. E. Gardner, Europhys. Lett. (epl) **4**, 481 (1987)
6. E. Gardner, J. Phys. A: Math. Gen. **21**, 257 (1988)
7. E. Gardner, B. Derrida, J. Phys. A **22**(12), 1983 (1989)
8. W. Krauth, M. Mézard, J. Phys. **50**(20), 3057 (1989)
9. A.L. Blum, R.L. Rivest, Neural Netw. **5**(1), 117 (1992)
10. H. Horner, Z. Phys. B **86**(2), 291 (1992)
11. A. Engel, C.V. den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001)
12. G. Gyorgyi, Phys. Rev. A **41**(12), 7097 (1990)
13. H. Sompolinsky, N. Tishby, H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990)
14. T. Uezu, K. Nokura, Prog. Theor. Phys. **95**(2), 273 (1996)

15. M. Mézard, G. Parisi, M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987)
16. M. Mézard, A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009)
17. H. Huang, H. Zhou, J. Stat. Mech.: Theory Exper. **2010**(8), 8014 (2010)
18. H. Huang, H. Zhou, EPL **96**(5), 58003 (2011)
19. S. Franz, G. Parisi, J. Phys. I **5**(11), 1401 (1995)
20. C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, J. Stat. Mech.: Theory Exper. **2016**(2), 23301 (2016)
21. H. Huang, K.Y.M. Wong, Y. Kabashima, J. Phys. A: Math. Theor. **46**, 375002 (2013)
22. P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, R. Zecchina, J. Stat. Mech.: Theory Exper. **2019**(12), 124018 (2019)
23. N.S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P.T.P. Tang, in *ICLR* (2017). arXiv:1609.04836
24. L. Dinh, R. Pascanu, S. Bengio, Y. Bengio, in *ICML* (2017). arXiv:1703.04933
25. E. Barkai, D. Hansel, I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990)
26. E. Barkai, I. Kanter, Europhys. Lett. (EPL) **14**(2), 107 (1991)
27. E. Barkai, D. Hansel, H. Sompolinsky, Phys. Rev. A **45**(6), 4146 (1992)

# Chapter 14
# Mean-Field Model of Multi-layered Perceptron


Deep learning has already become a powerful tool in the areas such as image classification and speech recognition [1, 2]. Deep learning with many layers has been proved to be a universal approximator [3]. However, compared with its achievement, the mechanism of deep networks is still challenging to understand. Redundancy is one of the characteristics of deep neural networks, which means that the deep network is robust under the removal perturbation of connections between layers [4]. In other words, the generalization ability of deep network does not significantly change until a large number of connections (a threshold) between layers are removed. In this chapter, we introduce a random active path (RAP) model on a multilayer perceptron network to study the redundancy property [5]. In the RAP model, the paths are randomly activated, and the weights along the paths are constrained by the corresponding inputs, and therefore a p-weight glass model is naturally introduced. By applying mean-field methods, we analyze the statistical properties of the model under the removal perturbation of connections between layers to different extents. A critical value of the perturbation is revealed, separating a paramagnetic phase from the spin glass phase, where the paramagnetic phase shows a poor generalization ability, which is a non-robust regime of the deep networks. The RAP model still relies on assumptions amenable for a theoretical analysis, which should inspire future refinement. In this chapter, we also introduce mean-field training algorithms for multilayer perceptrons with discrete weights, and moreover, the ensemble backpropagation to understand the credit assignment problem in deep neural networks.

## 14.1 Random Active Path Model

Redundancy is known to be one of the characteristics of deep neural networks. From the perspective of statistical physics, we try to study how the statistical property of a deep neural network changes with respect to the removal perturbation of synapses. To address this question qualitatively, we propose a random active path model on a multilayer perceptron network with binary synapses.



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_14

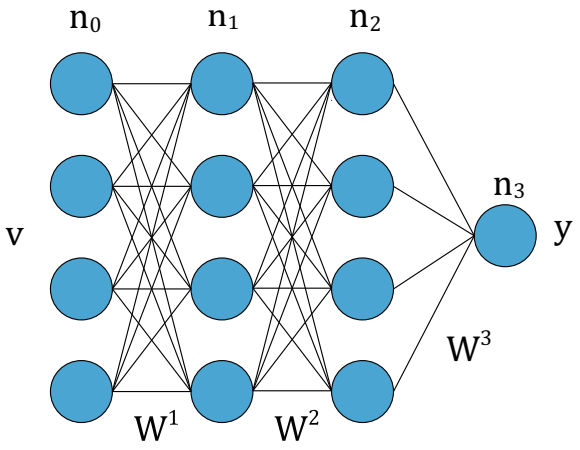



**Fig. 14.1** The structure of the multilayer perceptron network, considered in the RAP model, where a single output with an identity transfer function is considered

We consider a four-layered perceptron network (Fig. 14.1). Each layer has $n_l (l = 0, 1, 2, 3)$ units. The input is an $n_0$-dimensional vector $\boldsymbol{v}$ with binary($\pm 1$) element $v_i$, and the weight matrix $\boldsymbol{W}^l$ with binary($\pm 1$) element $w_{ij}^l$ specifies connections between layer $l$ and layer $l-1$. The non-linear function $f(\cdot)$ is chosen to be ReLU function, which is defined as $f(u_i) = \max(0, u_i)$. Finally, we can obtain the form of the output $y$ as

$$y = f_{L-1}(\boldsymbol{W}^{L-1} f_{L-2}(\boldsymbol{W}^{L-2} \cdots f_1(\boldsymbol{W}^1 \boldsymbol{v}))). \tag{14.1}$$

Note that $L = 4$ here. To propose the random active path model, we should define the active path first. An active path refers to the path from one input unit to the output unit and finally contributes to the output value. Thus, an active path must meet two demands: First, all units along the path are activated (the activation values of the units are positive) because of the ReLU activation function; Second, each connection on the path is present, while each synapse is deleted with a dilution probability. Therefore, the form of the output can be re-expressed as

$$y = \sum_{a=1}^{\Psi} v^a \prod_{k=1}^{L-1} \boldsymbol{W}_a^k, \tag{14.2}$$

where $\Psi$ denotes the total number of the active path, $v^a$ denotes the input node in the $a$th path and $W_a^k$ denotes the entry of $\boldsymbol{W}^k$ that is present in the $a$th path.

In addition to the dilution, whether a path is active depends also on the data-driven layered representations for a multilayer perceptron network performing real-world tasks. However, for simplicity, we assume that the activation of each path is independent of the input in our model, where the units of each layer are activated independently with a layer-dependent unit activation probability $\xi_l$, which can be

empirically estimated from a practical training. The Hamiltonian of the model can thus be written as

$$H(\boldsymbol{W}) = -\sum_{a=1}^{n^3} A_a v^a \prod_{i\in\partial a} W_i, \tag{14.3}$$

where $n$ denotes the number of the units at each layer except the last one. $A_a$ is a binary value indicating activation ($A_a = 1$) or silence ($A_a = 0$). The probability of a path activation is $P(A_a = 1) = \prod_l \xi_l(1 - p_l)$, where $\xi_l$ is the unit activation probability and $p_l$ is the dilution probability.

In statistical physics, an equilibrium system always has a relatively low energy (Hamiltonian in our RAP model), while in the deep learning, a practical network always has a relative low training loss. Thus, to build an intuitive relationship between the Hamiltonian and the loss function used in training, we assume that the true labels are $Y_t = \pm\Lambda$ ($\Lambda > 0$), where $\Lambda$ is the maximal output of the network. Moreover, we assume that the true output $Y_t$ is a random variable such that $P(Y_t = \pm\Lambda) = \frac{1}{2}$, and the input $v^a$ is also a random variable such that $P(v^a = \pm 1) = \frac{1}{2}$. Hence, $\mathrm{sgn}(Y_t)$ can be absorbed into the input ($\mathrm{sgn}(Y_t)v^a \to v^a$), and the model is statistically invariant. Then, the Hamiltonian and the loss function can be written, respectively, as

$$H = -\mathrm{sgn}(Y_t)y, \tag{14.4a}$$

$$C = |Y_t - y|. \tag{14.4b}$$

For simplicity, we choose the absolute error loss. It is easy to verify that minimizing the loss function between the target and the actual output $y$ is equivalent to finding the minimal value of Hamiltonian in the RAP model.

### *14.1.1 Results from Cavity Method*

Given the form of the Hamiltonian, we apply cavity method in mean-field theory to approximately acquire the statistical properties of the RAP model. First, we consider that the weight configuration $\boldsymbol{W}$ follows a Boltzmann distribution $P(\boldsymbol{W}) = e^{-\beta H(\boldsymbol{W})}/Z$, where $\beta$ is the inverse temperature, and $Z$ is the partition function. Under the cavity approximation, we could derive a set of self-consistent equations which are called message passing equations:

$$\begin{aligned} m_{i\to a} &= \tanh\left(\sum_{b\in\partial i\setminus a} u_{b\to i}\right), \\ u_{b\to i} &= \tanh^{-1}\left(\tanh\beta v^b \prod_{j\in\partial b\setminus i} m_{j\to b}\right). \end{aligned} \tag{14.5}$$

Note that, the form of Eq. (14.5) is similar to the standard message passing equations in Chap. 2 except for two crucial differences. First, weights along the paths are constrained by the corresponding inputs, where $v^a = +1$ denotes a ferromagnetic interaction, and $v^a = -1$ denotes an anti-ferromagnetic interaction. Thus, a factor graph (Fig. 14.2) can be naturally constructed, where two types of nodes can be connected to the deep network function. Second, weights configuration $\boldsymbol{W}$ is a subset of total weights $\{\boldsymbol{W}^l\}$ unless all the paths are activated. By recursively solving these equations, the iteration will converge to a fixed point $\{m_{i\to a}, u_{b\to i}\}$, which corresponds to a local or global minimum of the Bethe free energy (see Chap. 3). Therefore, we can acquire the statistical properties of RAP model, including magnetization and entropy.

To characterize the potential phase transitions in the model, we further define an order parameter $Q = \frac{1}{N_w}\sum_i {m_i}^2$, where $N_w$ is the total number of weights in the model. $Q$ measures the responses of the network to the input data, and thus a high $Q$ refers to a biased inference of weights (indicating an effective learning process). As shown in Fig. 14.3, when the dilution probability of the second layer $p_1$ increases (more units are deleted), $Q$ decreases slowly at first and sharply drops at a threshold $p_1$, which is a critical value separating a paramagnetic regime with poor generalization performance from spin glass regime with good generalization performance. Entropy here represents the number of candidate weight configurations, which increases as the dilution probability $p_1$ increases, indicating that the deep network becomes less constrained, like in a paramagnetic phase. Entropy also displays a slight jump at the same critical value of $p_1$, which is a characteristic of the first-order phase transition. Overall, by applying the mean-field cavity method, we can reveal that increasing the magnitude of the removal perturbation (dilution probability) will trigger a first-order transition to an undesired paramagnetic regime, which has poor generalization performances as expected.

### *14.1.2 An Infinite Depth Analysis*

Since modern deep networks can have thousands of layers, we then ask whether an infinite depth limit exists in our current model, and whether the joint energy level distribution becomes factorized, and a frozen phase can be identified when the temperature is lowered down [6]. Applying similar methods to those in Chap. 7, we derive the form of the joint energy-level distribution of our model in the infinite depth limit. First, the Hamiltonian can be re-expressed as

$$H(\boldsymbol{W}) = -\sum_{a=1}^{n^p} \tilde{A}_a \prod_{i\in\partial a} W_i, \tag{14.6}$$

where $p = L - 1$, $\tilde{A}_a = A_a v^a$, and its probability distribution is re-defined as follows:

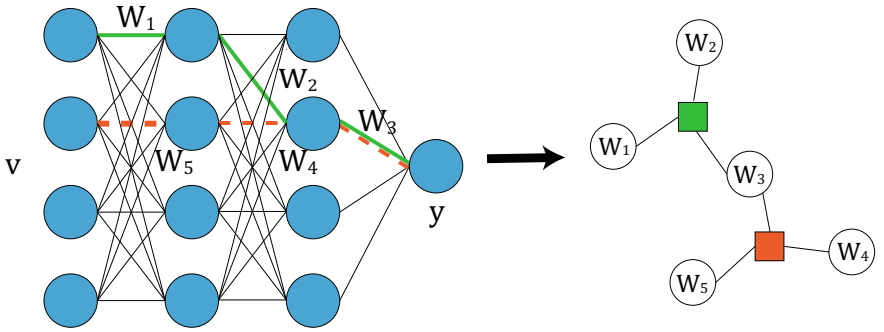



**Fig. 14.2** Schematic illustration of a factor graph of the RAP model. Left panel: a four layers perceptron, where two paths are activated. Right panel: a factor graph of the four-layered perceptron, where variable nodes (circle nodes) are weights to be estimated, constraint nodes (square nodes) are the active paths, and the lines between the variable nodes and the constraint nodes can be interpreted as connection strengths which are specified by the input in our current RAP model

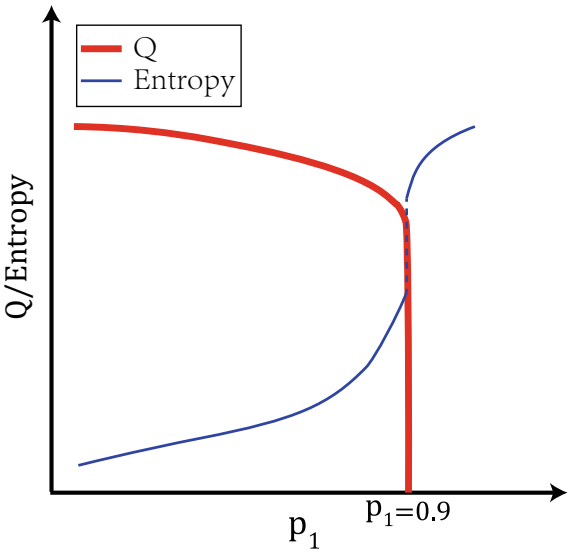



**Fig. 14.3** $Q$ and entropy versus $p_1$. $p_1$ denotes the dilution probability applied to weights between the first and second layers. In numerical stimulations, we set $\xi_1 = 0.5$, $\xi_2 = 0.1$ and $p_2 = 0$, $p_3 = 0$. By increasing $p_1$, $Q$ will sharply drop at $p_1 = 0.9$, which is a first-order transition, characterized by the entropy gap (dashed line) as well. This plot is a schematic one of that published in the recent work [5]

$$
\begin{aligned}
P(\tilde{A}_a = 0) &= 1 - \prod_l \xi_l (1 - p_l), \\
P(\tilde{A}_a = +1) &= \frac{1}{2} \prod_l \xi_l (1 - p_l), \\
P(\tilde{A}_a = -1) &= \frac{1}{2} \prod_l \xi_l (1 - p_l).
\end{aligned}
\tag{14.7}
$$

Then, we denote $P_0 = P(\tilde{A}_a = 0)$ for simplicity. The joint distribution of $\mathcal{N} = 2^{N_w}$ energy levels can be written as

$$\begin{aligned} P(E_1, E_2, \ldots, E_{\mathcal{N}}) &= \langle \delta(E_1 - H(\boldsymbol{W}^1)) \times \delta(E_2 - H(\boldsymbol{W}^2)) \times \cdots \times \delta(E_{\mathcal{N}} - H(\boldsymbol{W}^{\mathcal{N}})) \rangle_{\tilde{A}_a} \\ &= \prod_{\alpha=1}^{\mathcal{N}} \int \frac{d\hat{E}_\alpha}{2\pi} \, e^{\mathrm{i}\hat{E}_\alpha E_\alpha} \left\langle e^{-\mathrm{i}\hat{E}_\alpha H(\boldsymbol{W}^\alpha)} \right\rangle_{\tilde{A}_a} . \end{aligned} \tag{14.8}$$

We define $A = \prod_{\alpha=1}^{\mathcal{N}} \left\langle e^{\mathrm{i}\hat{E}_\alpha H(\boldsymbol{W}^\alpha)} \right\rangle_{\tilde{A}_a}$, then we have:

$$\begin{aligned} A &= \prod_{a=1}^{n^p} \left\langle \prod_{\alpha=1}^{\mathcal{N}} e^{\mathrm{i}\hat{E}_\alpha \tilde{A}_a \prod_{i\in\partial a} W_i^\alpha} \right\rangle_{\tilde{A}_a} \\ &= \prod_{a=1}^{n^p} \left[ P_0 + \left\langle \prod_{\alpha=1}^{\mathcal{N}} e^{\mathrm{i}\hat{E}_\alpha \tilde{A}_a \prod_{i\in\partial a} W_i^\alpha} \right\rangle_{\tilde{A}_a=\pm 1} \right]. \end{aligned} \tag{14.9}$$

According to the identity: $e^{a\sigma} = \cosh(a)[1 + \sigma \tanh(a)]$ (valid only for $\sigma = \pm 1$), we derive that

$$A = \prod_{a=1}^{n^p} \left[ P_0 + \left\langle \prod_{\alpha=1}^{\mathcal{N}} \cosh(\mathrm{i}\hat{E}_\alpha)[1 + \tilde{A}_a \prod_{i\in\partial a} W_i^\alpha \tanh(\mathrm{i}\hat{E}_\alpha)] \right\rangle_{\tilde{A}_a=\pm 1} \right]. \tag{14.10}$$

Now, we perform the gauge transformation: $\tilde{A}_a \to \tilde{A}_a \prod_{i\in\partial a} W_i^\gamma$, leading to

$$\begin{aligned} A = \prod_{a=1}^{n^p} \Bigg[ P_0 + \frac{1 - P_0}{2} \prod_{\alpha=1}^{\mathcal{N}} \cosh(\mathrm{i}\hat{E}_\alpha) \Big[ \prod_{\alpha=1}^{\mathcal{N}} (1 + \tanh(\mathrm{i}\hat{E}_\alpha) \prod_{i\in\partial a} W_i^\gamma W_i^\alpha) \\ + \prod_{\alpha=1}^{\mathcal{N}} (1 - \tanh(\mathrm{i}\hat{E}_\alpha) \prod_{i\in\partial a} W_i^\gamma W_i^\alpha) \Big] \Bigg]. \end{aligned} \tag{14.11}$$

Then, we replace $W_i^\gamma W_i^\alpha$ by its mean $\langle W_i^\gamma W_i^\alpha \rangle$ neglecting the thermal fluctuation, and we have immediately $\prod_{i\in\partial a} W_i^\gamma W_i^\alpha \simeq q^p$, where $q$ is clearly the overlap function between two configurations. We have further $\prod_{\alpha=1}^{\mathcal{N}} (1 + \tanh(\mathrm{i}\hat{E}_\alpha) q^p) \approx 1 + q^p \sum_{\alpha=1}^{\mathcal{N}} \tanh(\mathrm{i}\hat{E}_\alpha)$. Note that $q^p$ is a negligible term in a large-$p$ limit. We finally complete the calculation of the A part.

$$
\begin{aligned}
A &= \prod_{a=1}^{n^p}\Big[P_0 + \frac{1-P_0}{2}\prod_{\alpha=1}^{\mathcal{N}}\cosh(\mathrm{i}\hat{E}_\alpha)\Big[1+q^p\sum_{\alpha=1}^{\mathcal{N}}\tanh(\mathrm{i}\hat{E}_\alpha)+1-q^p\sum_{\alpha=1}^{\mathcal{N}}\tanh(\mathrm{i}\hat{E}_\alpha)\Big]\Big] \\
&= \prod_{a=1}^{n^p}\Big[P_0+(1-P_0)\prod_{\alpha=1}^{\mathcal{N}}\cosh(\mathrm{i}\hat{E}_\alpha)\Big], \\
&= \prod_{a=1}^{n^p}\Big[1-\prod_l \xi_l(1-p_l)+\prod_l \xi_l(1-p_l)\prod_{\alpha=1}^{\mathcal{N}}\cosh(\mathrm{i}\hat{E}_\alpha)\Big], \\
&= e^{\sum_{a=1}^{n^p}\ln\Big[1-\prod_l \xi_l(1-p_l)+\prod_l \xi_l(1-p_l)\prod_{\alpha=1}^{\mathcal{N}}\cosh(\mathrm{i}\hat{E}_\alpha)\Big]}, \\
&\approx e^{-n^p\Big[\prod_l \xi_l(1-p_l)(1-\prod_{\alpha=1}^{\mathcal{N}}\cosh(\mathrm{i}\hat{E}_\alpha))\Big]}.
\end{aligned}
\tag{14.12}
$$

To sum up, we can obtain the form of the joint distribution of the energy levels:

$$
P(E_1, E_2, \ldots, E_{\mathcal{N}}) = \int\prod_{\alpha=1}^{\mathcal{N}}\frac{d\hat{E}_\alpha}{2\pi}e^{\sum_{\alpha=1}^{\mathcal{N}}\mathrm{i}\hat{E}_\alpha E_\alpha}e^{n^p[\prod_l \xi_l(1-p_l)(\prod_{\alpha=1}^{\mathcal{N}}\cosh(\mathrm{i}\hat{E}_\alpha)-1)]}.
\tag{14.13}
$$

Let us finally discuss the small conjugated energy limit. $\prod_{\alpha=1}^{\mathcal{N}}\cosh(\mathrm{i}\hat{E}_\alpha) = \prod_{\alpha=1}^{\mathcal{N}}(1+\frac{(\mathrm{i}\hat{E}_\alpha)^2}{2}) \approx 1+\sum_{\alpha=1}^{\mathcal{N}}\frac{(\mathrm{i}\hat{E}_\alpha)^2}{2}$ where $\hat{E}_\alpha$ is a negligible term by the limit. Thus, we have the following result:

$$
\begin{aligned}
P(E_1, E_2, \ldots, E_{\mathcal{N}}) &= \int\prod_{\alpha=1}^{\mathcal{N}}\frac{d\hat{E}_\alpha}{2\pi}e^{\sum_{\alpha=1}^{\mathcal{N}}\mathrm{i}\hat{E}_\alpha E_\alpha - n^p\prod_l \xi_l(1-p_l)\sum_{\alpha=1}^{\mathcal{N}}\frac{(\hat{E}_\alpha)^2}{2}}, \\
&= \prod_{\alpha=1}^{\mathcal{N}}\frac{1}{\sqrt{2\pi n^p\prod_l \xi_l(1-p_l)}}e^{-\frac{E_\alpha^2}{2n^p\prod_l \xi_l(1-p_l)}}.
\end{aligned}
\tag{14.14}
$$

In the small conjugated energy limit, the energy levels become independent random variables. More precisely, single energy level follows a Gaussian distribution with a fluctuation of the order of $[n^p\prod_l \xi_l(1-p_l)]^{\frac{1}{2}}$ around zero, where $n^p\prod_l \xi_l(1-p_l)$ is exactly the number of active paths in the model. However, it is not correct to conclude that the energy levels for the RAP model are generally independent random variables, as we use the small conjugated energy limit, whose physics remains elusive. It is also not excluded that energy levels maybe organized into non-trivial structures in the infinite depth limit. Therefore, more systematic studies are required, including confirmation of the first-order transition in a practical training as well.

## 14.2 Mean-Field Training Algorithms

The mean-field method to train a deep supervised network with binary synapses was first introduced in the previous work [7]. In our current setting, each weight $w_{ij}^l$ is sampled from a Bernoulli distribution $P(w_{ij}^l)$ parametrized by an external field $\theta_{ij}^l$ as follows:

$$P(w_{ij}^l) = \sigma(\theta_{ij}^l)\delta_{w_{ij}^l,1} + [1-\sigma(\theta_{ij}^l)]\delta_{w_{ij}^l,-1}, \tag{14.15}$$

with mean $\mu_{ij}^l = 2\sigma(\theta_{ij}^l) - 1$ and variance $(\sigma_{ij}^l)^2 = -4\sigma^2(\theta_{ij}^l) + 4\sigma(\theta_{ij}^l)$. $\sigma(x)$ is a sigmoid function. According to the central limit theorem, the feedforward transformation can be re-parametrized as

$$z_j^l = m_j^l + v_j^l \cdot \epsilon_j^l, \tag{14.16a}$$

$$a_j^l = \frac{1}{\sqrt{N_{l-1}}}\,\mathrm{ReLU}(z_j^l), \tag{14.16b}$$

where $N_{l-1}$ is the number of neurons at the previous layer, $\epsilon_j$ is a standard Gaussian random variable, $m_j^l = \sum_i \mu_{ij}^l a_i^{l-1}$, and $v_j^l = \sqrt{\sum_i (\sigma_{ij}^l)^2 (a_i^{l-1})^2}$. We use the ReLU function $[\max(0, x)]$ here.

During the error backpropagation phase, we need to compute the gradient of the loss function $\mathcal{L}$ (e.g., cross-entropy for classification problems) with respect to the external field $\boldsymbol{\theta}$, which proceeds as follows:

$$\frac{\partial \mathcal{L}}{\partial \theta_{ij}^l} = \frac{\partial \mathcal{L}}{\partial z_j^l}\frac{\partial z_j^l}{\partial \theta_{ij}^l} = \frac{\partial \mathcal{L}}{\partial z_j^l}\left(\frac{\partial m_j^l}{\partial \theta_{ij}^l} + \epsilon_j^l \frac{\partial v_j^l}{\partial \theta_{ij}^l}\right). \tag{14.17}$$

We then define $\Delta_j^l \equiv \frac{\partial \mathcal{L}}{\partial z_j^l}$. On the top layer, $\Delta_j^l = y_j^L - \hat{y}_j^L$, where $y_j^L = \frac{e^{z_j^L}}{\sum_i e^{z_i^L}}$ is the softmax output, and $\hat{y}_j^L$ is the (one-hot) label of the input. On the lower layers, given $\Delta_k^{l+1}$, we can iteratively compute $\Delta_j^l$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial z_j^l} &= \sum_k \frac{\partial \mathcal{L}}{\partial z_k^{l+1}}\frac{\partial z_k^{l+1}}{\partial z_j^l} \\ &= \sum_k \Delta_k^{l+1} f'(z_j^l)\left(\mu_{jk}^{l+1} + \epsilon_k^{l+1}\frac{(\sigma_{jk}^{l+1})^2 a_j^l}{v_k^{l+1}}\right).\end{aligned} \tag{14.18}$$

Finally, we compute $\frac{\partial m_j^l}{\partial \theta_{ij}^l}$ and $\frac{\partial v_j^l}{\partial \theta_{ij}^l}$, respectively. It then proceeds as follows:

$$\frac{\partial m_j^l}{\partial \theta_{ij}^l} = \frac{\partial m_j^l}{\partial \mu_{ij}^l}\frac{\partial \mu_{ij}^l}{\partial \theta_{ij}^l} = 2a_i^{l-1}\sigma'(\theta_{ij}^l), \tag{14.19a}$$

$$\frac{\partial v_j^l}{\partial \theta_{ij}^l} = \frac{\partial v_j^l}{\partial (\sigma_{ij}^l)^2}\frac{\partial (\sigma_{ij}^l)^2}{\partial \theta_{ij}^l} = -2\frac{(a_i^{l-1})^2\mu_{ij}^l\sigma'(\theta_{ij}^l)}{v_j^l}. \tag{14.19b}$$

Note that $\boldsymbol{\epsilon}$ is sampled and quenched for both forward and backward computations in a single mini-batch gradient descent. After the learning is terminated, an effective network with binary weights can be constructed by sampling the Bernoulli distribution parametrized by external fields.

## 14.3 Spike and Slab Model

Deep learning has achieved impressive performance in a variety of scientific and industrial fields. Nevertheless, little has been known about the mechanism of the black box of deep neural networks, e.g., how much credit should be assigned to each network-parameter after learning. For a specific task, the backpropagation method has long been applied to train a feedforward neural network [8]. In the process of learning, the neural network is capable of coordinating a large number of parameters and makes an accurate decision at the output layer. The traditional backpropagation method provides only point estimates of the network parameters, which could not capture the decision uncertainty caused by noisy sensory inputs. In contrast, from the ensemble perspective of candidate networks accomplishing a task, our recent work [9] proposed a spike and slab (SaS) model to learn the credit assignment, bridging the gap between microscopic interactions of components and macroscopic behavior, thereby identifying key parameters capturing informative and nuisance factors in the sensory inputs connected to the output behavior of the network.

### *14.3.1 Ensemble Perspective*

In this section, we derive the ensemble backpropagation algorithm for feedforward neural networks with $L$ layers ($L-2$ hidden layers in addition to the input and output layers). We remark that it is straightforward to adapt the following method to other network architectures, such as CNNs. The depth of the network $L$ can be designed arbitrarily large. For each layer $l$, the width of the corresponding layer is denoted as $N_l$. Therefore, $N_1$ and $N_L$ are determined by the number of pixels in an input image and the number of output classes, respectively, for a classification task. The weight matrix of our model can be written as $\mathbf{w}$, whose element $w_{ij}^l$ denotes the connection from neuron $i$ at the upstream layer $l$ to neuron $j$ at the downstream layer $l+1$. The activation of the neuron $j$ at the layer $l+1$ is a non-linear func-

tion of the pre-activation $z_j^{l+1} = \frac{1}{\sqrt{N_l}} \sum_i w_{ij}^l h_i^l$, where the scaling factor $\frac{1}{\sqrt{N_l}}$ ensures that the weighted sum is independent of the upstream layer width. The rectified linear unit (ReLU) function is applied to create the non-linearity, which preserves the positive pre-activation values while setting the negative values to zero. The output is transferred to the probabilities over all classes by using the softmax function $h_j^L = \frac{e^{z_j^L}}{\sum_i e^{z_i^L}}$, which can be used by the network to make a decision. For simplicity, a categorization task is considered here, and we denote $\hat{\boldsymbol{h}}$ as the corresponding target label which is in the one-hot form. Meanwhile, we use cross-entropy as the loss function $C = -\sum_i \hat{h}_i^L \ln h_i^L$, which requires the gradient descent method to minimize the cross-entropy. For the categorization task, the training data with the size $T$ is applied to train the network by adjusting all the connections to minimize the objective function until a satisfied accuracy is reached. To test the generalization ability of the network, the unseen data with the size $V$ is used.

The standard way to train a deep network is the well-known backpropagation algorithm. However, it can only lead to one point estimate of the connection weights after a single running of the algorithm. Here, we assume that there may exist a random ensemble of neural networks that fulfill the computational task given the width and depth of the deep network. This ensemble may occupy a tiny portion of the entire model space. In that case, we propose a theoretical model whose weight is characterized by a spike and slab (SaS) distribution as follows (Fig 14.4):

$$P(w_{ij}^l) = \pi_{ij}^l \delta(w_{ij}^l) + (1 - \pi_{ij}^l) \mathcal{N}(w_{ij}^l | m_{ij}^l, \Xi_{ij}^l), \tag{14.20}$$

where the spike probability $\delta(w_{ij}^l)$ has a mass at zero, and the slab is characterized by a Gaussian distribution with mean $m_{ij}^l$ and variance $\Xi_{ij}^l$ over a continuous support. These two parts also have their physics interpretations, respectively. The spike is associated with the concept of network compression, while the slab is related to the uncertainty of decision making [9].

### 14.3.2 Training Equations

In this section, we apply the mean-field method to train the SaS model and learn the parameters $\boldsymbol{\theta}_{ij}^l \equiv (\pi_{ij}^l, m_{ij}^l, \Xi_{ij}^l)$ for all the layers. To begin with, we derive the first and second moments of the weight $w_{ij}^l$ as follows:

$$\mu_{ij}^l \equiv \mathbb{E}[w_{ij}^l] = m_{ij}^l (1 - \pi_{ij}^l), \tag{14.21a}$$

$$\varrho_{ij}^l \equiv \mathbb{E}[(w_{ij}^l)^2] = (1 - \pi_{ij}^l)[\Xi_{ij}^l + (m_{ij}^l)^2]. \tag{14.21b}$$

As mentioned before, the pre-activation can be written as $z_j^{l+1} = \frac{1}{\sqrt{N_l}} \sum_i w_{ij}^l h_i^l$. Given a large width of the layer, the central limit theorem indicates that the pre-

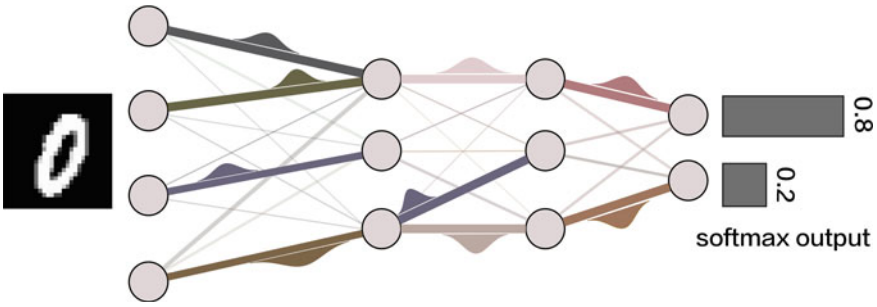



**Fig. 14.4** The schematic illustration of the model learning credit assignment. A deep neural network of four layers including two hidden layers is used to recognize a handwritten digit, say zero, with the softmax output indicating the probability of the categorization. Each connection is specified by a spike and slab distribution, where the spike indicates the probability of the absence of this connection, and the slab is modeled by a Gaussian distribution of weight values as pictorially shown only on strong connections with different means and variances. Other weak connections indicate nearly unit spike probabilities, although they also carry a slab distribution (not shown in the illustration for simplicity). The figure is adapted from the work [9]

activation follows a Gaussian distribution with mean $G_i^l$ and variance $(\Delta_i^l)^2$ as follows:

$$G_i^l = \frac{1}{\sqrt{N_{l-1}}} \sum_k \mu_{ki}^{l-1} h_k^{l-1}, \tag{14.22a}$$

$$(\Delta_i^l)^2 = \frac{1}{N_{l-1}} \sum_k (\varrho_{ki}^{l-1} - (\mu_{ki}^{l-1})^2)(h_k^{l-1})^2. \tag{14.22b}$$

According to this statistics, the pre-activation can be re-parametrized by

$$z_i^l = G_i^l + \epsilon_i^l \Delta_i^l, \tag{14.23a}$$

$$h_i^l = f(z_i^l), \tag{14.23b}$$

where the transfer function $f(z)$ used here is RELU for $l < L$, and softmax function for $l = L$. $\epsilon_i^l$ is a standard Gaussian variable randomly generated for each component in every layer. Meanwhile, $\boldsymbol{\epsilon}^l$ is quenched for every single mini-batch, and the same value is used in both feedforward and backward computations. To train the model, we apply the gradient descent method to minimize the objective function, which can be written as follows:

$$\Delta \boldsymbol{\theta}_{ki}^l = -\eta \mathcal{K}_i^{l+1} \frac{\partial z_i^{l+1}}{\partial \boldsymbol{\theta}_{ki}^l}, \tag{14.24}$$

where $\mathcal{K}_i^{l+1} \equiv \frac{\partial \mathcal{C}}{\partial z_i^{l+1}}$, and $\eta$ indicates the learning rate. The gradients are evaluated over mini-batches which are obtained by dividing the training data into subsets (so-called mini-batches). To calculate the gradients in Eq. (14.24), we first derive the derivative for each hyper-parameter based on Eq. (14.23) as follows:

$$\frac{\partial z_i^{l+1}}{m_{ki}^l} = \frac{(1-\pi_{ki}^l)h_k^l}{\sqrt{N_l}} + \frac{\mu_{ki}^l \pi_{ki}^l (h_k^l)^2 \epsilon_i^{l+1}}{N_l \Delta_i^{l+1}}, \tag{14.25a}$$

$$\frac{\partial z_i^{l+1}}{\partial \pi_{ki}^l} = -\frac{m_{ki}^l h_k^l}{\sqrt{N_l}} - \frac{((2\pi_{ki}^l - 1)(m_{ki}^l)^2 + \Xi_{ki}^l)(h_k^l)^2 \epsilon_i^{l+1}}{2N_l \Delta_i^{l+1}}, \tag{14.25b}$$

$$\frac{\partial z_i^{l+1}}{\partial \Xi_{ki}^l} = \frac{(1-\pi_{ki}^l)(h_k^l)^2 \epsilon_i^{l+1}}{2N_l \Delta_i^{l+1}}. \tag{14.25c}$$

The above derivatives characterize how sensitive the pre-activation is under the change of the hyper-parameters $\boldsymbol{\theta}_{ij}^l \equiv (\pi_{ij}^l, m_{ij}^l, \Xi_{ij}^l)$. Then, we have to calculate the derivative $\mathcal{K}_i^{l+1}$. For $l = L$, $\mathcal{K}_i^L$ can be directly estimated as $\mathcal{K}_i^L = h_i^L - \hat{h}_i^L$. For other layers, $\mathcal{K}_i^l$ can be estimated by using the chain rule, which results in the equations below

$$\mathcal{K}_i^l = \delta_i^l f'(z_i^l), \tag{14.26a}$$

$$\delta_i^l = \sum_k \mathcal{K}_k^{l+1} \frac{\partial z_k^{l+1}}{\partial h_i^l}, \tag{14.26b}$$

where $f'(z)$ denotes the derivative of the transfer function, and $\delta_i^l \equiv \frac{\partial \mathcal{C}}{\partial h_i^l}$. Eq. (14.26b) shows clearly how the gradient signal flows from the output layer down to any intermediate one. To proceed, we have to compute $\frac{\partial z_k^{l+1}}{\partial h_i^l}$, which shows how sensitive the pre-activation at the deeper layer is under the change of the input neural activity to that layer. This part is derived as follows:

$$\frac{\partial z_k^{l+1}}{\partial h_i^l} = \frac{\mu_{ik}^l}{\sqrt{N_l}} + \frac{(\varrho_{ik}^l - (\mu_{ik}^l)^2) h_i^l \epsilon_k^{l+1}}{N_l \Delta_k^{l+1}}. \tag{14.27}$$

Based on the above mean-field method, the hyper-parameters of the model can be updated, and the SaS model naturally captures the fluctuation of the hypothesis space, which significantly differs from the standard backprop [8, 10]. Particularly, if we enforce $\boldsymbol{\pi} = 0$ and $\boldsymbol{\Xi} = 0$, $\boldsymbol{m}$ becomes identical to a single weight configuration. The training method immediately recovers the standard backprop. Hence, the training protocol mentioned above can be thought of as a generalized backpropagation (gBP) at the weight distribution level. The model can separate the deterministic part ($\pi = 0, 1$) from the uncertainty part ($\pi \in (0, 1)$, and $\Xi \neq 0, m \neq 0$). Note that the uncertainty part may capture nuisance factors in sensory inputs. These factors are not informative to the computation task. The gBP can reveal that a U-shaped

$\pi$-distribution, and an L-shaped $\Xi$-distribution, a peak model entropy (derived from the SaS distribution, and assuming that the joint distribution factorizes) in the central part of the network, matching an encoding-recoding-decoding paradigm. We refer interested readers for more details to the original work [9]. In particular, the VIP weights ($\pi = 0$, $\Xi = 0$) play a vital role in determining the final decision-making behavior, which can be quantified by the SaS model.

## References

1. J. Schmidhuber, Neural Netw. **61**, 85 (2015)
2. Y. LeCun, Y. Bengio, G. Hinton, Nature **521**(7553), 436 (2015)
3. K. Hornik, Neural Netw. **4**(2), 251 (1991)
4. Y. LeCun, J.S. Denker, S.A. Solla, in *Advances in Neural Information Processing Systems 2*, vol. 2 (1989), pp. 598–605
5. H. Huang, A. Goudarzi, Phys. Rev. E **98**, 042311 (2018)
6. B. Derrida, Phys. Rev. B **24**(5), 2613 (1981)
7. O. Shayer, D. Levi, E. Fetaya, in *ICLR 2018 : International Conference on Learning Representations* (2018)
8. David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, Nature **323**, 533 (1986)
9. C. Li, H. Huang, Phys. Rev. Lett. **125**, 178301 (2020)
10. L. Bottou, F.E. Curtis, J. Nocedal, SIAM Rev. **60**(2), 223 (2018)

# Chapter 15
# Mean-Field Theory of Dimension Reduction


The sensory cortex in the brain has long been proposed to learn hidden features of sensory inputs in a way called unsupervised learning, which requires no labels or rewards from the data, just by gradually creating better representations of the sensory inputs along a hierarchy of information flow to extract the intrinsic features hidden in the data. Both in the fields of artificial intelligence and neuroscience, the sensory inputs are physically high-dimension data. To extract the latent features in the input data, the process of creating more abstract representations along the hierarchy (e.g., the ventral visual stream of primates) is realized through a non-linear dimensionality reduction of high-dimensional data. Nevertheless, these results have been empirically revealed, which makes computation along hierarchy in deep neural networks extremely nontransparent. In this chapter, we introduce a framework based on mean-field theory to analyze the dimension reduction of data representation across layers (Huang in Phys. Rev. E 98:062313, 2018 [1]; Zhou and Huang in Phys. Rev. E 103:012315, 2021 [2]).

## 15.1 Mean-Field Model

A multilayer feedforward neural network with non-linear transformations of sensory inputs is considered here for the purpose of simplicity (Fig. 15.1). The number of hidden layers is denoted as the depth of this network; the network can be arbitrarily deep. The number of units at each layer is defined as the width of the corresponding layer; we assume that the width of each layer has equal value ($N$) for simplicity. The input data vector can be represented by $\boldsymbol{v}$, and the non-linear transformed representations of the pre-activation $\tilde{a}_i = \sum_j w_{ij}^l h_j^{l-1}$ are denoted as ($\boldsymbol{h}^1$, $\boldsymbol{h}^2$, $\boldsymbol{h}^3$, …, $\boldsymbol{h}^d$). More specifically, $h_i^l = \phi(\tilde{a}_i + b_i^l)$, and we choose the non-linear function as $\phi(x) = \tanh(x)$ without loss of generality. Weights connecting the $(l-1)$th to the $l$th layers are specified by a matrix $\boldsymbol{w}^l$. Biases of neurons at layer $l$ are defined as $\boldsymbol{b}^l$. To facilitate further analytic studies, we make the random weight assumption. Weight



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_15

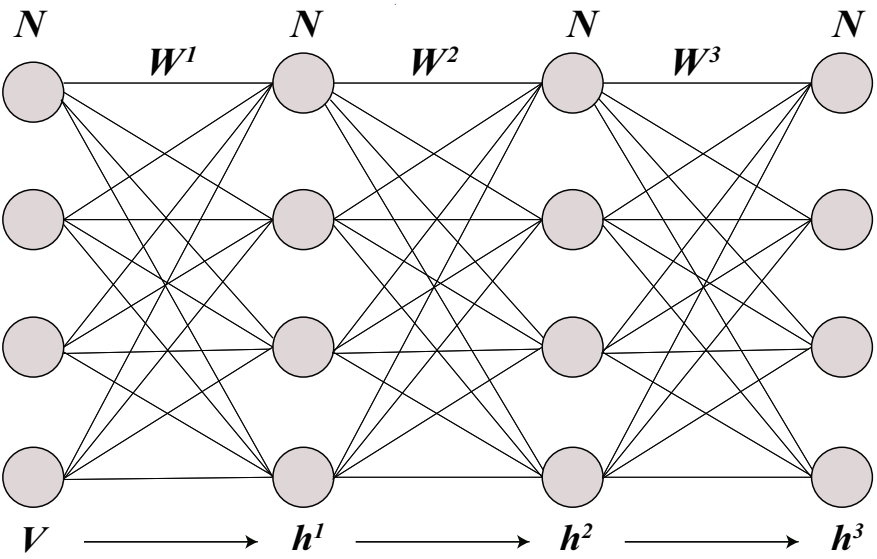



**Fig. 15.1** Schematic illustration of a deep neural network. Here, we introduce the feedforward network with three hidden layers with the input $\boldsymbol{v}$, and the internal representation output of each layer is denoted as $(\boldsymbol{h}^1, \boldsymbol{h}^2, \boldsymbol{h}^3)$. There are N units in each layer

here follows a normal distribution $\mathcal{N}(0, \frac{g}{N})$, and the bias follows another normal distribution with different variance $\mathcal{N}(0, \sigma_b)$. $g$ characterizes the weight strength, while $\sigma_b$ characterizes the bias strength.

To generate the input data, a high-dimensional point in N-dimensional input space, we consider a point-cloud with a maximal correlation strength $\rho$. We assume that each point follows the multivariate Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$, where the covariance entry is defined as $\langle v_i v_j \rangle = \frac{r_{ij}}{\sqrt{N}}$ for all $i \neq j$ ($r_{ij}$ is a random variable uniformly distributed from $-\rho$ to $\rho$), and $\langle v_i^2 \rangle = 1$. In the following derivations, we define the deviation of pre-activation $\tilde{a}_i$ from its mean over the input ensemble as $a_i^l = \sum_j w_{ij}^l (h_j^{l-1} - \langle h_j^{l-1} \rangle)$. It is evident that $a_i^l$ has zero mean, which shows a great convenience in the following analysis.

The covariance of $\boldsymbol{a}^l$ can also be derived by its definition as $\Delta_{ij}^l = \langle a_i^l a_j^l \rangle$. We can get the exact form based on the mean-subtracted activation from the following procedure:

$$\begin{aligned}\Delta_{ij}^{l} &= \langle a_i^l a_j^l \rangle \\ &= \left\langle \sum_k w_{ik}^l (h_k^{l-1} - \langle h_k^{l-1} \rangle) \times \sum_m w_{jm}^l (h_m^{l-1} - \langle h_m^{l-1} \rangle) \right\rangle \\ &= \sum_{km} w_{ik}^l w_{jm}^l \langle (h_k^{l-1} - \langle h_k^{l-1} \rangle)(h_m^{l-1} - \langle h_m^{l-1} \rangle) \rangle \\ &= \sum_{km} w_{ik}^l w_{jm}^l (\langle h_k^{l-1} h_m^{l-1} \rangle - \langle h_k^{l-1} \rangle \langle h_m^{l-1} \rangle) \\ &= \sum_{km} w_{ik}^l w_{jm}^l C_{km}^{l-1} \\ &= [\boldsymbol{w}^l \mathbf{C}^{l-1} (\boldsymbol{w}^l)^{\mathrm{T}}]_{ij}.\end{aligned} \tag{15.1}$$

From the form of $\Delta_{ij}^l$, we know that the covariance of $\boldsymbol{a}^l$ is related to the covariance matrix of neural activity at the $(l-1)$th layer. We further define the data average of neural activity at the $l$th layer $\langle \boldsymbol{h}^l \rangle$ as $\boldsymbol{m}^l$. The elements of $\boldsymbol{a}^l$ can thus be written as $a_i^l = \sum_j w_{ij}^l (h_j^{l-1} - m_j^{l-1})$. When $N$ is large, each neuron at an intermediate layer receives a large number of inputs, which indicates the applicability of the central limit theorem. As a result, the pre-activation $(a_i^l + \sum_j w_{ij}^l m_j^{l-1} + b_i^l)$ follows a normal distribution with the mean of $(\sum_j w_{ij}^l (m_j^{l-1}) + b_i^l)$ and variance of $(\Delta_{ii}^l)$, which results in the following approximate from of $m_i^l$.

$$m_i^l = \langle h_i^l \rangle = \int Dt \phi(\sqrt{\Delta_{ii}^l} t + [\boldsymbol{w}^l \boldsymbol{m}^{l-1}]_i + b_i^l), \tag{15.2}$$

where $Dt = \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$.

Following the same spirit, we obtain the analytic form of $C_{ij}^l$ by the central limit theorem. First, we unfold $C_{ij}^l$ by its definition:

$$\begin{aligned}C_{ij}^l &= \langle h_i^l h_j^l \rangle - \langle h_i^l \rangle \langle h_j^l \rangle \\ &= \langle h_i^l h_j^l \rangle - m_i^l m_j^l,\end{aligned} \tag{15.3}$$

where the part $\langle h_i^l h_j^l \rangle = \langle \phi(a_i^l + \sum_j w_{ij}^l m_j^{l-1} + b_i^l) \phi(a_j^l + \sum_k w_{jk}^l m_k^{l-1} + b_j^l) \rangle$ has to be parametrized by two standard Gaussian variables $x$ and $y$ because of the covariance $\langle a_i^l a_j^l \rangle$. Based on the statistical structure of $a_i^l$ and $a_j^l$, these two activations can be first parametrized as

$$a_i^l = \sqrt{\Delta_{ii}^l} x, \tag{15.4a}$$

$$a_j^l = \sqrt{\Delta_{jj}^l} (\Psi x + \sqrt{1 - \Psi^2} y), \tag{15.4b}$$

where $\Psi = \frac{\Delta_{ij}^l}{\sqrt{\Delta_{ii}^l \Delta_{jj}^l}}$. In this way, we finally obtain $C_{ij}^l$ as follows:

$$C_{ij}^l = \int Dx Dy \phi(\sqrt{\Delta_{ii}^l} x + b_i^l + [\boldsymbol{w}^l \boldsymbol{m}^{l-1}]_i) \phi(\sqrt{\Delta_{jj}^l}(\Psi x + \sqrt{1-\Psi^2} y) + b_j^l + [\boldsymbol{w}^l \boldsymbol{m}^{l-1}]_j) - m_i^l m_j^l. \tag{15.5}$$

However, the form of $C_{ij}^l$ is still very complicated for a theoretical analysis to gain underlying mechanisms of dimensionality reduction. According to equilibrium statistical physics in the thermodynamic limit, $C_{ij}^l$ is of the order $O(\frac{1}{\sqrt{N}})$ for $i \neq j$, and therefore $\langle (\Delta_{ij}^l)^2 \rangle = \sum_{k,m} \langle (w_{ik}^l)^2 \rangle \langle (w_{jm}^l)^2 \rangle (C_{km}^{l-1})^2 = N^2 \frac{g}{N} \frac{g}{N} \frac{1}{N} \sim O(\frac{1}{N})$. Hence, $\Delta_{ij}^l$ is also of the order $O(\frac{1}{\sqrt{N}})$. Meanwhile, we can also analyze the magnitude of $\Delta_{ii}^l$. $\langle \Delta_{ii}^l \rangle = \sum_k \langle (w_{ik}^l)^2 \rangle C_{kk}^{l-1} = \frac{g}{N} \sum_k C_{kk}^{l-1} \sim O(g)$. In this sense, $\Delta_{ij}^l$ is a very small variable in a large-width limit. Hence, we can carry out a Taylor expansion of $C_{ij}^l$ around $\Delta_{ij}^l = 0$:

$$C_{ij}^l \simeq \int Dx Dy \phi(\sqrt{\Delta_{ii}^l} x + z_i^l) \phi'(\sqrt{\Delta_{jj}^l} x + z_j^l) \frac{x \Delta_{ij}}{\sqrt{\Delta_{ii}^l}}, \tag{15.6}$$

where $z_{i,j}^l = b_{i,j}^l + [\boldsymbol{w}^l \boldsymbol{m}^{l-1}]_{i,j}$. Based on the identity $\int Dz \tanh(z) z = \int Dz \tanh'(z)$, we can simplify Eq. (15.6) as follows:

$$C_{ij}^l \simeq \int Dx Dy \phi'(\sqrt{\Delta_{ii}^l} x + z_i^l) \phi'(\sqrt{\Delta_{jj}^l} y + z_j^l) \Delta_{ij}^l. \tag{15.7}$$

Nevertheless, the form of $C_{ij}^l$ still contains an integral part, which makes it inconvenient in the further theoretical analysis. Considering the magnitudes of $\Delta_{ii}^l$ and $\Delta_{jj}^l$, if we make another assumption that the parameter $g$ is also small, the parameter $\Delta_{ii}^l$ and $\Delta_{jj}^l$ can also be seen as small physics quantities, which suggests another Taylor expansion of $C_{ij}^l$ around $\Delta_{ii}^l = 0$ and $\Delta_{jj}^l = 0$. Hence, we obtain

$$\begin{aligned} C_{ij}^l &\approx \int Dx Dy [\phi'(z_i^l) + \phi''(z_i^l) \sqrt{\Delta_{ii}^l} x][\phi'(z_j^l) + \phi''(z_j^l) \sqrt{\Delta_{jj}^l} y] \Delta_{ij}^l \\ &= \phi'(z_i^l) \phi'(z_j^l) \Delta_{ij}^l \\ &= K_{ij}^l \Delta_{ij}^l, \end{aligned} \tag{15.8}$$

where $K_{ij}^l = \phi'(z_i^l) \phi'(z_j^l)$, and we only retain the first-order Taylor expansion of $C_{ij}^l$. Equation (15.7) holds in the large-width limit, while Eq. (15.8) requires additionally the assumption of the small-coupling strength in deep networks.

## 15.2 Linear Dimensionality and Correlation Strength

In this section, we define two important physics quantities: linear dimensionality of the presentation ($D^l$) and covariance strength $\Sigma^l$.

To characterize the collective property of the entire hidden representation, we define the linear dimensionality of the representation at layer $l$ as $D^l = \frac{(\sum_{i=1}^N \lambda_i)^2}{\sum_{i=1}^N \lambda_i^2}$, where $\{\lambda_i\}$ is the eigenspectrum of the covariance matrix $\mathbf{C}^l$. According to the Cauchy inequality formula

$$\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i\right)^2 \leq \frac{1}{N}\sum_{i=1}^{N}\lambda_i^2, \tag{15.9}$$

from which, we derive that a normalized dimensionality $\tilde{D}^l = D^l/N$ is generally upper-bounded by one. If the eigenvalues of $\mathbf{C}^l$ are all equal, which implies that each component of the representation is generated independently with the same variance, then $D^l = N$. However, if there exist non-trivial correlations in the representation, the linear dimensionality $D^l$ will be smaller than $N$, which will be theoretically and numerically revealed in our model.

Based on our mean-field framework, we first study the dimension reduction process. The theoretical results are computed based on the large-$N$ limit, as shown in Eq. (15.7). The simulation results are computed by a direct propagation of the inputs in our feedforward network. Both the theoretical and simulation results (Fig. 15.2) show that the representation dimensionality progressively decreases along the hierarchy, and these two results agree with each other perfectly, which validates our mean-field derivations.

To get deeper insights about the hidden representation, we have to analyze the overall strength of covariance at layer $l$, i.e., $\Sigma^l$. Because of the symmetry property of the covariance matrix $\mathbf{C}^l$, we define the overall covariance strength as $\Sigma^l = \frac{2}{N(N-1)}\sum_{i<j}(C_{ij}^l)^2$. In fact, these two key parameters of our model, $\Sigma^l$ and $D^l$, are closely related. According to the definition, we have

$$\begin{aligned}\widetilde{D}^l &= \frac{1}{N}\frac{(\sum_{i=1}^N \lambda_i)^2}{\sum_{i=1}^N \lambda_i^2} \\ &= \frac{(\mathrm{Tr}\mathbf{C}^l)^2}{N\mathrm{Tr}(\mathbf{C}^l)^2} \\ &= \frac{(\frac{1}{N}\sum_i C_{ii}^l)^2}{\frac{2}{N}\sum_{i<j}(C_{ij}^l)^2 + \frac{1}{N}\sum_i (C_{ii}^l)^2},\end{aligned} \tag{15.10}$$

where $\mathrm{Tr}(\mathbf{C}^l)$ denotes the trace of the matrix $\mathbf{C}^l$. As the overall covariance strength $\Sigma^l = \frac{1}{N(N-1)}\sum_{i<j}(C_{ij}^l)^2$, we can build the relationship between the normalized dimension $\widetilde{D}^l$ and the covariance strength as follows:

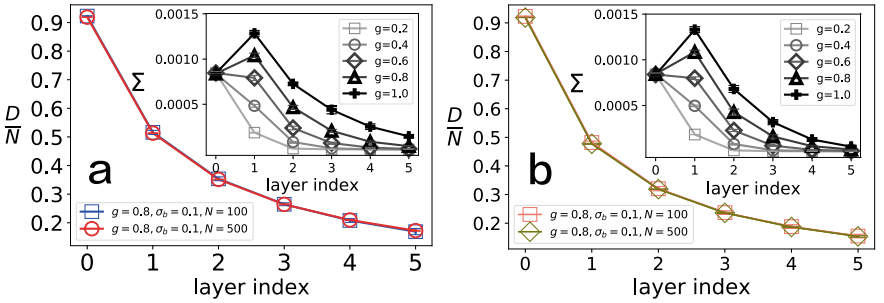



**Fig. 15.2** The numerical simulation of $\Sigma^l$ and $\widetilde{D}^l$ in comparison with theoretical predictions. Input data are generated with $\frac{\rho}{\sqrt{N}} = 0.05$. The left panel shows the feedforward simulation based on $10^5$ samples, and the right panel shows the theoretical results based on the large-$N$ limit assumption. The inset shows how the overall strength of covariance changes with depth and connection strength $g$ when $\sigma_b = 0.1$, $N = 100$. Ten network realizations are considered for numerical simulations

$$\Sigma^l = \frac{1}{N-1}\left[\frac{(\frac{1}{N}\sum_i C_{ii}^l)^2}{\widetilde{D}^l} - \frac{1}{N}\sum_i (C_{ii}^l)^2\right]. \tag{15.11}$$

In physics, strong-enough connection strength maintains the weakly correlated neural activities at further stages of the hierarchy, which facilitates the signal propagation through layers of deep networks by minimal (or maximally compressed) representations (Fig. 15.2).

We also explore how the parameter $\sigma_b$ affects the overall covariance strength and the dimensionality of representations (Fig. 15.3). $\sigma_b$ strongly affects the overall covariance strength $\Sigma^l$ across layers, i.e., a higher $\sigma_b$ induces a lower $\Sigma^l$, yet yielding little impact on the dimensionality. It is intuitive that strong bias would freeze the neural firing pattern, thereby reducing correlations among neural activities. The competition with the coupling effect leads to the observed dimensionality.

### *15.2.1 Iteration Equations for Correlation Strength*

Next, we try to understand the mechanisms of dimension reduction and neural decorrelation. We first discuss the iterative form of the strength $\Sigma^l$. As we already know, the elements of the correlation matrix $\mathbf{C}^l$ can be approximated as $C_{ij}^l \approx K_{ij}^l \Delta_{ij}^l$, where $K_{ij}^l = \phi'(z_i^l)\phi'(z_j^l)$ in the large $N$ and small-$g$ limits. According to this assumption, $\Sigma^l$ can be obtained as

$$\begin{aligned}\Sigma^l &= \frac{2}{N(N-1)}\sum_{i<j}(C_{ij}^l)^2 \\ &= \frac{2}{N(N-1)}\sum_{i<j}(K_{ij}^l)^2(\Delta_{ij}^l)^2,\end{aligned} \tag{15.12}$$

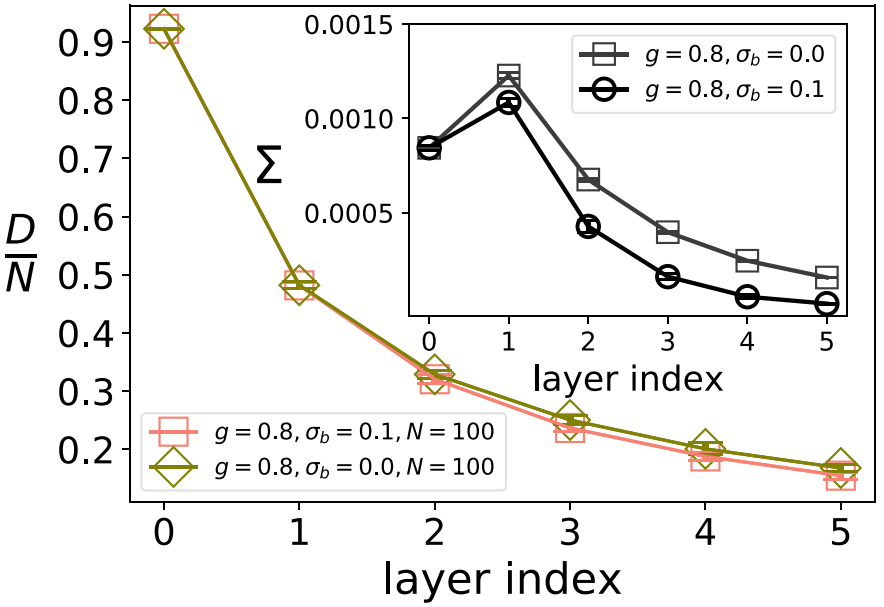



**Fig. 15.3** The numerical simulation of $\Sigma^l$ and $\widetilde{D}^l$ (in the large-$N$ limit assumption). Data are generated with $\frac{\rho}{\sqrt{N}} = 0.05$. The inset shows how the overall strength of covariance $\Sigma^l$ changes with depth and $\sigma_b$, when $g = 0.8$ and $N = 100$. Ten network realizations are considered for each network width and $\sigma_b$

where $(\Delta_{ij}^l)^2 = \sum_{km} (w_{ik}^l)^2 (C_{km}^{l-1})^2 (w_{jm}^l)^2 \simeq \frac{g^2}{N^2} \sum_{km} (C_{km}^{l-1})^2$ because of the statistical structure of $\boldsymbol{w}^l$. Hence, Eq. (15.12) can be rewritten as

$$\begin{aligned}
\Sigma^l \simeq & \frac{2}{N(N-1)} \sum_{i<j} (K_{ij}^l)^2 (\Delta_{ij}^l)^2 \\
\simeq & \frac{2}{N(N-1)} \sum_{i<j} (K_{ij}^l)^2 \frac{g^2}{N^2} (2 \sum_{k<m} (C_{km}^{l-1})^2 + \sum_k C_{kk}^{l-1}) \\
= & \frac{2}{N(N-1)} \sum_{i<j} (K_{ij}^l)^2 \frac{g^2}{N^2} (N(N-1)\Sigma^{l-1} + \sum_k C_{kk}^{l-1}) \\
= & g^2 \Sigma^{l-1} \overline{(K_{ij}^l)^2} + \frac{g^2 \overline{(K_{ij}^l)^2}}{N^2} \sum_k (C_{kk}^l)^2,
\end{aligned} \tag{15.13}$$

where $\overline{(K_{ij}^l)^2} = \frac{2}{N(N-1)} \sum_{i<j} (K_{ij}^l)^2$. Based on the fact that in the thermodynamic limit, the correlation between different weights is negligible and that the covariance of the mean pre-activations of different units is negligible as well, $\overline{(K_{ij}^l)^2}$ can be

approximated by $\overline{(K_{ij}^l)^2} \simeq \overline{[\phi'(z_i^l)]^2}^2 \equiv (\kappa^l)^2$. Hence, Eq. (15.13) can be simplified as

$$\Sigma^l \simeq g^2(\kappa^l)^2\Sigma^{l-1} + \frac{g^2(\kappa^l)^2}{N^2}\sum_i (C_{ii}^{l-1})^2. \tag{15.14}$$

According to our model setting, $N\Sigma^1 = g^2(\kappa^1)^2(N\Sigma^0 + 1)$, there exists a critical point where the strength $\Sigma^1 = \Sigma^0$. We can then arrive at the critical point as $N\Sigma_* = \frac{g^2(\kappa^1)^2}{1-g^2(\kappa^1)^2}$. The quantity $\kappa^l$ can be derived as follows:

$$\begin{aligned}\kappa^l =& \overline{[\phi'(b_i^l + [\boldsymbol{w}^l\boldsymbol{m}^{l-1}]_i)]^2} \\ =& \int Du\,Dt[\phi'(\sqrt{\sigma_b}u + \sqrt{gQ^{l-1}}t)]^2,\end{aligned} \tag{15.15}$$

where $Q^l = \frac{1}{N}\sum_i (m_i^l)^2$, and $m_i^l = \langle h_i^l\rangle = \int Dt\phi(\sqrt{\Delta_{ii}^l}t + [\boldsymbol{w}^l\boldsymbol{m}^{l-1}]_i + b_i^l)$. Note that the quenched-disorder average has been performed over the network parameter statistics. In addition, we can recursively update $Q^l$ as follows:

$$Q^l = \int Du\,Dt\phi^2[\sqrt{\sigma_b}u + \sqrt{gQ^{l-1}}t]. \tag{15.16}$$

Note that the initial $Q^0 = 0$ by the construction of the random model. Given the above theoretical analysis, we can calculate $\kappa^l$ iteratively, i.e., in a layer-by-layer manner. $\kappa^1 = \int Du[\phi'(\sqrt{\sigma_b}u)]^2$, and the critical point $\Sigma_*$ of the first layer is shown in Fig. 15.4.

As shown in Fig. 15.4, we have determined the critical point $\Sigma_*$ (so-called operating point [1]) of the first layer. In fact, this critical point $\Sigma_*$ defines the condition where the overall strength $\Sigma^1 = \Sigma^0$. It also means that if $\Sigma^0 < \Sigma_*$, there will be a boost of $\Sigma^1$, and decrease otherwise, as shown in (Fig. 15.4). Furthermore, the correlation strength $\Sigma^l$ at every layer always has a layer-dependent operating point, which determines the correlation level of neural activations, i.e., either growing or decreasing.

### *15.2.2 Mechanism of Dimension Reduction*

As we mentioned before, the normalized dimension $\widetilde{D}^l$ has already been proven to be reduced across layers in (Fig. 15.2). Next, we will see why dimension reduction is possible in our toy model. The normalized dimension $\widetilde{D}^{l+1}$ is defined as

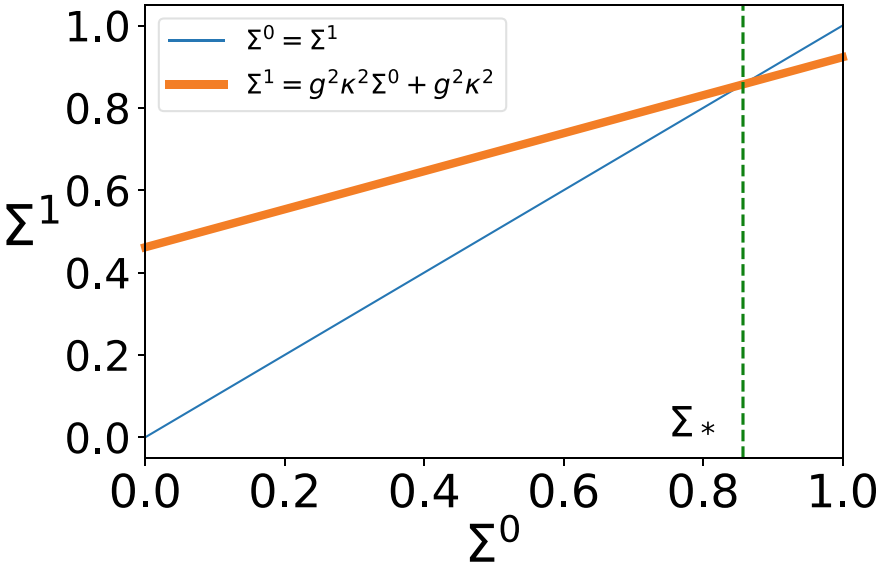



**Fig. 15.4** Illustration of the operating point controlling the magnitude of neural correlation level. The correlation strength $\Sigma^{0,1}$ (including also the critical one) has been scaled by the network width $N$. $\kappa$ here indicates its value at the first layer (see details in the main text)

$$\begin{aligned}\widetilde{D}^{l+1} &= \frac{1}{N}\frac{(\sum_{i=1}^{N}\lambda_i)^2}{\sum_{i=1}^{N}\lambda_i^2} \\ &= \frac{(\mathrm{TrC}^{l+1})^2}{N\mathrm{Tr}(\mathrm{C}^{l+1})^2} \\ &= \frac{(\frac{1}{N}\sum_i C_{ii}^{l+1})^2}{\frac{2}{N}\sum_{i<j}(C_{ij}^{l+1})^2 + \frac{1}{N}\sum_i (C_{ii}^{l+1})^2}.\end{aligned} \tag{15.17}$$

To compare $\widetilde{D}^{l+1}$ and $\widetilde{D}^{l}$, we have to substitute the physics quantities of layer $(l+1)$ in $\widetilde{D}^{l+1}$ by their counterparts of layer $l$. By definition, $\frac{2}{N}\sum_{i<j}(C_{ij}^{l+1})^2 = (N-1)\Sigma^{l+1}$. Note that $\Sigma^{l+1} = g^2(\kappa^{l+1})^2\Sigma^l + \frac{g^2(\kappa^{l+1})^2}{N^2}\sum_i (C_{ii}^l)^2$. We can then get the form of $\frac{2}{N}\sum_{i<j}(C_{ij}^{l+1})^2$ as $(N-1)(g^2(\kappa^{l+1})^2\Sigma^l + \frac{g^2(\kappa^{l+1})^2}{N^2}\sum_i (C_{ii}^l)^2)$. Moreover, $C_{ii}^{l+1} = K_{ii}^{l+1}\Delta_{ii}^{l+1}$, where the part $\Delta_{ii}^{l+1} = \langle a_i^{l+1}a_i^{l+1}\rangle$ can be approximated by $\Delta_{ii}^{l+1} \approx \frac{g}{N}\sum_k C_{kk}^l$, and we can thus obtain the simplified $C_{ii}^{l+1}$ as

$$C_{ii}^{l+1} = K_{ii}^{l+1}\Delta_{ii}^{l+1} \approx gK_{ii}^{l+1}k_1^l, \tag{15.18}$$

where $k_1^l \overset{\text{def}}{=} \frac{1}{N}\sum_i C_{ii}^l$. For further simplicity, we define $\overline{K_{ii}^{l+1}} = \frac{1}{N}\sum_i K_{ii}^{l+1}$, $\overline{(K_{ii}^{l+1})^2} = \frac{1}{N}\sum_i (K_{ii}^{l+1})^2$, and $k_2^l \overset{\text{def}}{=} \frac{1}{N}\sum_i (C_{ii}^l)^2$. Due to the i.i.d assumption of network parameter distribution in our model, $\widetilde{D}^{l+1}$ can be rewritten as

$$
\begin{aligned}
\widetilde{D}^{l+1} &= \frac{(\frac{1}{N}\sum_i C_{ii}^{l+1})^2}{\frac{2}{N}\sum_{i<j}(C_{ij}^{l+1})^2 + \frac{1}{N}\sum_i (C_{ii}^{l+1})^2} \\
&= \frac{g^2\overline{K_{ii}^{l+1}}^2 (k_1^l)^2}{(N-1)\Sigma^{l+1} + g^2(k_1^l)^2\overline{(K_{ii}^{l+1})^2}} \\
&= \frac{g^2\overline{K_{ii}^{l+1}}^2 (k_1^l)^2}{(N-1)(g^2(\kappa^{l+1})^2\Sigma^l + \frac{g^2(\kappa^{l+1})^2}{N^2}\sum_i (C_{ii}^l)^2) + g^2(k_1^l)^2\overline{(K_{ii}^{l+1})^2}} \\
&= \frac{g^2\overline{K_{ii}^{l+1}}^2 (k_1^l)^2}{(N-1)(g^2\overline{(K_{ij}^{l+1})^2}\Sigma^l + \frac{g^2\overline{(K_{ij}^{l+1})^2}}{N^2}\sum_i (C_{ii}^l)^2) + g^2(k_1^l)^2\overline{(K_{ii}^{l+1})^2}} \\
&= \frac{(k_1^l)^2}{(N-1)\Sigma^l + k_2^l + \frac{\overline{(K_{ii}^{l+1})^2}}{\overline{K_{ii}^{l+1}}^2}(k_1^l)^2},
\end{aligned}
\tag{15.19}
$$

where we have used the fact that $\overline{(K_{ij}^{l+1})^2} = \overline{K_{ii}^{l+1}}^2$, thanks to the i.i.d setting.

To compare $\widetilde{D}^{l+1}$ with $\widetilde{D}^l$, we write down the definition of $\widetilde{D}^l$ as follows:

$$
\begin{aligned}
\widetilde{D}^l &= \frac{(\frac{1}{N}\sum_i C_{ii}^l)^2}{\frac{2}{N}\sum_{i<j}(C_{ij}^l)^2 + \frac{1}{N}\sum_i (C_{ii}^l)^2} \\
&= \frac{(k_1^l)^2}{(N-1)\Sigma^l + k_2^l}.
\end{aligned}
\tag{15.20}
$$

Comparing Eqs. (15.20) and (15.19), we can easily draw the conclusion that because the additive term $\frac{\overline{(K_{ii}^{l+1})^2}}{\overline{K_{ii}^{l+1}}^2}(k_1^l)^2$ is always positive, the dimension reduction as $\widetilde{D}^{l+1} < \widetilde{D}^l$ is guaranteed mathematically. Hence, Eqs. (15.20) and (15.19) explain the dimensionality reduction across layers.

To get an explicit form of the additive term, we have to use the large-$N$ limit assumption. Note that

$$
\kappa^l = \int Dt\,Du[\phi'(\sqrt{\sigma_b}u + \sqrt{g\,Q^{l-1}}t)]^2 = \overline{K_{ii}^l}, \tag{15.21a}
$$

$$
Q^{l-1} = \int Du\,Dt\,\phi^2[\sqrt{\sigma_b}u + \sqrt{g\,Q^{l-2}}t], \tag{15.21b}
$$

$$
\overline{(K_{ii}^l)^2} = \int Dt\,Du[\phi'(\sqrt{\sigma_b}u + \sqrt{g\,Q^{l-1}}t)]^4. \tag{15.21c}
$$

According to the definition of $k_1^l$, we can get $k_1^l = \frac{1}{N}\sum_i C_{ii}^l = \overline{\langle h_i^l h_i^l\rangle - \langle h_i^l\rangle\langle h_i^l\rangle}$ $= \overline{\langle h_i^l h_i^l\rangle} - Q^l$, where $Q^l = \frac{1}{N}\sum_i (m_i^l)^2$. An iterative form is thus given by

$$k_1^l = \overline{\int Dx\phi\left[\sqrt{\Delta_{ii}^l}x + b_i^l + [\boldsymbol{w}^l\boldsymbol{m}^{l-1}]_i\right]^2} - Q^l, \tag{15.22}$$

whose quenched average can be performed explicitly as follows:

$$\begin{aligned} k_1^l &= \int Du \int Dt \int Dx\phi^2[\sqrt{gk_1^{l-1}}x + \sqrt{gQ^{l-1}}t + \sqrt{\sigma_b}u] - Q^l, \\ &= \int Du \int Dy\phi^2[\sqrt{gk_1^{l-1} + gQ^{l-1}}y + \sqrt{\sigma_b}u] - Q^l. \end{aligned} \tag{15.23}$$

Taken together, we arrive at the final form of the additive term $\frac{\overline{(K_{ii}^{l+1})^2}}{\overline{K_{ii}^{l+1}}^2}(k_1^l)^2$:

$$\begin{aligned} \frac{\overline{(K_{ii}^{l+1})^2}}{\overline{K_{ii}^{l+1}}^2}(k_1^l)^2 = &\frac{\int Dt\,Du[\phi'(\sqrt{\sigma_b}u + \sqrt{gQ^l}t)]^4}{[\int Dt\,Du[\phi'(\sqrt{\sigma_b}u + \sqrt{gQ^l}t)]^2]^2} \\ &\times \left[\int Du \int Dy\phi^2[\sqrt{gk_1^{l-1} + gQ^{l-1}}y + \sqrt{\sigma_b}u] - Q^l\right]^2. \end{aligned} \tag{15.24}$$

Finally, according to both theory and finite-size system simulation, we find that the additive positive term tends to be a very small value as the number of layers increases (Fig. 15.5), which is consistent with the observation in a finite-$N$ system. This indicates that, because of the property of the addictive term, the estimated dimensionality becomes nearly a constant in the deep layers.

To conclude, the mean-field theory reproduces the key features of dimensionality reduction and neural decorrelation process, which is also compatible with the redundancy reduction hypothesis [3] put forward in neuroscience. Whether the mechanisms revealed by the simple i.i.d setting are robust against taking more network details (e.g., learning effect) deserves future studies.

## 15.3 Dimension Reduction with Correlated Synapses

In the previous part, by a mean-field theory, the dimensionality of layered representations in neural networks whose synaptic weights are independently and identically distributed was calculated. However, in real cortical circuits, synaptic weights among neurons, even in the same layer, may not be ideally independent with each other [4]. Therefore, to understand the mechanism underlying how the weakly correlated synapses affect the neural representations is important. In this part, we will calculate the dimensionality of layered representations under the weakly correlated case by a mean-field theory [2].

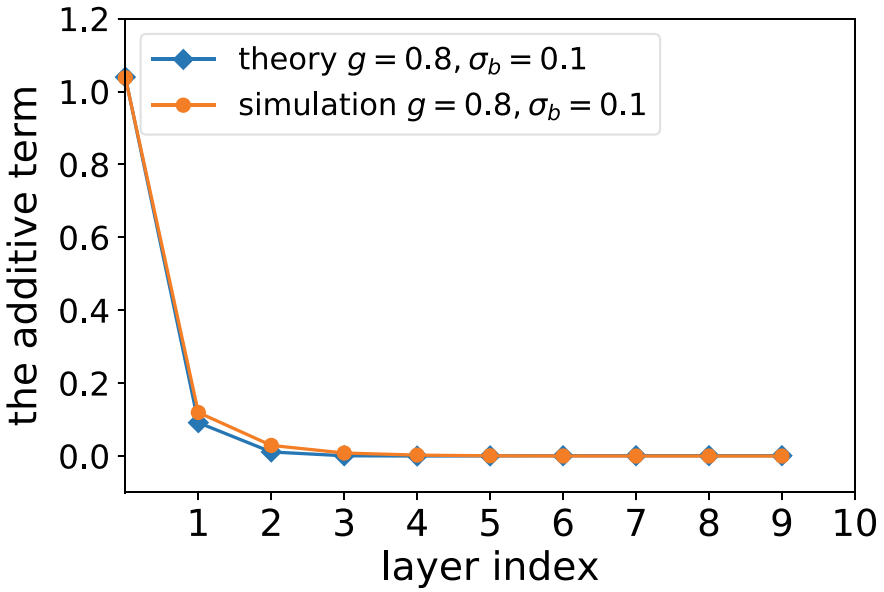



**Fig. 15.5** The behavior of the addictive term as a function of the network depth. The theory is based on the large-$N$ limit assumption, whereas the simulation part is carried out in a finite-$N$ system ($N = 100$)

### *15.3.1 Model Setting*

We consider a deep random neural network with $d$ hidden layers and $N$ neurons at each layer. The weight matrices are defined as $\mathbf{w}^l$ ($l$ is a layer index) whose $i$th row corresponds to incoming connections to the neuron $i$ at the higher layer (so-called the receptive field (RF) of the neuron $i$). Throughout this part, we just consider binary weights ($\pm 1$). The analysis of continual weights is straightforward [2]. The biases of neurons at the $l$th layer are denoted by $\mathbf{b}^l$. The pre-activations are $z_i^l \equiv g[\mathbf{w}^l \mathbf{h}^{l-1}]_i / N + b_i^l$ and the activations are $h_i^l = \phi(z_i^l)$. In this part, we use the non-linear transfer function $\phi(x) = \tanh(x)$.

The specific covariance structure we consider here is Fig. 15.6.

$$\overline{w_{ij}^l w_{ks}^l} = \delta_{js} q + \delta_{ik} \delta_{js} (1 - q) \ . \tag{15.25}$$

The weights have a zero mean. The biases follow a Gaussian distribution $\mathcal{N}(0, \sigma_b)$. Here, we do not enforce any scaling constraint a priori to the correlation level $q$, and it will be determined in a self-consistent way.

We consider random inputs which are independently sampled from a multivariate Gaussian distribution with zero mean and the covariance matrix $\Lambda = \frac{1}{N} \boldsymbol{\xi} \boldsymbol{\xi}^{\mathrm{T}}$, where $\boldsymbol{\xi}$ is an $N \times P$ matrix whose components follow a normal distribution of zero mean

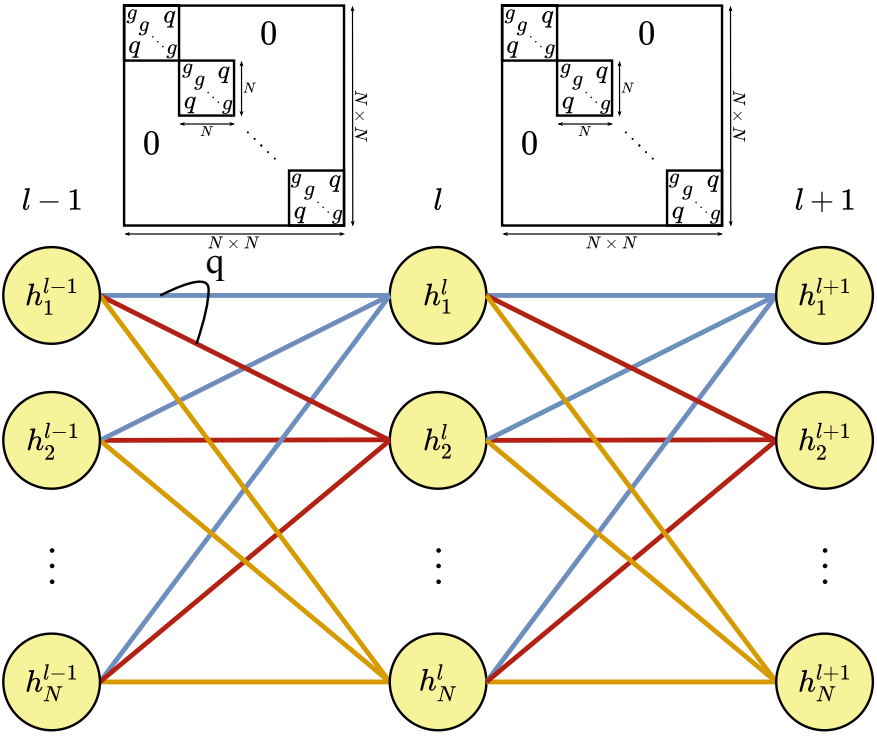



**Fig. 15.6** Schematic illustration of a deep neural network with correlated synapses. The deep neural network carries out a layer-wise transformation of a sensory input. During the transformation, a cascade of internal representations ($\{\mathbf{h}^l\}$) are generated by the correlated synapses, with the covariance structure specified by the matrix above the layer. $g$ characterizes the variance of synaptic weights, while the diagonal block characterizes the inter-receptive-field correlation among corresponding synapses (different line colors), and $q$ specifies the synaptic correlation strength. We do not know a priori the exact scaling form of $q$, which is self-consistently determined by our theory. The figure is adapted from the paper [2]

and variance $\sigma^2$ ($\sigma = 0.5$ here). The ratio $\alpha = P/N$ controls the spectral density of the covariance matrix (see Chap. 17).

## *15.3.2 Mean-Field Calculation*

### 15.3.2.1 Mean-Field Iteration of Activity Moments

In this section, we derive the mean-field iteration of activity moments. We first derive the mean-field equation for the mean activity $m_i^l$ as follows:

$$\begin{aligned} m_i^l &= \langle h_i^l \rangle \\ &= \left\langle \phi\left( \frac{g}{\sqrt{N}} \left[ \mathbf{w}^l \mathbf{h}^{l-1} \right]_i + b_i^l \right) \right\rangle \\ &= \left\langle \phi\left( a_i^l + \frac{g}{\sqrt{N}} \left[ \mathbf{w}^l \mathbf{m}^{l-1} \right]_i + b_i^l \right) \right\rangle , \end{aligned} \tag{15.26}$$

where the average $\langle \cdot \rangle$ is defined over the activity statistics throughout this section, and we define the mean-subtracted weighted-sum (or pre-activation) $a_i^l = \frac{g}{\sqrt{N}} \sum_j w_{ij}^l \left( h_j^{l-1} - \left\langle h_j^{l-1} \right\rangle \right)$, then its expectation is zero, and variance is given by $\Delta_{ij}^l = \left\langle a_i^l a_j^l \right\rangle = \frac{g^2}{N} \left[ \mathbf{w}^l \mathbf{C}^{l-1} \left( \mathbf{w}^l \right)^T \right]_{ij}$, where $\mathbf{C}$ denotes the covariance matrix of the neural activity. Because $a_i^l$ is the sum of $N$ nearly independent random terms, as $N \to \infty$, we apply the central limit theorem, and obtain

$$m_i^l = \int Dt \phi \left( \sqrt{\Delta_{ii}^l} t + \frac{g}{\sqrt{N}} \sum_j w_{ij}^l m_j^{l-1} + b_i^l \right) , \tag{15.27}$$

where $Dt = e^{-t^2/2} dt / \sqrt{2\pi}$. Then, we consider the covariance of activities. Note that the Gaussian random variable $a_i^l$ has a variance $\Delta_{ii}^l$. The activity covariance is then given by

$$\begin{aligned} C_{ij}^l &= \langle h_i^l h_j^l \rangle - \langle h_i^l \rangle \langle h_j^l \rangle \\ &= \left\langle \phi\left( a_i^l + \frac{g}{\sqrt{N}} \left[ \mathbf{w}^l \mathbf{m}^{l-1} \right]_i + b_i^l \right) \phi\left( a_j^l + \frac{g}{\sqrt{N}} \left[ \mathbf{w}^l \mathbf{m}^{l-1} \right]_j + b_j^l \right) \right\rangle - m_i^l m_j^l \\ &= \int Dx Dy \phi\left( \sqrt{\Delta_{ii}^l} x + b_i^l + \frac{g}{\sqrt{N}} \left[ \mathbf{w}^l \mathbf{m}^{l-1} \right]_i \right) \phi\left( \sqrt{\Delta_{jj}^l} (\psi x + y\sqrt{1-\psi^2}) \right. \\ &\quad \left. + b_j^l + \frac{g}{\sqrt{N}} \left[ \mathbf{w}^l \mathbf{m}^{l-1} \right]_j \right) - m_i^l m_j^l , \end{aligned} \tag{15.28}$$

where $Dx = e^{-x^2/2} dx / \sqrt{2\pi}$, and $\psi = \Delta_{ij}^l / \sqrt{\Delta_{ii}^l \Delta_{jj}^l}$. $a_i^l$ and $a_j^l$ have been parametrized by two independent standard Gaussian random variables, say $x$ and $y$, respectively. The pre-activation correlation has been captured by the correlation coefficient $\psi$ ($|\psi| \leq 1$).

With the activity moments, we can then evaluate the dimensionality of the $l$th layer by

$$D^l = \frac{\left( \sum_i \lambda_i \right)^2}{\sum_i \lambda_i^2} = \frac{\left( \operatorname{Tr} \mathbf{C}^l \right)^2}{\operatorname{Tr}(\mathbf{C}^l)^2} = \frac{\left( \sum_i C_{ii}^l \right)^2}{\sum_{i,j} (C_{ij}^l)^2}, \tag{15.29}$$

where $\{\lambda_i\}$ is the eigenspectrum of the covariance matrix $\mathbf{C}^l$. Then we can define the normalized dimensionality as $\tilde{D}^l = \frac{(\operatorname{Tr} \mathbf{C}^l)^2}{N \operatorname{Tr}(\mathbf{C}^l)^2}$, which is then independent of the

network width $N$. To derive the recursion of dimensionality for each layer, we define additionally $\mathcal{K}_1^l = \frac{1}{N}\sum_i C_{ii}^l, \mathcal{K}_2^l = \frac{1}{N}\sum_i (C_{ii}^l)^2$, and $\Sigma^l = \frac{2}{N^2}\sum_{i<j}(C_{ij}^l)^2$ for a large value of $N$. The normalized dimensionality of the $l$th layer is thus expressed as

$$\tilde{D}^l = \frac{(\mathcal{K}_1^l)^2}{N\Sigma^l + \mathcal{K}_2^l}, \tag{15.30}$$

which is useful for the following theoretical analysis.

### 15.3.2.2 Expansion of Two-Point Correlations

In the mean-field limit, we can assume $C_{ij}^l \sim O(1/\sqrt{N})$ for $i \neq j$ [1]. We first analyze the off-diagonal part of the covariance matrix. First, we notice that $\overline{\Delta_{ij}^2} = \frac{g^4}{N^2}\sum_{k,l}\overline{w_{ik}^2 w_{jl}^2}C_{kl}^2 = N^2\frac{g^4}{N^2}\frac{1}{N} \sim O\left(\frac{g^4}{N}\right)$, which means that $\Delta_{ij} \sim O(\frac{g^2}{\sqrt{N}})$. The overline here denotes the disorder average over the network parameters. In other words, when $N$ is sufficiently large, $\Delta_{ij}$ is very small. Then, we execute a Taylor expansion with respect to a small $\Delta_{ij}$ whose layer index is added below

$$\begin{aligned}\phi\left(\sqrt{\Delta_{jj}^l}(\psi x + y\sqrt{1-\psi^2}) + z_j^0\right) &= \phi\left(\sqrt{\Delta_{jj}^l}y + z_j^0\right) \\ &\quad + \phi'\left(\sqrt{\Delta_{jj}^l}y + z_j^0\right)\frac{x\Delta_{ij}^l}{\sqrt{\Delta_{ii}^l}} + O\left(\left(\Delta_{ij}^l\right)^2\right),\end{aligned} \tag{15.31}$$

where we define $z_j^0 = b_j^l + \frac{g}{\sqrt{N}}\left[\mathbf{w}^l\mathbf{m}^{l-1}\right]_j$. By noting that $m_i^l = \int Dt\phi\left(\sqrt{\Delta_{ii}^l}t + z_i^0\right)$, we obtain

$$\begin{aligned}C_{ij}^l &= \int Dx\,Dy\phi\left(\sqrt{\Delta_{ii}^l}x + z_i^0\right)\phi'\left(\sqrt{\Delta_{jj}^l}y + z_j^0\right)\frac{x\Delta_{ij}^l}{\sqrt{\Delta_{ii}^l}} + O\left(\left(\Delta_{ij}^l\right)^2\right) \\ &= \int Dx\,Dy\phi'\left(\sqrt{\Delta_{ii}^l}x + z_i^0\right)\phi'\left(\sqrt{\Delta_{jj}^l}y + z_j^0\right)\Delta_{ij}^l + O\left(\left(\Delta_{ij}^l\right)^2\right).\end{aligned} \tag{15.32}$$

Therefore, we can write $C_{ij}^l \simeq \langle\phi'\left(\sqrt{\Delta_{ii}^l}x + z_i^0\right)\rangle_x\langle\phi'\left(\sqrt{\Delta_{jj}^l}y + z_j^0\right)\rangle_y\Delta_{ij}^l$, where the linear coefficient is an average over standard normal variables, and is called hereafter $K_{ij}^l$ for the following analysis.

We next remark that $\overline{\Delta_{ii}} \simeq \frac{g^2}{N}\sum_k\overline{w_{ik}^2}C_{kk} = g^2\mathcal{K}_1 \sim O(g^2)$. In the small-$g$ limit, we can carry out an expansion in $\sqrt{\Delta_{ii}}$ whose layer index is added below, and get

$$\begin{aligned} C_{ij}^{l} &\simeq \int Dx Dy \left[\phi'\left(z_i^0\right) + \phi''\left(z_i^0\right)\sqrt{\Delta_{ii}^{l}}x\right]\left[\phi'\left(z_j^0\right) + \phi''\left(z_j^0\right)\sqrt{\Delta_{jj}^{l}}y\right]\Delta_{ij}^{l} \\ &= \phi'(z_i^0)\phi'(z_j^0)\Delta_{ij}^{l}. \end{aligned} \tag{15.33}$$

We then analyze the diagonal part of the covariance matrix,

$$\begin{aligned} C_{ii}^{l} &= \left\langle h_i^l h_i^l \right\rangle - \left\langle h_i^l \right\rangle \left\langle h_i^l \right\rangle \\ &= \left\langle \phi^2\left(a_i^l + z_i^0\right)\right\rangle - m_i^l m_i^l \\ &= \int Dx \phi^2\left(\sqrt{\Delta_{ii}^{l}}x + z_i^0\right) - \int Dx \phi\left(\sqrt{\Delta_{ii}^{l}}x + z_i^0\right)\int Dy \phi\left(\sqrt{\Delta_{ii}^{l}}y + z_i^0\right). \end{aligned} \tag{15.34}$$

We expand the above formula in the small $\Delta_{ii}^{l}$, i.e., $\phi\left(a_i^l + z_i^0\right) = \phi\left(z_i^0\right) + \phi'\left(z_i^0\right)\sqrt{\Delta_{ii}^{l}}x$, and obtain

$$\begin{aligned} C_{ii}^{l} &\simeq \int Dx \left[\phi\left(z_i^0\right) + \phi'\left(z_i^0\right)\sqrt{\Delta_{ii}^{l}}x\right]^2 - \left[\int Dx\left(\phi\left(z_i^0\right) + \phi'\left(z_i^0\right)\sqrt{\Delta_{ii}^{l}}x\right)\right]^2 \\ &= \left[\phi'(z_i^0)\right]^2 \Delta_{ii}^{l}. \end{aligned} \tag{15.35}$$

Therefore, we can write $C_{ii}^{l} \simeq K_{ii}^{l}\Delta_{ii}^{l}$, where $K_{ii}^{l}$ is the shorthand for the linear coefficient. To improve the prediction accuracy, one needs to include high-order terms into this approximation. We observe that if we use Eq. (15.32) by setting $i = j$, the theoretical prediction can match the numerical simulation results even in a relatively large value of $g$. This may be due to the fact that the contribution of $\Delta_{ii}^{l}$ is taken into account when computing $K_{ii}^{l}$.

#### 15.3.2.3 Iteration of the Correlation Strength $\boldsymbol{\Sigma}^{l}$

First, we calculate $\mathcal{K}_1^l$, and in the large $N$ and small $g$ limits, we obtain

$$\begin{aligned} \mathcal{K}_1^l &= \overline{\left(\phi'\left(\frac{g}{\sqrt{N}}\sum_{j=1}^{N} w_{ij}^{l} m_j^{l-1} + b_i^l\right)\right)^2}\,\Delta_{ii}^{l} \\ &\simeq \overline{\left(\phi'\left(\frac{g}{\sqrt{N}}\sum_{j=1}^{N} w_{ij}^{l} m_j^{l-1} + b_i^l\right)\right)^2} g^2 \mathcal{K}_1^{l-1}, \end{aligned} \tag{15.36}$$

where $\overline{\cdot}$ means an average over the distribution of network parameters, and $\Delta_{ii}^{l}$ is approximated by

$$\Delta_{ii}^{l} \simeq \frac{g^2}{N}\overline{\sum_{k,j}^{N} w_{ik}^{l} w_{ij}^{l} C_{kj}^{l-1}} = \frac{g^2}{N}\sum_{k=1}^{N} C_{kk}^{l-1} = g^2 \mathcal{K}_1^{l-1} \,. \tag{15.37}$$

Note that the argument of $\phi'(\cdot)$ is a sum of a large number of nearly independent random variables. It is then easy to write that $\overline{\frac{g}{\sqrt{N}}\sum_j w_{ij}^{l} m_j^{l-1} + b_i^{l}} = 0$, and

$$\overline{\left(\frac{g}{\sqrt{N}}\sum_j w_{ij}^{l} m_j^{l-1} + b_i^{l}\right)^2} = \frac{g^2}{N}\sum_j \left(m_j^{l-1}\right)^2 + \sigma_b \,.$$

According to the central limit theorem, we obtain

$$\overline{\left(\phi'\left(\frac{g}{\sqrt{N}}\sum_{j=1}^{N} w_{ij}^{l} m_j^{l-1} + b_i^{l}\right)\right)^2} = \int Dx \left(\phi'(x\sqrt{g^2 Q^{l-1} + \sigma_b})\right)^2 \stackrel{\text{def}}{=} \overline{K_{ii}^{l}} \,, \tag{15.38}$$

where we have defined $Q^{l-1} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\left(m_i^{l-1}\right)^2$. The recursion of $Q^l$ becomes

$$\begin{aligned} Q^{l} &= \frac{1}{N}\sum_i \left[\int Dt\phi\left(\sqrt{\Delta_{ii}^{l}}t + \frac{g}{\sqrt{N}}\sum_j w_{ij} m_j^{l-1} + b_i^{l}\right)\right]^2 \\ &= \int Dx \left[\int Dt\phi\left(\sqrt{g^2\mathcal{K}_1^{l-1}}t + x\sqrt{g^2 Q^{l-1} + \sigma_b}\right)\right]^2 , \end{aligned} \tag{15.39}$$

where we have used $\Delta_{ii}^{l} = g^2 \mathcal{K}_1^{l-1}$.

Finally, we obtain the recursion for $\mathcal{K}_1^{l}$,

$$\mathcal{K}_1^{l} = g^2 \overline{K_{ii}^{l}} \mathcal{K}_1^{l-1}. \tag{15.40}$$

Note that $\mathcal{K}_1^{l}$ can also be calculated recursively without the small-$g$ assumption as follows:

$$\begin{aligned} \mathcal{K}_1^{l} &= \int Dx\, Dt\, \phi^2\left(\sqrt{g^2\mathcal{K}_1^{l-1}}t + x\sqrt{\sigma_b + g^2 Q^{l-1}}\right) - Q^{l} \\ &= \int Dx\, \phi^2\left(\sqrt{g^2\mathcal{K}_1^{l-1} + \sigma_b + g^2 Q^{l-1}}\, x\right) - Q^{l}. \end{aligned} \tag{15.41}$$

Next, the recursion of $\mathcal{K}_2^{l}$ can be calculated by definition as follows:

$$\begin{aligned}\mathcal{K}_2^l &= \frac{1}{N}\sum_i \left(C_{ii}^l\right)^2 \\ &= \frac{1}{N}\sum_i \left(K_{ii}^l g^2 \mathcal{K}_1^{l-1}\right)^2 \\ &= \overline{(K_{ii}^l)^2} g^4 (\mathcal{K}_1^{l-1})^2 \,,\end{aligned} \tag{15.42}$$

where we have assumed that $K_{ii}^l$ in the large-$N$ limit does not depend on the specific site index, and thus

$$\overline{(K_{ii}^l)^2} = \overline{\left(\phi'\left(\frac{g}{\sqrt{N}}\sum_j w_{ij}^l m_j^{l-1} + b_i^l\right)\right)^4} = \int Dx \left(\phi'(x\sqrt{g^2 Q^{l-1} + \sigma_b})\right)^4 \,. \tag{15.43}$$

Note that $\mathcal{K}_2^l$ can be evaluated recursively without the small-$g$ assumption as

$$\mathcal{K}_2^l = \langle \left[\langle \phi^2(f)\rangle_z - \langle \phi(f)\rangle_z^2\right]^2\rangle_{u,t}, \tag{15.44}$$

where $z$, $u$ and $t$ are all standard normal variables, and $f \overset{\text{def}}{=} \sqrt{g^2 \mathcal{K}_1^{l-1}} z + \sqrt{\sigma_b} u + \sqrt{g^2 Q^{l-1}} t$.

We finally derive the recursion of $\Sigma^l$. First, for the binary weights, to compute $\Delta_{ij}^2$, where the layer index can be added later, we have

$$\begin{aligned}\left(\sum_{k,l} w_{ik} w_{jl} C_{kl}\right)^2 &= \left(\sum_{k\neq l} w_{ik} w_{jl} C_{kl} + \sum_k w_{ik} w_{jk} C_{kk}\right)^2 \\ &\simeq \sum_{k\neq l; k'\neq l'} w_{ik} w_{ik'} w_{jl} w_{jl'} C_{kl} C_{k'l'} + \sum_{k,k'} w_{ik} w_{jk} w_{ik'} w_{jk'} C_{kk} C_{k'k'} \\ &\simeq (1+q^2)\sum_{k\neq l} C_{kl}^2 + (1-q^2)\sum_k C_{kk}^2 + q^2\left(\sum_k C_{kk}\right)^2 ,\end{aligned} \tag{15.45}$$

where the cross-term vanishes in statistics to derive the second equality, due to the vanishing intra-RF correlation for one hidden neuron. The third equality is derived by considering the inter-RF correlation in our current setting. Finally, we arrive at

$$\begin{aligned}N\Sigma^{l+1} &= \frac{2}{N}\sum_{i<j} \overline{(K_{ij}^{l+1})^2}\frac{g^4}{N^2}\left[(1+q^2)N^2\Sigma^l + (1-q^2)N\mathcal{K}_2^l + q^2 N^2\left(\mathcal{K}_1^l\right)^2\right] \\ &= \overline{(K_{ij}^{l+1})^2} g^4\left[(1+q^2)N\Sigma^l + q^2 N(\mathcal{K}_1^l)^2 + (1-q^2)\mathcal{K}_2^l\right] \\ &\simeq \overline{(K_{ij}^{l+1})^2} g^4\left[N\Sigma^l + \mathcal{K}_2^l + r^2(\mathcal{K}_1^l)^2\right] \,.\end{aligned} \tag{15.46}$$

A unique scaling for $q$ must then be $q = \frac{r}{\sqrt{N}}$, resulting in $q^2 N = r^2$ where $r \sim O(1)$, and thus Eq. (15.46) is self-consistent in physics as well. Besides, $\overline{(K_{ij}^{l+1})^2}$ is used to replace $(K_{ij}^{l+1})^2$ in the mean-field approximation and can be computed recursively as follows:

$$\begin{aligned}\overline{(K_{ij}^{l+1})^2} = &\left\langle\left\langle \phi'\left(\sqrt{g^2\mathcal{K}_1^l}x + \sqrt{g^2 Q^l}z_1 + \sqrt{\sigma_b}u_1\right)\right\rangle_x^2 \right. \\ &\left. \times \left\langle \phi'\left(\sqrt{g^2\mathcal{K}_1^l}y + \sqrt{g^2 Q^l}\left(\rho z_1 + \sqrt{1-\rho^2}z_2\right) + \sqrt{\sigma_b}u_2\right)\right\rangle_y^2 \right\rangle_{z_1,z_2,u_1,u_2},\end{aligned} \tag{15.47}$$

where $x, y, z_1, z_2, u_1, u_2$ are all standard Gaussian random variables, capturing both thermal and disorder average (inner and outer ones, respectively). The correlation coefficient is given by

$$\rho \overset{\text{def}}{=} \frac{\overline{(z_i^0 - b_i)(z_j^0 - b_j)}}{\sqrt{\overline{(z_i^0 - b_i)^2} \cdot \overline{(z_j^0 - b_j)^2}}} = q. \tag{15.48}$$

We finally remark that the synaptic correlation is able to boost the neural correlation level when transmitting signal via hidden representations. From the linear relationship between $\Sigma^{l+1}$ and $\Sigma^l$ [see Eq. (15.46)], one derives for the binary weights that the operating point is given by

$$\Sigma_*^l = \frac{\Upsilon \mathcal{K}_2^l}{1-\Upsilon} + \frac{\Upsilon r^2 (\mathcal{K}_1^l)^2}{1-\Upsilon}, \tag{15.49}$$

where $\Sigma_*^l$ has been multiplied by $N$, and $\Upsilon \overset{\text{def}}{=} g^4 \overline{(K_{ij}^{l+1})^2}$. Equation (15.49) implies that the operating point is increased by the synaptic correlations (the last term in the equation). The intercept of the linear relationship is also increased by a positive amount $\Upsilon r^2 (\mathcal{K}_1^l)^2$. Note that the slope of the linear relationship under the orthogonal-weight and correlated-weight cases are the same. These phenomenons are shown in Fig. 15.7.

#### 15.3.2.4 Iteration of the Dimensionality Across Layers

According to the definition, with the help of Eqs. (15.40) and (15.42) and the recursion equation for $\Sigma^l$, we obtain

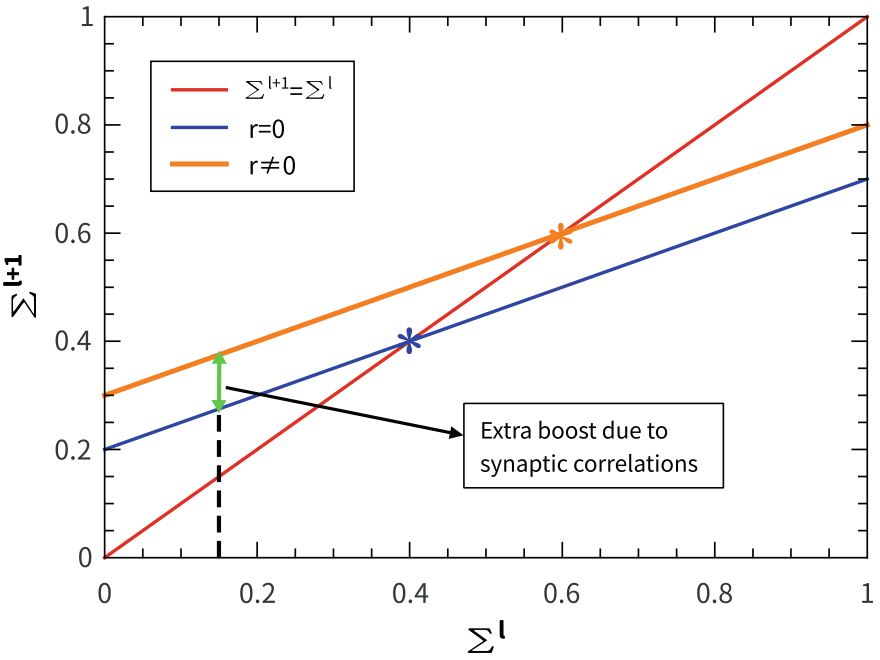



**Fig. 15.7** The schematic illustration showing how synaptic correlations elevate the neural correlation level (multiplied by $N$) and the operating point in hidden representations of deep neural networks. The boost is indicated by the double arrow for an example in which the input $\Sigma^l$ is below the operating point (indicated by star-symbols) where $\Sigma^{l+1} = \Sigma^l$. The figure is adapted from the paper [2]

$$\begin{aligned}\tilde{D}^l &= \frac{\left(\mathcal{K}_1^l\right)^2}{N\Sigma^l + \mathcal{K}_2^l} \\ &= \frac{(\mathcal{K}_1^{l-1})^2}{\gamma_1(N\Sigma^{l-1} + \mathcal{K}_2^{l-1}) + \left(\gamma_1 r^2 + \gamma_2\right)\left(\mathcal{K}_1^{l-1}\right)^2}\,,\end{aligned} \tag{15.50}$$

where $\gamma_1 = \overline{K_{ij}^2}/\overline{K_{ii}}^2$, $\gamma_2 = \overline{K_{ii}^2}/\overline{K_{ii}}^2$ and $r = qN^{\frac{1}{2}}$. When the superscripts of layer index for $K_{ij}$ and $K_{ii}$ are clear, the superscripts are omitted. Here, we manage to use the activity statistics at previous layers to estimate the dimensionality of the current layer, rather than the original formula [Eq. (15.30)]. Thus, the mechanism for dimensionality change can be revealed. The output dimensionality is tuned by a multiplicative factor $\gamma_1$ and an additive term [the last term in the denominator of Eq. (15.50)].

Note that to evaluate $\gamma_1$ and $\gamma_2$, we need to compute the following quantities,

$$\begin{aligned}
\overline{K_{ii}} &= \left\langle \left(\phi'\left(x\sqrt{g^2Q^{l-1}+\sigma_b}\right)\right)^2 \right\rangle_x , \\
\overline{K_{ii}^2} &= \left\langle \left(\phi'\left(y\sqrt{g^2Q^{l-1}+\sigma_b}\right)\right)^4 \right\rangle_y , \\
\overline{K_{ij}^2} &= \left\langle \left\langle \phi'\left(\sqrt{g^2\mathcal{K}_1^{l-1}}x+\sqrt{g^2Q^{l-1}}z_1+\sqrt{\sigma_b}u_1\right)\right\rangle_x^2 \right. \\
&\quad \left. \times \left\langle \phi'\left(\sqrt{g^2\mathcal{K}_1^{l-1}}y+\sqrt{g^2Q^{l-1}}\left(\rho z_1+\sqrt{1-\rho^2}z_2\right)+\sqrt{\sigma_b}u_2\right)\right\rangle_y^2 \right\rangle_{z_1,z_2,u_1,u_2} ,
\end{aligned} \tag{15.51}$$

where $\rho = q$. $Q^l$, $\mathcal{K}_1^l$ and $\mathcal{K}_2^l$ can also be computed recursively by following the iterative equations mentioned before.

#### 15.3.2.5 Closed-Form Mean-Field Iterations for Estimating the Dimensionality

The equations of the mean-field iteration are given by

$$\Delta_{ij}^l = \frac{g^2}{N}\sum_{k,k'} w_{ik}^l C_{kk'}^{l-1} w_{jk'}^l, \tag{15.52}$$

$$m_i^l = \int Dt\phi\left(\sqrt{\Delta_{ii}^l}t + \frac{g}{\sqrt{N}}\sum_{j=1}^{N} w_{ij}^l m_j^{l-1} + b_i^l\right) , \tag{15.53}$$

and

$$\begin{aligned}
C_{ij}^l = \int Dx\,Dy\phi\left(\sqrt{\Delta_{ii}^l}x + b_i^l + \frac{g}{\sqrt{N}}\sum_k w_{ik}^l m_k^{l-1}\right)\cdot \\
\phi\left(\sqrt{\Delta_{jj}^l}(\psi x + y\sqrt{1-\psi^2}) + b_j^l + \frac{g}{\sqrt{N}}\sum_{k'} w_{jk'}^l m_{k'}^{l-1}\right) - m_i^l m_j^l .
\end{aligned} \tag{15.54}$$

### *15.3.3 Numerical Results Compared with Theory*

#### 15.3.3.1 The Generation of Weights and Synthetic Data

We consider a five-layer fully connected neural network with one input layer and four hidden layers. The number of neurons in each layer is specified by $N$. The parameters of the network are generated by following the procedure below, and after the

initialization, all parameters remain unchanged during the simulation of dimension estimation, and then the result is averaged over many independent realizations of the same statistics of network parameters.

The binary weight ($w_{ij} = \pm 1$) follows a statistics of zero mean and the covariance specified by

$$\overline{w_{ij}^l w_{ks}^l} = \delta_{js} q + \delta_{ik}\delta_{js}\,(1-q) = q\delta_{js}\,(1-\delta_{ik}) + \delta_{ik}\delta_{js}\,. \tag{15.55}$$

Diagonalization of the full covariance matrix of binary weights is challenging. However, no correlation occurs within each RF. Then, we can generate the network weights for each diagonal block in Fig. 15.8 independently by a dichotomized Gaussian (DG) process [5]. In the DG process, the binary weights can be generated by $w_{ij}^l = \mathrm{sign}(x_{ij}^l)$, where

$$\mathrm{sign}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}, \tag{15.56}$$

where $x_{ij}^l$ is sampled from a multivariate Gaussian distribution of zero mean (due to $\overline{w_{ij}^l} = 0$) and the following covariance, as also shown in a schematic illustration in Fig. 15.8,

$$\overline{x_{ij}^l x_{ks}^l} = \delta_{js}\,\Sigma + \delta_{ik}\delta_{js}\,(1-\Sigma) = \Sigma\delta_{js}\,(1-\delta_{ik}) + \delta_{ik}\delta_{js}\,. \tag{15.57}$$

The relation between $q$ and $\Sigma$ can be established by matching the covariance of the DG process with our prescribed correlation level $q$, i.e.,

$$q = \iint Dx\,Dy\,\mathrm{sign}(x)\,\mathrm{sign}\left(\Sigma x + \sqrt{1-\Sigma^2}\,y\right) = \frac{2}{\pi}\arcsin\Sigma. \tag{15.58}$$

Then, we have

$$\Sigma = \sin\frac{\pi q}{2}. \tag{15.59}$$

A sample of the multivariate Gaussian distribution with the $N \times N$ covariance matrix $\mathbf{\Sigma}$ (diagonal blocks in Fig. 15.8) can be obtained by first carrying out a Cholesky decomposition of the covariance, i.e., $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower-triangular matrix. A sample is then obtained as $\mathbf{z} = \mathbf{L}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$. $\mathbb{I}$ denotes an identity matrix. The parameter $b_i^l$ follows $\mathcal{N}(0, \sigma_b)$ independently.

#### 15.3.3.2 Results

In this section, we make comparisons between theoretical predictions and numerical results. The experimental details are shown in the caption of the figures. We highlight that the theoretical predictions derived in this section provide a principled

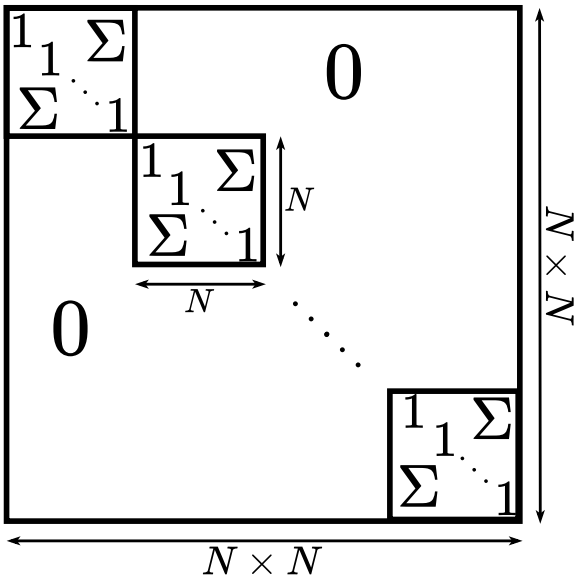



**Fig. 15.8** The schematic illustration of the covariance matrix of $\boldsymbol{x}^l$ is used to generate correlated binary weights. The figure is adapted from the paper [2]

understanding of heuristic tricks of weight and neural decorrelation widely used in machine learning community [6–8].

We find that the weak correlation among synapses is able to reduce further the hidden-representation dimensionality across layers compared to the case of orthogonal weights [Fig. 15.9a]. Moreover, the synaptic correlation $r$ can also boost the correlation strength $\Sigma$ [Fig. 15.9b]. The boost is larger at earlier layers of deep networks. We can draw a conclusion that the weak synaptic correlation accelerates the dimension reduction, while reducing the decay speed of the neural correlation strength.

In Fig. 15.9c, we show that the change of $g$ and $\sigma_b$ has no (or negligible) effect on the dimension reduction. In contrast, the weight strength elevates the correlation level, playing the similar role to the synaptic correlation [Fig. 15.9b]. Besides, increasing the firing bias would further decorrelate the hidden representation.

The output dimensionality is tuned by a multiplicative factor $\gamma_1$ and an additive term [the last term in the denominator of Eq. (15.50)] [Fig. 15.9d]. We observe that the multiplicative factor $\gamma_1$ grows until arriving at the unity; this factor always equals the unity at $q = 0$. The additive term is always positive and decreases with the network depth, thereby contributing an additional reduction of dimensionality. Those two terms overall make the dimension reduction weaker at deeper layers.

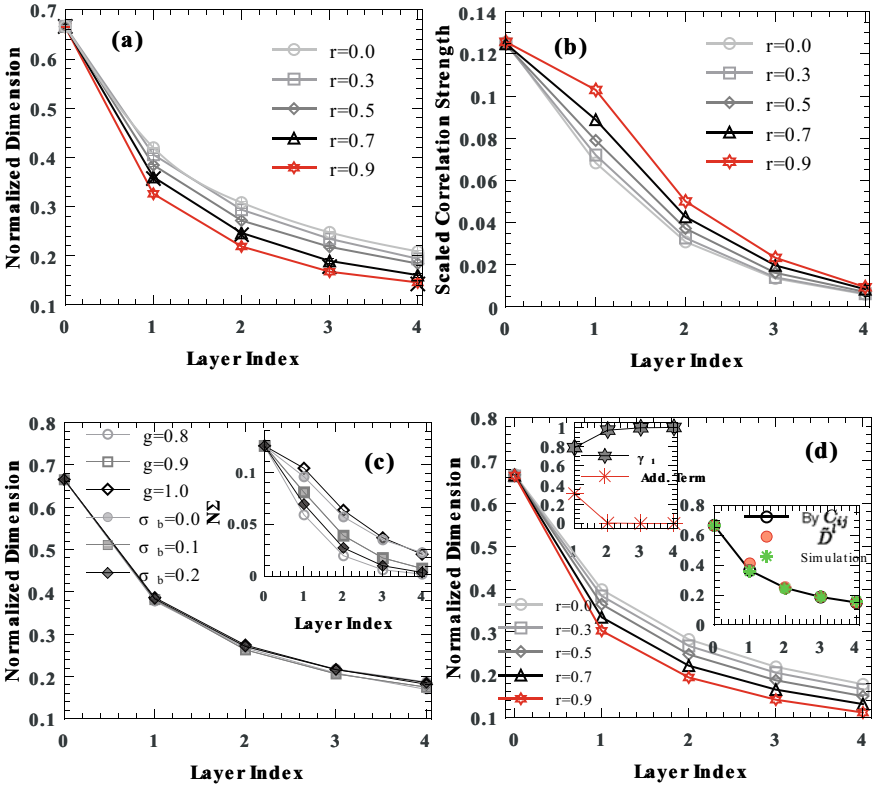



**Fig. 15.9** Typical behavior of dimension reduction in networks of binary weights. Simulations were carried out on networks of finite size $N = 200$, and averaged over ten instances with negligible error bars. **a** Layer-wise dimension reduction with different correlation level $r$. $g = 0.9$, $\alpha = 2$, and $\sigma_b = 0.1$. The covariance is obtained by Eqs. (15.53) and (15.54). The cross symbol indicates the simulation result obtained by layer-wise propagating $10^5$ samples. **b** Layer-wise decorrelation with $r$. Other parameters are the same as in (**a**). The neural correlation strength has been scaled by $N$. **c** Dimension reduction and decorrelation with different values of $g$ and $\sigma_b$. $r = 0.5$. $g = 0.9$ when $\sigma_b$ varies, and $\sigma_b = 0.1$ when $g$ varies. **d** Large-$N$ limit behavior for $g = 0.4$. The left inset shows the behavior of $\gamma_1$ and the additive term. The right inset shows a comparison of the estimated dimensions between theory and simulation ($N = 200$). In both insets, $r = 0.5, \sigma_b = 0.1$ and $\alpha = 2$. The figure is adapted from the paper [2]

## References

1. H. Huang, Phys. Rev. E **98**, 062313 (2018)
2. J. Zhou, H. Huang, Phys. Rev. E **103**, 012315 (2021)
3. H. Barlow, in *Sensory Communication*, ed. by W. Rosenblith (MIT Press, Cambridge, 1961), pp. 217–234
4. K. Harris, G. Shepherd, Nat. Neurosci. **18**, 170 (2015)
5. J.H. Macke, P. Berens, A.S. Ecker, A.S. Tolias, M. Bethge, Neural Comput. **21**(2), 397 (2009)

6. M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, D. Batra, in *ICLR 2016* (2015).arXiv:1511.06068
7. P. Rodriguez, J. Gonzalez, G. Cucurull, J.M. Gonfaus, X. Roca, in *ICLR 2017* (2016). arXiv:1611.01967
8. G. Jin, X. Yi, L. Zhang, L. Zhang, S. Schewe, X. Huang, in *NeurIPS 2020* (2020). arXiv:2010.05983

# Chapter 16
# Chaos Theory of Random Recurrent Neural Networks


In the context of computational neuroscience, analyzing the model of recurrent neural networks (RNNs) is a promising frontier to reveal dynamical computation principles underlying cognitive functions, e.g., working memory, decision-making and learning (Wulfram Gerstner et al. in Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition. Cambridge University Press, Cambridge, 2014 [1]). This is usually achieved by simulating a spiking neural network whose dynamics is described by the evolution of neuronal membrane potentials. The idea is that a neuron in a spiking neural network is not captured by a single activation value (e.g., 0 or 1, thereby unlike the standard Hopfield network). Only when the membrane potential reaches a threshold value, a spike is emitted; after that, a silent period of short duration is maintained. An abstraction of this spiking dynamics is the firing rate model, whose dynamics properties can be analyzed by statistical mechanics tools. In this chapter, we will introduce the dynamical mean-field theory to draw a complete picture about how fixed-point dynamics shifts to chaotic states, and how experimentally observed irregular asynchronous cortical activity can be explained by a mean-field argument.

## 16.1 Spiking and Rate Models

The information in the neural networks is represented by the neural firing activities, including spiking activities as well, and thus the spatio-temporal evolution of these activity patterns is a manifestation of neural information processing. We first briefly describe the spiking model. When a neuron is activated, it produces a discrete spiking signal that is transmitted to other neighboring neurons, increasing or decreasing neighbors' membrane potential via inhibitory or excitatory connections (depending on the cell type of the spiking neuron). In contrast to traditional artificial neural networks, the spiking models process spatio-temporal information, in terms of the following leaky-integrated firing (LIF) equation [2]:



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_16

$$\frac{dV_i(t)}{dt} = -\frac{V_i(t) - V_{\text{rest}}}{\tau_m} + \sum_{j,n} J_{ij}\delta\left(t - t_{jn} - \Delta_{ij}\right) + J_{\text{ext}} \sum_n \delta(t - \widetilde{t}_{in}), \quad (16.1)$$

where $V_i(t)$ can be seen as the membrane potential of the $i$th neuron at the moment $t$, usually at the order of millivolt in a biological neural network. Here, $J_{ij}$ is the synaptic efficacy which couples the output of the (presynaptic) $j$th neuron to the target (post-synaptic) $i$th neuron, and $J_{ii} = 0$. The coupling unit here is millivolt per second. The positive and negative properties of $J_{ij}$ depend on the cell type of neurons, namely excitatory or inhibitory neurons. In a neural circuit, excitatory neurons produce positive outward synapses, while inhibitory neurons produce negative outward synapses. $\tau_m$ (e.g., 20 ms) stands for the membrane time constant, determining the time scale that the membrane potential decays from $V_i(t)$ to the resting voltage $V_{\text{rest}}$. In other words, it specifies the time scale of the membrane potential dynamics. $\Delta_{ij}$ is the signal-transmission delay from $j$th neuron to $i$th neuron. $t_{jn}$ is the $n$th spiking time of the $j$th neuron. Hence, $\sum_{j,n} J_{ij}\delta(t - t_{jn} - \Delta_{ij})$ is the sum of contributions from neighboring spiking neurons of the $i$th neuron. The last part $J_{\text{ext}} \sum_n \delta(t - \widetilde{t}_{in})$ characterizes the external contribution (e.g., a stimulus) to the internal recurrent dynamics. If $J_{\text{ext}} = 0$, the dynamics is called the spontaneous dynamics, or the autonomous dynamics.

In a more biological reality, the recurrent synaptic input $I_s(t) = \sum_{j,n} J_{ij}\delta(t - t_{jn})$ (the delay neglected here) can be also described by an instantaneous jump and exponential decay process [3]:

$$\tau_s \frac{dI_s}{dt} = -I_s(t) + \sum_{j,n} J_{ij}\delta(t - t_{jn}), \quad (16.2)$$

where $\tau_s$ captures the synaptic relaxation time scale. Therefore, Eq. (16.1) assumes the delta-function post-synaptic currents, i.e., the synaptic time constant can be neglected compared to the neuronal time constant.

Figure 16.1 shows a schematic illustration of the membrane potential of representative neurons. The membrane potential evolves from an initial value; the membrane potential fluctuates until it reaches the threshold value, which sends out immediately a spike. After the spike, the membrane potential decays rapidly to the resting voltage, and then stays there in a total of a few milliseconds (defined as the refractory period). During the period ($\tau_{\text{ref}}$, e.g., 2 ms), all the input signals are ignored. A waking animal cortex always shows asynchronous irregular activities with low firing frequency. The dynamical system theory can be applied to analyze the spiking model to get insights about the collective properties of the network. We recommend interested readers the seminal paper [2] and many recent works citing this seminal paper. We would not explore statistical analysis of the LIF model here, which is explored in-depth in the book [1].

We would rather study a simpler system, called the firing rate model. Whether a spiking dynamics is related to a rate model at the macroscopic level is still under heated debated [4]. For simplicity, we use a firing rate to describe the dynamics

of the neurons in a population, instead of spikes. In other words, the firing rate is interpreted as the firing frequency (or probability) in a specified temporal interval. The model has $N$ neurons, whose states are characterized by their local currents (e.g., summed and filtered synaptic current inputs): $x_i(t), i = 1, \ldots, N$; the firing rate can be expressed as $r_i = \tanh(x_i)$; other transfer functions can also be applied. Note that a sigmoid function is physically consistent with the firing probability definition. Here, we would not put much biological reality, and instead focus on the mathematical analysis. Each pair of neuron $i$, $j$ is connected by a synapse of weight $J_{ij}$. The rate description of Eq. (16.1) is simplified as follows:

$$\frac{dx_i}{dt} = -x_i + \eta_i, \tag{16.3}$$

where $\eta_i = \sum_{j=1}^{N} J_{ij}\phi\left(x_j\right)$, and we omit the external drive. We choose $\phi(x) = \tanh(x)$ as the non-linear transfer function of each neuron. We further assume that each synapse is independently sampled from a zero-mean Gaussian distribution $J_{ij} \sim \mathcal{N}\left(0, \frac{g^2}{N}\right)$, where $g$ characterizes the coupling (or recurrent feedback) strength of the rate model. The scaling of $\frac{1}{N}$ in the variance is to ensure the weighted-sum input of each neuron is of $O(1)$ in the large-$N$ limit.

## 16.2 Dynamical Mean-Field Theory

Dynamic mean-field theory is inspired from a generating function formalism of the rate dynamics (see a review [5]). This theory was first applied to the recurrent neural network in 1988 [6]. The theory is also called the path integral approach, having a long history in computing the disorder average of all dynamics trajectories [7]. This approach has been originally designed to study stochastic dynamics in spin systems [8–11]. The path integral method provides a systematic analysis of the high-dimensional dynamics in a complex system, allowing for a thorough analysis of the fluctuations around the saddle point of the action function [11].

### *16.2.1 Dynamical Mean-Field Equation*

Here, we derive intuitively the steady state of the rate dynamics, i.e., the dynamics of the system can be reduced to the dynamics of a single representative neuron, driven by a Gaussian noise. In this sense, $\eta_i$ can be thought of as a time-dependent Gaussian variable. The driving force of the dynamics fluctuates around the zero mean. The statistics is given by

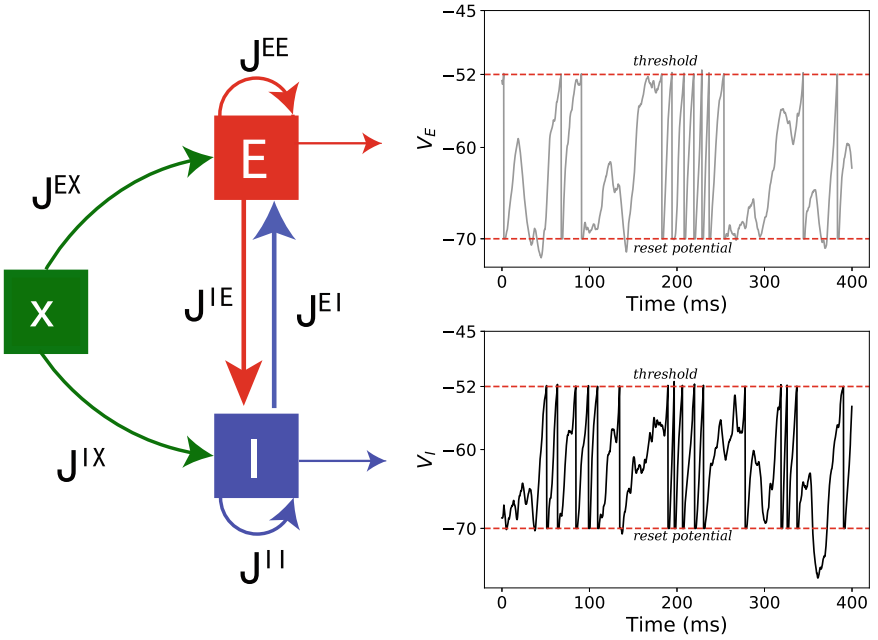



**Fig. 16.1** An illustration of spiking dynamics in a recurrent neural network. (Left) Network architecture. The randomly connected recurrent model is composed of three populations—excitatory one indicated by E, inhibitory one indicated by I, and the external population X simulating an uncorrelated Poisson process. (Right) Spike trains of one randomly selected E neuron and one I neuron from the spiking model following the dynamics equations [Eqs. (16.1) and (16.2)]. The spiking threshold is set to be $-52$ mV. Once the membrane potential of one neuron reaches the threshold, the membrane potential drops to the reset potential $-70$ mV and remains unchanged for a duration of $\tau_{\text{ref}} = 2$ ms. Other time scales are $\tau_m = 20$ ms, and $\tau_s = 10$ ms for both E and I neurons. The population size is $N = 16000$, in which 4000 inhibitory neurons are present. The connection probability is set to $p = 0.1$ for a sparse network

$$\langle \eta_i(t) \rangle = \sum_{j=1}^{N} \left[ J_{ij} \phi\left(x_j\right) \right]_J \approx 0, \tag{16.4}$$

where $[\dots]_J$ denotes the disorder average over the coupling distribution, and $\langle \dots \rangle$ denotes the temporal average of the dynamics. In Eq. (16.4), $\langle \dots \rangle$ is replaced by $[\dots]$ because of the assumption in statistical physics that the ensemble average is equivalent to the temporal average in the long time limit. The time-delayed correlation is defined by

$$\begin{aligned}\left\langle \eta_i(t)\eta_j(t+\tau)\right\rangle &= \left[\sum_{l=1}^{N} J_{il} \sum_{k=1}^{N} J_{jk}\phi\left(x_l(t)\right)\phi\left(x_k(t+\tau)\right)\right] \\ &= \delta_{ij}\frac{g^2}{N}\sum_{k}[\phi(x_k(t))\phi(x_k(t+\tau))] \\ &= \delta_{ij}g^2\langle\phi(x_k(t))\phi(x_k(t+\tau))\rangle \\ &= \delta_{ij}g^2 C(\tau),\end{aligned} \tag{16.5}$$

where the autocorrelation function is introduced as follows:

$$C(\tau) = \langle\phi(x_k(t))\phi(x_k(t+\tau))\rangle, \tag{16.6}$$

which measures the similarity between the state of the system at the time step $t$ and the state after a temporal separation $\tau$. In the long time limit, the autocorrelation depends only on the temporal separation $\tau$. In other words, the dynamics is time-translation invariant.

Applying the Fourier transformation to both sides of Eq. (16.3), we have

$$(1+\mathrm{i}\omega)\hat{x}(\omega) = \hat{\eta}(\omega), \tag{16.7}$$

$$(1-\mathrm{i}\omega)\hat{x}(-\omega) = \hat{\eta}(-\omega), \tag{16.8}$$

where $\hat{x}(\omega)$ is the Fourier transformation of $x(t)$, and $\hat{x}(-\omega)$ is the conjugated quantity of $\hat{x}(\omega)$. Multiplying both sides of Eqs. (16.7) and (16.8), we have

$$\left(1+\omega^2\right)\hat{x}(-\omega)\hat{x}(\omega) = \hat{\eta}(\omega)\hat{\eta}(-\omega). \tag{16.9}$$

Performing an inverse Fourier transform to the right-hand side of Eq. (16.9), we have

$$\begin{aligned}\frac{1}{2\pi}\int \hat{\eta}(\omega)\hat{\eta}(-\omega)e^{\mathrm{i}\omega\tau}d\omega &= \frac{1}{2\pi}\iint \eta(t)e^{-\mathrm{i}\omega t}dt\int \eta\left(t'\right)e^{\mathrm{i}\omega t'}\times e^{\mathrm{i}\omega\tau}dt'd\omega \\ &= \frac{1}{2\pi}\int\int \eta(t)\eta\left(t'\right)dtdt'\int e^{\mathrm{i}\omega\left(t'+\tau-t\right)}d\omega \\ &= \iint \eta(t)\eta\left(t'\right)dtdt'\delta\left(t-t'-\tau\right) \\ &= \langle\eta(t)\eta(t+\tau)\rangle.\end{aligned} \tag{16.10}$$

The similar inverse Fourier transformation applies to the left-hand side of Eq. (16.9):

$$\begin{aligned}&\frac{1}{2\pi}\int\left(1+\omega^2\right)\hat{x}(\omega)\hat{x}(-\omega)e^{\mathrm{i}\omega\tau}d\omega\\=&\frac{1}{2\pi}\int\left(1-(\mathrm{i}\omega)^2\right)\hat{x}(\omega)\hat{x}(-\omega)e^{\mathrm{i}\omega\tau}d\omega\\=&\left(1-\frac{d^2}{d\tau^2}\right)\Delta(\tau),\end{aligned}\tag{16.11}$$

where the local field (or current) autocorrelation $\Delta(\tau)=\langle x_i(t)x_i(t+\tau)\rangle$.

Collecting Eqs. (16.5), (16.10) and (16.11), we arrive at a motion equation describing the dynamics of $\Delta(\tau)$:

$$\Delta-\ddot{\Delta}=g^2C(\tau),\tag{16.12}$$

where $\ddot{\Delta}$ indicates the second-order derivative of $\Delta$ with respect to time. To solve Eq. (16.12), we can write $C(\tau)$ as a function of $\Delta(\tau)$. Equation (16.6) tells us that $C(\tau)$ is a function of $x(t)$. $\Delta(\tau)$ depends also on $x(t)$. Furthermore, $x(t)$ can be approximated by a Gaussian distribution according to the CLT; the mean and covariance are given, respectively, by

$$\begin{aligned}&\langle x(t)\rangle=\langle x(t+\tau)\rangle=0;\\&\langle x(t)x(t+\tau)\rangle=\Delta(\tau).\end{aligned}\tag{16.13}$$

We then use the following parametrization of the random local current $x(t)$:

$$\begin{aligned}&x(t)=\alpha y+\beta z;\\&x(t+\tau)=\alpha y'+\beta z.\end{aligned}\tag{16.14}$$

To satisfy Eq. (16.13), $\alpha=\sqrt{\Delta(0)-|\Delta(\tau)|}$, and $\beta=\sqrt{|\Delta(\tau)|}$. Then $C(\tau)$ can be written in the following form:

$$C(\tau)=\int DyDy'Dz\phi(\alpha y+\beta z)\phi\left(\alpha y'+\beta z\right)=\int Dz\left[\int Dy\phi(\alpha y+\beta z)\right]^2.\tag{16.15}$$

The form of Eq. (16.12) suggests the existence of a potential energy $V$, which satisfies

$$\ddot{\Delta}=-\frac{\partial V}{\partial\Delta}.\tag{16.16}$$

The underlying physics is that Eq. (16.12) can be thought of as a particle moving in a potential well. We then have the following form of the potential:

$$V=-\frac{\Delta^2}{2}+g^2V_2;\quad\frac{\partial V_2}{\partial\Delta}=C(\tau).\tag{16.17}$$

The exact form of $V_2$ can be derived as follows:

$$V_2 = \int Dz \left[ \int Dy \Phi(\alpha y + \beta z) \right]^2, \tag{16.18}$$

where $\frac{d\Phi(x)}{dx} = \phi(x)$, or $\Phi(x) = \int_0^x dy \phi(y)$. One can prove that Eq. (16.18) meets the constraint [Eq. (16.17)]. A detailed proof of Eq. (16.18) is left as an exercise for interested readers. [Hint: Price's Theorem]

Finally, we summarize the dynamical mean-field equation of the rate dynamics:

$$\begin{aligned} \ddot{\Delta} &= -\frac{\partial V}{\partial \Delta}, \\ V(\Delta) &= -\frac{\Delta^2}{2} + g^2 \int Dz \left[ \int Dy \Phi(\sqrt{\Delta(0) - |\Delta|} y + \sqrt{|\Delta|} z) \right]^2, \end{aligned} \tag{16.19}$$

where we omit the time-dependence of $\Delta$ when writing $\Delta$.

### 16.2.2 Regimes of Network Dynamics

As mentioned in the previous section, Eq. (16.19) describes the motion of a particle in a potential well—$V(\Delta)$ with an initial velocity $\dot{\Delta}(0)$. $\Delta(\tau)$ records the coordinate of the particle at time $\tau$. The shape of $V(\Delta)$ depends on the strength of synapse $g$ and the initial position $\Delta(0)$. There are two physical constraints for Eq. (16.19):

- $\Delta$ is bounded like $\Delta(0) \geq |\Delta(t)|$, and $\Delta(0) > 0$.
- $\Delta(t)$ is a differentiable even function, i.e., $\Delta(t) = \Delta(-t)$, because of the time-translation invariance in the long time limit. In addition, $\dot{\Delta}(0) = 0$ (a maximal value reached at $t = 0$), and thus the initial kinetic energy is zero.

Equation (16.19) indicates a mutual transformation between kinetic energy and potential energy, and thus the total energy of the particle is conserved, which means that $\frac{1}{2}\dot{\Delta}(t)^2 + V(\Delta(t)) = V(\Delta(0))$ at every time $t$. Since the kinetic energy is always positive, we have $V(\Delta(t)) \leq V(\Delta(0))$. If we can determine the shape of $V(\Delta)$ at any given $g$ and initial state $\Delta(0)$, we can fully characterize the trajectory of $\Delta$ over time by using Eq. (16.19) and the physical constraints. Let us calculate the derivatives of the potential:

$$\frac{\partial V}{\partial \Delta} = -\Delta + g^2 \int Dz \left[ \int Dy \phi(\sqrt{\Delta(0) - |\Delta|} y + \sqrt{|\Delta|} z) \right]^2, \tag{16.20a}$$

$$\frac{\partial^2 V}{\partial \Delta^2} = -1 + g^2 \int Dz \left[ \int Dy \phi'(\sqrt{\Delta(0) - |\Delta|} y + \sqrt{|\Delta|} z) \right]^2. \tag{16.20b}$$

Note that Eq. (16.20a) is consistent with Eqs. (16.12) and (16.17).

It is easy to get $\frac{\partial V}{\partial \Delta}\big|_{\Delta=0} = 0$ for an odd transfer function. Because $0 < \phi' \leq 1$ for the considered transfer function, the integral in Eq. (16.20b) must be smaller than

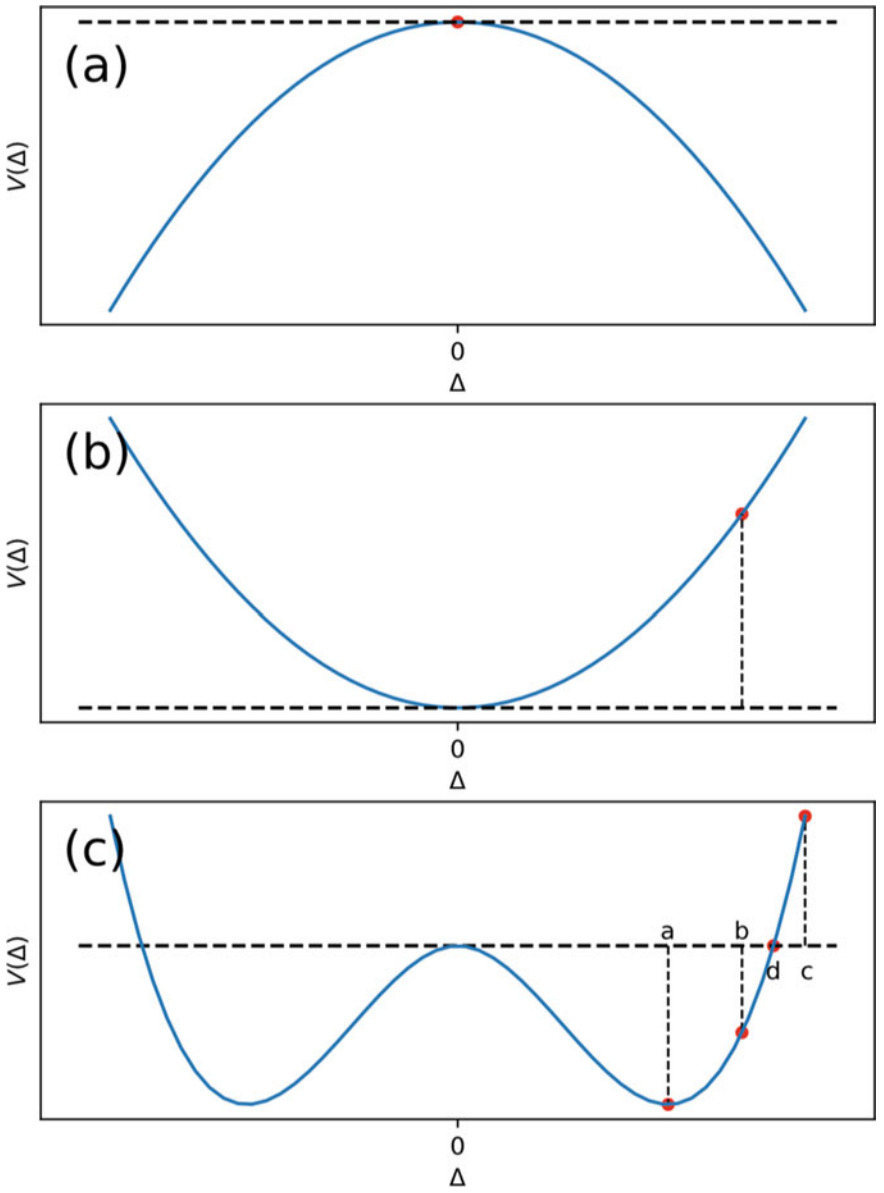



**Fig. 16.2** Dynamics regimes in recurrent rate neural networks [6]. The solid points on the curve show possible initial positions. **a** $g < 1$. **b** $g > 1$ and a small $\Delta(0)$. **c** $g > 1$ and a large $\Delta(0)$. Discussions are presented in the main text

one, and thus we have $\frac{\partial^2 V}{\partial \Delta^2} \leq g^2 - 1$. If $g < 1$, $\frac{\partial^2 V}{\partial \Delta^2}$ will always be negative, and thus $V$ is strictly concave, suggesting that the maximum appears at $\Delta = 0$. If $g > 1$, $\frac{\partial^2 V}{\partial \Delta^2}$ may be positive, but the second derivative also depends on $\Delta(0)$. When $\Delta(0)$ is quite small, $\frac{\partial^2 V}{\partial \Delta^2}|_{\Delta=0} = -1 + g^2 \left[\int Dy \phi'(\sqrt{\Delta(0)}y)\right]^2$, and thus the second-order derivative can be positive, suggesting a convex part for the potential. Therefore, the shape of $V(\Delta)$ can be either a single well or a double well, depending on the sign of $\frac{\partial^2 V}{\partial \Delta^2}$ at $\Delta = 0$.

The potential-shape determines the characteristics of the network dynamics, classified into the following types:

- `Concave shape, for` $g < 1$.
  Because of the concave shape of $V$, $\Delta(t)$ starting from $\Delta(0)$ will tend to grow, violating the physics bound. In addition, due to the energy conservation, the other solution is given by $\Delta(t) = \Delta(0) = 0$, suggesting that $\Delta(t)$ must always stay at the initial point. This indicates an all-silent dynamics state, a trivial fixed-point solution of the dynamical mean-field equation [Fig. 16.2a].
- `Convex shape, for` $g > 1$, `and small` $\Delta(0)$.
  The trajectory of $\Delta$ oscillates from $\Delta(0)$ to $-\Delta(0)$, indicating a limit-cycle solution for the dynamics, as shown in Fig. 16.2b.
- `Double well shape, for` $g > 1$, `and a relatively large` $\Delta(0)$.
  The dynamics now depends on the initial value of $\Delta$. As shown in Fig. 16.2c, we have the following observations: (i) When $\Delta(0)$ is at the bottom of one well (point a), the particle will stay there, which is called a static solution of the dynamics. (ii) $\Delta$ will oscillate around the bottom of the well, provided that $\Delta(0)$ is slightly away from the bottom of one well (point b). (iii) $\Delta$ will oscillate from $\Delta(0)$ to $-\Delta(0)$ in Fig. 16.2 (point c). (iv) At the point d, $V(\Delta(0)) = V(0)$, and the initial energy can exactly bring the particle to $\Delta = 0$. $\Delta$ thus decays monotonically with time, i.e., $\Delta(\tau)$ decays to zero as $\tau \to \infty$. This solution represents a chaotic state of the network, characterized also by a positive value of the maximal Lyapunov exponent (see the next section). The decay rate of $\Delta$ can be characterized by the relaxation time scale $\tau_e$. Let $\Delta(t) \sim \Delta(0) \exp(-t/\tau_e)$, $\tau_e$ can then be derived as $\tau_e = \left[-\partial^2 V(0)/\partial \Delta^2\right]^{-1/2}$ [see Eq. (16.16)].

To sum up, the steady state of the network dynamics can be captured by different types of solutions: fixed points, limit cycles and chaos. The stability of these solutions can be verified by the Hessian of fluctuations around the saddle point of the action based on the path integral representation of the dynamics [11]. The static solution below $g = 1$ is stable, while for $g > 1$, all the oscillatory solutions are unstable, but the only stable solution is the chaotic one. Note that for a finite-size network, the network may display oscillatory patterns of activity, whereas the oscillations will vanish with increasing network sizes; in other words, the chaos transition with $g$ becomes sharper as $N$ becomes larger.

## 16.3 Lyapunov Exponent and Chaos

Next, we show how the chaotic state emerges, i.e., we study how infinitesimal perturbations grow or shrink along the dynamics evolution. This criticality is mathematically characterized by a Lyapunov exponent (the maximal one). A positive exponent implies that nearby trajectories (e.g., starting from nearly the same condition) diverge exponentially fast with time. In other words, chaos depends on the initialization condition.

We first derive the dynamics of perturbations, i.e., we add an infinitesimal fluctuation $\delta x_i(t)$ to Eq. (16.3), and get

$$\frac{dx_i(t)}{dt} + \frac{d\delta x_i(t)}{dt} = -\left(x_i(t) + \delta x_i(t)\right) + \sum_j J_{ij}\left[\phi(x_j(t)) + \phi'(x_j(t))\delta x_j(t)\right]. \tag{16.21}$$

Comparing Eq. (16.21) with Eq. (16.3), we obtain the equation that describes how the perturbation changes with time. That is,

$$(\partial_t + 1)\delta x_i(t) = \sum_j J_{ij}\phi'(x_j(t))\delta x_j(t). \tag{16.22}$$

By making a time translation to Eq. (16.22), we get

$$(\partial_{t+\tau} + 1)\,\delta x_k(t+\tau) = \sum_l J_{kl}\phi'\left(x_l(t+\tau)\right)\delta x_l(t+\tau). \tag{16.23}$$

Multiplying Eq. (16.22) with Eq. (16.23), and taking the average over $\mathbf{J}$ on both sides, we have

$$(\partial_t + \partial_{t+\tau} + \partial_t\partial_{t+\tau} + 1)\Delta_g(t,\tau) = g^2 C_{\phi'}\Delta_g(t,\tau), \tag{16.24}$$

where $C_{\phi'} = \langle\phi'(x(t))\phi'(x(t+\tau))\rangle$, $\Delta_g(t,\tau) = \langle\delta x(t)\delta x(t+\tau)\rangle$, where $\langle\ldots\rangle$ refers to the temporal average.

Performing the variable transformation: $T = t + \tau + t$; $\tau = t + \tau - t$, and using the chain rule of the partial differential, i.e., $\partial_t(f(T,\tau)) = \partial_\tau(f(T,\tau))\frac{\partial\tau}{\partial t} + \partial_T(f(T,\tau))\frac{\partial T}{\partial t}$ (the chain rule for the time derivative w.r.t $t+\tau$ is similar), we recast the left-hand side of Eq. (16.24) into the form $\left[(1+\partial_T)^2 - \partial_\tau^2\right]\Delta_g(T,\tau)$. Notice that $C_{\phi'}(\tau) = \frac{\partial C(\tau)}{\partial\Delta(\tau)} = \frac{\partial^2 V_2}{\partial\Delta^2(\tau)}$, we have $g^2 C_{\phi'}(\tau) = \frac{\partial^2 V}{\partial\Delta^2(\tau)} + 1$, after using Eq. (16.20b). Then, Eq. (16.24) can be reduced to

$$\left[(1+\partial_T)^2 - \partial_\tau^2\right]\Delta_g(T,\tau) = \left(\frac{\partial^2 V}{\partial\Delta^2(\tau)} + 1\right)\Delta_g(T,\tau). \tag{16.25}$$

Next, we are going to study the maximal Lyapunov exponent of the perturbation dynamics. If $|\delta\mathbf{x}(t)| \sim |\delta\mathbf{x}(0)|e^{\lambda t}$, and the maximum exponent $\lambda_{\max}$ of $\boldsymbol{\lambda}$ is positive,

the difference between the original trajectory and the trajectory under the infinitesimal initial deviation will be amplified, thereby leading to a chaotic state. $\lambda_{\max}$ is then given by

$$\begin{aligned}\lambda_{\max} &= \lim_{t\to\infty}\frac{1}{t}\log\left(\frac{\|\delta\mathbf{x}(t)\|_2}{\|\delta\mathbf{x}(0)\|_2}\right)\\ &= \lim_{t\to\infty}\frac{1}{2t}\log\left(\sum_i(\delta x_i(t))^2\right)\\ &= \lim_{t\to\infty}\frac{1}{2t}\log\left[N\Delta_g(t,\tau=0)\right]\\ &= \lim_{t\to\infty}\frac{1}{2t}\log\left[\Delta_g(t,\tau=0)\right],\end{aligned} \tag{16.26}$$

where $\|\delta\mathbf{x}(0)\|_2 = 1$ is assumed.

We further assume a time-separation ansätz for $\Delta_g(t,\tau) \equiv e^{tk}\psi(\tau)$, and thus $\lambda_{\max} = k/2$. Substituting $\Delta_g(T,\tau) = e^{kT/2}\psi(\tau)$ into Eq. (16.25), we have

$$\left(-\partial_\tau^2 - \frac{\partial^2 V}{\partial\Delta^2(\tau)}\right)\psi(\tau) = \left(1-(1+k/2)^2\right)\psi(\tau). \tag{16.27}$$

Equation (16.27) is exactly a one-dimensional time-independent Schrödinger equation. $\tau$ is now interpreted as the spatial coordinate. $-\frac{\partial^2 V}{\partial\Delta^2(\tau)}$ is the quantum potential $W(\tau)$, and $\left(1-(1+k/2)^2\right)$ is the energy $E$. The eigenvalues (or energies) $E_n$ determine the exponential growth rate $k_n$, like $\Delta_g(2t,0) = e^{k_n t}\psi_n(0)$, where $\tau = 0$ leads to $T = 2t$. The rate $k_n$ is given by

$$k_n^{\pm} = 2(-1\pm\sqrt{1-E_n}). \tag{16.28}$$

Denoting the ground state energy as $E_0$, we have immediately:

$$\lambda_{\max} = \frac{k_n^+}{2} = -1+\sqrt{1-E_0}. \tag{16.29}$$

In the case of zero-fixed point, a constant quantum potential is expected. Therefore, $E_0 = W(\Delta=0) = 1-g^2$. The critical coupling strength is then set by $g=1$, above which the zero-fixed point (trivial solution) is destabilized, replaced by a chaotic state. Therefore, once the lowest energy $E_0$ becomes negative, the chaotic state (very sensitive to small changes of the initial condition) appears. An important property of the transition to fluctuating activity is the divergent time scale $\tau_e$ of the fluctuations at a critical coupling $g=1$ ($\tau_e = \left[-\partial^2 V(0)/\partial\Delta^2\right]^{-1/2}$). The edge-of-chaos hypothesis in RNNs' training suggests that the very slow dynamics around the transition regime is very useful for processing long-term temporal dependence of input sequences [12–16]. In addition, a recent theoretical work shows that the proliferation of stationary

points (topological complexity) is coupled with the appearance of a chaotic attractor (dynamical complexity) [17].

## 16.4 Excitation-Inhibition Balance Theory

Neurons in the cortex of behaving animals show temporally irregular spiking patterns. We consider the hypothesis that this irregularity is caused by the balance of excitatory and inhibitory currents into the cortical cells [18–20]. In a biological brain, local cortical circuit is composed of thousands of neurons, with each neuron receiving approximately the order of $O(10^3)$ inputs from other neurons (from the same or different cortical layers, some of them may be long-ranged). We introduce a network model with excitatory and inhibitory populations of simple binary units, whose connectivity profile is random and sparse. Excitatory inputs drive a regular firing, which must be counteracted by local inhibition to yield a low rate irregular cortical firing pattern [20, 21]. In this balanced network, a balance between the excitatory and inhibitory inputs emerges dynamically for a wide range of parameters. When synaptic weights are scaled like $O(1/\sqrt{N})$, where $N$ is the network size, the balanced state is thus achieved by canceling mean excitatory and inhibitory inputs (Fig. 16.3), and thus the fluctuations drive the asynchronous activity [22]. This balance is thus achieved dynamically rather than a fine-tuning of synaptic strength.

We consider a firing rate model of $N_E$ excitatory cells and $N_I$ inhibitory cells, where $K$ excitatory, $K$ inhibitory and $K$ external neurons project to each neuron in the network on average. Although the average number of projections $K$ is large, it is still much smaller than the subpopulation size, i.e., $1 \ll K \ll N_{E,I}$. The connection between the $i$th post-synaptic neuron belonging to the population $k$ and the $j$th presynaptic neuron belonging to the population $l$ is denoted by $J_{kl}^{ij}$, where $k = 1$ or $l = 1$ represents the excitatory subpopulation, while $k = 2$ or $l = 2$ represents the inhibitory subpopulation. Because of the sparse network, $J_{kl}^{ij} = \frac{J_{kl}}{\sqrt{K}}$ with a probability $K/N_l$, where the synaptic constant $J_{k1}$ is positive, and $J_{k2}$ is negative.

We denote the binary variable $\sigma_i^k(t)$ as the state of the neuron $i$ in the population $k$. Therefore, the corresponding total synaptic input $u_i^k(t)$ can be expressed as $u_i^k(t) = \sum_{l=1}^{2} \sum_{j=1}^{N_l} J_{kl}^{ij} \sigma_j^l(t) + u_k^0$, where $u_k^0$ is the external input to any neuron in the $k$th subpopulation, and is defined as $u_k^0 = E_k \sqrt{K} m_0$, where $E_k \sim O(1)$, and $m_k \in [0, 1]$ represents the mean activity of neurons in different subpopulation including the external one. The new state of the $i$th neuron at time $t$ is determined by

$$\sigma_i^k(t) = \Theta(u_i^k(t) - \theta_k), \tag{16.30}$$

where $\theta_k$ is the firing threshold, and $\Theta$ is a step function. Since the model neurons are threshold-type units, the absolute scale of $u_i^k$ is irrelevant. We thus set the synaptic strength as follows:

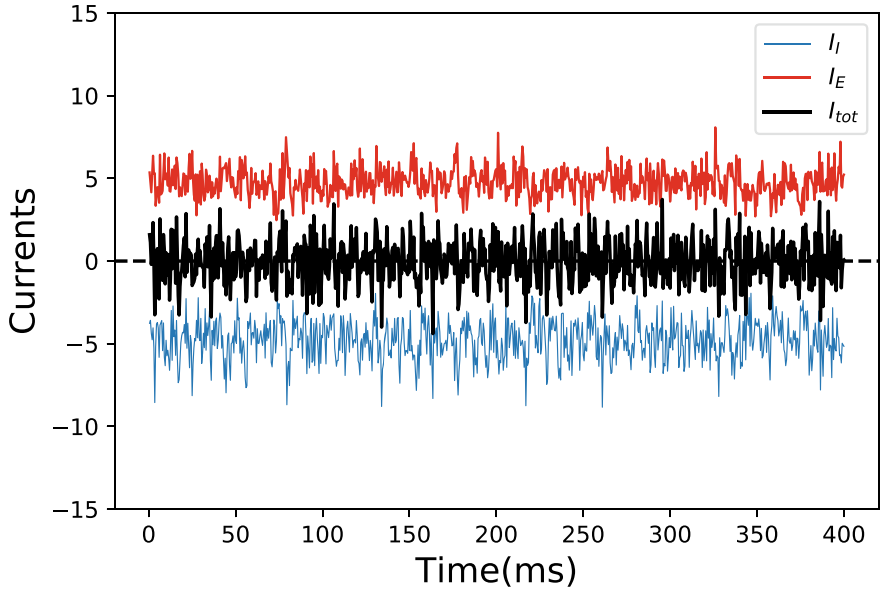



**Fig. 16.3** An illustration of the balanced state in spiking dynamics of a recurrent neural network [Eqs. (16.1) and (16.2)]. The balanced state is characterized by the mean (population averaged) synaptic current of an excitatory contribution $I_E$, an inhibitory contribution $I_I$ and the total current $I_{tot}$. The total current fluctuates around zero, showing the characteristic of the dynamic balance

$$J_{EE} = J_{IE} = 1; \tag{16.31a}$$

$$J_E = -J_{EI} > 0; \tag{16.31b}$$

$$J_I = -J_{II} > 0. \tag{16.31c}$$

Next, we consider the population-averaged inputs of the excitatory and inhibitory cells $u_k(t)$ as

$$u_k(t) = [u_i^k(t)] = \sum_{l=1}^{2} \sum_{j=1}^{N_l} [J_{kl}^{ij}][\sigma_j^l(t)] + u_k^0 = \sqrt{K} \left( \sum_{l=1}^{2} J_{kl} m_l(t) + E_k m_0 \right), \tag{16.32}$$

where the population average $[\dots]$ is defined as the quenched average over the connectivity statistics, and is calculated as $\frac{\sqrt{K} J_{kl}}{N_l}$, and the population-averaged firing rates are defined as $m_l(t) = [\sigma_i^l(t)] = \frac{1}{N_l} \sum_{i=1}^{N_l} \sigma_i^l(t)$. To do the average in Eq. (16.32), we have neglected the correlations between the random fluctuation in the neural activity and the particular realization of the connectivity.

Similarly, we derive the variance $\alpha_k$ as

$$\alpha_k(t) = [(\delta u_i^k(t))^2] = \sum_{l,l'}^{2} \sum_{j,j'}^{N_l} [\delta(J_{kl}^{ij}\sigma_j^l(t))\delta(J_{kl'}^{ij'}\sigma_{j'}^{l'}(t))] = \sum_{l=1}^{2} (J_{kl})^2 m_l(t), \tag{16.33}$$

where the symbol $\delta u = u - [u]$ denoting the fluctuation around the mean. Note that to derive the above variance, we use the result $[(J_{kl}^{ij}\sigma_l^j(t))^2] = J_{kl}^2 m_l/N_l$, and we neglect the small term $[J_{kl}^{ij}\sigma_l^j(t)]^2 = J_{kl}^2 m_l^2 K/N_l^2$ because of $K \ll N_l$. In a balanced state, the temporal fluctuations in the inputs are of the same order with the populations-averaged inputs. By matching the order of magnitude of the population averaged mean and variance, we derive a necessary condition for a balanced state, i.e., both the excitatory and the inhibitory inputs cancel each other in the large-$K$ limit, more precisely being of the order $O(1/\sqrt{K})$. This leads to the following equations:

$$Em_0 + m_I - J_E m_I = 0; \tag{16.34a}$$

$$Im_0 + m_E - J_I m_I = 0, \tag{16.34b}$$

where $E$, $I$ represent the strength of excitatory and inhibitory external inputs, respectively. Then, we obtain a solution:

$$m_E = \frac{J_I E - J_E I}{J_E - J_I} m_0; \tag{16.35a}$$

$$m_I = \frac{E - I}{J_E - J_I} m_0. \tag{16.35b}$$

To have a reasonable solution (not pathological state), we require that $0 < m_l < 1$. Therefore, the following constraints for the model parameters must be obeyed:

$$\frac{E}{I} > \frac{J_E}{J_I} > 1; \quad J_E > 1. \tag{16.36}$$

When $\frac{E}{I} < \frac{J_E}{J_I}$, there exists a solution with $m_l = 0$. This is because, if $m_E = 0$, then $m_I = \frac{Im_0}{J_I}$, and we will have

$$u_E = \sqrt{K}\left(E - \frac{J_E}{J_I} I\right) m_0 < 0. \tag{16.37}$$

On the other hand, when $J_E < 1$ and $J_I < 1$, there appears a solution with $m_E = m_I = 1$ even for $m_0 = 0$. In this case, we have:

$$u_k = \sqrt{K}(1 - J_k) > 0. \tag{16.38}$$

$m_k = 0, 1$ can be thought of as the pathological state of the cortical dynamics, in that all-silent and all-active states are not preferred.

Under this excitation-inhibition balance theory, the neural firing event is purely driven by fluctuations, producing asynchronous irregular patterns, as observed in awake cortex [21]. Recent studies argued that the residual input can be comparable to the excitatory input. But the excitation and inhibition still cancel, yet not as tight as the above balance theory. This scenario is called the loosely balanced setting [23]. The response of a loosely balanced network can be non-linear function of input activity. In contrast, the tightly balanced network responds linearly to its input (see Eq. (16.35a), and Fig. 16.4). The slope is related to the inverse of the mean-recurrent-strength matrix.

We finally remark that the dynamical regime of the recurrent population plays an important role in non-linear computations a neural circuit can implement. Therefore, to provide a mechanistic understanding via theoretical arguments is still promising in current research of theoretical neuroscience, in particular, bridging the gap between models and experimental data.

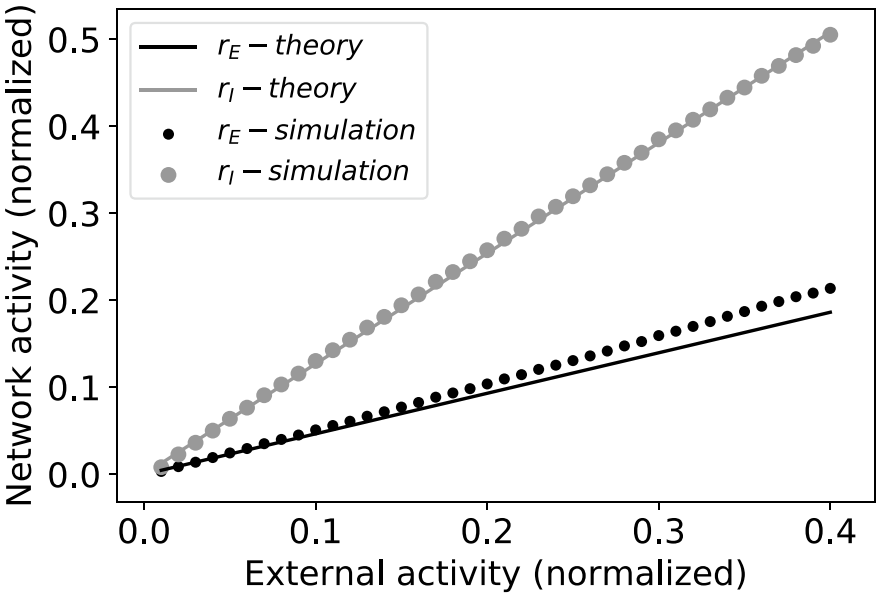



**Fig. 16.4** Linear input tracking of excitation-inhibition balanced network. The mean-field theory predictions are compared with the simulations of a spiking network. Parameters are the same as in Fig. 16.3. We denote $r_x$ as the population($x$)-averaged firing rate. The rate is normalized (scaled) by the maximal value

## 16.5 Training Recurrent Neural Networks

### *16.5.1 Force-Training*

The chaotic activity near the edge of chaos can be used for computation tasks, such as generating oscillating activity, and simulating a decision-making process of a cognitive task [15]. The computational goal can be achieved by modifying only the output weight, maintaining the randomly connected pool of neurons. The algorithm that realizes this type of learning is called FORCE-learning [15]. FORCE is used for the shorthand of first order reduced and controlled error learning. Here, we briefly introduce the training details.

First we have a target output $f_i(t)$, and the readout is obtained as $z(t) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{r}$, where $\boldsymbol{w}$ is the readout weight, and $\boldsymbol{r}$ is the internal dynamics of the RNN. Readout weights can be updated by the following local least mean squared rule:

$$\Delta w_{ij}(t) = -\eta(t) e_i(t) r_j(t), \tag{16.39}$$

where $\eta$ is the learning rate, and the error $e_i(t) = z_i(t) - f_i(t)$. This rule can be further revised by taking into account the correlation function of the rate dynamics:

$$\Delta w_{ij}(t) = -e_i(t)[\mathbf{C}(t)\boldsymbol{r}(t)]_j, \tag{16.40}$$

where $\mathbf{C}(t)$ is a running estimate of the inverse of the correlation matrix of the network activity plus a regularization term:

$$\mathbf{C}(t) = \left( \sum_{t'=t_0}^{t} \boldsymbol{r}(t')\boldsymbol{r}^T(t') + \alpha \mathbf{I} \right)^{-1}, \tag{16.41}$$

where $t_0$ is the starting time, and $\mathbf{C}(0) = \frac{\mathbf{I}}{\alpha}$, and an iterative solution is given by

$$\mathbf{C}(t) = \mathbf{C}(t - \Delta t) - \frac{\mathbf{C}(t - \Delta t)\boldsymbol{r}(t)\boldsymbol{r}^{\mathrm{T}}(t)\mathbf{C}(t - \Delta t)}{1 + \boldsymbol{r}^{\mathrm{T}}(t)\mathbf{C}(t - \Delta t)\boldsymbol{r}(t)}, \tag{16.42}$$

which follows the Sherman–Morrison formula. This learning rule can be adapted to learning the recurrent weight as well [24], and to supervised learning in spiking networks [25, 26].

### *16.5.2 Backpropagation Through Time*

Recurrent neural networks (RNNs) is able to implement tasks involving time-dependent signals, such as natural language processing and time sequence forecast.

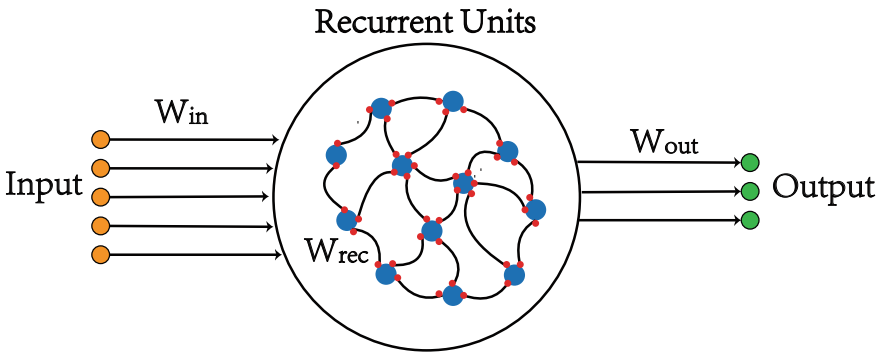



**Fig. 16.5** Schematic illustration of a recurrent neural network. $W_{\text{in}}$, $W_{\text{rec}}$ and $W_{\text{out}}$ are corresponding connection matrices

Moreover, RNNs can also be used to model brain dynamics of any cognition tasks. In this subsection, we introduce a widely used training method for RNNs, which is backpropagation through time (BPTT). We first consider a canonical RNN structure. Then, the derivation of BPTT is carried out in detail based on the chain rule. Finally, we use RNN to perform a classification task on the MNIST benchmark dataset to verify the effectiveness of the RNN model trained by the BPTT algorithm.

#### 16.5.2.1 Dynamics Equation

We consider a discrete-time RNN model (Fig. 16.5), where $N_{\text{rec}}$ recurrent units connected to each other are described by the recurrent activity vector $\mathbf{h}(t)$. For simplicity, we consider at every time step, an input vector $\mathbf{x}(t)$ of $N_{\text{in}}$ dimension enters the network to provide signals to recurrent activities, which are read out to form a time-dependent output $\mathbf{y}(t)$ of $N_{\text{out}}$ dimensions. In practice, the time step for turning on an input or reading out the decision signal depends on specified settings of a task. The dynamics equation of the model reads

$$h_i(t+1) = h_i(t) + \frac{1}{\tau}\left[-h_i(t) + \phi\left(u_i(t+1)\right)\right], \tag{16.43a}$$

$$u_i(t+1) = \sum_{j=1}^{N_{\text{rec}}} W_{ij}^{\text{rec}} h_j(t) + \sum_{j=1}^{N_{\text{in}}} W_{ij}^{\text{in}} x_j(t+1), \tag{16.43b}$$

$$y_k(t) = \sum_{i=1}^{N_{\text{out}}} W_{ki}^{\text{out}} h_i(t), \tag{16.43c}$$

where $\phi(\cdot)$ is a non-linear function, $u_i(t+1)$ is the input current to the unit $i$ at a time step $t+1$, and $\tau$ is the time constant characterizing how fast the RNN dynamics

is. Alternatively, it can be compared to its continuous version—$\tau dh_i/dt = -h_i + \phi(u_i)$. There are only three sets of weight matrices in our setting, which are the input weight $\mathbf{W}_{\text{in}} = \{W_{ij}^{\text{in}}\}$, recurrent weight $\mathbf{W}_{\text{rec}} = \{W_{ij}^{\text{rec}}\}$, and output weight $\mathbf{W}_{\text{out}} = \{W_{ij}^{\text{out}}\}$. These matrices are all time-independent, but need to be adjusted during learning. Our goal is to train the network to produce a desired output $\mathbf{y}(t)$ at each time step, given a time-dependent input $\mathbf{x}(t)$ and an initial activity vector $\mathbf{h}(0)$. Then, the loss function that measures the difference between the target output $\mathbf{y}^*(t)$ and the actual output $\mathbf{y}(t)$ can be defined by the mean square error integrated over time:

$$\mathcal{L} = \frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{N_{out}}[\varepsilon_k(t)]^2, \tag{16.44a}$$

$$\varepsilon_k(t) = y_k(t) - y_k^*(t), \tag{16.44b}$$

where $T$ is the total number of time steps, $\varepsilon_k(t)$ is defined as the error of the output unit $k$ at a time step $t$.

#### 16.5.2.2 Derivations of BPTT

Backpropagation through time is a standard training algorithm for RNNs [27]. In this section, we introduce an easy way to derive BPTT based on the chain rule. More precisely, we first derive the explicit forms of the derivatives of the loss function $\mathcal{L}$ with respect to the input weight $\mathbf{W}_{\text{in}}$, recurrent weight $\mathbf{W}_{\text{rec}}$ and output weight $\mathbf{W}_{\text{out}}$. As the error back-propagates from the output, we first consider $\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{out}}}$. According to the chain rule,

$$\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{out}}} = \sum_{t=t_0}^{T}\sum_{k=1}^{N_{\text{out}}}\frac{\partial \mathcal{L}}{\partial y_k(t)}\frac{\partial y_k(t)}{\partial W_{ab}^{\text{out}}} = \sum_{t=t_0}^{T}\frac{\partial \mathcal{L}}{\partial y_a(t)}\frac{\partial y_a(t)}{\partial W_{ab}^{\text{out}}} = \sum_{t=t_0}^{T}\varepsilon_a(t)h_b(t). \tag{16.45}$$

Then, we derive $\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{rec}}}$ based on the chain rule, and we can easily obtain

$$\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{rec}}} = \sum_{t}\frac{\partial \mathcal{L}}{\partial h_a(t)}\frac{\partial h_a(t)}{\partial W_{ab}^{\text{rec}}} = \frac{1}{\tau}\sum_{t}\frac{\partial \mathcal{L}}{\partial h_a(t)}\phi'(u_a(t))h_b(t-1), \tag{16.46}$$

where we define $z_a(t) \equiv \frac{\partial \mathcal{L}}{\partial h_a(t)}$ as the error of the recurrent unit $a$ at a time step $t$, which backpropagates through the network during training. At the last time step $T$, we have

$$\frac{\partial \mathcal{L}}{\partial h_a(T)} = \sum_{k=1}^{N_{\text{out}}}\frac{\partial \mathcal{L}}{\partial y_k(T)}\frac{\partial y_k(T)}{\partial h_a(T)} = \sum_{k=1}^{N_{\text{out}}}\varepsilon_k(T)W_{ka}^{\text{out}}. \tag{16.47}$$

At the other time steps $t = 0, 1, \ldots, T-1$, the derivation proceeds as follows:

$$\frac{\partial \mathcal{L}}{\partial h_a(t)} = \sum_{k=1}^{N_{\text{out}}} \frac{\partial \mathcal{L}}{\partial y_k(t)} \frac{\partial y_k(t)}{\partial h_a(t)} + \sum_{j=1}^{N_{\text{rec}}} \frac{\partial \mathcal{L}}{\partial h_j(t+1)} \frac{\partial h_j(t+1)}{\partial h_a(t)}$$
$$= \sum_{k=1}^{N_{\text{out}}} \varepsilon_k(t) W_{ka}^{\text{out}} + \frac{1}{\tau} \sum_{j=1}^{N_{\text{rec}}} \phi'(u_j(t+1)) W_{ja}^{\text{rec}} z_j(t+1) + (1 - \frac{1}{\tau}) z_a(t+1). \tag{16.48}$$

Finally, $\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{in}}}$ is derived as

$$\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{in}}} = \sum_t \frac{\partial \mathcal{L}}{\partial h_a(t)} \frac{\partial h_a(t)}{\partial W_{ab}^{\text{in}}} = \sum_t z_a(t) \phi'(u_a(t)) x_b(t), \tag{16.49}$$

which is easy to obtain from the Eq. (16.46) by replacing $h_b(t-1)$ by $x_b(t)$.

With the derivation of the above three derivatives, we can summarize the BPTT algorithm in three steps. First, following the dynamics described by Eq. (16.43), recurrent activity $\mathbf{h}(t)$ and output $\mathbf{y}(t)$ evolve over time. Thus, the error $\boldsymbol{\varepsilon}(t)$ can be directly computed. Second, the gradient term caused by error, namely $\boldsymbol{z}(t)$ is integrated backwards in time described by Eqs. (16.47) and (16.48), and we can finally obtain the gradients for three sets of weights,

$$\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{out}}} = \sum_{t=t_0}^{T} \varepsilon_a(t) h_b(t), \tag{16.50a}$$

$$\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{rec}}} = \sum_{t=t_0}^{T} z_a(t) \phi'(u_a(t)) h_b(t-1), \tag{16.50b}$$

$$\frac{\partial \mathcal{L}}{\partial W_{ab}^{\text{in}}} = \sum_{t=t_0}^{T} z_a(t) \phi'(u_a(t)) x_b(t). \tag{16.50c}$$

We remark that the learning can be implemented by applying different kinds of optimizers, such as vanilla SGD or Adam.

In the above BPTT, the gradient flows back from every time step $t$ to every time step $t' < t$, which may be very deep in the temporal domain, causing gradients to explode or vanish. However, in practice, a truncated version can be designed, where gradients do not flow from $t$ to $t'$ if the temporal distance $|t - t'|$ exceeds a truncation window size [28].

We finally show an example of training a RNN with the BPTT derived above. MNIST is a benchmark classification dataset, containing handwritten digits patterns from 0 to 9, which are $28 \times 28$ grayscale images. We input the training patterns row by row at each time step, which means that the input dimension $N_{\text{in}}$ for the RNN is 28, and the total number of time steps for processing one image is also 28. The output dimension $N_{\text{out}}$ should be equal to the total classes which is 10. 150 recurrent units are used in this example. For the MNIST classification task, we only consider the error generated from the last time step, when the whole image has been shown

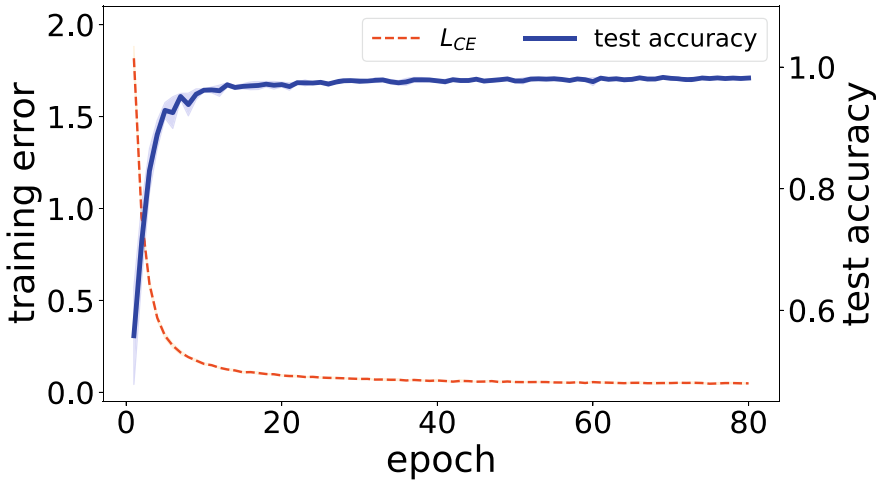



**Fig. 16.6** Training trajectories of a RNN performing the MNIST classification. The lines are mean results from the five independent runs, where the shadows indicate standard deviations

to the RNN. During each training epoch, 12800 images randomly selected from the total training data (60000 images) are divided into 100 mini-batches with their size equal to 128. The loss function is the cross-entropy, and Adam is used to optimize the gradient with the learning rate of 0.01. In addition, we apply the gradient clipping in which each gradient element for the weight matrices is clipped to the absolute value of one.

The training performance is shown in Fig. 16.6. The training error and test accuracy saturate in tens of epochs, which verify the effectiveness of the BPTT algorithm. Advanced algorithms taking the weight distribution into account are proposed in the recent work [29], showing advantages of revealing the weight uncertainty and temporal credit assignments underlying the network output behavior, in both engineering tasks and computational cognition tasks.

## References

1. R.N. Wulfram Gerstner, W.M. Kistler, L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition* (Cambridge University Press, Cambridge, 2014)
2. N. Brunel, J. Comput. Neurosci. **8**(3), 183 (2000)
3. N. Fourcaud, N. Brunel, Neural Comput. **14**(9), 2057 (2002)
4. R. Engelken, F. Farkhooi, D. Hansel, C. van Vreeswijk, F. Wolf, F1000Research **5**(2043) (2016). https://doi.org/10.12688/f1000research.9144.1
5. M. Helias, D. Dahmen (2019). arXiv:1901.10416
6. H. Sompolinsky, A. Crisanti, H.J. Sommers, Phys. Rev. Lett. **61**(3), 259 (1988)
7. J.A. Hertz, Y. Roudi, P. Sollich, J. Phys. A **50**(3), 33001 (2017)
8. P.C. Martin, E.D. Siggia, H.A. Rose, Phys. Rev. A **8**, 423 (1973)

9. C. De Dominicis, Phys. Rev. B **18**, 4913 (1978)
10. H. Sompolinsky, A. Zippelius, Phys. Rev. B **25**, 6860 (1982)
11. A. Crisanti, H. Sompolinksy, Phys. Rev. E **98**(6), 62120 (2018)
12. N. Bertschinger, T. Natschlager, Neural Computation **16**(7), 1413 (2004)
13. H. Jaeger, H. Haas, Science **304**(5667), 78 (2004)
14. W. Maass, P. Joshi, E.D. Sontag, PLOS Comput. Biol. **3**(1), 15 (2007)
15. D. Sussillo, L. Abbott, Neuron **63**(4), 544 (2009)
16. D.V. Buonomano, W. Maass, Nature Rev. Neurosci. **10**(2), 113 (2009)
17. G. Wainrib, J. Touboul, Phys. Rev. Lett. **110**(11), 118101 (2013)
18. C. van Vreeswijk, H. Sompolinsky, Science **274**(5293), 1724 (1996)
19. C. van Vreeswijk, H. Sompolinsky, Neural Comput. **10**(6), 1321 (1998)
20. M. Okun, I. Lampl, Nature Neurosci. **11**(5), 535 (2008)
21. M.N. Shadlen, W.T. Newsome, J. Neurosci. **18**(10), 3870 (1998)
22. A. Renart, J.D.L. Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, K.D. Harris, Science **327**(5965), 587 (2010)
23. Y. Ahmadian, K.D. Miller (2019). arXiv:1908.10101
24. R. Laje, D.V. Buonomano, Nature Neurosci. **16**(7), 925 (2013)
25. D. Thalmeier, M. Uhlmann, H.J. Kappen, R.M. Memmesheimer, PLOS Comput. Biol. **12**(6), 29 (2016)
26. W. Nicola, C. Clopath, Nature Commun. **8**(1), 2208 (2017)
27. R. Pascanu, T. Mikolov, Y. Bengio, in *Proceedings of The 30th International Conference on Machine Learning* (2013), pp. 1310–1318
28. C. Tallec, Y. Ollivier (2017). arXiv:1705.08209
29. W. Zou, C. Li, H. Huang (2021). arXiv: 2102.03740

# Chapter 17
# Statistical Mechanics of Random Matrices


Random matrix theory plays an important role in neural network research, especially in characterizing the stability of the collective behavior of the network, which is related to phase transitions (e.g., in the Hopfield model), or dynamical modes in recurrent neural networks. The asymptotic properties of random matrices whose entries follow a pre-defined distribution can be connected to the thermodynamic behavior in statistical physics (Edwards and Jones in J. Phys. A: Math. Gen. 9(10):1595, 1976 [1]; Sommers et al. in Phys. Rev. Lett. 60:1895, 1988 [2]). Therefore, the eigenspectrum of a random matrix ensemble can be reduced to calculating the free energy function of a two-body spin interaction model, in which the spin could be continuous. In this chapter, we will introduce statistical mechanics calculations of the spectral density for random matrix ensembles, and its connection to neural networks (Rajan and Abbott in Phys. Rev. Lett. 97(18):188104, 2006 [3]; Rogers et al. in Phys. Rev. E 78(3):31116, 2008 [4]).

## 17.1 Spectral Density

Considering an $N \times N$ symmetric matrix $\mathbf{J}$, whose entries follow a distribution, e.g., a Gaussian with zero mean and variance $g/N$, we then write the spectral density intuitively:

$$\rho(\lambda) = \frac{1}{N} \sum_i \delta(\lambda - \lambda_i), \tag{17.1}$$

where $\lambda_i$ is a specific eigenvalue of the matrix $\mathbf{J}$. To transform the definition to an analytic form, we first introduce the Sokhotski–Plemelj formula:

$$\lim_{\epsilon \to 0^+} \frac{1}{x \pm \mathrm{i}\epsilon} = \mathrm{P}\left(\frac{1}{x}\right) \mp \mathrm{i}\pi\delta(x), \tag{17.2}$$



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_17

where the Cauchy principal value integral is defined as

$$\mathrm{P}\int_{-\infty}^{\infty}\frac{\varphi(x)dx}{x}\equiv\lim_{\delta\to 0^+}\left\{\int_{-\infty}^{-\delta}\frac{\varphi(x)dx}{x}+\int_{\delta}^{\infty}\frac{\varphi(x)dx}{x}\right\},\tag{17.3}$$

where $\varphi(x)$ is a real-valued test function. If we further define a resolvent $(\lambda\mathbb{I}-\mathbf{J})^{-1}$, then we have the following Green's function:

$$G_N(\lambda)=\frac{1}{N}\mathrm{Tr}(\lambda\mathbb{I}-\mathbf{J})^{-1}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda-\lambda_i}.\tag{17.4}$$

Note that the resolvent can be disorder-averaged by using the replica method, and is thus a very useful quantity for the analysis of random matrix. When $N\to\infty$, we have

$$\mathbb{E}G_N(\lambda)=G_\infty(\lambda)=\int dx'\frac{\rho(x')}{\lambda-x'}\tag{17.5}$$

where $\mathbb{E}$ means the expectation with respect to the random realization of the matrix. This is the so-called Stieltjes transform of $\rho(x)$. We assume here that the spectral density of a random matrix almost surely converges in the large-$N$ limit. By using Eq. (17.2), one can then prove that

$$\rho(x)=\frac{1}{\pi}\lim_{\epsilon\to 0^+}\mathrm{Im}\,G_\infty(x-\mathrm{i}\epsilon)=\frac{1}{\pi}\mathbb{E}\lim_{\epsilon\to 0^+}\mathrm{Im}\,G_N(x-\mathrm{i}\epsilon),\tag{17.6}$$

where we have used the representation of the delta function:

$$\delta(x)=\frac{1}{\pi}\lim_{\epsilon\to 0^+}\frac{\epsilon}{x^2+\epsilon^2}.\tag{17.7}$$

To sum up, we have the following analytic form to retrieve the spectral density $\rho(\lambda)$:

$$\begin{aligned}\rho(\lambda)&=\frac{1}{N\pi}\mathbb{E}\lim_{\epsilon\to 0^+}\mathrm{Im}\sum_{i=1}^{N}\frac{1}{\lambda-\mathrm{i}\epsilon-\lambda_i}\\&=\frac{1}{N\pi}\mathbb{E}\lim_{\epsilon\to 0^+}\mathrm{Im}\sum_{i=1}^{N}\frac{\partial}{\partial\lambda}\ln(\lambda-\mathrm{i}\epsilon-\lambda_i)\\&=\frac{1}{N\pi}\mathbb{E}\lim_{\epsilon\to 0^+}\mathrm{Im}\frac{\partial}{\partial\lambda}\ln\left[\prod_{i=1}^{N}(\lambda-\mathrm{i}\epsilon-\lambda_i)\right]\\&=\frac{1}{N\pi}\mathbb{E}\lim_{\epsilon\to 0^+}\mathrm{Im}\frac{\partial}{\partial\lambda}\ln\left[\det\left((\lambda-\mathrm{i}\epsilon)\mathbb{I}-\mathbf{J}\right)\right],\end{aligned}\tag{17.8}$$

where we have used the relation—

$$\det(\lambda\mathbb{I} - \mathbf{J}) = \prod_{i=1}^{N}(\lambda - \lambda_i). \tag{17.9}$$

Because we are interested in a random matrix ensemble, the spectral density must be averaged over the statistics of the ensemble. Therefore, we first transform the spectral density in a form of the partition function, thanks to the fact that the determinant can be transformed back to its integral representation, using either multivariate Fresnel integral or Gaussian integral:

$$\frac{1}{\sqrt{\det(\mathbf{A})}} = \left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^N \int_{-\infty}^{\infty} \prod_i dx_i \exp\left[-\mathrm{i}\sum_{i,j} x_i \mathbf{A}_{ij} x_j\right], \tag{17.10}$$

or,

$$\frac{1}{\sqrt{\det(\mathbf{A})}} = \frac{1}{(2\pi)^{\frac{N}{2}}} \int d^N\mathbf{x} \exp\left[-\frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{2}\right], \tag{17.11}$$

which holds as long as $\mathbf{A}$ is positive definite. As we apply in other chapters, the disorder average can be carried out by the replica method:

$$\begin{aligned}\rho(\lambda) &= -\frac{2}{N\pi}\mathbb{E}\lim_{\epsilon\to 0^+}\mathrm{Im}\frac{\partial}{\partial\lambda}\lim_{n\to 0}\frac{1}{n}\left\{\left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^{Nn}\int_{-\infty}^{\infty}\prod_{i,\alpha}dx_i^\alpha\left[\exp\left(-\mathrm{i}\sum_{i,j,\alpha}x_i^\alpha(\lambda\delta_{ij}-J_{ij})x_j^\alpha\right)\right]-1\right\}\\ &:= -\frac{2}{N\pi}\mathbb{E}\lim_{\epsilon\to 0^+}\mathrm{Im}\frac{\partial}{\partial\lambda}\lim_{n\to 0}\frac{1}{n}\left(Z^n-1\right)\\ &= -\frac{2}{N\pi}\lim_{\epsilon\to 0^+}\mathrm{Im}\frac{\partial}{\partial\lambda}\lim_{n\to 0}\frac{\ln(\mathbb{E}Z^n)}{n},\end{aligned} \tag{17.12}$$

where $\alpha$ indicates the replica index, $\lambda$ in the partition function should be replaced by $\lambda - \mathrm{i}\epsilon$ [according to Eq. (17.8)], and

$$Z^n := \left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^{Nn}\int_{-\infty}^{\infty}\prod_{i,\alpha}dx_i^\alpha\exp\left(-\mathrm{i}\sum_{i,j,\alpha}x_i^\alpha(\lambda\delta_{ij}-J_{ij})x_j^\alpha\right). \tag{17.13}$$

Taken together, we obtain a two-body interaction Hamiltonian for estimating the spectral density of random matrix ensembles, which can be read off from the definition of the replicated partition function.

## 17.2 Replica Method and Semi-circle Law

We assume that the random matrix statistics is specified by $J_{ij} \sim \mathcal{N}(0, J^2/N)$ and $J_{ij} = J_{ji}$ (the so-called Wigner ensemble). In the following derivation, we set $\lambda$ to the one with a small imaginary part $\epsilon$. In other words, $\lambda$ in the following expressions should be replaced by $\lambda - \mathrm{i}\epsilon$. According to the previous section, we have

$$
\begin{aligned}
\mathbb{E}Z^n &= \mathbb{E}\left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^{Nn} \int_{-\infty}^{\infty} \prod_{i,\alpha} dx_i^\alpha \exp\left(-\mathrm{i}\lambda \sum_{i,\alpha} (x_i^\alpha)^2 + \mathrm{i} \sum_{i,j,\alpha} x_i^\alpha x_j^\alpha J_{ij}\right) \\
&= \left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^{Nn} \int_{-\infty}^{\infty} \prod_{i,\alpha} dx_i^\alpha \exp\left(-\mathrm{i}\lambda \sum_{i,\alpha} (x_i^\alpha)^2\right) \prod_{i<j} \mathbb{E} \exp\left(2\mathrm{i} \sum_{\alpha} x_i^\alpha x_j^\alpha J_{ij}\right) \\
&= \left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^{Nn} \int_{-\infty}^{\infty} \prod_{i,\alpha} dx_i^\alpha \exp\left[-\mathrm{i}\lambda \sum_{i,\alpha} (x_i^\alpha)^2\right] \exp\left[-\frac{J^2}{N} \sum_{i\neq j} \left(\sum_{\alpha} x_i^\alpha x_j^\alpha\right)^2\right] \\
&= \left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^{Nn} \int_{-\infty}^{\infty} \prod_{i,\alpha} dx_i^\alpha \exp\left[-\mathrm{i}\lambda \sum_{i,\alpha} (x_i^\alpha)^2\right] \exp\left[\frac{J^2}{N} \sum_{i} \left(\sum_{\alpha} (x_i^\alpha)^2\right)^2\right] \\
&\quad \times \exp\left[-\frac{J^2}{N} \sum_{i,j} \left(\sum_{\alpha} x_i^\alpha x_j^\alpha\right)^2\right],
\end{aligned}
\tag{17.14}
$$

where we have used the integral identity: $\mathbb{E}_z e^{az} = e^{\sigma^2 a^2/2}$ for $z \sim \mathcal{N}(0, \sigma^2)$. We have neglected the diagonal entries of $\mathbf{J}$. Note that

$$
\begin{aligned}
\frac{J^2}{N} \sum_{i,j} \left(\sum_{\alpha} x_i^\alpha x_j^\alpha\right)^2 &= \frac{J^2}{N} \sum_{i,j} \sum_{\alpha,\beta} x_i^\alpha x_j^\alpha x_i^\beta x_j^\beta \\
&= \frac{J^2}{N} \sum_{\alpha} \left(\sum_{i} (x_i^\alpha)^2\right)^2 + \frac{J^2}{N} \sum_{\alpha\neq\beta} \sum_{i,j} x_i^\alpha x_j^\alpha x_i^\beta x_j^\beta,
\end{aligned}
\tag{17.15}
$$

where we need only retain the terms of $\alpha = \beta$. In other words, $\{x_i^\alpha\}$ are in the replica space mutually orthogonal. Moreover, the remaining term $\frac{J^2}{N} \sum_i \left(\sum_\alpha (x_i^\alpha)^2\right)^2$ is of the order $n^2$ and thus neglected as well. Consequently, we have

$$
\mathbb{E}Z^n = \left\{\left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^{N} \int_{-\infty}^{\infty} \prod_{i} dx_i \exp\left[-\mathrm{i}\lambda \sum_{i} (x_i)^2 - \frac{J^2}{N} \left(\sum_{i} (x_i)^2\right)^2\right]\right\}^n.
\tag{17.16}
$$

By applying the Hubbard–Stratonovich transform (let $a = 2J^2/N$):

$$
\exp\left(-\frac{ax^2}{2}\right) = \int \frac{\mathrm{d}s}{\sqrt{2\pi a}} \exp\left(-\frac{s^2}{2a} \pm \mathrm{i}xs\right),
\tag{17.17}
$$

we have

$$\exp\left[\frac{-J^2}{N}\left(\sum_i (x_i)^2\right)^2\right] = \left(\frac{N}{2\pi}\right)^{1/2} \frac{1}{(2J^2)^{1/2}} \int_{-\infty}^{\infty} ds \exp\left(\frac{-N}{4J^2}s^2\right) \exp\left(-\mathrm{i}s \sum_i (x_i)^2\right)$$
$$= \left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{(2J^2)^{1/2}} \int_{-\infty}^{\infty} ds \exp\left(\frac{-\lambda^2}{4J^2}Ns^2\right) \exp\left(-\mathrm{i}\lambda s \sum_i (x_i)^2\right), \tag{17.18}$$

where we have rescaled the variable $s \to \lambda s$. It then follows that

$$\begin{aligned}
\mathbb{E}Z^n &= \left\{\left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^N \left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{(2J^2)^{1/2}} \int_{-\infty}^{\infty} ds \prod_i dx_i \exp\left[-\mathrm{i}\lambda(1+s)(x_i)^2\right] \exp\left[-\frac{\lambda^2 N s^2}{4J^2}\right]\right\}^n \\
&= \left\{\left[\frac{e^{\frac{\mathrm{i}\pi}{4}}}{\pi^{\frac{1}{2}}}\right]^N \left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{(2J^2)^{1/2}} \int_{-\infty}^{\infty} ds \left[e^{-\mathrm{i}\pi/4}\sqrt{\frac{\pi}{\lambda(1+s)}}\right]^N \exp\left[-\frac{\lambda^2 N s^2}{4J^2}\right]\right\}^n \\
&= \left\{\left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{(2J^2)^{1/2}} \int_{-\infty}^{\infty} ds \ [\lambda(1+s)]^{-\frac{N}{2}} \exp\left[-\frac{\lambda^2 N s^2}{4J^2}\right]\right\}^n \\
&= \left\{\left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{(2J^2)^{1/2}} \int_{-\infty}^{\infty} ds \ \exp\left[-\frac{N}{2}\ln(\lambda(1+s))\right] \exp\left[-\frac{\lambda^2 N s^2}{4J^2}\right]\right\}^n \\
&= \left\{\left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{(2J^2)^{1/2}} \exp\left[-\frac{N}{2}\ln\lambda\right] \int_{-\infty}^{\infty} ds \ \exp\left[-\frac{N}{2}\ln(1+s) - \frac{\lambda^2 N s^2}{4J^2}\right]\right\}^n \\
&:= \left\{\left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{(2J^2)^{1/2}} \exp\left[-\frac{N}{2}\ln\lambda\right] \int_{-\infty}^{\infty} ds \ \exp\left[-Ng(s)\right]\right\}^n,
\end{aligned} \tag{17.19}$$

where we have defined $g(s)$ as

$$g(s) = \frac{1}{2}\ln(1+s) + \frac{\lambda^2 s^2}{4J^2}. \tag{17.20}$$

Now we use the Laplace method due to the large-$N$ limit, i.e.,

$$\int_{-\infty}^{\infty} ds \ \exp\left[-Ng(s)\right] \approx \exp\left[-Ng(s^\star)\right] \sqrt{\frac{2\pi}{N|g''(s^\star)|}}, \tag{17.21}$$

where $g'(s^\star) = 0$. Thus, the saddle-point solution $s^\star$ is obtained by solving the following equation:

$$s^2 + s + \frac{J^2}{\lambda^2} = 0. \tag{17.22}$$

A solution is given by

$$s = \frac{1}{2}\left[-1 \pm \sqrt{\Delta}\right] = \frac{1}{2}\left[-1 \pm \sqrt{1 - \frac{4J^2}{\lambda^2}}\right], \tag{17.23}$$

where $\Delta := 1 - \frac{4J^2}{\lambda^2}$. For $|\lambda| < 2J$, i.e., $\Delta < 0$, the saddle points occur at

$$s_0^{\pm} = \frac{1}{2}\left[-1 \pm \mathrm{i}\sqrt{\frac{4J^2}{\lambda^2} - 1}\right], \tag{17.24}$$

while for $|\lambda| > 2J$, we have

$$s^{\pm} = \frac{1}{2}\left[-1 \pm \sqrt{1 - \frac{4J^2}{\lambda^2}}\right]. \tag{17.25}$$

Only the solution $s_0^-$ or $s^-$ can make the saddle-point approximation reasonable (considering the contour integration in the complex plane, see details of proof in Ref. [1]). Hence,

$$\begin{aligned}\mathbb{E}Z^n &= \left\{\left(\frac{N}{2\pi}\right)^{1/2} \frac{\lambda}{\left(2J^2\right)^{1/2}} \exp\left[-\frac{N}{2}\ln\lambda\right] \exp\left[-Ng(s_0^-)\right]\right\}^n \\ &= \left(\frac{N}{2\pi}\right)^{n/2} \frac{\lambda^n}{\left(2J^2\right)^{n/2}} \exp\left[-\frac{nN}{2}\ln\lambda\right] \exp\left[-nNg(s_0^-)\right].\end{aligned} \tag{17.26}$$

Finally, we take the limit $n \to 0$, and get

$$\begin{aligned}\mathcal{F}(\lambda) \equiv \lim_{n\to 0} \frac{\ln(\mathbb{E}Z^n)}{n} &= \lim_{n\to 0} \frac{\partial}{\partial n} \ln(\mathbb{E}Z^n) \\ &= \lim_{n\to 0} \frac{\partial}{\partial n}\left[\frac{n}{2}\ln\left(\frac{N\lambda^2}{4\pi J^2}\right) - \frac{nN}{2}\ln\lambda - nNg(s_0^-)\right] \\ &= \frac{1}{2}\ln\left(\frac{N\lambda^2}{4\pi J^2}\right) - \frac{N}{2}\ln\lambda - Ng(s_0^-) \\ &\simeq -\frac{N}{2}\ln\lambda - Ng(s_0^-) \quad , \text{as } N \to \infty.\end{aligned} \tag{17.27}$$

Note that in the above equation, $\lambda$ should be replaced by $\lambda - \mathrm{i}\epsilon$. We thus derive the eigenvalue spectrum when $|\lambda| < 2J$:

$$\begin{aligned}\rho(\lambda) &= -\frac{2}{N\pi} \lim_{\epsilon\to 0^+} \mathrm{Im} \frac{\partial}{\partial\lambda} \mathcal{F}(\lambda - \mathrm{i}\epsilon) \\ &= \frac{1}{\pi} \lim_{\epsilon\to 0^+} \mathrm{Im} \frac{\partial}{\partial\lambda}\left[\ln(\lambda - \mathrm{i}\epsilon) + \ln(1 + s_0^-(\lambda - \mathrm{i}\epsilon)) + \frac{\lambda^2 (s_0^-(\lambda - \mathrm{i}\epsilon))^2}{2J^2}\right] \\ &= \frac{1}{2\pi J^2}\sqrt{4J^2 - \lambda^2}.\end{aligned} \tag{17.28}$$

For $|\lambda| > 2J$, $g(s^-)$ is real, and thus $\rho(\lambda) = 0$. A comparison of the theory to the numerical eigenvalue spectrum is shown in Fig. 17.1. Interested readers can check if a simple annealed approximation of the free energy leads to the same result.

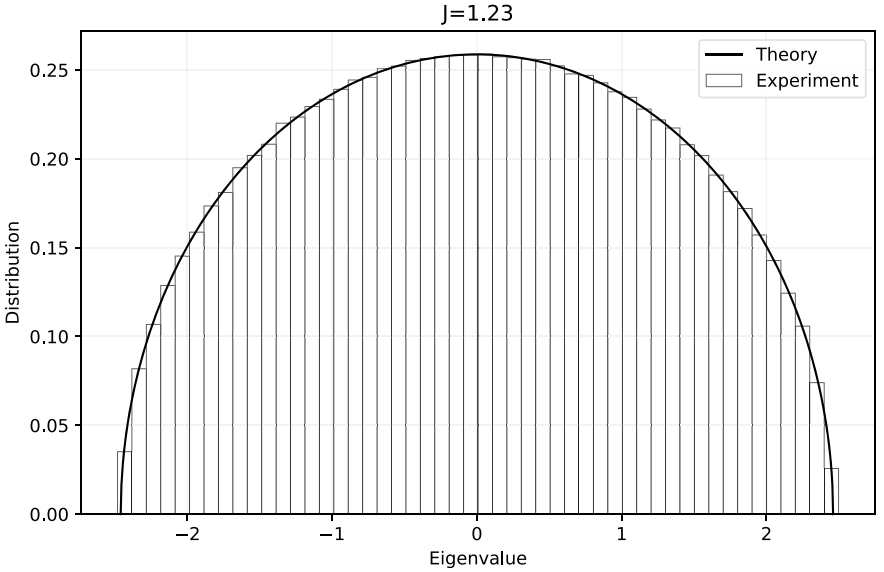



**Fig. 17.1** The semi-circle law. Comparison between theory and numerical simulations are shown. The random matrix has a size $N = 1000$. 100 random realizations of the random matrix are considered

## 17.3 Cavity Approach and Marchenko–Pastur Law

In this section, we introduce the cavity approach to estimate the asymptotic spectral density. Given the partition function defined as in the previous section, we can write the spectral density:

$$\rho_{\mathbf{A}}(\lambda) = -\frac{2}{\pi N} \lim_{\epsilon \to 0^+} \mathrm{Im} \left( \frac{\partial}{\partial z} \ln Z_{\mathbf{A}}(z) \right)_{z=\lambda - \mathrm{i}\epsilon}, \tag{17.29}$$

where

$$Z_{\mathbf{A}}(z) = \int \left( \prod_i \frac{dx_i}{\sqrt{2\pi}} \right) \exp \left[ -\frac{1}{2} \sum_{i,j}^{N} x_i (z\mathbb{I} - \mathbf{A})_{ij} x_j \right]. \tag{17.30}$$

Thus, we derive the corresponding Hamiltonian:

$$\mathcal{H}_{\mathbf{A}}(\mathbf{x}, z) = \frac{1}{2} \sum_{i,j}^{N} x_i (z\mathbb{I} - \mathbf{A})_{ij} x_j, \tag{17.31}$$

where we consider $A_{ii} = 0 \, \forall i$. Finally, the spectral density can be calculated as

$$\rho_{\mathbf{A}}(\lambda) = \lim_{\epsilon \to 0^+} \frac{1}{\pi N} \sum_i \mathrm{Im}\left[\langle x_i^2 \rangle_z\right]_{z=\lambda - \mathrm{i}\epsilon}, \tag{17.32}$$

where $\langle \cdots \rangle$ denotes the average w.r.t the Boltzmman distribution under the Hamiltonian.

The cavity iteration can be written out explicitly as follows:

$$P_{i\to j}(x_i) = \frac{e^{-z x_i^2/2}}{Z_{i\to j}} \int d\mathbf{x}_{\partial i \setminus j} \exp\left( x_i \sum_{k \in \partial i \setminus j} A_{ik} x_k \right) \prod_{k \in \partial i \setminus j} P_{k\to i}(x_k). \tag{17.33}$$

However, the integral is hard to work out analytically. We make a Gaussian ansätz [4], without a rigorous proof. More precisely,

$$P_{i\to j}(x) = \frac{1}{\sqrt{2\pi \Delta_{i\to j}}} e^{-\frac{x^2}{2\Delta_{i\to j}}}. \tag{17.34}$$

Then, the cavity iteration is transformed to

$$\Delta_{i\to j}(z) = \frac{1}{z - \sum_{k\in\partial i\setminus j} A_{ik}^2 \Delta_{k\to i}(z)}. \tag{17.35}$$

The variance for the marginal probability can be derived immediately as

$$\Delta_i(z) = \frac{1}{z - \sum_{k\in\partial i} A_{ik}^2 \Delta_{k\to i}(z)}. \tag{17.36}$$

According to Eq. (17.32), one has

$$\rho_{\mathbf{A}}(\lambda) = \lim_{\epsilon \to 0^+} \frac{1}{\pi N} \sum_i \mathrm{Im}\left[\langle \Delta_i(z) \rangle_z\right]_{z=\lambda - \mathrm{i}\epsilon}. \tag{17.37}$$

The Hamiltonian can be defined on a tree-like pairwise-interaction graph if the matrix $\mathbf{A}$ is sparse, and the cavity approximation is valid. We also assume that $A_{ij} \sim \mathcal{N}(0, J^2/c)$, where $c$ is the average connectivity of the graph. In the large connectivity limit (denoted as $c \to \infty$), we can define

$$\Delta = \lim_{c\to\infty} \frac{1}{c} \sum_i \Delta_i. \tag{17.38}$$

Therefore, we have

$$\lim_{c\to\infty} \sum_k A_{ik}^2 \Delta_{k\to i} = J^2 \Delta, \tag{17.39}$$

where we have used $\Delta_{k\to i} \simeq \Delta_k$. Hence, we derive that

$$\Delta = \frac{1}{z - J^2\Delta}, \tag{17.40}$$

which gives the semi-circle law as derived by the replica method.

Now we consider the random matrix is the interaction matrix of the Hopfield model:

$$A_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu, \tag{17.41}$$

where $\boldsymbol{\xi}$ is an $N \times P$ matrix with entries subject to the binomial distribution with equal probabilities for two peaks. This matrix ensemble is called the Wishart ensemble. The Hamiltonian can be written in the form of

$$\mathcal{H}_{\mathbf{A}}(\mathbf{x}, z) = \frac{z}{2}\sum_i x_i^2 - \frac{1}{2}\sum_\mu \Big[m_\mu(\mathbf{x}_{\partial\mu})\Big]^2, \tag{17.42}$$

where the auxiliary quantity $\boldsymbol{m}$ is defined by

$$m_\mu(\mathbf{x}_{\partial\mu}) = \frac{1}{\sqrt{N}}\sum_{i\in\partial\mu} \xi_i^\mu x_i. \tag{17.43}$$

The belief propagation equation reads as follows:

$$P_{i\to\mu}(x_i) \propto e^{-zx_i^2/2}\int d\mathbf{m}_{\partial i\setminus\mu} \exp\left[\frac{1}{2}\sum_{\nu\in\partial i\setminus\mu}\Big(m_{\nu\to i} + \frac{\xi_i^\nu x_i}{\sqrt{N}}\Big)^2\right]\prod_{\nu\in\partial i\setminus\mu} Q_{\nu\to i}(m_{\nu\to i}), \tag{17.44a}$$

$$Q_{\nu\to i}(m_{\nu\to i}) \propto \int d\mathbf{x}_{\partial\nu\setminus i}\,\delta\Big(m_{\nu\to i} - \frac{1}{\sqrt{N}}\sum_{j\in\partial\nu\setminus i}\xi_j^\nu x_j\Big)\prod_{j\in\partial\nu\setminus i} P_{j\to\nu}(x_j). \tag{17.44b}$$

Assuming that the cavity distribution $P_{i\to\mu}(x_i) \sim \mathcal{N}(0, \Delta_{i\to\mu})$ and $Q_{\mu\to i}(m_\mu) \sim \mathcal{N}(0, \Gamma_{\mu\to i})$, we can derive the recursive equation for these two variances [4]. The marginal one is given by

$$\Delta_i(z) = \frac{1}{z - \frac{1}{N}\sum_\mu (\xi_i^\mu)^2 \frac{1}{1-\Gamma_{\mu\to i}}}, \tag{17.45}$$

where $(\xi_i^\mu)^2 = 1$. In the large-$N$ limit, we can further define

$$\Delta = \lim_{N\to\infty}\frac{1}{N}\sum_i \Delta_i. \tag{17.46}$$

Equation (17.45) suggests that

$$\frac{1}{\Delta} - z + \alpha \frac{1}{1-\Delta} = 0, \tag{17.47}$$

where $\alpha = P/N$, and we have used $\Gamma_{\mu\to i}(z) = \frac{1}{N}\sum_{j\in\partial\mu\setminus i}(\xi_j^\mu)^2 \Delta_{j\to\mu}(z) \simeq \Delta$. This equation is exactly the saddle-point equation if we use the replica method to compute the disorder average [5]. Next, we show how to solve this equation to get the Marcenko–Pastur law [6], an asymptotic spectral density for sample covariance matrix.

Solving Eq. (17.47) (we change the notation $z$ to $\lambda$ in the following derivation), we have

$$\Delta = \frac{-(\alpha - \lambda_\epsilon - 1) \pm \sqrt{(\alpha - \lambda_\epsilon - 1)^2 - 4\lambda_\epsilon}}{2\lambda_\epsilon}, \tag{17.48}$$

where $\lambda_\epsilon = \lambda - \mathrm{i}\epsilon$. We then have

$$\mathrm{Im}\,\Delta = \frac{-(\alpha - \lambda_\epsilon - 1)\epsilon}{2(\lambda^2 + \epsilon^2)} \pm \mathrm{Im}\,\frac{(\lambda + \mathrm{i}\epsilon)\sqrt{(\alpha - \lambda_\epsilon - 1)^2 - 4\lambda + 4\mathrm{i}\epsilon}}{2(\lambda^2 + \epsilon^2)}. \tag{17.49}$$

We now have to solve the following equation of complex values:

$$\sqrt{c + \mathrm{i}d} = s + \mathrm{i}t, \tag{17.50}$$

where $c = (\alpha - \lambda_\epsilon - 1)^2 - 4\lambda$. From Eq. (17.49), we have a solution:

$$t = \pm \frac{4\epsilon}{\sqrt{2(\sqrt{c^2 + 16\epsilon^2} + c)}}, \tag{17.51}$$

and,

$$s = \pm \frac{\sqrt{\sqrt{c^2 + 16\epsilon^2} + c}}{\sqrt{2}}. \tag{17.52}$$

Then, we have

$$\begin{aligned}
\lim_{\epsilon\to 0^+} \mathrm{Im}\,\Delta &= \lim_{\epsilon\to 0^+} \frac{-(\alpha - \lambda_\epsilon - 1)\epsilon}{2(\lambda^2 + \epsilon^2)} \pm \lim_{\epsilon\to 0^+} \mathrm{Im}\,\frac{(\lambda + \mathrm{i}\epsilon)\sqrt{(\alpha - \lambda_\epsilon - 1)^2 - 4\lambda + 4\mathrm{i}\epsilon}}{2(\lambda^2 + \epsilon^2)} \\
&= \frac{-(\alpha - \lambda_\epsilon - 1)}{2}\pi\delta(\lambda) \pm \lim_{\epsilon\to 0^+} \mathrm{Im}\,\frac{(\lambda + \mathrm{i}\epsilon)(s + \mathrm{i}t)}{2(\lambda^2 + \epsilon^2)} \\
&= \frac{-(\alpha - 1)}{2}\pi\delta(\lambda) \pm \lim_{\epsilon\to 0^+} \frac{\lambda t}{2(\lambda^2 + \epsilon^2)} \pm \lim_{\epsilon\to 0^+} \frac{\epsilon s}{2(\lambda^2 + \epsilon^2)},
\end{aligned} \tag{17.53}$$

where Eq. (17.7) is used to get the delta function.

Finally, if $\alpha \geq 1$, the first and third terms in Eq. (17.53) cancel, and the second term gives rise to

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\epsilon\to 0^+} \mathrm{Im}\,\Delta = \frac{\sqrt{|c|}}{2\pi\lambda}\mathbb{I}_{\lambda_-,\lambda_+}(\lambda). \tag{17.54}$$

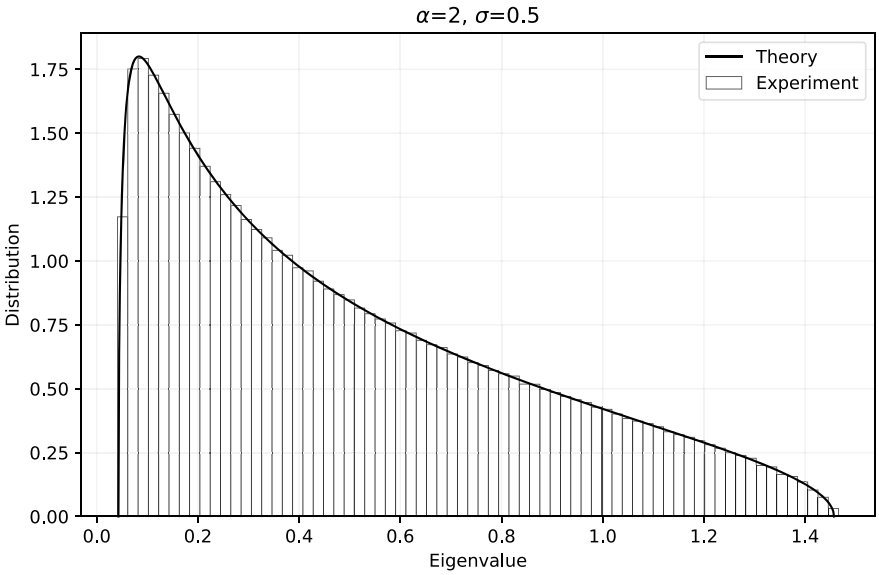



**Fig. 17.2** The eigenvalue distribution of the covariance matrix **A**. The pattern entry follows a Gaussian distribution with zero mean and variance $\sigma^2$ (see details in the work [7]). Here we set $\alpha = 2$, $\sigma = 0.5$ and $N = 1000$. 100 random instances of the matrix ensemble are considered

The indicator function $\mathbb{I}_{\lambda_-,\lambda_+}(\lambda)$ reports one if $\lambda$ falls within the interval $[\lambda_-, \lambda_+]$, and zero otherwise, which guarantees that the value of $c$ is negative. In the case of $\alpha < 1$, the first and third terms in Eq. (17.53) give rise to a delta peak. We thus have

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathrm{Im}\, \Delta = \frac{\sqrt{|c|}}{2\pi\lambda} \mathbb{I}_{\lambda_-,\lambda_+}(\lambda) + (1-\alpha)\delta(\lambda), \tag{17.55}$$

where $\lambda_\pm = (1 \pm \sqrt{\alpha})^2$. In sum, we derive the Marcenko–Pastur law:

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathrm{Im}\, \Delta = \frac{\sqrt{(\lambda-\lambda_-)(\lambda_+-\lambda)}}{2\pi\lambda} \mathbb{I}_{\lambda_-,\lambda_+}(\lambda) + (1-\alpha)\delta(\lambda)\mathbb{I}_{0,1}(\alpha). \tag{17.56}$$

A comparison between theory and numerical results is shown in Fig. 17.2. We finally remark that the Marcenko–Pastur law could also be derived using the annealed or quenched computation of the replica method [5, 7].

## 17.4 Spectral Densities of Random Asymmetric Matrices

If the matrix **A** is a non-Hermitian matrix (e.g., asymmetric interaction matrix in recurrent neural networks), the eigenvalues are complex. Then the spectral density must be defined as follows:

$$\rho_{\mathbf{A}}(z) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - \mathrm{Re}(\lambda_i))\delta(y - \mathrm{Im}(\lambda_i)). \tag{17.57}$$

We first define a complex variable $z = x + \mathrm{i}y$, where $x$ and $y$ are real. $z^*$ denotes the complex conjugate of $z$. The Wirtinger derivatives can be defined as follows:

$$\partial_z = \frac{1}{2}(\partial_x - \mathrm{i}\partial_y), \tag{17.58a}$$

$$\partial_{z^*} = \frac{1}{2}(\partial_x + \mathrm{i}\partial_y). \tag{17.58b}$$

The Wirtinger derivative has the following properties: $\partial_z(z) = \partial_{z^*}(z^*) = 1$, and $\partial_z(z^*) = \partial_{z^*}(z) = 0$. We then have the following identity:

$$\partial_{z^*}(1/z) = \partial_z(1/z^*) = \pi\delta(x)\delta(y). \tag{17.59}$$

To interpret the above mathematical identity, we imagine a two-dimensional classical electrostatic field ($E$) generated by a unit charge. The Gauss law implies that

$$2\pi r E = 1/\varepsilon_0, \tag{17.60}$$

where $r$ denotes the distance from the charge on the plane, and $\varepsilon_0$ is a physical constant. Therefore, the Gauss law reads as well

$$\nabla \cdot \frac{\mathbf{r}}{r^2} = 2\pi\delta(\mathbf{r}), \tag{17.61}$$

which leads to the Poisson equation $\nabla^2 \ln(|\mathbf{r}|) = 2\pi\delta(\mathbf{r})$. $\nabla$ denotes the gradient operator. The Laplacian operator $\nabla^2$ is defined by $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ (in two-dimensional space). Therefore, by defining $E_x = \mathrm{Re}(1/z)$ and $E_y = -\mathrm{Im}(1/z)$, Eq. (17.61) turns out to be Eq. (17.59).

Then, we have

$$\begin{aligned}\rho_{\mathbf{A}}(z) &= \frac{1}{\pi}\partial_{z^*}\frac{1}{N}\sum_i \frac{1}{z - \lambda_i} \\ &= \frac{1}{\pi}\partial_{z^*}\frac{1}{N}\mathrm{Tr}(z\mathbb{I}_N - \mathbf{A})^{-1}.\end{aligned} \tag{17.62}$$

Therefore, in the large $N$ limit, the empirical density of different random realizations of the matrix converges to the average density:

$$\rho(z) = \frac{1}{\pi}\partial_{z^*}\left\langle\frac{1}{N}\mathrm{Tr}(z\mathbb{I}_N - \mathbf{A})^{-1}\right\rangle, \tag{17.63}$$

where the average Green function $G(z) = \left\langle \frac{1}{N} \operatorname{Tr}\left[(z\mathbb{I}_N - \mathbf{A})^{-1}\right]\right\rangle$ where the average is done w.r.t the random realizations of $\mathbf{A}$, and $\mathbb{I}_N$ denotes an $N \times N$ identity matrix.

We further define the average Green function on the complex plane:

$$G(z) = \left\langle \frac{1}{N} \operatorname{Tr}\left[(z\mathbb{I}_N - \mathbf{A})^{-1}\right]\right\rangle = \left\langle \frac{1}{N} \sum_{\lambda} \frac{1}{z-\lambda} \right\rangle = \int d^2\lambda \frac{\rho(\lambda)}{z-\lambda}, \tag{17.64}$$

where $\int d^2\lambda \bullet$ indicates an integral over the complex plane, we have used $\operatorname{Tr}[\mathbf{P}^{-1}\mathbf{A}\mathbf{P}] = \operatorname{Tr}\mathbf{A}$ for any invertible $\mathbf{P}$, and if $\mathbf{A}$ is invertible, $\lambda(\mathbf{A}^{-1}) = 1/\lambda(\mathbf{A})$, where $\lambda$ indicates the matrix's eigenvalues. Note also that adding a diagonal matrix $z\mathbb{I}_N$ to $\mathbf{A}$ just increases each eigenvalue of $\mathbf{A}$ by $z$. Considering a contour integral around a closed path $\mathcal{C}$, we can use the residue theorem to prove that

$$\frac{1}{2\pi \mathrm{i}} \int_{\mathcal{C}} dz G(z) = \int_{\mathcal{S}} d^2\lambda \rho(\lambda), \tag{17.65}$$

where $\mathcal{S}$ indicates the region bounded by the closed path (the eigenvalue is not on the path). In addition, a complex form of Gauss law implies that

$$\frac{1}{2\pi} \int_{\mathcal{S}} d^2 z \left[\frac{\partial G}{\partial x} + \mathrm{i}\frac{\partial G}{\partial y}\right] = \int_{\mathcal{S}} d^2\lambda \rho(\lambda), \tag{17.66}$$

which requires that

$$\frac{\partial \operatorname{Re} G}{\partial x} - \frac{\partial \operatorname{Im} G}{\partial y} = 2\pi\rho, \tag{17.67a}$$

$$\frac{\partial \operatorname{Im} G}{\partial x} + \frac{\partial \operatorname{Re} G}{\partial y} = 0. \tag{17.67b}$$

Equation (17.67a) implies that $E_x = 2 \operatorname{Re} G$ and $E_y = -2 \operatorname{Im} G$ [2], relating the distribution of an electric charge to the electric field, while Eq. (17.67b) corresponds to $\nabla \times \mathbf{E} = 0$, suggesting a scalar potential $\Phi$, i.e., $\mathbf{E} = -\nabla\Phi$. Therefore, we have the Poisson equation for the two-dimensional electrostatics $\nabla^2\Phi = -4\pi\rho$, corresponding to the spectral density problem. Altogether, the spectral density for a non-Hermitian random matrix can be obtained through finding a potential:

$$\Phi(z, z^*) = -\frac{1}{N} \langle \ln \det\left[(z^*\mathbb{I}_N - \mathbf{A}^{\mathrm{T}})(z\mathbb{I}_N - \mathbf{A})\right]\rangle, \tag{17.68}$$

which can be shown to be consistent with Eq. (17.64) and the electrostatics representation, using $\det(\mathbf{AB}) = \det\mathbf{A}\det\mathbf{B}$ and $\det(\mathbf{A}^{\mathrm{T}}) = \det\mathbf{A}$. The Green function is given by

$$G(z) = \partial_z \Phi(z, z^*), \tag{17.69a}$$

$$G^*(z) = \partial_{z^*} \Phi(z, z^*). \tag{17.69b}$$

Equation (17.67a) also implies that $\mathrm{Re}\left[\partial_{z^*} G\right] = 2\pi\rho$, which means that if $G$ is only the function of $z$ in a region, the eigenvalue density must be zero in that region. Therefore, the non-zero spectral density is related to the non-holomorphic behavior of Green's function [8]. This property can be used to determine the boundary separating holomorphic and non-holomorphic solutions of the spectral problem.

To sum up, we have

$$\rho(x, y) = \frac{-1}{4\pi N} \nabla^2 \langle \ln \det \left[ (z^* \mathbb{I}_N - \mathbf{A}^{\mathrm{T}})(z\mathbb{I}_N - \mathbf{A}) \right] \rangle, \tag{17.70}$$

where $\nabla^2 = 4\partial_z \partial_{z^*}$. The determinant can be transformed to the Gaussian integral representation over complex variable. Then, the spectral density problem is reduced to a disorder system composed of a large number of interacting particles. Thus, the cavity approach or replica method can be applied to derive the analytic form of the asymptotic spectral density. In some specific problems, $n$ replicas decouple in the thermodynamic limit, then an annealed calculation can be performed [2].

$$\Phi(z) = \frac{1}{N} \ln \left\langle \int \prod_i \frac{d^2 z_i}{\pi} \exp\Big( -\epsilon \sum_i |z_i|^2 - \sum_{i,j,k} z_i^* (z^* \delta_{ik} - \mathbf{A}_{ik}^{\mathrm{T}})(z\delta_{kj} - \mathbf{A}_{kj}) z_j \Big) \right\rangle, \tag{17.71}$$

where a positive infinitesimal quantity $\epsilon$ is usually introduced to avoid singularities caused by $z{=}\lambda_i$. More precisely, $\Phi(z){=} -\frac{1}{N} \langle \ln \det \left[ (z^* \mathbb{I}_N - \mathbf{A}^{\mathrm{T}})(z\mathbb{I}_N - \mathbf{A}) + \epsilon \mathbb{I}_N \right] \rangle$. For example, the derivation of the (Girko's) circular law for the fully asymmetric random matrix falls within this class [9]. When the matrix is dense, diagrammatic expansion techniques (Feynman diagrams) are also useful for deriving the asymptotic spectrum of non-Hermitian matrices [10, 11]. This method can also derive the eigen-spectrum of the E-I interaction neural populations, where cell types are distinguished and thus Dale's law is respected [3].

To apply the diagrammatic method, we introduce the following Hermitization process. We first construct a $2N \times 2N$ Hermitian matrix [12, 13]:

$$H = \begin{pmatrix} 0 & \mathbf{A} - z\mathbb{I}_N \\ \mathbf{A}^\dagger - z^* \mathbb{I}_N & 0 \end{pmatrix}. \tag{17.72}$$

$\mathbf{A}^\dagger$ denotes the transpose conjugate of a matrix. The Green function reads then

$$\mathcal{G}(\omega) = \frac{1}{\omega \mathbb{I}_N - H} = \begin{pmatrix} \mathcal{G}^{11} & \mathcal{G}^{12} \\ \mathcal{G}^{21} & \mathcal{G}^{22} \end{pmatrix}, \tag{17.73}$$

in a block structure, and $\omega$ is a constant. Note that each of four blocks (e.g., $\mathcal{G}^{11}$) is an $N \times N$ matrix. We then have the following matrix identity:

$$\begin{pmatrix} \omega\mathbb{I}_N & z\mathbb{I}_N - \mathbf{A} \\ z^*\mathbb{I}_N - \mathbf{A}^\dagger & \omega\mathbb{I}_N \end{pmatrix} \begin{pmatrix} \mathcal{G}^{11} & \mathcal{G}^{12} \\ \mathcal{G}^{21} & \mathcal{G}^{22} \end{pmatrix} = \begin{pmatrix} \mathbb{I}_N & 0 \\ 0 & \mathbb{I}_N \end{pmatrix}. \tag{17.74}$$

Inspecting the upper left block, we have immediately

$$\omega\mathcal{G}^{11} + (z\mathbb{I}_N - \mathbf{A})\mathcal{G}^{21} = \mathbb{I}_N. \tag{17.75}$$

Then $G(z) = (z\mathbb{I}_N - \mathbf{A})^{-1} = \mathcal{G}^{21}$ when $\omega = 0$, and thus the spectral density is obtained by

$$\rho(x, y) = \frac{1}{\pi}\partial_{z^*}\left\langle \frac{1}{N}\operatorname{Tr}\mathcal{G}^{21}(\omega = 0, z, z^*)\right\rangle. \tag{17.76}$$

To compute $\mathcal{G}(\omega)$, we first write $\omega\mathbb{I}_N - H = \mathcal{G}_0^{-1} - \mathcal{J}$, where

$$\mathcal{G}_0^{-1} = \begin{pmatrix} \omega\mathbb{I}_N & z\mathbb{I}_N \\ z^*\mathbb{I}_N & \omega\mathbb{I}_N \end{pmatrix} \text{ and } \mathcal{J} = \begin{pmatrix} 0 & \mathbf{A} \\ \mathbf{A}^\dagger & 0 \end{pmatrix}. \tag{17.77}$$

We assume that $\mathcal{J}$ has a zero mean, and thus $\mathcal{G}_0$ is just $\mathcal{G}$ with $\mathbf{A} = 0$. $\mathcal{G}$ can then be expanded in $\mathcal{G}_0$ as follows:

$$\mathcal{G} = \left\langle \frac{1}{\mathcal{G}_0^{-1} - \mathcal{J}} \right\rangle = \sum_{n=0}^{\infty} \mathcal{G}_0 \langle (\mathcal{J}\mathcal{G}_0)^n \rangle. \tag{17.78}$$

By applying the Wick contraction (supposed that the distribution over $\mathcal{J}$ is Gaussian), and noting that only planar diagrams remain, one can re-organize the above expansion as follows:

$$\mathcal{G} = \sum_{n=0}^{\infty} \mathcal{G}_0(\Sigma\mathcal{G}_0)^n = \frac{1}{\mathcal{G}_0^{-1} - \Sigma}, \tag{17.79}$$

where the self-energy matrix $\Sigma$ is introduced as $\Sigma = \langle \mathcal{J}\mathcal{G}\mathcal{J} \rangle$ (i.e., Dyson–Schwinger relation), which is the sum of all contributions coming from all one-particle irreducible diagrams [13]. The key equation $\mathcal{G}^{-1} = \mathcal{G}_0^{-1} - \Sigma[\mathcal{G}]$ is also called Dyson equation in physics. The Dyson equation gives the self-consistent way to compute the spectra density of random non-Hermitian matrices.

## References


1. S.F. Edwards, R.C. Jones, J. Phys. A: Math. Gen. **9**(10), 1595 (1976)
2. H.J. Sommers, A. Crisanti, H. Sompolinsky, Y. Stein, Phys. Rev. Lett. **60**, 1895 (1988)
3. K. Rajan, L.F. Abbott, Phys. Rev. Lett. **97**(18), 188104 (2006)
4. T. Rogers, I.P. Castillo, R. Kuhn, K. Takeda, Phys. Rev. E **78**(3), 31116 (2008)
5. J. Zhou, Z. Jiang, T. Hou, Z. Chen, K.Y.M. Wong, H. Huang (2021). arXiv:2103.14324
6. V. Marcenko, L. Pastur, Math. USSR-Sb. **1**, 457 (1967)

7. J. Zhou, H. Huang, Phys. Rev. E **103**, 012315 (2021)
8. F. Haake, F. Izrailev, N. Lehmann, D. Saher, H.J. Sommers, Zeitschrift fur Physik B Condensed Matter **88**(3), 359 (1992)
9. V. Girko, Theory Probab. Appl. **29**, 694 (1985)
10. Y. Wei, Phys. Rev. E **85**(6), 66116 (2012)
11. Y. Ahmadian, F. Fumarola, K.D. Miller, Phys. Rev. E **91**(1), 12820 (2015)
12. R. Janik, M. Nowak, G. Papp, I. Zahed, Nucl. Phys. **501**, 603 (1997)
13. J. Feinberg, A. Zee, Nucl. Phys. **504**(3), 579 (1997)

# Chapter 18
# Perspectives


This book introduces basics of statistical mechanics and its relationship to current theoretical studies of neural networks (including deep neural networks), mainly focusing on an overview of main tools to deal with non-linearity intrinsic in neural computation, and detailed illustration of deep insights provided by physics analysis in a few typical examples (most of them were proposed by the authors' own works). In Marr's viewpoint [1], understanding a neural system can be divided into three levels: computation (which task the brain solves), algorithms (how the brain solves the task, i.e., information processing level) and implementation (neural circuit level). In artificial neural networks, researchers build a naive mapping of the first two levels into a toy model level (especially for theoretical studies). Even the first two levels are now turned into ideas to solve challenging real-world problems, driven by deep learning [2, 3]. However, biological details are also being incorporated into standard models of neural networks [4–6]. Indeed, neuroscience researches about the biological mechanisms of perception, cognition, memory and action have already provided a variety of fruitful insights inspiring the empirical/scientific studies of artificial neural networks, which in turn inspires the neuroscience researchers to design mechanistic models to understand the brain [7, 8]. Therefore, it is promising to integrate physics, statistics, computer science, psychology, neuroscience and engineering to provide theoretical predictions, and reveal inner workings of deep (biological) networks and even intelligence.

The goal of providing a unified framework for neural computation is very challenging. Due to re-boosted interests in neural networks, there appear a lot of important yet unsolved scientific questions. We shall list some of them below, and provide our personal viewpoints towards a statistical mechanics theory of these fundamental questions.

*Representation Learning*

From a viewpoint of unsupervised learning aiming at extracting statistical regularities from raw data, one can ask what a good representation is and how the meaningful



H. Huang, *Statistical Mechanics of Neural Networks*,
https://doi.org/10.1007/978-981-16-7570-6_18

representation is achieved. We have not yet satisfied answers for these questions. A promising argument is that entangled manifold at earlier layers of a deep hierarchy is gradually disentangled into linearly separable features at output layers [9–13]. The manifold perspective is also promising in system neuroscience studies of associative learning by separating overlapping patterns of neural activities [14]. A coherent theory of manifold transformation is still lacking, prohibiting us from an understanding of which key network parameters control the geometry of manifold, and even affects the learning process, for which there may exist a few factors having their origin from biological contexts, e.g., normalization, attention, homeostatic control [15, 16]. Another argument from information-theoretic viewpoints demonstrates that the input information is maximally compressed into a hidden representation whose task-related information should be maximally retrieved at the output layers, according to the information bottleneck theory [17]. However, this theory is still under debate [18].

*Generalization*

Intelligence can be considered to some extent as the ability of generalization, especially given very few examples for learning. Therefore, generalization is also a hot topic in current studies of deep learning. Traditional statistical learning theory claims that over-fitting effects should be strong when the number of examples is much less than the number of parameters to learn, which thereby could not explain the current success of deep learning. A promising perspective is to study the causal connection between the loss landscape and the generalization properties [19–21]. For a single layered perceptron, a statistical mechanics theory can be systematically derived [22, 23]. In contrast to the classical bias-variance trade-off (U-shaped curve of the test error versus increasing model complexity) [24], deep learning achieves the state-of-the-art performance in the over-parameterized regime [20, 25]. However, for a multi-layered perceptron model, how to provide an analytic argument about the over-fitting effects versus different parameterization regimes (e.g., under-, over- and even super-parameterization) becomes a non-trivial task [26]. Furthermore, clarifying which of lazy-learning (or neural tangent kernel limit) and feature-learning (or mean-field limit) may explain the success of deep supervised learning remains open and challenging [27, 28].

*Adversarial vulnerability*

Adversarial examples are defined with those inputs with human-imperceptible modifications yet leading to unexpected errors in a deep learning decision-making system. This adversarial vulnerability of deep neural networks poses a significant challenge in the practical applications of both real-world problems and scientific studies. In physics, systems with a huge number of degrees of freedom is able to be captured by a low-dimensional macroscopic description. In this sense, a low-dimensional explanation with a few order parameters about the origin of the adversarial vulnerability is lacking so far. Although some recent efforts were devoted to this direction [29–31], more exciting results are expected in near future works.

*Continual Learning*

A biological brain is good at adapting the acquired knowledge from similar tasks to domains of new tasks, even if only handful examples are available in the new

domain. In contrast, neural networks are in general poor at the multi-task learning, although impressive progresses have been achieved in recent years. For example, during learning, a diagonal Fisher information term is computed to measure importances of weights (then a rapid change is not allowed) for previous tasks [32]. A later refinement by allowing synapses accumulating task relevant information over time was also proposed [33]. More machine learning techniques to reduce the catastrophic forgetting effects are summarized in the review [34]. However, we still do not know the exact mechanisms for mitigating the catastrophic forgetting effects in a principled way, which calls for theoretical studies of deep learning in terms of adaptation to domain-shift training, i.e., connection weights trained in a solution to one task are transformed to benefit learning on a related task. Furthermore, it remains unclear how the related knowledge contained in a source task can be transferred effectively to boost the performance in a target task, suppose that both tasks share common semantics in the latent space.

*Causal Learning*

Deep learning is criticized as being nothing but a fancy curve-fitting tool, making a naive association between inputs and outputs. In other words, this tool could not distinguish correlation from causation. A human-like AI must be good at retrieving causal relationship among feature components in sensory inputs, thereby carving relevant information from a sea of irrelevant noise [35, 36]. Therefore, understanding cause and effects in deep learning systems is particularly important for a next-generation artificial intelligence. The question whether the current deep learning algorithm is able to do causal reasoning remains elusive. Consequently, designing theory-accessible toy models becomes a key to address this question, although it would be very challenging to identify causes for observed effects by simple physics equations.

*Internal Model of the Brain*

The brain is argued to learn to build an internal model of the outside world, reflected by spontaneous neural activities as a reservoir for computing (e.g., sampling) [37]. The agreement between spontaneous activity and stimulus-evoked one increases during development especially for natural stimuli [38], while the spontaneous activity outlines the regime of evoked neural responses [39]. The stimuli were shown to carve a clustered neural space [40]. Then, an interesting question is what the spontaneous neural space looks like, and how it dynamically evolves. Furthermore, how sensory inputs combined with the ongoing cortical activity to determine animal behavior remains open and challenging. On the other hand, reinforcement learning was used to build world models of structured environments [41]. In reinforcement learning, data are used to drive actions which are evaluated based on reward signals the agent receives from the environment. It is thus interesting which kind of internal models the agent establishes through learning from interactions with the environments. Moreover, a recent work shows a connection between the reinforcement learning and statistical physics [42], which suggesting that a statistical mechanics theory could potentially be established to understand the internal model, with potential impacts on studying neural computations in the brain.

*Theory of Consciousness*

One of the most controversial question is the origin of consciousness—whether the consciousness is an emergent behavior of a highly heterogeneous neural circuit with various carefully designed regions (e.g., a total of $10^{14}$ connections for human brain). The subjectivity of the conscious experience is in contradiction with the objectivity of a scientific explanation. According to Damasio's model [43], the ability to identify one's self in the world and its relationship with the world is considered a central characteristic of conscious state. Whether a machine algorithm can achieve the self-awareness remains elusive. There are other two major cognitive theories of consciousness: one is the global workspace framework [44], which relates consciousness to the widespread and sustained propagation of cortical neural activities by demonstrating that consciousness arises from information-processing computations of specialized modules. The other is the integrated information theory that provides a quantitative characterization of conscious state by integrated information [45]. Both theories follow a top-down approach, which is in stark contrast to the statistical mechanics approach following a bottom-up manner building the bridge from microscopic interactions to macroscopic behavior. These hypotheses are still under intensive criticism despite some cognitive experiments they can explain [46]. From an information-theoretic argument, the conscious state may require a diverse range of configurations of interactions between brain networks, which can be linked to the entropy concept in physics [47]. The large entropy leads to optimal segregation and integration of information. Taken together, whether the consciousness can be created from an interaction of local dynamics within complex neural substrate is still unsolved [48]. A statistical mechanics theory, if possible, is always promising in the sense that one can ask theoretical predictions from just a few physics parameters.

To sum up, in this chapter, we provide some naive thoughts about some fundamental important questions related to neural networks, for which building a good theory[1] is far from being completed. In physics, we have the principle of least action, from which we can deduce the classical mechanics or electrodynamics laws. We are not sure whether in physics of neural networks (and even the brain) there exists a general principle that can be expressed in a concise form of mathematics. Readers interested in the interplay between physics theory and neural computations are encouraged to promote advances along these exciting yet risky research lines.

## References

1. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, Cambridge, 1982)
2. J. Schmidhuber, Neural Netw. **61**, 85 (2015)
3. Y. LeCun, Y. Bengio, G. Hinton, Nature **521**(7553), 436 (2015)
4. L.F. Abbott, B. DePasquale, R.M. Memmesheimer, Nat. Neurosci. **19**(3), 350 (2016)

[1] "Nothing is more practical than a good theory"—Ludwig Boltzmann.

5. B.A. Richards, T.P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R.P. Costa, A. de Berker, S. Ganguli, C.J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G.W. Lindsay, K.D. Miller, R. Naud, C.C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A.C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, K.P. Kording, Nat. Neurosci. **22**(11), 1761 (2019)
6. T.P. Lillicrap, A. Santoro, L. Marris, C.J. Akerman, G. Hinton, Nature Reviews Neuroscience **21**(6), 335 (2020)
7. L.K. Yamins Daniel, J.J. DiCarlo, Nat. Neurosci. **19**(3), 356 (2016)
8. A. Saxe, S. Nelli, C. Summerfield, Nat. Rev. Neurosci. **22**, 55 (2020)
9. J.J. DiCarlo, D.D. Cox, Trends Cognit. Sci. **11**(8), 333 (2007)
10. Y. Bengio, A. Courville, P. Vincent, IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798 (2013)
11. P.P. Brahma, D. Wu, Y. She, IEEE Trans. Neural Netw. Learn. Syst. **27**(10), 1997 (2016)
12. H. Huang, Phys. Rev. E **98**, 062313 (2018)
13. U. Cohen, S. Chung, D.D. Lee, H. Sompolinsky, Nat. Commun. **11**(1), 1 (2020)
14. N.A. Cayco-Gajic, R.A. Silver, Neuron **101**(4), 584 (2019)
15. G.G. Turrigiano, S.B. Nelson, Nat. Rev. Neurosci. **5**(2), 97 (2004)
16. J.H. Reynolds, D.J. Heeger, Neuron **61**(2), 168 (2009)
17. R. Shwartz-Ziv, N. Tishby (2017). arXiv:1703.00810
18. A.M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B.D. Tracey, D.D. Cox, J. Stat. Mech.: Theory Exper. **2019**(12), 124020 (2019)
19. C. Baldassi, C. Borgs, J.T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, Proc. Natl. Acad. Sci. **113**(48), E7655 (2016)
20. S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, M. Wyart, J. Phys. A: Math. Theor. **52**, 474001 (2019)
21. W. Zou, H. Huang (2020). arXiv:2007.08093
22. G. Gyorgyi, Phys. Rev. A **41**(12), 7097 (1990)
23. H. Sompolinsky, N. Tishby, H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990)
24. P. Mehta, M. Bukov, C.H. Wang, A.G. Day, C. Richardson, C.K. Fisher, D.J. Schwab, Phys. Rep. **810**, 1 (2019)
25. M. Belkin, D. Hsu, S. Ma, S. Mandal, Proc. Natl. Acad. Sci. U.S.A. **116**(32), 15849 (2019)
26. B. Adlam, J. Pennington, in *ICML 2020: 37th International Conference on Machine Learning* (2020)
27. A. Jacot, F. Gabriel, C. Hongler, in *Advances in Neural Information Processing Systems*, vol. 31 (2018), pp. 8571–8580
28. C. Fang, H. Dong, T. Zhang, Proc. IEEE 1–21 (2021). https://doi.org/10.1109/JPROC.2020.3048020
29. R. Kenway (2018). arXiv:1803.06111
30. L. Bortolussi, G. Sanguinetti (2018). arXiv:1811.03571
31. Z. Jiang, J. Zhou, H. Huang, Chin. Phys. B **30**, 048702 (2021)
32. J. Kirkpatrick, R. Pascanu, N.C. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Proce. Natl. Acad. Sci. U.S.A. **114**(13), 3521 (2017)
33. F. Zenke, B. Poole, S. Ganguli, in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 (2017), pp. 3987–3995
34. G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, S. Wermter, Neural Netw. **113**, 54 (2019)
35. B. Schölkopf (2019). arXiv:1911.10500
36. J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, New York, 2018)
37. D.L. Ringach, Curr.Opin. Neurobiol. **19**(4), 439 (2009)
38. P. Berkes, G. Orban, M. Lengyel, J. Fiser, Science **331**, 83 (2011)
39. L. Artur, B. Peter, K.D. Harris, Neuron **62**, 413 (2009)
40. H. Huang, T. Toyoizumi, Phys. Rev. E **93**, 062416 (2016)
41. D. Ha, J. Schmidhuber (2018). arXiv: 1803.10122
42. J. Rahme, R.P. Adams (2019). arXiv:1906.10228

43. A. Damasio, Nature **413**, 781 (2001)
44. S. Dehaene, M. Kerszberg, J.P. Changeux, Proc. Natl. Acad. Sci. U.S.A. **95**(24), 14529 (1998)
45. G. Tononi, BMC Neurosci. **5**(1), 42 (2004)
46. C. Koch, M. Massimini, M. Boly, G. Tononi, Nat. Rev. Neurosci. **17**(5), 307 (2016)
47. R.G. Erra, D.M. Mateos, R. Wennberg, J.L.P. Velazquez, Phys. Rev. E **94**(5), 52402 (2016)
48. P. Krauss, A. Maier (2020). arXiv:2003.14132