Replica method and asymptotic equivalence

Minh-Toan Nguyen

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : EEATS - Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)
Spécialité : Signal Image Parole Télécoms
Unité de recherche : Grenoble Images Parole Signal Automatique

La méthode des répliques et l'équivalence asymptotique

Replica method and asymptotic equivalence

Présentée par :

Minh-Toan NGUYEN

Direction de thèse :

Romain COUILLET PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Directeur de thèse

Rapporteurs :

Marc LELARGE
DIRECTEUR DE RECHERCHE, INRIA CENTRE DE PARIS
Walid HACHEM
DIRECTEUR DE RECHERCHE, CNRS ILE-DE-FRANCE VILLEJUIF

Thèse soutenue publiquement le **5 décembre 2023**, devant le jury composé de :

Olivier MICHEL, PROFESSEUR DES UNIVERSITES, GRENOBLE INP	Président
Romain COUILLET, PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Directeur de thèse
Marc LELARGE, DIRECTEUR DE RECHERCHE, INRIA CENTRE DE PARIS	Rapporteur
Walid HACHEM, DIRECTEUR DE RECHERCHE, CNRS ILE-DE-FRANCE VILLEJUIF	Rapporteur
Abla KAMMOUN, SENIOR SCIENTIST, Université des sciences et technologies du roi Abdallah	Examinatrice
Florent KRZAKALA, FULL PROFESSOR, Ecole Polytechnique Fédérale de Lausanne	Examinateur

THÈSE
Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale: EEATS
Spécialité: Signal Image Parole Telecoms
Unité de recherche: Grenoble Images Parole Signal Automatique (GIPSA)

Replica method and asymptotic equivalence

Présentée par: Minh-Toan NGUYEN

Thèse soutenue publiquement le 5 December 2023, devant le jury composé de:

Romain COUILLET	*Professeur, UGA*	Directeur de thèse
Walid HACHEM	*Directeur de recherche, CNRS*	Rapporteur
Marc LELARGE	*Directeur de recherche, INRIA*	Rapporteur
Abla KAMMOUN	*Senior Scientist, KAUST*	Examinatrice
Florent KRZAKALA	*Professeur, EPFL*	Examinateur
Olivier MICHEL	*Professeur, UGA*	Président
Steeve ZOZOR	*Directeur de recherche, CNRS*	Invité

Replica method and asymptotic equivalence

Acknowledgments

I would like to thank my advisor Romain Couillet for his guidance, his support and confidence in me throughout my research journey. I rarely need to wait more than 15 minutes for an email response from him, no matter if it is a research question or an administrative issue. He is also a considerate advisor who anticipates my needs and provides me with suggestions before I even ask, especially when he granted me three-month extension so that I can write up the thesis at my own pace. From trying to recover his results, I developed ideas and improved my technical skills. I am especially grateful of him for introducing me to the method of deterministic equivalent, which largely inspired this thesis.

I would like to thank the two reporters, Marc Lelarge and Walid Hachem, for their feedback and their patience in listening to me explaining my work.

I would like to thank the jury for their time, consideration and interest in my work.

I would like to thank Nicolas Le Bihan and Gerard Besson for serving as my Comité de Suivi Individuel during my PhD.

Thanks to GIPSA-lab and LIG for providing me with the necessary resources during my PhD. I am grateful to MIAI for financial support.

Lastly, I would like to thank my friends and family for their care, love and support.

Contents

Notations

- $A \leftrightarrow \bar{A}$: A is asymptotically equivalent to $\bar{A}$
- $\langle x, y \rangle$: inner product of $x, y \in \mathbb{K}^n$, which is $x^\top y = \sum_i x_i y_i$ for $\mathbb{K} = \mathbb{R}$ and $x^\dagger y = \sum_i \bar{x}_i y_i$ for $\mathbb{K} = \mathbb{C}$.
- $\lfloor x \rfloor$: largest $n \in \mathbb{Z}$ such that $n \leq x$.
- $x^{\downarrow n} := x(x-1)\dots(x-n+1)$
- $x^{\uparrow n} := x(x+1)\dots(x+n-1)$
- $[n] := \{1, \dots, n\}$
- $\mathbb{N} := \{0, 1, 2, \dots\}$ (set of natural numbers), $\mathbb{N}_+ = \{1, 2, \dots\}$.
- D_x: the diagonal matrix with diagonal elements given by the vector x

Résumé

La méthode des répliques est un outil préféré des physiciens pour étudier les grands systèmes désordonnés. Le terme «réplique» vient du fait que la méthode implique des copies indépendantes du système, autrement dit les «répliques». Bien qu'elle soit très peu rigoureuse, la méthode des répliques peut résoudre des problèmes difficiles dans divers domaines : théorie des matrices aléatoires, optimisation convexe, optimisation combinatoire, inférence bayésienne, etc. La méthode a été utilisée avec succès pour analyser des modèles théoriques en communication, traitement du signal et apprentissage automatique.

L'équivalence asymptotique est omniprésente dans les systèmes de grand dimension. L'un des exemples les plus simples de ce phénomène est qu'un vecteur choisi uniformément dans une sphère de grande dimension se comporte comme des variables aléatoires gaussiennes indépendantes. Ce phénomène est également évident dans la méthode des équivalents déterministes en théorie des matrices aléatoires, la méthode objective en optimisation combinatoire et le CGMT (Convex Gaussian min-max theorem en anglais) en optimisation convexe aléatoire. Ces méthodes montrent que le système étudié se comporte asymptotiquement comme un système plus simple. En conséquence, de nombreux calculs difficiles sur le système d'origine peuvent être effectués plus facilement sur le système équivalent.

Dans cette thèse, nous montrons comment calculer l'équivalent asymptotique d'un système désordonné avec les répliques. Ceci est différent de la méthode des répliques habituelle, qui calcule une quantité à la fois. Après avoir développé un cadre théorique, nous calculerons les équivalents déterministes de certaines matrices aléatoires, dériverons formellement le CGMT et un nouveau résultat similaire pour les matrices orthogonales aléatoires, et montrerons comment la symétrie des répliques implique que certains problèmes d'inférence en grande dimension se comportent comme des canaux gaussiens indépendants. De plus, nous montrerons comment des structures de probabilité telles que la distribution de Poisson-Dirichlet et la coalescente de Bolthausen-Sznitman émergent directement de l'ansatz de Parisi.

Abstract

Replica method is a favorite tool of physicists for studying large disordered systems. The term 'replica' comes from the fact that the method involves independent copies of the system, also referred to as 'replicas'. Despite being highly non-rigorous, replica method can solve difficult problems across various domains: random matrix theory, convex optimization, combinatorial optimization, Bayesian inference, etc. The method has been successfully used to analyze theoretical models in communication, signal processing and machine learning.

Asymptotic equivalence is ubiquitous in high dimensional systems. One of the simplest examples of this phenomenon is that a vector chosen uniformly from a high dimensional sphere behaves like independent Gaussian random variables. This phenomenon is also evident in the method of deterministic equivalents in random matrix theory, the objective method in combinatorial optimization, and the CGMT (convex Gaussian min-max theorem) in random convex optimization. These methods shows that the system under study behaves asymptotically like a simpler system. As a result, many difficult computations on the original system can be done more easily on the equivalent system.

In this thesis, we show how to compute the asymptotic equivalent of a disordered system with replicas. This is different from the usual replica method, which computes one quantity at a time. After developing some theoretical framework, we will compute the deterministic equivalents of some random matrices, formally derive the CGMT and a similar new result for random orthogonal matrices, and show how replica symmetry implies that some high dimensional inference problems behave like independent Gaussian channels. Moreover, we will show how probability structures such as the Poisson-Dirichlet distribution and the Bolthausen-Sznitman coalescent directly emerge from Parisi's replica symmetry breaking ansatz.

Chapter 1

Introduction

The phenomenon of asymptotic equivalence is illustrated by various examples in Section 1.1. The replica method will be presented in Section 1.3 after basic terminologies of disordered systems being introduced in Section 1.2. The rest of the chapter presents the structure and highlights the contributions of the thesis.

1.1 Asymptotic equivalence through examples

We give here examples of simple yet non-trivial problems that illustrate the asymptotic equivalence phenomenon.

Example 1.1. (*GOE matrix and spikes*) Let X be a square matrix of size n with independent standard Gaussian entries and consider $A = \frac{X+X^\top}{\sqrt{2n}}$. The random matrix A is said to be sampled from the *Gaussian Orthogonal Ensemble* (GOE). For large values of n ($\sim 10^3$ for example), the distribution of the eigenvalues of A is very close to a semi-circular shape described by the following density

$$\mu(dx) = \frac{1}{2\pi}\sqrt{4-x^2}1_{[-2,2]}(x)dx$$

Now let us consider matrix $Y = A + \lambda uu^\top$, where u is an arbitrary unit vector and $\lambda \geq 0$. Plotting the eigenvalues of Y, we observe that as λ increases from 0 to 1, the spectral density of Y remains the same, i.e. consisting of a single semi-circular bulk. However, as soon as $\lambda > 1$, the largest eigenvalue $\hat{\lambda}$ separates from the bulk with asymptotic position

$$\hat{\lambda} \to \lambda + \frac{1}{\lambda}, \quad n \to \infty.$$

Moreover, let $\hat{u}$ be the unit eigenvector associated with $\hat{\lambda}$, then

$$\langle u, \hat{u}\rangle^2 \to 1 - \frac{1}{\lambda^2}.$$

If we consider A as noises and u as an unknown signal to be estimated from observing $\lambda uu^\top + A$, then the top eigenvector $\hat{u}$ gives an estimate that is correlated with u when $\lambda > 1$.

In general, for a large random symmetric or Hermitian matrix M, we are mainly interested in the following problems

- Computing the *limiting spectral density* of M, i.e. the deterministic shape that emerges when plotting the distribution of its eigenvalues.
- Let us consider a *spiked model* M, i.e. a low-rank perturbation of some random matrix model with known behavior. In most cases, its limiting spectral density is the same as the non-perturbed model and there exists a phase transition in which isolated eigenvalues appear. We are interested in the threshold of this phase transition, the position of the spikes as well as the corresponding eigenvectors and how they are related to the perturbation.

The key to answer these questions is the resolvent matrix

$$Q(z) = (M - zI)^{-1}$$

where $z \in \mathbb{C}_+$. In many cases, $Q(z)$ behaves like a deterministic matrix $\bar{Q}(z)$, called the *deterministic equivalent* of $Q(z)$. Computing $\bar{Q}(z)$ will be our main objective. The method of deterministic equivalents, which goes beyond the classic Stieltjes transform, is one of the main technical tools behind recent applications of random matrix theory in communication [24] and machine learning [27]. From the deterministic equivalent of the resolvent $Q(z)$ we can study not only the limiting behaviors of the eigenvalues of M but also its eigenvectors, which often contains information about the signal hidden behind the data of matrix M.

Example 1.2. (*A linear regression on linear model*) Consider the following simple model of linear regression studied in [38], [55], in which n data points $(x_i, y_i)_{i=1}^n$ are generated by the model

$$y_i = w_\star^\top x_i + \xi_i$$

where x_i are drawn independently from $\mathcal{N}(0, I_d)$, ξ_i are independent noises, each following $\mathcal{N}(0, \sigma^2)$. We want to estimate the hidden parameter $w_\star$, which is a unit vector in $\mathbb{R}^d$, by solving the following optimization problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^n \ell(y_i - w^\top x_i) + R(w) \tag{1.1}$$

in which the *loss function* ℓ is convex, reaching its minimum at zero and the *regularization function* R is also convex. For example we can take $\ell(x) = x^2$ and $R(w) = \lambda \|w\|_1$ for some $\lambda \geq 0$. The minimizer of the cost function, denoted as $\hat{w}$, is taken as an estimator of $w_\star$. We consider the high dimensional limit where $n, d \to \infty$ with a fixed *sampling*

ratio $\lim_{n\to\infty} n/d = \alpha \in (0,\infty)$. It turns out that in this limit the minimum cost divided by n as well as of the distance $\|\hat{w} - w_\star\|$ converge to deterministic values. By computing these values we can see how the quality of the estimator and the minimum cost depend on the noise level σ, the sampling ratio α as well as the choice of regularization method.

This as well as many other problems in random convex optimization can be tackled by the Gaussian min-max theorem (CGMT) [105], which states that the following optimization problem

$$\min_{x\in K_x} \max_{y\in K_y} x^\top W y + f(x,y),$$

where K_x, K_y are compact subsets of $\mathbb{R}^m, \mathbb{R}^n$, W is a $m \times n$ matrix with independent standard Gaussian entries, and f satisfies some convexity criteria, is equivalent in a certain sense to the problem

$$\min_{x\in K_x} \max_{y\in K_y} f(x,y) + \|x\|\langle h, y\rangle + \|y\|\langle g, x\rangle$$

in the limit $m, n \to \infty$, where the vectors g and h are independent, following the normal distributions $\mathcal{N}(0, I_m)$ and $\mathcal{N}(0, I_n)$ respectively. Combining this theorem with the method of Lagrange multipliers, we can transform the initial problem into a simplified, decoupled form that is amenable to theoretical analysis.

Example 1.3. (*Rank-1 matrix factorization*) We consider here the following model studied in [59]

$$Y = \sqrt{\frac{\lambda}{n}} x^0 x^{0\top} + Z \tag{1.2}$$

where $x^0 \in \mathbb{R}^n$ is an unknown signal with i.i.d. entries generated from a known probability distribution P_X with finite second moment and Z has independent, standard Gaussian entries, representing the noises. We want to infer the vector x^0 from the observation Y. The fundamental object of our study is the posterior $P_{x^0|Y}$ which contains all the information about x^0 that can be extracted from Y.

It turns out that $P_{x^0|Y}$ behaves like the posterior law of x^0 given

$$\tilde{Y} = \sqrt{2\lambda q_\star} x^0 + \xi, \tag{1.3}$$

where $\xi \sim \mathcal{N}(0, I_n)$ is independent of x^0, and $q_\star$ is obtained from maximizing a certain function. The value of $q_\star$ depends on the parameter λ. There exists a value λ_c such that $q_\star = 0$ if $\lambda \le \lambda_c$ and $q_\star > 0$ if $\lambda > \lambda_c$. When $q_\star = 0$, the asymptotically equivalent problem (1.3) means estimating x^0 from pure noise, which is impossible. In other words, λ_c is the threshold of the transition between unrecoverable and the recoverable phase.

Example 1.4. (*Minimal matching*) Consider the problem of assigning n jobs to n machines in a way that minimizes the total cost of doing the jobs. Suppose that the cost of

finishing job i on machine j is c_{ij}. In the simplest case, c_{ij}'s are assumed to be independent random variables with the uniform distribution on $[0,1]$. The minimal cost is given by

$$A_n = \min_{\pi} \sum c_{i\pi(i)}$$

where the minimum is taken over all permutations π of $[n]$. Using the replica method, [67] obtained

$$\lim_{n\to\infty} A_n = \frac{\pi^2}{6} \tag{1.4}$$

This result is rigorously proved in [2]. The key insight is that matching on the bipartite graph is asymptotically equivalent in some sense to matching on a randomly weighted infinite tree. The author then constructed the optimal matching on this equivalent structure and obtained (1.4) along with many other results, giving a much more detailed description of the minimal matching.

Example 1.5. (*SK model*) Sherrington-Kirkpatrick (SK) model is considered as the holy grail of the physics of disordered systems. This model considers the following probability distribution on $\{-1,1\}^n$

$$P(\sigma) \propto e^{-\beta E(\sigma)}$$

where $\beta > 0$, $\sigma = (\sigma_i) \in \{-1,1\}^n$ and

$$E(\sigma) = \frac{1}{\sqrt{n}} \sum_{i<j} J_{ij}\sigma_i\sigma_j \tag{1.5}$$

and J_{ij} are random parameters drawn independently from the law $\mathcal{N}(0,1)$.

In contrast with the previous models, the SK model has a much more complex behavior. When β is above a certain threshold, the measure P^H asymptotically behaves like a mixture of an infinite number of *pure states*, each corresponds to a probability measure in which the coordinates are asymptotically uncorrelated. The pure states can be organized into a tree-like structure that has deep connections with various mathematical objects such as the Poisson-Dirichlet distribution, the Ruelle probability cascade and the Bolthausen-Sznitman coalescent. We refer curious readers to Parisi's Nobel Lecture [83] on this topic.

Despite coming from different fields, these examples can be put into the framework of statistical physics, whose basic definitions and questions are presented in the next section.

1.2 Statistical physics and disordered systems

Let $(\mathcal{X}, \mu)$ be a measure space on which we define the following probability measure

$$P^H(dx) \propto e^{H(x)}\mu(dx). \tag{1.6}$$

where H is a function from $\mathcal{X}$ to $\mathbb{R}$ such that $\int \mu(dx)e^{H(x)}$ is finite. H is called the **Hamiltonian** of the system defined by the **Gibbs measure** P^H over the **configuration space** $\mathcal{X}$ equipped with measure μ. The **free energy** of H is defined as

$$F^H = \log \int \mu(dx)e^{H(x)}. \tag{1.7}$$

This definition of Hamiltonian and free energy follows the mathematical convention and does not match the usual physical meanings due to some differences in sign and scale. However, this does not affect the mathematical meaning of our discussion.

If H_1, H_2 are Hamiltonians on measure spaces $(\mathcal{X}_1, \mu_1)$ and $(\mathcal{X}_2, \mu_2)$, then the Hamiltonian $H_1(x_1)+H_2(x_2)$ on the measure space $(\mathcal{X}_1 \times \mathcal{X}_2, \mu_1 \times \mu_2)$ has free energy $F^{H_1}+F^{H_2}$. This simple result will be useful for computing free energy of large systems consisting of many independent components.

Note that for any constant c, $F^{H+c} = F^H + c$ and $P^{H+c} = P^H$. While H uniquely determines P^H, the reverse is not true: there are many choices of Hamiltonian leading to the same probability measure.

If $H(x)$ is generated by a probability measure over functions, then it describes a **disordered system**. By this definition, the system defined by a deterministic H is technically a disordered system. We will use $\langle\cdot\rangle$ for the expectation with respect to the Gibbs measure and $\mathbb{E}[\cdot]$ for the expectation with respect to the randomness of H. The definition of H can be parametric, i.e. $H(x) = H(x, \Theta)$ for some random parameters Θ or non-parametric, for example when $H(x)$ is a Gaussian random field. We use the word **disorders** to indicate the randomness of H. If we fix a randomly generated Hamiltonian H, then $X^1, X^2, \ldots$ i.i.d. as P^H are called **replicas** of the system.

We will be interested in the behavior of P^H in the **thermodynamic limit**, or **infinite size limit**, where the **size** n of the configuration space $\mathcal{X}$ goes to infinity. In most cases $\mathcal{X} = S^n$ for some set S, however we will not write explicitly the dependence of the model on n. Although we mainly focus on large systems, fixed-size systems are also important for us as they are the building blocks of many large systems.

For a disordered system with Hamiltonian H, we are interested in the following problems.

- To compute the leading term of F^H in the infinite size limit. The free energy is typically **self averaged**, meaning $F^H \simeq \mathbb{E}[F^H]$, so the leading term is given by $\mathbb{E}[F^H]$. For some problems, it is also important to compute the the corrections of the free energy beyond the leading term, which is not treated in this thesis.

- To describe the Gibbs measure P^H, i.e. to find a tractable system $\bar{P}$ that behaves like P^H. The system $\bar{P}$ is called the **asymptotic equivalent** of P^H. Asymptotic equivalence is a broad concept that is difficult to capture in one definition. We only give a rough description here. Two disordered systems $P, \bar{P}$ are asymptotically equivalent if

$$\mathbb{E}\int m(x)P(dx) \simeq \mathbb{E}\int m(x)\bar{P}(dx)$$

for all $m(x)$ in a class of functions wide enough to include many functions of interest.

The definition of asymptotic equivalence can be extended for Hamiltonians. Two Hamiltonians H and $\bar{H}$ are asymptotically equivalent, or $H \leftrightarrow \bar{H}$ in notation, if $P^H \leftrightarrow P^{\bar{H}}$ and $F^H \simeq F^{\bar{H}}$. Since two different Hamiltonians can give rise to the same probability measure, the second requirement is necessary.

While every probability measure can be written in the form of Gibbs measure, not all of these writings are meaningful. We give here some examples of Gibbs measures.

Example 1.6. Let $\mathcal{X} = \{-1, 1\}$ with counting measure and $H(x) = \lambda x$ for some $\lambda \in \mathbb{R}$. Then $F^H = \log(2\cosh\lambda)$

Example 1.7. Let $\mathcal{X} = \mathbb{N}$ with counting measure and $H(n) = -\beta n$ for some $\beta > 0$, then $F^H = -\log(1 - e^{-\beta})$ and P^H corresponds to the geometric random variable with parameter e^{β}.

Example 1.8. Let $\mathcal{X} = \mathbb{R}^n$ with $\mu(dx) = (2\pi)^{-n/2} dx_1 \dots dx_n$ and

$$H(x) = -\frac{1}{2} x^\top A x$$

where A is a symmetric positive definite matrix. Then

$$F^H = -\frac{1}{2} \log\det A$$

Example 1.9. Let $\mathcal{X} = \mathbb{C}^n$ with

$$\mu(dz) = \prod_{i=1}^{n} \frac{d\,\mathrm{Re}(z_i) d\,\mathrm{Im}(z_i)}{\pi}$$

and

$$H(x) = -x^\dagger A x$$

where A is a positive Hermitian matrix. Then

$$F^H = -\log\det A$$

The examples in Section 1.1 can be put into the framework of disordered systems.

- For the case of random matrices, if a large random real symmetric matrix A has bounded norm, then the resolvent $Q(z) = (A - zI)^{-1}$ is positive definite when z is less than a certain number. We can define the following disordered system

  $$P(dx) \propto e^{-\frac{1}{2} x^\top (A - zI) x} dx, \quad x \in \mathbb{R}^n,$$

  in which A plays the role of disorders. Note that this is simply the Gaussian measure $\mathcal{N}(0, Q(z))$. It turns out that in high dimensional limit, P behaves like $\mathcal{N}(0, \bar{Q}(z))$ for some deterministic $\bar{Q}(z)$ that coincides with the deterministic equivalent of $Q(z)$. The deterministic equivalent of $Q(z)$ when $z \in \mathbb{C}^+$ can be obtained by analytic continuation.

- In the case of random optimization over either continuous or discrete spaces, to study the problem of maximizing a random function $E(x)$ over some configuration space $\mathcal{X}$, we consider the disordered system with Hamiltonian $H(x) = \beta E(x)$. As $\beta \to \infty$, P^H will concentrate at the global maxima of H, and the maximum is given by $\lim_{\beta\to\infty} F^H/\beta$.
- For Bayesian inference problems where the signal x^0 is generated from a known distribution P_X, all the information about the signal that we can extract from the observation Y is contained in the posterior distribution $P(x|Y)$. This posterior distribution can be seen as a disordered system with random parameters Y. Studying this disordered systems brings insights into the inference problem.

1.3 Replica method

The replica method is based on the following identity

$$\log Z = \partial_{r=0} Z^r$$

This simple trick can be used to get rid of the logarithms in some difficult calculations, at the expense of introducing a new variable r. The variable r is then treated as if it were an integer. This makes the calculation easier but also makes it non-rigorous. In some cases, this heuristic can be viewed as first obtaining a formula for r in $\mathbb{N}$ (or a subset of $\mathbb{N}$) and then analytically continuing the result to a real or complex domain. However, this interpretation does not fit all replica computations. For example, in the replica computation for the SK model, G. Parisi worked directly with $r \to 0$ and did not derive any result for any $r \in \mathbb{N}_+$. On this, M. Talagrand also had a similar remark, saying that it is difficult to see Parisi's calculation as an extrapolation of the case $r \in \mathbb{N}_+$. In verbatim, in page 2 of [100] he wrote "... *it seems very difficult to justify this value as an extrapolation of the case* $a \in \mathbb{N}^*$".

Let us consider a quick application of the replica trick. Suppose $Z \sim \mathcal{N}(0,1)$ and we want to calculate $\mathbb{E}\left[\log Z^2\right]$. We have

$$\mathbb{E}\left[Z^{2r}\right] = \frac{(2r)!}{2^r r!}$$

for $r \in \mathbb{N}$, so we guess

$$\mathbb{E}\left[Z^{2r}\right] = \frac{\Gamma(1+2r)}{2^r \Gamma(1+r)}$$

for $r \in \mathbb{R}_+$. By the replica trick, we have

$$\mathbb{E}\left[\log Z^2\right] = \partial_{r=0} \mathbb{E}\left[Z^{2r}\right] = \partial_{r=0} \frac{\Gamma(1+2r)}{2^r \Gamma(1+r)} = -\gamma - \log 2$$

where γ is the Euler constant. The last equality follows from $\Gamma(1) = 1, \Gamma'(1) = -\gamma$.

The replica method is often used to compute free energies of large disordered systems. The computation becomes more difficult than in the example above since it involves an infinite-dimensional limit. For a disordered system with Hamiltonian H, let us recall that the free energy is given by $F^H = \log Z$, where $Z = \int e^{H(x)}\mu(dx)$. We assume that F^H is self-averaged, so its leading term is given by $\mathbb{E}[F^H]$. By the replica trick we have

$$\mathbb{E}[F^H] = \mathbb{E}[\log Z] = \partial_{r=0} \log \mathbb{E}[Z^r]. \tag{1.8}$$

Next, treating r as if it were an integer, we write

$$Z^r = \int e^{H(x_1)+\cdots+H(x^r)}\mu(dx^1)\dots\mu(dx^r). \tag{1.9}$$

From (1.8) and (1.9), by exchanging the integral with the expectation, we have

$$\mathbb{E}[F^H] = \partial_{r=0} \log \int \mathbb{E}e^{H(x_1)+\cdots+H(x^r)}\mu(dx^1)\dots\mu(dx^r), \tag{1.10}$$

The remaining task is to handle the integral in the last expression. When this is finished, we obtain the leading term of the free energy.

The replica method can also be used to compute the quantities of the form $\mathbb{E}[\langle q(x)\rangle]$ for some function q. To do this we need the following trick:

$$\langle f(x)\rangle = \partial_{\lambda=0} \log \int e^{H(x)+\lambda f(x)}\mu(dx).$$

By exchanging $\mathbb{E}[\cdot]$ with the derivative, we have

$$\mathbb{E}[\langle q(x)\rangle] = \partial_{\lambda=0}\mathbb{E} \log \int e^{H(x)+\lambda q(x)}\mu(dx)$$

and now we can use the replica method to compute then $\mathbb{E}\log(\cdot)$ term, which has almost the same expression as the free energy, except for the perturbation term $\lambda f(x)$.

This way of using replicas allows us to compute various quantities related to the system. From these quantities, with some keen observations we may be able reach the ultimate insight that the system behaves like a simpler one, called the asymptotic equivalent of the system. We will show in Chapter 4 how to directly find the asymptotic equivalent with replicas. From this asymptotic equivalent, many important quantities related to the system can be computed easily.

1.4 Structure of the thesis

Chapter 2 is based on the preprint [76] submitted to IEEE Transactions on Information Theory. We will use the replica trick to compute higher derivatives of the mutual information in Gaussian channels. The obtained result is remarkably similar to the cumulant-moment formula, which can also be derived by the replica trick.

Chapter 3, based on the preprint [77], will show that random exchangeable structures such as the Polya urn model and the Chinese restaurant process can be constructed from sets with real cardinalities, or *formal sets*. The usual calculations, which involves induction, integrals and Jacobi determinants, now can be done with only combinatorial arguments. The chapter will only be used later in Chapter 8. We decide to put this chapter near the beginning because it is entirely elementary and we hope it will familiarize readers with the way of thinking when doing replica computations.

Chapter 4 contains the main theoretical contribution of the thesis. We will start by presenting Gibbs' principle, a fundamental result in statistical physics. From this, we will derive Result 4.2, which concerns systems with rather simple Hamiltonians. After this, the replicas come into the scene as we define the replicated Hamiltonian and replica density. Result 4.2 is then formally applied to these formal objects to derive the asymptotic equivalent of disordered systems.

The remaining chapters demonstrate the applications of the theoretical framework developed in Chapter 4. Chapter 5 will apply the replica method to compute asymptotic equivalents of some random matrices. We provide a detailed replica computation for the GOE model, whereas the other replica computations in the thesis are more concise. Chapter 6 will be about random convex optimization problems. Using replicas, we will derive the CGMT and a similar result for isotropically orthonormal matrices. We will also show that the CGMT holds greater power than previously believed through several examples from classic to recent literature. Chapter 7 will study Bayes-optimal inference problems with the replica method, with a focus on the asymptotic equivalence aspect rather than the information theoretic aspect. A section of this chapter presents our paper [78]. Based on Chapter 3, Chapter 8 will show how the tree-like structure that organizes the pure states of the SK model can be derived directly from Parisi ansatz.

1.5 Contributions

We highlight here the contributions and novelties of the thesis in each chapter

- Chapter 2. A generalization of the I-MMSE formula to higher derivatives and a new derivation the classic cumulant-moment formula using replicas. We discover a form τ similar in properties to the joint cumulant κ.

- Chapter 3. The derivations from scratch, based on the formal set constructions, of some important properties of the the Polya urn model, Dirichlet distribution, Chinese restaurant process, Poisson-Dirichlet distribution and exchangeable nested partitions related to Bolthausen-Sznitman coalescent. This perspective will serve a greater purpose in translating Parisi ansatz into meaningful probability objects, however, it is also interesting in its own right, as it shows how the usual calculations, which involves induction, integrals and Jacobian determinants, can be done by enumerative combinatorics in a much simpler way. A portion of this chapter (without the nested partitions) can be found in our preprint [77].

- Chapter 4. The use of replicas to compute asymptotic equivalents of disordered systems, which is the main theoretical contribution of the thesis. This is different from the usual replica computations with compute one quantity related to the system at a time. Result 4.2 is used for all replica computations in this manuscript. It replaces the conventional way of using Dirac delta functions in replica computations found in the literature.

- Chapter 5. A connection between replicas and deterministic equivalents in random matrix theory. Replica method has long been used to compute Stieltjes transforms and spectra of large random matrices. The deterministic equivalent, a more flexible tool compared to the Stieltjes transform, appeared quite recently. Its connection with the replica method was made in [19], although the formulation presented there is more complicated. The chapter also contains some technical contributions, specifically the 'right' choice of Hamiltonian in some examples (Section 5.3.2 and 5.3.3) that allows computations to be done elegantly.

- Chapter 6. A formal derivation of the CGMT from the replicas. This establishes the connection between two seemingly unrelated techniques. A further contribution is the observation, without proof, that the CGMT holds in a much more general context than explicitly stated, enabling a more straightforward application than typically seen in the literature. Especially, we derive Result 6.2 similar to the CGMT, concerning isotropically random orthogonal matrices (IRO) matrices. This result will be important for analyzing numerous random optimization problems that involve IRO matrices, for example the signal recovering problems where the measurement matrix is IRO instead of Gaussian.

- Chapter 7. The analysis of Bayes-optimal inference problem from the standpoint of asymptotic equivalence instead of the conventional information-theoretic perspective. The chapter also presents our paper [78] published in AISTATS 2023.

- Chapter 8. A direct probabilistic translation of Parisi ansatz, which results in the picture of pure states organized by a tree-like structure. In the literature, this structure is known to be deeply connected to the Parisi ansatz in some way, although no direct connection has been established.

Chapter 2

An application of the replica trick: derivatives of mutual information in Gaussian channels

The I-MMSE formula connects two important quantities in information theory and estimation theory. It states that in a Gaussian channel, the derivative of the mutual information is one-half of the minimum mean-squared error. Higher derivatives of the mutual information is related to estimation errors of higher moments, however a general formula is unknown. In this paper, we derive a general formula for the derivatives of mutual information between inputs and outputs of multiple Gaussian channels with respect to the signal-to-noise ratios. The obtained result is remarkably similar to the classic cumulant-moment relation.

This chapter is based on the preprint [76].

2.1 Introduction

Consider the following Gaussian channel

$$Y = \sqrt{\lambda} X + Z \tag{2.1}$$

in which X, Y are respectively the input and the output, Z is a standard Gaussian noise independent of X and λ is a non-negative parameter called the signal-to-noise ratio (SNR). Let $I_X(\lambda) = I(X;Y)$ be the mutual information between the input and output of the channel. This quantity is linked to the minimum mean-squared error (MMSE) for estimating X from Y by the fundamental I-MMSE formula [46]

$$I'_X(\lambda) = \frac{1}{2}\mathrm{MMSE}(\lambda) \tag{2.2}$$

where

$$\mathrm{MMSE}(\lambda) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2].$$

Higher derivatives of $I_X(\lambda)$ are also given in [47]. The key idea is the incremental channel approach, which reduces the calculation of $I_X^{(k)}(\lambda)$ for any positive λ to that of $I_X^{(k)}(0)$. Then the derivatives $I_X^{(k)}(0)$ are obtained by Taylor expansion. This method can compute the k-th derivatives for any given k. However, a general formula for all k is currently unknown.

Computing the derivatives $I_X^{(k)}(0)$ is a special case of the problem where we need to compute $H(Z_\lambda) - H(Z)$, called the *neg-entropy* of Z_λ, up to a certain error, where Z_λ converges in law to the standard normal random variable Z when $\lambda \to 0$. Some examples of Z_λ are

$$\begin{aligned} \sqrt{\lambda}X + Z, \\ \sqrt{\lambda}X + \sqrt{1-\lambda}Z, \\ (X_1 + \cdots + X_n)/\sqrt{n}. \end{aligned}$$

where X is a random variable with zero mean and unit variance, $X_1, \ldots, X_n$ are i.i.d as X and $\lambda = n^{-1/2}$ in the last example. Results of this type are used to approximate the neg-entropy in Independent Component Analysis [22] [52] and to analyze the leakage of a protected message in [91].

In this work, we consider multiple scalar Gaussian channels, each has its own SNR. We derive a general formula for the derivatives of the input-output mutual information with respect to the SNRs. We obtain a formula that is remarkably similar to the classic cumulant-moment relation. Our work relies on two key components: the compression of Gaussian channels and the non-rigorous replica method originated from the physics of disordered systems [69].

In Section 2.2 we will state the main result and some of its consequences. After presenting the main tools in Section 2.3, in Section 2.4 we will derive the results. We also give a new derivation of the classic cumulant-moment formula in Section 2.5.

Notation. We will reserve the bold letters for vectors. A vector $(x_1, \ldots, x_n)$ will be denoted as $\boldsymbol{x}$. Other notations include

- $[n] = \{1, \ldots, n\}$
- $k \circ x = (x, \ldots, x)$ where x is repeated k times
- $\boldsymbol{x} \odot \boldsymbol{y} = (x_1 y_1, \ldots, x_n y_n)$
- $x^{\downarrow k} = x(x-1)\ldots(x-k+1)$.
- 0 denotes both the number zero and the vector zero in any dimension.
- $\sqrt{\boldsymbol{\lambda}} = (\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$

2.2 Statement of results

Let $\boldsymbol{X}$ be a random variable in $\mathbb{R}^n$ with finite joint moments. The function $I_{\boldsymbol{X}}$ from $\mathbb{R}^n_+$ to $\mathbb{R}$ is defined as

$$I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = I(\boldsymbol{X}; \sqrt{\boldsymbol{\lambda}} \odot \boldsymbol{X} + \boldsymbol{Z}) \tag{2.3}$$

where $\boldsymbol{Z}$ is independent of $\boldsymbol{X}$ and follows the standard normal distribution in $\mathbb{R}^n$. The square root is applied element-wise. We will give a general formula for the derivatives of $I_{\boldsymbol{X}}$.

Our result will be stated in terms of multisets and partitions. A *multiset* is a collection of elements in which repetitions are allowed. A *partition* of a multiset is a way of dividing it into parts, or *blocks*. A partition π consisting of blocks $B_1, \ldots, B_k$ is written as $\pi = (B_1, \ldots, B_k)$. A partition is *diverse* if each of its blocks contains distinct elements. For example, the partition $(\{1,2\}, \{1,2\})$ of the multiset $\{1,1,2,2\}$ is diverse while the partition $(\{1,1\}, \{2,2\})$ is not.

For any random variables $X_1, \ldots, X_n$ with $n \geq 1$, define

$$\tau(X_1, \ldots, X_n) = \sum_{\pi} \frac{(-1)^{k-1}(k-2)!}{2^{s(\pi)}} \mathbb{E}[X_{B_1}] \ldots \mathbb{E}[X_{B_k}] \tag{2.4}$$

where $X_B = \prod_{i \in B} X_i$. The sum is taken over all diverse partitions $\pi = (B_1, \ldots, B_k)$ of the multiset $\{1, 1, \ldots, n, n\}$ and $s(\pi)$ is the number of pairs of identical blocks in π. For example, if $\pi = (\{1,2\}, \{1,2\})$ then $s(\pi) = 1$.

The equation (2.4) resembles the classic cumulant-moment relation [97], which states that

$$\kappa(X_1, \ldots, X_n) = \sum_{\pi} (-1)^{k-1}(k-1)! \, \mathbb{E}[X_{B_1}] \ldots \mathbb{E}[X_{B_k}] \tag{2.5}$$

in which the sum is over all partitions $\pi = (B_1, \ldots, B_k)$ of $[n]$ and $\kappa(X_1, \ldots, X_n)$ is the joint cumulant of the random variables $X_1, \ldots, X_n$ and is defined as

$$\kappa(X_1, \ldots, X_n) = \partial_{\lambda_1} \ldots \partial_{\lambda_n} \psi_{\boldsymbol{X}}(0)$$

where

$$\psi_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \log \mathbb{E} e^{\langle \boldsymbol{\lambda}, \boldsymbol{X} \rangle}. \tag{2.6}$$

The joint cumulant is multilinear and $\kappa((X_i)_{i \in [n]}) = 0$ if $[n]$ can be divided into two non-empty sets I and J such that $(X_i)_{i \in I}$ and $(X_j)_{j \in J}$ are independent.

Like κ, the form τ does not depend on the order of its arguments. Beside this, the properties of τ resemble that of κ, as stated in the following result:

Proposition 2.1.

a) τ is multiquadratic.

b) $\tau((X_i)_{i \in [n]}) = 0$ if $[n]$ can be divided into two disjoint, non-empty sets I and J such that $(X_i)_{i \in I}$ and $(X_j)_{j \in J}$ are independent.

Here, a function f from a vector space V to $\mathbb{R}$ is *quadratic* if

$$f(\lambda x) = \lambda^2 f(x),$$
$$2f(x) + 2f(y) = f(x-y) + f(x+y),$$

for all $\lambda \in \mathbb{R}$ and $x, y \in V$. A multivariate function is said to be *multiquadratic* if it is quadratic in each of its argument.

As the final piece of definition, we define $\tau(\,\cdot\,|Y)$, where Y is a random variable or an event, by replacing the expectations $\mathbb{E}[\,\cdot\,]$ in the definition of τ by $\mathbb{E}[\,\cdot\,|Y]$. We are now ready to state the main result:

Theorem 2.1. *For the function $I_{\boldsymbol{X}}(\boldsymbol{\lambda})$ defined in (2.3):*
a) The first order derivatives of the mutual information are given by

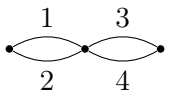
$$\partial_{\lambda_i} I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \mathbb{E}[(X_i - \mathbb{E}[X_i|\boldsymbol{Y}])^2]. \tag{2.7}$$

b) For higher derivatives,

$$\partial_{\lambda_1}^{k_1}, \ldots, \partial_{\lambda_n}^{k_n} I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \mathbb{E}[\tau(k_1 \circ X_1, \ldots, k_n \circ X_n|\boldsymbol{Y})] \tag{2.8}$$
$$= \mathbb{E}[\bar{\tau}(k_1 \circ \bar{X}_1, \ldots, k_n \circ \bar{X}_n|\boldsymbol{Y})] \tag{2.9}$$

where $\bar{X}_i = X_i - \mathbb{E}[X_i|\boldsymbol{Y}]$ and the form $\bar{\tau}$ is defined by the same formula (2.4), except that the sum is over all diverse partitions with blocks of size larger than one.

There is a convenient way to list the diverse partitions of $\{1, 1, \ldots, n, n\}$ in the expansion of τ or $\bar{\tau}$, by drawing graphs. These partitions are in bijection with the graphs that has no loop (edge that connects a vertex to itself), with n edges labeled by $[n]$. The bijection is as follows. Given a partition, we can construct a graph whose vertices represent the blocks of the partition, by connecting two blocks by the edge $i \in [n]$ if the element i belongs to these blocks. For example, we obtain from the partition $(\{1,2\}, \{3,4\}, \{1,2,3,4\})$ the following graph:



Conversely, given a graph, we can obtain the corresponding partition by looking at the edges that connect to each vertex, thereby complete the bijection.

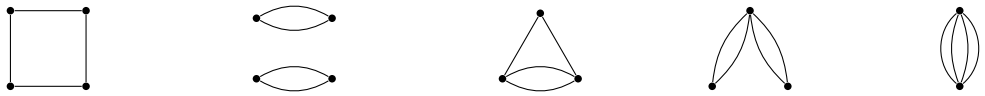
We obtain the following expressions

$$\bar{\tau}(X_1, X_2) = -\frac{1}{2}\mathbb{E}[X_1 X_2]^2 \tag{2.10}$$

$$\bar{\tau}(X_1, X_2, X_3) = -\frac{1}{2}\mathbb{E}[X_1 X_2 X_3]^2 + \mathbb{E}[X_1 X_2]\mathbb{E}[X_2 X_3]\mathbb{E}[X_3 X_1] \tag{2.11}$$

$$\begin{aligned}\bar{\tau}(X_1, X_2, X_3, X_4) = &-2(\mathbb{E}[X_1 X_2]\mathbb{E}[X_2 X_3]\mathbb{E}[X_3 X_4]\mathbb{E}[X_4 X_1] + \text{two other terms}) \\ &-\frac{1}{2}(\mathbb{E}[X_1 X_2]^2\mathbb{E}[X_3 X_4]^2 + \text{two other terms}) \\ &+\mathbb{E}[X_1 X_2]\mathbb{E}[X_1 X_3 X_4]\mathbb{E}[X_2 X_3 X_4] + \text{five other terms} \\ &+\mathbb{E}[X_1 X_2 X_3 X_4]\mathbb{E}[X_1 X_2]\mathbb{E}[X_3 X_4] + \text{two other terms} \\ &-\frac{1}{2}\mathbb{E}[X_1 X_2 X_3 X_4]^2\end{aligned} \tag{2.12}$$

The lines in the last equation correspond to the following graphs:



Moreover, the terms in each of those lines correspond to different ways of labeling the edges of the corresponding graph.

From equations (2.10-2.12), we recover the following results of [47] for the scalar Gaussian channel (2.1):

$$\begin{aligned}I_X^{(2)}(\lambda) &= \frac{1}{2}\mathbb{E}\left[-M_2^2\right] \\ I_X^{(3)}(\lambda) &= \frac{1}{2}\mathbb{E}\left[2M_2^3 - M_3^2\right] \\ I_X^{(4)}(\lambda) &= \frac{1}{2}\mathbb{E}\left[-15M_2^4 + 12M_3^2 M_2 + 6M_4 M_2^2 - M_4^2\right]\end{aligned}$$

where

$$M_k = \mathbb{E}[(X - \mathbb{E}[X|Y])^k|Y].$$

Moreover, from equations (2.10-2.12), we can compute any derivative up to the fourth order of the function $I_{\boldsymbol{X}}$ defined in (2.3). For example, when $\boldsymbol{X}$ has more than one coordinate, we have

$$\partial_{\lambda_1}^2 \partial_{\lambda_2} I_{\boldsymbol{X}}(0) = \tau(X_1, X_1, X_2) = -\frac{1}{2}\mathbb{E}[X_1^2 X_2]^2 + \mathbb{E}[X_1^2]\mathbb{E}[X_1 X_2]^2$$

2.3 Tools

We present here the tools for deriving the results: the compression of Gaussian channels and the replica method. The compression theorem implies that we can compute any

derivative of $I_{\boldsymbol{X}}$ if we know how to compute $\partial_{\lambda_1} \dots \partial_{\lambda_n} I_{\boldsymbol{X}}(0)$. The replica method, on the other hand, gives a combinatorial formula for $\partial_{\lambda_1} \dots \partial_{\lambda_n} I_{\boldsymbol{X}}(0)$.

2.3.1 Compression of Gaussian channels

Proposition 2.2. *A set of Gaussian channels with the same signal X and independent noises is equivalent to a single Gaussian channel with signal X and SNR equal to the sum of individual SNRs.*

We say that two inference problems are the equivalent if they perform exactly the same. More precisely, any performance metric (mutual information, MMSE) gives the same result on these problems.

Proof. Suppose the channels are

$$Y_i = \sqrt{\lambda_i} X + Z_i, i = 1, \dots, n$$

We can check that $S = \sum_i \sqrt{\lambda_i} Y_i$ is a sufficient statistics for estimating X from the outputs. The proposition follows from the fact that $S/\sqrt{\lambda}$, where $\lambda = \sum_i \lambda_i$, can be written as $\sqrt{\lambda} X + \xi$, where ξ is independent of X and follows the standard normal distribution. □

2.3.2 Mutual information and replicas

Let X, Y be random variables with values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are equipped with measure μ and ν, called the *underlying measure*. Let $p_X, p_Y, p_{X,Y}$ be the density functions of the random variables X, Y, (X, Y) with respect the underlying measures $\mu, \nu, \mu \otimes \nu$. Denote $p(y|x) = p_{Y|X}(y|x)$ for simplicity. By definition, the mutual information between X and Y is

$$I(X, Y) = \mathbb{E} \log p(Y|X) - \mathbb{E} \log p_Y(Y) \tag{2.13}$$

It is important to note that the mutual information does not depend on the choice of underlying measures.

Next, we have

$$\begin{aligned} \mathbb{E} \log p_Y(Y) &= \int \nu(dy) p_Y(y) \log p_Y(y) \\ &= \partial_{r=1} \int \nu(dy)\, p_Y(y)^r \end{aligned}$$

The integral in the last expression is difficult to evaluate when r is a real number. However, for $r \in \mathbb{N}$,

$$p_Y(y)^r = \mathbb{E}[p(y|X)]^r = \mathbb{E}[p(y|X^1) \dots p(y|X^r)] \tag{2.14}$$

where X^a for $a \in [r]$ are independent and identically distributed as X. We call these random variables *replicas* of X and call the indexes $a \in [r]$ *replica indexes*. We will

perform the calculations for $r \in \mathbb{N}$ and assume the same result applies for $r \in \mathbb{R}_+$, before taking the derivative. In summary, we have the following replica representation of $I(X;Y)$

$$\mathbb{E} \log p(Y|X) - \partial_{r=1} \mathbb{E} \int \nu(dy) p(y|X_1) \dots p(y|X_r) \tag{2.15}$$

2.4 Proofs

Instead of using equation (2.4) as definition for τ, let us define

$$\tau(X_1, \dots, X_n) = \begin{cases} I'_{X_1}(\lambda_1) - \frac{1}{2}\mathbb{E}[X_1^2], & n = 1 \\ \partial_{\lambda_1} \dots \partial_{\lambda_n} I_{\boldsymbol{X}}(0), & n \geq 2 \end{cases} \tag{2.16}$$

and recover the formula (2.4), along with other properties of τ.

Lemma 2.1. *If $X_1, \dots, X_n$ are random variables and $k_1, \dots, k_n$ are non-negative integers whose sum is greater than 1, then*

$$\partial_{\lambda_1}^{k_1} \dots \partial_{\lambda_n}^{k_n} I_{\boldsymbol{X}}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\mathbf{0}} = \tau\left(k_1 \circ X_1, \dots, k_n \circ X_n\right)$$

Proof. Without loss of generality, suppose that $k_1 \geq 1$ for all i. Let $m = \sum_i k_i$. For any $\boldsymbol{\lambda} \in \mathbb{R}_+^m$, let us divide its entries into consecutive blocks of sizes $k_1, \dots, k_n$ and let $s_1, \dots, s_n$ be the sum of elements in each of these blocks. By the compression of Gaussian channels, we have

$$I_{X_1,\dots,X_n}(s_1, \dots, s_n) = I_{k_1 \circ X_1, \dots, k_n \circ X_n}(\lambda_1, \dots, \lambda_m),$$

The result is obtained by applying $\partial_{\lambda_1} \dots \partial_{\lambda_m}$ on both sides of this equation at $\boldsymbol{\lambda} = 0$. □

Next, using the replica method, we obtain the following result

Lemma 2.2. *For any random variable $X_1, \dots, X_n$,*

$$\tau(X_1, \dots, X_n) = \sum_{\pi} \frac{(-1)^{k-1}(k-2)!}{2^{s(\pi)}} \mathbb{E}[X_{B_1}] \dots \mathbb{E}[X_{B_k}] \tag{2.17}$$

where $X_B = \prod_{i \in B} X_i$ and the sum is over all diverse partitions $\pi = (B_1, \dots, B_k)$ of the multiset $\{1, 1, \dots, n, n\}$ and $s(\pi)$ is the number of pairs of identical blocks in π. Moreover, let $\bar{\tau}$ be defined in the same way as τ, except that the sum is over diverse partitions with all blocks of size greater than one, then

$$\tau(X_1, \dots, X_n) = \bar{\tau}(\bar{X}_1, \dots, \bar{X}_n) \tag{2.18}$$

where $\bar{X}_i = X_i - \mathbb{E}[X_i]$

Remark 2.1. Apply the lemma for $n = 1$ we have

$$\tau(X) = -\frac{1}{2}\mathbb{E}[X]^2$$

which implies $I'_X(0) = \mathrm{Var}(X)$

Proof. For the Gaussian channels given by (2.3), we have $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$. By choosing the standard Gaussian measure as the underlying measure of $\mathcal{Y}$ and Lebesgue measure as the underlying measure of $\mathcal{X}$, we have

$$p(\boldsymbol{y}|\boldsymbol{x}) = \exp\Big(\sum_{i=1}^n \sqrt{\lambda_i} y_i x_i - \frac{1}{2}\lambda_i x_i^2\Big)$$

From this and (2.15), we obtain the following replica representation of $I_{\boldsymbol{X}}(\boldsymbol{\lambda})$

$$\frac{1}{2}\sum_{i=1}^n \lambda_i \mathbb{E}[X_i^2] - \partial_{r=1}\mathbb{E}\exp\Big(\sum_{i=1}^n \lambda_i \sum_{1\le a<b\le r} X_i^a X_i^b\Big) \tag{2.19}$$

where $\boldsymbol{X}^1, \dots, \boldsymbol{X}^r$ are i.i.d. as $\boldsymbol{X}$. From this, we have as $n \ge 2$

$$\tau(X_1, \dots, X_n) = -\partial_{r=1}\mathbb{E}\Big[\prod_{i=1}^n \sum_{1\le a<b\le r} X_i^a X_i^b\Big].$$

By expanding the product and exchanging the sum with the expectation, we have

$$\tau(X_1, \dots, X_n) = -\partial_{r=1} \sum_{a_1<b_1,\dots,a_n<b_n} \mathbb{E}[X_1^{a_1} X_1^{b_1} \dots X_n^{a_n} X_n^{b_n}]$$

Since two random variables with different replica indexes are independent,

$$\mathbb{E}[X_1^{a_1} X_1^{b_1} \dots X_n^{a_n} X_n^{b_n}] = \mathbb{E}[X_{B_1}] \dots \mathbb{E}[X_{B_k}],$$

where $\pi = (B_1, \dots, B_k)$ is the partition of the multiset $\{1, 1, \dots, n, n\}$ such that i and j are in the same block if and only if the corresponding replica indexes are equal. Since $a_i < b_i$ for all $i \in [n]$, the partition π is diverse. On the other hand, each diverse partition π with k blocks corresponds to $2^{-s(\pi)} r^{\downarrow k}$ ways of choosing the replica indexes, where $s(\pi)$ is the number of pairs of identical blocks in π and $r^{\downarrow k}$ is the number of ways of assigning different replica indexes in $[r]$ to k different blocks. The factor $2^{-s(\pi)}$ accounts for the fact that some of the blocks of π are identical. As a result,

$$\tau(X_1, \dots, X_n) = -\partial_{r=1} \sum_{\pi} \frac{r^{\downarrow k}}{2^{s(\pi)}} \mathbb{E}[X_{B_1}] \dots \mathbb{E}[X_{B_k}]$$

where the sum is over all diverse partitions $\pi = (B_1, \dots, B_k)$ of the multiset $\{1, 1, \dots, n, n\}$. Consider $P_{r,k}$ as a polynomial in real variable r, we have

$$\partial_{r=1} r^{\downarrow k} = (-1)^k (k-2)!$$

from which we obtain the combinatorial expansion of τ.

The equation (2.18) follows from the fact that mutual information is invariant by invertible transformations, so

$$I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = I_{\boldsymbol{X}-\boldsymbol{c}}(\boldsymbol{\lambda})$$

for any $\boldsymbol{c} \in \mathbb{R}^n$. The choice $\boldsymbol{c} = \mathbb{E}[\boldsymbol{X}]$ eliminates the partitions that contain blocks of size one in the expansion of τ. □

Proof of Theorem 2.1. We will prove the equation (2.8) of the theorem. The other two equations can be obtained in the same way. Consider the Gaussian channels given in (2.3). Suppose the following data is given in addition to $\boldsymbol{Y}$,

$$\boldsymbol{Y}' = \sqrt{\boldsymbol{\delta}} \odot \boldsymbol{X} + \boldsymbol{Z}'$$

The noise $\boldsymbol{Z}'$ is standard Gaussian, independent of all other random variables. By the compression of Gaussian channels, we have

$$I(\boldsymbol{X}; \boldsymbol{Y}, \boldsymbol{Y}') = I_{\boldsymbol{X}}(\boldsymbol{\lambda} + \boldsymbol{\delta}) \tag{2.20}$$

Thus,

$$\begin{aligned} I_{\boldsymbol{X}}(\boldsymbol{\lambda} + \boldsymbol{\delta}) - I_{\boldsymbol{X}}(\boldsymbol{\lambda}) &= I(\boldsymbol{X}; \boldsymbol{Y}, \boldsymbol{Y}') - I(\boldsymbol{X}; \boldsymbol{Y}) \\ &= I(\boldsymbol{X}; \boldsymbol{Y}' | \boldsymbol{Y}) \\ &= \int P_{\boldsymbol{Y}}(d\boldsymbol{y}) I(\boldsymbol{X}; \boldsymbol{Y}' | \boldsymbol{Y} = \boldsymbol{y}) \end{aligned} \tag{2.21}$$

Now taking the derivative $\partial_{\delta_1}^{k_1} \dots \partial_{\delta_n}^{k_n}$ at $\boldsymbol{\delta} = 0$ on both sides of this equation and exchange the derivative with the integral, we obtain

$$\partial_{\lambda_1}^{k_1} \dots \partial_{\lambda_n}^{k_n} I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \int P_{\boldsymbol{Y}}(d\boldsymbol{y})\, \partial_{\delta_1}^{k_1} \dots \partial_{\delta_n}^{k_n} I(\boldsymbol{X}; \boldsymbol{Y}' | \boldsymbol{Y} = \boldsymbol{y})|_{\boldsymbol{\delta}=\boldsymbol{0}} \tag{2.22}$$

Let $\boldsymbol{X}^{\boldsymbol{y}}$ be the random variable $\boldsymbol{X}$ conditioned on the event $\boldsymbol{Y} = \boldsymbol{y}$. Since $\boldsymbol{Z}'$ is independent of $\boldsymbol{Y}$, we have

$$I(\boldsymbol{X}; \boldsymbol{Y}' | \boldsymbol{Y} = \boldsymbol{y}) = I(\boldsymbol{X}^{\boldsymbol{y}}; \sqrt{\boldsymbol{\delta}} \odot \boldsymbol{X}^{\boldsymbol{y}} + \boldsymbol{Z}')$$

From this and (2.22), we have

$$\partial_{\lambda_1}^{k_1} \dots \partial_{\lambda_n}^{k_n} I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \int P_{\boldsymbol{Y}}(d\boldsymbol{y})\, \partial_{\delta_1}^{k_1} \dots \partial_{\delta_n}^{k_n} I(\boldsymbol{X}^{\boldsymbol{y}}; \sqrt{\boldsymbol{\delta}} \odot \boldsymbol{X}^{\boldsymbol{y}} + \boldsymbol{Z}')|_{\boldsymbol{\delta}=\boldsymbol{0}} \tag{2.23}$$

Since $\boldsymbol{Z}'$ is independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$, it is also independent of $\boldsymbol{X}^{\boldsymbol{y}}$. By Lemma 2.1, we have

$$\begin{aligned} \partial_{\delta_1}^{k_1} \dots \partial_{\delta_n}^{k_n} I(\boldsymbol{X}^{\boldsymbol{y}}; \sqrt{\boldsymbol{\delta}} \odot \boldsymbol{X}^{\boldsymbol{y}} + \boldsymbol{Z}')|_{\boldsymbol{\delta}=\boldsymbol{0}} &= \tau(k_1 \circ X_1^{\boldsymbol{y}}, \dots, k_n \circ X_n^{\boldsymbol{y}}) \\ &= \tau(k_1 \circ X_1, \dots, k_n \circ X_n | \boldsymbol{Y} = \boldsymbol{y}) \end{aligned}$$

From this and (2.23) we obtain the equation (2.8) in the theorem.

Proof of Proposition 2.1. a) The claim is obvious when $n = 1$. For $n \geq 2$, consider the following channels

$$Y_i = \sqrt{\lambda_i} X_i + Z_i,$$

for $i = 2, \ldots, n$, and

$$Y_1 = \sqrt{2\lambda_1} X_1 + Z_1$$
$$Y_1' = \sqrt{2\lambda_1} X_1' + Z_1'.$$

Let $Y_1^{\pm} = (Y_1 \pm Y_1')/\sqrt{2}$, then we have

$$Y_1^{\pm} = \sqrt{\lambda_1}(X_1 \pm X_1') + Z_1^{\pm}$$

where Z_1^+, Z_1^- are independent standard Gaussian variables. Since mutual information is invariant by invertible transformations, we have

$$I_{X_1, X_1', X_2, \ldots, X_n}(2\lambda_1, 2\lambda_1, \lambda_2, \ldots, \lambda_n) = I_{X_1 + X_1', X_1 - X_1', X_2, \ldots, X_n}(\lambda_1, \lambda_1, \lambda_2, \ldots, \lambda_n)$$

Taking the derivative $\partial_{\lambda_1} \ldots \partial_{\lambda_n}$ at $\boldsymbol{\lambda} = 0$ on both sides of the previous equation, we obtain

$$2\tau(X_1, \cdot) + 2\tau(X_1', \cdot) = \tau(X_1 + X_1', \cdot) + \tau(X_1 - X_1', \cdot)$$

where $\cdot = (X_2, \ldots, X_n)$, from which we conclude that τ is multiquadratic.

b) If $[n]$ can be divided into two non-empty sets I and J such that $(X_i)_{i \in I}$ and $(X_j)_{j \in J}$ are independent, then

$$I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = I_{(X_i)_{i \in I}}((\lambda_i)_{i \in I}) + I_{(X_j)_{j \in J}}((\lambda_j)_{j \in J})$$

as the noises are independent. By taking the derivative $\partial_{\lambda_1} \ldots \partial_{\lambda_n}$ at $\boldsymbol{\lambda} = 0$ on both sides of the previous equation, we obtain

$$\tau(X_1, \ldots, X_n) = 0$$

□

2.5 The cumulant-moment formula

The replica method also offers a quick derivation of the classic cumulant-moment relation given in (2.5). For the function $\psi_{\boldsymbol{X}}$ given in (2.6), using the replica trick, we have

$$\psi_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \partial_{r=0} \Big[\mathbb{E} \exp\Big(\sum_i \lambda_i X_i \Big) \Big]^r$$

Let $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^r$ be i.i.d. as $\boldsymbol{X}$. We have

$$\psi_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \partial_{r=0} \mathbb{E} \exp\Big(\sum_i \lambda_i \sum_a X_i^a \Big)$$

Applying $\partial_{\lambda_1} \ldots \partial_{\lambda_n}$ at $\boldsymbol{\lambda} = 0$ on both side of this, we obtain

$$\kappa(X_1, \ldots, X_n) = \partial_{r=0} \mathbb{E} \prod_i \sum_a X_i^a$$

Expand the product and exchange the expectation with the sum, we have

$$\kappa(X_1, \ldots, X_n) = \partial_{r=0} \sum_{a_1, \ldots, a_n} \mathbb{E}[X_1^{a_1} X_2^{a_2} \ldots X_n^{a_n}]$$

Since $X_i^{a_i}$ and $X_j^{a_j}$ are independent if $a_i \neq a_j$, we have

$$\mathbb{E}[X_1^{a_1} \ldots X_n^{a_n}] = \mathbb{E}[X_{B_1}] \ldots \mathbb{E}[X_{B_k}]$$

where $(B_1, \ldots, B_k)$ is the partition of $[n]$ such that i and j are in the same block if $a_i = a_j$. On the other hand, each partition with k blocks corresponds to $r^{\downarrow k}$ ways of choosing $(a_1, \ldots, a_n)$. Therefore

$$\kappa(X_1, \ldots, X_n) = \partial_{r=0} \sum_{\pi} r^{\downarrow k} \, \mathbb{E}[X_{B_1}] \ldots \mathbb{E}[X_{B_k}]$$

where the sum runs over all partitions $\pi = (B_1, \ldots, B_k)$ of $[n]$. The cumulant-moment formula follows from $\partial_{r=0} r^{\downarrow k} = (-1)^{k-1}(k-1)!$.

2.6 Conclusion

We derived a general formula for the derivatives with respect to the SNRs of the mutual information between inputs and outputs of multiple Gaussian channels. The result can be expressed by a form τ that is remarkably similar to the joint cumulant κ. The similarity between κ and τ is summarized in the following table:

κ	τ
multilinear	multiquadratic
$\kappa(X_1, \ldots, X_n) = \partial_{\lambda_1} \ldots \partial_{\lambda_n} \psi_{\boldsymbol{X}}(\boldsymbol{\lambda})\|_{\boldsymbol{\lambda}=0}$	$\tau(X_1, \ldots, X_n) = \partial_{\lambda_1} \ldots \partial_{\lambda_n} I_{\boldsymbol{X}}(\boldsymbol{\lambda})\|_{\boldsymbol{\lambda}=0}$
sum over partitions of $\{1, \ldots, n\}$	sum over diverse partitions of $\{1, 1, \ldots, , n, n\}$
do not depend on the order of arguments	
vanish if the arguments can be divided into two independent parts	

Chapter 3

Exchangeability and formal sets

In this chapter, we show that exchangeable structures such as Polya urn model and Chinese restaurant process can be constructed from sets with a real number of elements. From this construction, the exchangeability of these structures becomes obvious and the calculations on them become extremely simple. Moreover, the usual calculations on these structures, which involve induction, integrals and Jacobian determinants, now can be done by simple combinatorial calculations. In this way, this chapter is similar to the previous one, as it provides non-rigorous yet convenient representations of complicated mathematical objects in order to simplify calculations and derive interesting results.

This chapter will not be used until Chapter 8. However, we decide to put it near the beginning as we hope it will familiarize readers with the replica method.

This chapter is an extended version of the preprint [77].

3.1 Introduction

The cardinality of any set is a natural number. We will break this rule by considering a general kind of sets, called *formal sets*, which have a real number of elements. We will not try to make sense of the formal sets, instead we will treat them like usual sets and build from them meaningful structures such as Polya urn model and Chinese restaurant process. From this viewpoint, we can see through some non-trivial properties of these structures and the related objects. One of these properties is exchangeability, satisfied when a random structure has its law unchanged when a finite number of its elements are permuted. The exchangeability, which comes as a surprise from the definition of these structures, becomes obvious in the formal constructions. Moreover, the formal construction greatly simplify calculations. The usual calculations on these structures, which involve induction, integrals and Jacobian determinants, now can be done by simple combinatorial calculations.

The Polya urn model and the Chinese restaurant process that we will build from the formal sets are important probabilistic structures. The Polya urn model is a classic example of infinite exchangeable sequences. It is closely related to Dirichlet distribution [62] frequently used in Bayesian statistics. The Chinese restaurant process is an

exchangeable random partition of $\mathbb{N}$. It gives rise to Ewens sampling formula [28] and Poisson-Dirichlet distribution, which appears in various mathematical problems such as random walk [30], Brownian motion [88], fragmentation and coalescent process [13] [14] and prime factorization of large integers [15] [37]. The Polya urn model and the Chinese restaurant process are members of a larger family of exchangeable structures widely used in Bayesian topic model, as they provide a flexible and elegant framework for modeling data without assuming a fixed number of clusters. Some notable members of this family are Latent Dirichlet model [17] [90], Indian buffet process [44], hierarchical Dirichlet process [102] and nested Chinese restaurant process [16] [43].

3.2 Exchangeability

A finite sequence $(Y_1, \dots, Y_n)$ of random variables is *exchangeable* if

$$(Y_1, \dots, Y_n) \stackrel{d}{=} (Y_{\sigma(1)}, \dots, Y_{\sigma(n)}) \tag{3.1}$$

for each permutation σ of $[n]$. An infinite sequence $(Y_i)_{i=1}^{\infty}$ is exchangeable if

$$(Y_1, Y_2, \dots) \stackrel{d}{=} (Y_{\sigma(1)}, Y_{\sigma(2)}, \dots) \tag{3.2}$$

for each finite permutation of $\mathbb{N}_+$, i.e. permutations such that $\{i : \sigma(i) \neq i\}$ is finite.

Exchangeability arises naturally from sampling. Consider a set with m elements, each with a label that is not necessarily unique. Then if we randomly draw $n < m$ elements from the set without replacement, the labels of these elements form an exchangeable sequence. More generally, consider a random vector $(X_1, \dots, X_m)$. If we select n distinct indices $i_1, \dots, i_n$ randomly from the set $[m]$, the resulting sequence $(X_{i_1}, \dots, X_{i_n})$ is exchangeable.

We can create an exchangeable sequence by picking a random probability measure and draw an i.i.d. sequence from it. De Finetti's theorem states that all infinite exchangeable sequence can be constructed in this way. In other words,

Theorem 3.1. *Every infinite exchangeable sequence is a mixture of i.i.d. sequences.*

Consider an exchangeable sequence $(Y_i)_{i=1}^{\infty}$ where each Y_i takes values in a discrete set $\mathcal{Y}$. By the law of large number and by de Finetti's theorem, each element of $\mathcal{Y}$ has a limiting proportion in $(Y_i)_{i=1}^{\infty}$. Moreover, these proportions vary for different realizations of $(Y_i)_{i=1}^{\infty}$.

3.3 Polya urn model

3.3.1 The model

In this model, at the beginning we have a set containing elements labeled by $1, \dots, k$. Let $\alpha_i \in \mathbb{N}_+$ be the number of elements with label i. At each step, we choose uniformly

randomly a element from the set, record its label, and put it back along with another element of the same label. Let $(Y_i)_{i=1}^{\infty}$ be the sequence of the recorded labels. This sequence is a stochastic process that satisfies

$$\mathbb{P}(Y_1 = i) = \frac{\alpha_i}{\alpha_1 + \cdots + \alpha_k} \tag{3.3}$$

and

$$\mathbb{P}(Y_{n+1} = i | Y_1, \ldots, Y_n) = \frac{\alpha_i + n_i}{\alpha_1 + \cdots + \alpha_k + n}, \tag{3.4}$$

where n_i is the number of occurrence of i in the sequence $Y_1, \ldots, Y_n$. The equations (3.3) and (3.4) uniquely determines a stochastic process for $\alpha_1, \ldots, \alpha_k$ that are not restricted to $\mathbb{N}_+$, but rather extend to $\mathbb{R}_+$. We call such process *Polya urn process* with parameters $(\alpha_1, \ldots, \alpha_k) \in \mathbb{R}_+^k$.

The Polya urn process has the remarkable property of being exchangeable, which is not at all trivial from the definition, since to prove it one has no other way than doing explicit calculations. In the next section, this property can be explained elegantly with no calculation.

3.3.2 Formal set construction

Imagine a set containing a total of $-\alpha_1 - \cdots - \alpha_k$ elements divided into k groups labeled from 1 to k, each with sizes $-\alpha_1, \ldots, -\alpha_k$ respectively, where $\alpha_1, \ldots, \alpha_k > 0$. Forgetting the fact that the cardinality of a set must be a non-negative integer, let us see what happens when we sample without replacement from this set. The probability that the first element has label i is

$$\frac{-\alpha_i}{-\alpha_1 - \cdots - \alpha_k} = \frac{\alpha_i}{\alpha_1 + \cdots + \alpha_k},$$

Suppose that after n steps, we have taken out n_i elements with label i. In the set there remains $-\alpha_i - n_i$ elements with label i. The probability of the $(n+1)$-th element having label i is

$$\frac{-\alpha_i - n_i}{-\alpha_1 - n_1 - \cdots - \alpha_k - n_k} = \frac{\alpha_i + n_i}{\alpha_1 + \cdots + \alpha_k + n},$$

Although the underlying set is ill-defined, the probabilities arising from the sampling process are well-defined and precisely match those of the Polya urn model. With formal sets, the Polya urn model is described more concisely and its exchangeability becomes trivial.

3.3.3 Joint probability

Let $(y_1, \ldots, y_n)$ be the sequence of labels in the first n samplings. Suppose that the label i appears n_i times in this sequence. From the formal set construction we have

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_n = y_n\right) = \frac{(-\alpha_1)^{\downarrow n_1} \ldots (-\alpha_k)^{\downarrow n_k}}{(-\alpha_1 - \cdots - \alpha_k)^{\downarrow n}}$$

Here the denominator counts the sequences of n different elements from the formal set and $(-\alpha_i)^{\downarrow n_i}$ counts the sequences of different n_i elements with label i. From the formula $(-x)^{\downarrow n} = (-1)^n x^{\uparrow n}$, we obtain

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_n = y_n\right) = \frac{\alpha_1^{\uparrow n_1} \ldots \alpha_k^{\uparrow n_k}}{(\alpha_1 + \cdots + \alpha_k)^{\uparrow n}} \tag{3.5}$$

Note that with the usual definition of Polya urn model, we prove (3.5) by induction and then conclude that the sequence $(Y_i)_{i=1}^n$ is exchangeable.

3.3.4 Dirichlet distribution

Consider a Polya urn sequence $(Y_i)_{i=1}^n$ with parameters $(\alpha_1, \ldots, \alpha_k)$. The probability of having n_i labels i in the sequence $(Y_1, \ldots, Y_n)$ is

$$p(n_1, \ldots, n_k) = \frac{n!}{n_1! \ldots n_k!} \cdot \frac{\alpha_1^{\uparrow n_1} \ldots \alpha_k^{\uparrow n_k}}{(\alpha_1 + \cdots + \alpha_k)^{\uparrow n}}$$

Let $x_i = n_i/n$. Using the fact that

$$\frac{\Gamma(m+r)}{\Gamma(m+s)} \simeq m^{r-s}, \quad m \to \infty, \tag{3.6}$$

we obtain the density of $(x_1, \ldots, x_k)$ as n tends to infinity:

$$f(x_1, \ldots, x_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1} 1_{\Delta_k}(x_1, \ldots, x_k)$$

where

$$\Delta_k = \left\{x_1, \ldots, x_k \geq 0 : x_1 + \cdots + x_k = 1\right\}.$$

This is the Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$. In summary, the proportions of $1, \ldots, k$ in an infinite Polya urn sequence with parameters $(\alpha_1, \ldots, \alpha_k)$ follow the Dirichlet distribution with the same parameters.

From the formal set construction of Polya urn model, the following properties of the Dirichlet distribution can be derived with very little effort.

Theorem 3.2. *Consider a random vector $(X_1, \ldots, X_n)$ following a Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_n)$. Then*

1. (Aggregation) *If $[n]$ is partitioned into subsets $B_1, \ldots, B_k$, then*

$$\Big(\sum_{i \in B_1} X_i, \ldots, \sum_{i \in B_k} X_i\Big) \sim \mathrm{Dir}\Big(\sum_{i \in B_1} \alpha_i, \ldots, \sum_{i \in B_k} \alpha_i\Big) \tag{3.7}$$

2. (Neutrality) *Let I be an ordered subset*[1] *of $[n]$, and $\tilde{X}_I = X_I / \sum_{i\in I} X_i$. Then $\tilde{X}_I$ follows* $\mathrm{Dir}(\alpha_I)$ *and is independent of X_{I^c}.*

Proof. $(X_1, \ldots, X_n)$ can be viewed as the proportion of $1, \ldots, n$ in a Polya urn sequence $Y = (Y_i)_{i=1}^{\infty}$ with parameters $(\alpha_1, \ldots, \alpha_n)$. Consider the formal set construction of this Polya urn model.

The property 1 can be easily proved by combining all the elements labeled by B_j in the formal set into a new group for each $j = 1, \ldots, k$.

To prove 2, imagine assigning a special label 0 to each element in the formal set that already has a label from I. This new label temporarily conceals the original one. Next, we sample without replacement from the formal set as usual. Finally, we remove the labels 0 to reveal the original labels. The sequence of revealed labels forms a Polya urn process with parameters α_I. This sequence is independent of the observations before the revealing, therefore independent of X_{I^c}. □

Corollary 3.1. *Let $(X_1, \ldots, X_n)$ be a Dirichlet random vector with parameters $\alpha_1, \ldots, \alpha_n > 0$. Then*

1. (Marginal law)

$$P_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \propto x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1}(1 - x_1 - \cdots - x_k)^{\alpha_{k+1} + \cdots + \alpha_n - 1}$$

In particular

$$X_i \sim \mathrm{Beta}\Big(\alpha_i, \sum_{j\neq i} \alpha_j\Big) \tag{3.8}$$

2. (Gamma construction) *Let $Z_i \sim \Gamma(\alpha_i, 1)$ be independent. Then*

$$(X_1, \ldots, X_n) \overset{d}{=} \Big(\frac{Z_i}{Z_1 + \cdots + Z_n}\Big)_{i=1}^{n} \tag{3.9}$$

3. (Stick-breaking construction) *Consider $(X_1^*, \ldots, X_n^*)$ constructed as follows*

$$\begin{aligned} X_1^* &= W_1 \\ X_2^* &= W_2(1 - W_1) \\ &\cdots \\ X_{n-1}^* &= W_{n-1}(1 - W_{n-2}) \ldots (1 - W_1) \end{aligned}$$

and $X_n^* = 1 - X_1^* - \cdots - X_n^*$, where W_i are independent with

$$W_i \sim \mathrm{Beta}(\alpha_i, \alpha_{i+1} + \cdots + \alpha_n)$$

Then

$$(X_1, \ldots, X_n) \overset{d}{=} (X_1^*, \ldots, X_n^*)$$

[1] An ordered set is simply a set with elements arranged in some order. For $I = (i_1, \ldots, i_k)$, denote $x_I = (x_{i_1}, \ldots, x_{i_k})$.

Proof. The claim 1 follows from

$$(X_1, \ldots, X_k, X_{k+1} + \cdots + X_n) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k, \alpha_{k+1} + \cdots + \alpha_n).$$

To prove the claim 2, consider

$$(V_1, \ldots, V_n, T) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n, t)$$

It is easy to check that

$$(tV_1, \ldots, tV_n) \xrightarrow{d} (Z_1, \ldots, Z_n), \quad t \to \infty$$

On the other hand, from Theorem 3.2, we have

$$\Big(\frac{tV_i}{tV_1 + \cdots + tV_n}\Big)_{i=1}^{n} \overset{d}{=} (X_1, \ldots, X_n)$$

By taking $t \to \infty$, we obtain the claim 2. Claim 3 follows from the neutrality of Dirichlet distribution. First we have $X_1 \overset{d}{=} W_1$. Then the relative size of X_2 in $(X_2, \ldots, X_n)$ is independent of X_1 and has the same law as W_2, so $X_2 \overset{d}{=} W_2(1 - W_1)$. The claim can be proved by repeating this argument. □

3.4 Chinese restaurant process

3.4.1 The model

Let α, θ be parameters that satisfies either one of the following cases

1. $\alpha < 0$ and $\theta = -k\alpha$ for some $k \in \mathbb{N}_+$ or
2. $0 \le \alpha \le 1$ and $\theta > -\alpha$.

The Chinese restaurant process with parameters (α, θ) is defined as follows. Imagine a restaurant with an infinite number of tables. At the beginning all tables are empty. The first customer arrives and sits at any table. If the first n customers occupy m tables, then the $(n+1)$-th customer will choose

- a table with $t \ge 1$ customers with probability $\frac{t-\alpha}{n+\theta}$
- an empty table with probability $\frac{m\alpha+\theta}{n+\theta}$

The first n customers form a partition of $[n]$ in which tables represent blocks and customers represent elements. By continuing this process infinitely we obtain a random partition of $\mathbb{N}$. This random partition has the remarkable property of exchangeability, meaning that its law remains unchanged under finite permutations of $\mathbb{N}$.

3.4.2 Formal set construction

Consider a set of $-\theta$ elements divided into $-\theta/\alpha$ groups, each containing α elements. Let us see what happens when we sample without replacement from this set. Suppose the first n sampled elements belong to m groups. After this, there are $\theta - n$ elements remaining in the set. For a group with t sampled elements, there are $\alpha - t$ elements of that group still in the set. Therefore, the probability of the $(n+1)$-th element coming from this group is:

$$\frac{\alpha - t}{-\theta - n} = \frac{t - \alpha}{n + \theta}.$$

Since all the probabilities add up to 1, the probability for the $(n+1)$-th element to be in a new group is

$$\frac{m\alpha + \theta}{n + \theta}.$$

These probabilities exactly match those of Chinese restaurant process. With formal sets, the Chinese restaurant process can be described more concisely and it becomes trivial that the resulting partition of $\mathbb{N}$ is exchangeable.

For the parameters (α, θ) in case 1, the formal set consists of k blocks of size α, where $k \in \mathbb{N}_+$. Thus, the Chinese restaurant process is the same as the Polya urn process with k labels and parameters $(-\alpha, \ldots, -\alpha)$. The resulting random partition on $\mathbb{N}$ always has k blocks. In contrast, for the parameters (α, θ) in case 2, we will see that the resulting partition of $\mathbb{N}$ contains an infinite number of blocks almost surely.

3.4.3 Ewens-Pitman distribution

Theorem 3.3. (Ewens-Pitman distribution) *Consider the Chinese restaurant process with parameters (α, θ). Let Π_n be the random partition of $[n]$ formed by the first n customers. Let π be a partition of $[n]$ with block sizes $n_1, \ldots, n_k$. Then*

$$\mathbb{P}(\Pi_n = \pi) = \frac{(\theta/\alpha)^{\uparrow k}}{\theta^{\uparrow n}} \prod_{i=1}^{k} -(-\alpha)^{\uparrow n_i} \tag{3.10}$$

This surely can be proved by induction, starting from the definition, although that would be tedious. We give here a very short derivation using formal sets.

Proof. Recall that the Chinese restaurant process is equivalent to sampling without replacement from a formal set of $-\theta$ elements divided into $-\theta/\alpha$ groups of size α. There are $(-\theta)^{\downarrow n}$ ways of choosing a sequence of n different elements from the formal set. Next, we will compute the number of such sequences that lead to the partition π. There are $(-\theta/\alpha)^{\downarrow k}$ ways of assigning groups of the formal set to k blocks of π. Afterwards, for each block of size n_i, there are $\alpha^{\downarrow n_i}$ ways to select its elements. Therefore

$$\mathbb{P}(\Pi_n = \pi) = \frac{(-\theta/\alpha)^{\downarrow k}}{(-\theta)^{\downarrow n}} \prod_{i=1}^{k} \alpha^{\downarrow n_i} \tag{3.11}$$

From the fact that $(-x)^{\downarrow n} = (-1)^n x^{\uparrow n}$, we obtain the result given by (3.10). □

Remark 3.1. In Theorem 3.3, for $\alpha = 0$, we have Ewens sampling formula

$$\mathbb{P}(\Pi_n = \pi) = \frac{\theta^k}{\theta^{\uparrow n}} \prod_{i=1}^{k} (n_i - 1)!$$

For $\theta = 0$, we have

$$\mathbb{P}(\Pi_n = \pi) = \frac{(k-1)!}{\alpha(n-1)!} \prod_{i=1}^{k} -(-\alpha)^{\uparrow n_i}$$

These formulas can be obtained by setting $\alpha \to 0$ or $\theta \to 0$ in the Ewens-Pitman distribution.

3.4.4 Block weights

Consider the random partition of $\mathbb{N}$ generated from the Chinese restaurant process with parameters (α, θ). Let B_1 be the block containing 1, B_2 be the the block containing the smallest element not in B_1, B_3 be the block containing the smallest element not in B_1, B_2, and so on. For each i, the *weight* of B_i is defined as

$$V_i = \lim_{n \to \infty} \frac{|B_i \cap [n]|}{n} \tag{3.12}$$

It is clear that $\sum_i V_i = 1$. We have the following result, called *stick-breaking construction* of (V_i):

Theorem 3.4. *The sequence $(V_1, V_2, \dots)$ in (3.12) is well-defined and has the same law as $(V_1^\star, V_2^\star, \dots)$, where*

$$\begin{aligned} V_1^\star &= W_1 \\ V_2^\star &= W_2(1 - W_1) \\ &\dots \\ V_k^\star &= W_k(1 - W_{k-1}) \dots (1 - W_1) \\ &\dots \end{aligned}$$

for $W_1, W_2, \dots$ independent and

$$W_j \sim \text{Beta}(1 - \alpha, \theta + j\alpha)$$

for each $j = 1, 2, \dots$. As a consequence, the random partition has an infinite number of blocks, each with a positive weight.

Proof. Consider the formal set that gives rise to the Chinese restaurant process with parameters (α, θ). Let us call x_1 the first element drawn from the formal set. After x_1 is drawn, in the formal set there remain $\alpha - 1$ elements within the same group as x_1, and $-\theta - \alpha$ other elements. From our discussion on Polya urn model, in the infinite sequence drawn from this remaining set, the proportion of elements of the same group as x_1 is $W_1 \sim \text{Beta}(1-\alpha, \theta+\alpha)$.

Now let us ignore the block containing x_1 and consider the remaining sequence. The blocks in this sequence have a total weight of $1 - W_1$ and are generated from the drawing on a formal set of $-\theta - \alpha$ elements divided into groups of size α. Let x_2 be the first element of this sequence. By the same argument as in the previous paragraph, the block containing x_2 has relative size $W_2 \sim \text{Beta}(1-\alpha, \theta+2\alpha)$, therefore its size is $W_2(1-W_1)$.

Ignoring blocks containing x_1, x_2, the rest has a total weight of $(1 - W_1)(1 - W_2)$. By repeating the same argument we obtain the result of the theorem. □

By rearranging the sequence (V_i) in decreasing order, we obtain the sequence (S_i). These two sequences are reorderings of each other. Specifically, (S_i) is the decreasing reordering of (V_i), and V_i is referred to as the *size-biased reordering* of (S_i).

The sequence $S_1 \geq S_2 \geq \ldots$ is a sample from the *Poisson-Dirichlet distribution* with parameters (α, θ), which is a probability distribution on the set of non-increasing sequences of positive numbers adding up to 1. Results on Poisson-Dirichlet distribution can be found in [51] [85] [87] [88].

While the weights (V_i) are rather simple, as they can be described by the stick-breaking process, the weights (S_i) are much more complicated as the density of S_i for a fixed i can have singular points [31].

Correlation functions. The set $\{S_i, i \in \mathbb{N}\}$ defines a point process in $\mathbb{R}$. The k-correlation function of a random point process $\{X_i\}$ in $\mathbb{R}$ is defined as a function ρ_k with variables $x_1, \ldots, x_k$ such that the probability of the process having one point in each of the intervals $[x_i, x_i + dx_i]$ for $i = 1, \ldots, k$ is $\rho_k(x_1, \ldots, x_k)dx_1 \ldots dx_k$ as $dx_i \to 0$. Equivalently, the k-correlation function ρ_k can be defined as the function such that

$$\mathbb{E}\Big[\sum_{i_1,\ldots,i_k(\neq)} f(X_{i_1}, \ldots, X_{i_k})\Big] = \int f(x_1, \ldots, x_k)\rho(x_1, \ldots, x_k)dx_1 \ldots dx_k, \tag{3.13}$$

for any non-negative measurable function f, where the sum is over k different indices. Since $f \geq 0$, the sum on the left hand side has always a limit in $[0, \infty]$. It follows from both definitions that ρ_k is invariant by permutations of its variables. Note that ρ_k is not a probability density: if $\{X_i\}$ has infinite points then the integral of ρ_k over $x_1, \ldots, x_k$ is infinite.

Let us compute the correlation function of the point process $\{S_i\}$. Imagine drawing n elements from a the formal set with $-\theta$ elements divided into groups of size α. We will compute the probability $P(n_1, \ldots, n_k)$ that there are blocks of sizes $n_1, \ldots, n_k$ in the partition formed by these n elements, where $n_1, \ldots, n_k \in \mathbb{N}_+$ such that $n_1 + \cdots + n_k \leq n$. There are

- $\binom{-\theta/\alpha}{k}$ ways of choosing k groups in the formal set.
- $\binom{\alpha}{n_1}\dots\binom{\alpha}{n_k}$ ways of choosing the elements for the blocks of sizes $n_1,\dots,n_k$.
- $\binom{-\theta-k\alpha}{n-n_1-\cdots-n_k}$ ways of choosing other elements.
- $\binom{-\theta}{n}$ ways of choosing n elements from the formal set.

$P(n_1,\dots,n_1)$ is obtained as the product of the first three numbers divided by the fourth number. Let $x_k = n_k/n$. As n tends to infinity, by using (3.6) and the formula

$$\binom{-x}{m} = \frac{(-1)^m\Gamma(x+m)}{\Gamma(x)\Gamma(m+1)}$$

we obtain the k-correlation function for the block weights [51]

$$\rho_k(x_1,\dots,x_k) = c_{k,\alpha,\theta}\, x_1^{-\alpha-1}\dots x_k^{-\alpha-1}(1-x_1-\cdots-x_k)^{k\alpha+\theta-1} \tag{3.14}$$

where

$$c_{k,\alpha,\theta} = \frac{\Gamma(\theta/\alpha+k)\Gamma(\theta)\alpha^k}{\Gamma(\theta+k\alpha)\Gamma(\theta/\alpha)\Gamma(1-\alpha)^k}. \tag{3.15}$$

For Poisson-Dirichlet process with parameters $(\alpha,0)$, the coefficient $c_{k,\alpha,0}$ is computed by taking the limit $\theta\to 0$. Using the fact that $\lim_{\theta\to 0}\frac{\Gamma(\theta)}{\Gamma(\theta/\alpha)} = \frac{1}{\alpha}$, we obtain the correlation function for this process as

$$\rho_n(x_1,\dots,x_k) = \frac{\Gamma(k)\alpha^{k-1}}{\Gamma(k\alpha)\Gamma(1-\alpha)^k}x_1^{-\alpha-1}\dots x_k^{-\alpha-1}(1-x_1-\cdots-x_k)^{k\alpha-1}$$

In particular,

$$\rho_1(x) = \frac{x^{-\alpha-1}(1-x)^{\alpha-1}}{\Gamma(\alpha)\Gamma(1-\alpha)}$$

For Poisson-Dirichlet process with parameters $(0,\theta)$, the coefficient $c_{k,\alpha,\theta}$ converges to θ^k when $\alpha\to 0$, so the correlation function in this case is

$$\rho_k(x_1,\dots,x_k) = \theta^k x_1^{-1}\dots x_k^{-1}(1-x_1-\cdots-x_k)^{\theta-1}$$

In particular,

$$\rho(x) = \theta x^{-1}(1-x)^{\theta-1}.$$

3.5 Nested partition of $\mathbb{N}$

We can extend the formal set construction for the Chinese restaurant process to obtain random nested partitions of $\mathbb{N}$. Let $k \geq 1$ and $0 \leq \alpha_0 < \cdots < \alpha_k < 1$. Consider a formal set with α_0 elements divided into groups of size α_1, each being divided into groups of size α_2, and so on. Drawing from this formal set will lead to a nested partition of $\mathbb{N}$, in which the level-0 block $\mathbb{N}$ is divided into level-1 blocks, each contains level-2 blocks, and so on. The following properties of this random nested partition follow immediately from the formal set construction:

1. It is exchangeable, i.e. its law is invariant by finite permutations of $\mathbb{N}$.

2. The nested partition of level $\ell+1$ is obtained by partitioning each undivided block of the nested partition of level ℓ according to the Chinese restaurant process with parameters $(\alpha_{\ell+1}, -\alpha_\ell)$. Consequently, The relative sizes of the $(\ell+1)$-th level blocks contained in any ℓ-th level block are given by $\mathrm{PD}(\alpha_{\ell+1}, -\alpha_\ell)$.

3. The sizes of the ℓ-th level blocks are given by $\mathrm{PD}(\alpha_\ell, -\alpha_0)$.

When $\alpha_0 = 0$, this structure of nested partitions is the same as the partitions observed at different times of the Bolthausen-Sznitman coalescent [86]. It will be important for describing the Gibbs measure of the Sherrington-Kirkpatrick model in Chapter 8.

Chapter 4

Gibbs principle, asymptotic equivalence and replicas

One of the simplest examples of asymptotic equivalence is that the uniform distribution on a high dimensional sphere behaves like independent Gaussian random variables. This is an example of much more general result in statistical physics, Gibbs principle, also known as the *equivalence of ensembles*. This principle states that calculations can be done with either the *microcanonical ensemble* or the *canonical ensemble*, yielding identical results. In this chapter, we will break down the meaning of this statement and explore its mathematical consequences. From Gibbs principle, we will derive Result 4.2 that allows us to compute the asymptotic equivalent of systems with rather simple Hamiltonian that can be expressed in terms of simple macroscopic functions. Finally, we will describe how to use replicas to compute the asymptotic equivalent of disordered systems, based on Result 4.2 and the concepts of replicated Hamiltonian and replica density.

A rigorous treatment of Gibbs principle can be found in [109], especially Section V.

4.1 Microcanonical and canonical ensembles

We will use the vocabularies of statistical physics already introduced in Section 1.2. Let us consider the configuration space $\mathcal{X}$ equipped with a measure μ. Let E be a function from $\mathcal{X}$ to $\mathbb{R}$ called *energy* function. If $\mathcal{X}$ is discrete, the microcanonical ensemble is defined as the uniform distribution (according to μ) on $\{E(x) = a\}$ for some fixed a. If $\mathcal{X}$ is continuous, the **microcanonical ensemble** is defined as the limit when $\delta \to 0$ of the uniform distribution (also according to μ) on

$$\{x : E(x) \in [a, a + \delta]\}$$

for a fixed a. This ensemble is denoted by $\mathcal{U}_a$, as we fix the energy function from the beginning and don't need to explicitly refer to it in each notation.

Remark 4.1. When $\mathcal{X} = \mathbb{R}^n$ with Lebesgue measure, the microcanonical ensemble $\mathcal{U}_a$ is generally different from the uniform distribution[1] on $\{x : E(x) = a\}$. While the latter is intrinsic, i.e. the distribution only depends on the level set and does not depend on the function E, the former is not: two different energy functions with the same level set at a can give rise to different microcanonical ensembles. The two definitions coincide when two level sets $\{x : E(x) = a\}$ and $\{x : E(x) = a + \delta\}$ are 'parallel' when $\delta \to 0$, meaning that the norm of the gradient is constant for any point on the level set at a. For example, the function $E(x) = x_1^2 + \cdots + x_n^2$ on $\mathbb{R}^n$ and $E(x) = x_1 + \cdots + x_n$ on $\mathbb{R}_+^n$ have this property (μ is chosen to be the Lebesgue measure), while the function $E(x) = x_1^p + \cdots + x_n^p$ on $\mathbb{R}^n$ with $p > 0$ and $p \notin \{1, 2\}$ does not.

Given a parameter λ, the **canonical ensemble** is defined as the following probability measure

$$P_\lambda(dx) = \frac{e^{\lambda E(x)}}{Z(\lambda)} \mu(dx) \tag{4.1}$$

where

$$Z(\lambda) = \int \mu(dx) e^{\lambda E(x)}.$$

Compared to the microcanonical ensemble, the canonical ensemble is more easy to work with. It turns out that the two ensembles are equivalent for a suitable choice of parameters, as we will see in the next section.

The **Boltzmann entropy** of the microcanonical ensemble $\mathcal{U}_a$ is defined as

$$S(a) = \log \frac{\mu(E(x) \in da)}{da}, \tag{4.2}$$

while the **Shannon entropy** of P_λ is given by

$$\mathrm{Ent}(\lambda) = -\int_{\mathcal{X}} \mu(dx) P_\lambda(x) \log P_\lambda(x). \tag{4.3}$$

4.2 Gibbs principle

Gibbs principle is a very general and fundamental result in statistical physics. It enables the study of some complicated microcanonical ensembles by replacing them with equivalent and more manageable canonical ensembles. The principle is applicable to most systems in statistical physics, except for some rather exotic ones such as the model considered in [39].

With the setting of Section 4.1, **let us make the important assumption that the Boltzmann entropy $S(a)$ is a concave function in a.** Gibbs principle states

[1] The uniform distribution on a surface $S \subset \mathcal{X}$ of lower dimension in $\mathbb{R}^n$ is the limit when $\delta \to 0$ of the uniform distribution on the set of points within a Euclidean distance δ from S.

that, under this assumption, the microcanonical ensemble $\mathcal{U}_a$ and the canonical ensemble P_λ are asymptotically equivalent if

$$\int E(x)P_\lambda(dx) = a. \tag{4.4}$$

This condition is equivalent to

$$F'(\lambda) = a, \tag{4.5}$$

where

$$F(\lambda) = \log \int \mu(dx)e^{\lambda E(x)}. \tag{4.6}$$

This equation has a unique solution since $F(\lambda)$ is strictly convex, which is easy to prove.

Gibbs principle can be generalized for vector-valued functions. Let E be a function from $\mathcal{X}$ to $\mathbb{R}^k$ with k fixed, then the uniform distribution on

$$\{x \in \mathcal{X} : E_i(x) \in [a_i, a_i + \delta] \text{ for } i = 1, \dots, k\}, \quad \delta \to 0,$$

is asymptotically equivalent to the Gibbs measure

$$P_\lambda(dx) \propto e^{\langle \lambda, E(x) \rangle} \mu(dx)$$

where $\lambda \in \mathbb{R}^k$ is such that $\int E(x)P_\lambda(dx) = a$.

Remark 4.2. The concavity of $S(a)$ is necessary, as there exist counterexamples of Gibbs principle when $S(a)$ is not concave [39].

4.3 Entropy equivalence

Always under the assumption that $S(a)$ is concave, the **equivalence of entropy** states that

> Suppose that the microcanonical ensemble $\mathcal{U}_a$ is equivalent to the canonical ensemble P_λ, then the Boltzmann entropy of $\mathcal{U}_a$ is equal to the Shannon entropy of P_λ (up to the leading term).

Derivation. Suppose that microcanonical ensemble $\mathcal{U}_{a_\star}$ is equivalent to the canonical ensemble $P_{\lambda_\star}$. The Shannon entropy of $P_{\lambda_\star}$ is given by

$$\begin{aligned} \mathrm{Ent}(\lambda_\star) &= -\int_{\mathcal{X}} \mu(dx) P_{\lambda_\star}(x) \log P_{\lambda_\star}(x) \\ &= -\int_{\mathcal{X}} \mu(dx) P_{\lambda_\star}(x)(\lambda_\star E(x) - F(\lambda_\star)) \\ &= F(\lambda_\star) - \lambda_\star a_\star \end{aligned} \tag{4.7}$$

On the other hand, for any fixed λ, we have

$$F(\lambda) = \int \mu(dx) e^{\lambda E(x)} \tag{4.8}$$

$$= \log \int \mu(E(x) \in da) e^{\lambda a}$$

$$= \log \int da\, e^{\lambda a + S(a)}$$

$$\simeq \max_a \{\lambda a + S(a)\} \tag{4.9}$$

where the last approximation is taken in the infinite size limit using the saddle point approximation, or Laplace's method. Note that we are considering a system with size growing to infinity, so the integral involved in this approximation is exponentially large, and the saddle point method is applicable. It follows from (4.8) and the concavity of $S(a)$ that $F(\lambda)$ is the convex conjugate of $-S(a)$. Therefore

$$S(a) = \min_\lambda \{F(\lambda) - \lambda a\}. \tag{4.10}$$

The minimum is achieved if $F'(\lambda) = a$, which is exactly the condition for the equivalence of P_λ and $\mathcal{U}_a$. Therefore

$$S(a_\star) = F(\lambda_\star) - \lambda_\star a_\star. \tag{4.11}$$

From (4.7) and (4.11) we conclude that $S(a_\star) = \mathrm{Ent}(\lambda_\star)$. □

Remark 4.3. The entropy equivalence is useful in computing the cardinality or volume of an exponentially large set, up to the leading order. This type of problems can be formulated as computing the Boltzmann entropy of a microcanonical ensemble. By the equivalence of entropy, we can instead compute the entropy of the corresponding canonical ensemble, which is simpler in many cases.

4.4 Applications

4.4.1 Uniform distribution on ℓ_2 spheres

Consider the configuration space $\mathcal{X} = \mathbb{R}^n$ equipped with the Lebesgue measure and energy function $E(x) = \|x\|^2$. Gibbs principle implies that the random vector $(X_1, \dots, X_n) \in \mathbb{R}^n$ sampled uniformly from the sphere $\left\{\|x\|^2 = n\right\}$ is asymptotically equivalent to $(Z_1, \dots, Z_n)$, where $Z_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. This result also follows from the fact that

$$(X_1, \dots, X_n) \overset{d}{=} \frac{\sqrt{n}(Z_1, \dots, Z_n)}{(Z_1^2 + \cdots + Z_n^2)^{\frac{1}{2}}} \simeq (Z_1, \dots, Z_n)$$

In more precise mathematical sense, for any fixed k, $(X_1, \dots, X_k)$ converge in law to k independent standard Gaussian variables. Moreover, it is shown in [36] that for any $k = O(\sqrt{n})$, the total variation distance from $(X_1, \dots, X_k)$ to $\mathcal{N}(0,1)^{\otimes k}$ converge to zero.

4.4.2 Uniform distribution on ℓ_1-spheres

Consider the configuration space $\mathcal{X} = \mathbb{R}_+^n$ equipped with the Lebesgue measure and energy function $E(x) = \|x\|_1$. Gibbs principle implies that the random vector $(X_1, \ldots, X_n) \in \mathbb{R}_+^n$ sampled uniformly from

$$\left\{x \in \mathbb{R}_+^n : x_1 + \cdots + x_n = n\right\}$$

is asymptotically equivalent to $(V_1, \ldots, V_n)$, where $V_i \overset{i.i.d}{\sim} \text{Exp}(1)$. On the other hand, this result also follows from

$$(X_1, \ldots, X_n) \overset{d}{=} n\text{Dir}(1, \ldots, 1) \overset{d}{=} \frac{(V_1, \ldots, V_n)}{V_1 + \cdots + V_n} \simeq (V_1, \ldots, V_n)$$

4.4.3 'Uniform' distribution on ℓ_p spheres

The fact that the uniform distribution on ℓ^1-norm and ℓ^2-norm unit spheres can be generated from i.i.d. exponential and Gaussian random variables can be generalized to

Theorem 4.1. *Let $(X_1, \ldots, X_n)$ be sampled uniformly from*

$$\{(x_1, \ldots, x_n) \in \mathbb{R}^n : |x_1|^p + \cdots + |x_n|^p = [1, 1+\delta]\}, \quad \delta \to 0. \tag{4.12}$$

Then

1. *$X \overset{d}{=} V/\|V\|_p$ where $V_1, \ldots, V_n$ are i.i.d random variables with density*

$$P_V(x) \propto \exp\left(-\frac{|v|^p}{p}\right)$$

2. *$(|X_1|^p, \ldots, |X_n|^p) \sim \text{Dir}\left(\frac{1}{p}, \ldots, \frac{1}{p}\right)$*

Proof. By symmetry it is sufficient to prove the results when the random variables are restricted on $\mathbb{R}_+^n$. The marginal density of $(X_1, \ldots, X_{n-1})$ is

$$\begin{aligned} f_{X_{1:n-1}}(x_1, \ldots, x_{n-1}) &\propto \partial_{a=1} \int_0^\infty dx_n 1(x_1^p + \cdots + x_n^p \le a) \\ &= \partial_{a=1}(a - x_1^p - \cdots - x_{n-1}^p)^{\frac{1}{p}} \\ &= (1 - x_1^p - \cdots - x_{n-1}^p)^{\frac{1}{p}-1} \end{aligned}$$

Let $Y = V/\|V\|_p$. Since $V_i^p \sim \Gamma\left(\frac{1}{p}, \frac{1}{p}\right)$, we have

$$(Y_1^p, \ldots, Y_n^p) = \frac{(V_1^p, \ldots, V_n^p)}{V_1^p + \cdots + V_n^p} \sim \text{Dir}\left(\frac{1}{p}, \ldots, \frac{1}{p}\right)$$

which implies that the joint density of $U = (Y_1^p, \dots, Y_{n-1}^p)$ is

$$f_U(u_1, \dots, u_{n-1}) \propto u_1^{\frac{1}{p}-1} \dots u_{n-1}^{\frac{1}{p}-1}(1 - u_1 - \cdots - u_{n-1})$$

from which we obtain the joint density of $(Y_1, \dots, Y_{n-1})$ as

$$f_{Y_{1:n-1}}(y_1, \dots, y_{n-1}) \propto (1 - y_1^p - \cdots - y_{n-1}^p)^{\frac{1}{p}-1}$$

which is the same as the density of $f_{X_{1:n-1}}$. □

It follows from Theorem 4.1 that for any $p > 0$, the uniform distribution on

$$\{x \in \mathbb{R}^n : |x_1|^p + \cdots + |x_n|^p \in [n, n + \delta]\}, \quad \delta \to 0.$$

is asymptotically equivalent to $\mu_p^{\otimes n}$, where μ_p is the probability measure such that

$$\mu_p(dx) \propto \exp\left(-\frac{|x|^p}{p}\right) dx$$

which is the same conclusion reached by the Gibbs principle, by considering the space $\mathbb{R}^n$ equipped with Lebesgue measure and energy function $H = \sum_i |x_i|^p$.

Remark 4.4. In [96], the density of $V/\|V\|_p$ in Theorem 4.1 is called the *uniform distribution* on the ℓ^p-norm unit sphere, although this distribution is truly uniform only if $p \in \{1, 2\}$ (Remark 4.1).

4.4.4 Random partition of large integers

A *partition* of a positive integer n is a way of writing it as sum of positive integers without regard to orders. Denote $p(n)$ be the number of partitions of n. For example, $n = 4$ has five partitions

$$4 = 3 + 1 = 2 + 2 = 2 + 1 + 1 = 1 + 1 + 1 + 1$$

so $p(4) = 5$.

One way of represent a partition π of n is by letting n_k be the number of k in π and write $\pi = (n_1, n_2, \dots)$. For example, the partition $4 = 2 + 1 + 1$ correspond to the sequence $(2, 1, 0, 0, \dots)$.

Let $\pi = (N_1, N_2, \dots)$ be a partition chosen uniformly randomly from the set of all partitions of a large positive integer n. In other words, $(N_1, N_2, \dots)$ follows the the uniform distribution on the finite set

$$\left\{(n_1, n_2, \dots) \in \mathbb{N}^n : \sum_{k \geq 1} k n_k = n\right\}$$

This can be seen as the microcanonical ensemble at energy level n, where the energy of the configuration $(n_1, n_2, \dots) \in \mathbb{N}^n$ is given by $\sum_k k n_k$. Instead of working directly with this microcanonical ensemble, it is more convenient to work with the canonical ensemble

$$P_\lambda(n_1, n_2, \dots) \propto e^{\lambda \sum_{k\geq 1} k n_k} \tag{4.13}$$

for some parameter $\lambda < 0$. The free energy is simply

$$F(\lambda) = \sum_{k=1}^{n} -\log(1 - e^{\lambda k})$$

To find the value $\lambda_\star$ at which the two ensembles are equivalent, we need to solve $F'(\lambda) = n$. As $n \to \infty$, we expect that $\lambda_\star \to 0$, since N_1 should go to infinity under $P_{\lambda_\star}$ and $P_{\lambda_\star}(N_1 = m) \propto e^{\lambda_\star m}$. Therefore, the free energy can be approximated by an integral as

$$F(\lambda) \simeq \frac{1}{\lambda} \int_0^\infty \log(1 - e^{-x})\, dx$$

This integral can be computed by performing the Taylor expansion for $-\log(1-t)$ where $t = e^{-x}$ and exchanging the sum with the integral, leading the the Euler sum $\sum_{n\geq 1} \frac{1}{n^2}$. We obtain

$$\int_0^\infty -\log(1 - e^{-x})\, dx = \frac{\pi^2}{6},$$

so

$$F(\lambda) \simeq -\frac{\pi^2}{6\lambda}. \tag{4.14}$$

Solving $F'(\lambda) = n$, we obtain $\lambda_\star \simeq -\pi/\sqrt{6n}$. By (4.11) we have $S(n) = F(\lambda_\star) - \lambda_\star n \simeq \pi\sqrt{2n/3}$. Therefore

$$\log p(n) = S(n) \simeq \pi \sqrt{\frac{2n}{3}},$$

which gives the exponential term in the Hardy-Ramanujan formula

$$p(n) \simeq \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right).$$

Moreover, from the expression of the canonical ensemble, $N_1, N_2, \dots$ are asymptotically independent with

$$P(N_k = m) \propto \exp\Big(-\frac{k\pi m}{\sqrt{6n}}\Big) \tag{4.15}$$

In other words, $\pi k N_k/\sqrt{6n}$ for $k = 1, 2, \dots$ behave like independent exponential random variables with parameter 1. The precise mathematical meaning and rigorous proof of this statement can be found in Theorem 2.1 and 2.2 of [40].

4.4.5 Exponential tilting

So far we have only considered spaces $(\mathcal{X}, \mu)$ where μ is Lebesgue measure or counting measure. We now take μ to be a probability measure. Then the microcanonical ensemble with energy function $E(x) = \sum_i x_i$ and energy level na is the probability law of $X_1, \dots, X_n \overset{i.i.d}{\sim} \mu$ conditioned on the event that $\sum_i X_i = na$. Let the conditional random variables be $\tilde{X}_1, \dots, \tilde{X}_n$.

By Gibbs principle, as n tends to infinity, $\tilde{X}_i$'s behave like independent random variables generated from the probability density $\tilde{\mu}$ such that

$$\tilde{\mu}_\lambda(dx) \propto \mu(dx) e^{\lambda x} \tag{4.16}$$

where λ satisfies

$$\int \tilde{\mu}_\lambda(dx) x = a.$$

For example, if $X_i \sim \mathcal{N}(0,1)$ then $\tilde{X}_i \sim \mathcal{N}(a,1)$.

The conditioning event is (exponentially) 'rare' when $a \neq \mathbb{E}[X_i]$ and is 'typical' when $a = \mathbb{E}[X_i]$. If the conditioning event is rare then $\tilde{\mu} \neq \mu$, otherwise $\tilde{\mu} = \mu$. The measure $\tilde{\mu}$ is known as *exponential tilting* of μ, which is often used for rare-event simulations.

4.4.6 Intersection of ℓ^1- and ℓ^2-spheres

In [21], the author considered a random vector $X = (X_1, \dots, X_n)$ following the uniform distribution on

$$K := \left\{ x = (x_1, \dots, x_n) \in \mathbb{R}_+^n : \sum_i x_i = n, \sum_i x_i^2 = nb \right\}$$

By Gibbs principle, as $n \to \infty$, $(X_1, \dots, X_n)$ is asymptotically equivalent to $(Z_1, \dots, Z_n)$, where Z_i are independent random variables following the density

$$P_{r,s}(dx) \propto e^{-rx^2 - sx} dx$$

where $r > 0$ and s satisfies $\int x P_{r,s}(dx) = 1$ and $\int x^2 P_{r,s}(dx) = b$. These equations has a solution for $1 < b \leq 2$, in which case the paper showed that for any fixed k, the random vector $X_{1:k}$ converges in law to $Z_{1:k}$ as $n \to \infty$. Moreover, all joint moments of $X_{1:k}$ converges to the corresponding moments of $Z_{1:k}$.

The paper also discussed the case $b > 2$ where Gibbs principle is not applicable. In this case, it is showed that if $Z_1, \dots, Z_n$ are independent random variables following Exp(1), then $X_{1:k}$ converges in law to $Z_{1:k}$, but the convergence in moments does not hold.

4.5 A useful result

Gibbs principle and the free energy-entropy relation can be combined into the following result

Result 4.1. *Consider a measure space $(\mathcal{X}, \mu)$. Let E be a function from $\mathcal{X}$ to $\mathbb{R}$ such that $S(a)$ defined in (4.2) is concave. Then microcanonical ensemble $\mathcal{U}(\{E(x) \in da\})$ is asymptotically equivalent to a system with Hamiltonian of the following form*

$$H = \lambda(E(x) - a)$$

for some $\lambda \in \mathbb{R}$. The equivalence is achieved at

$$\lambda(a) = \arg\min_{\lambda} \psi(\lambda, a)$$

where $\psi(\lambda, a)$ be the free energy of the H. Moreover, the entropy function $S(a)$ defined in (4.2) satisfies

$$S(a) = \min_{\lambda} \psi(\lambda, a)$$

The following result, obtained from Result 4.1, will be used by all replica computations in this thesis.

Result 4.2. *Consider the Hamiltonian $H = H(x, q(x))$ on a measure space $(\mathcal{X}, \mu)$, where $q(x)$ is a function from $\mathcal{X}$ to $\mathbb{R}^k$ that satisfies a certain condition that will be stated later in Remark 4.6. The coordinates of $q(x)$ are called **macroscopic functions**. Define the following parametrized Hamiltonian associated with $q(x)$,*

$$\bar{H} = H(x, a) + \langle \lambda, q(x) - a \rangle, \quad \lambda, a \in \mathbb{R}^k.$$

*where a and λ are respectively called **constraint parameters** and **multipliers**. Let $F(\lambda, a)$ be the free energy of $\bar{H}$, called **free energy function**. Consider all solutions $(\lambda_\alpha, a_\alpha)$ of*

$$\max_{a} \min_{\lambda} F(\lambda, a).$$

*These solutions are called **dominant extremal points** of the free energy function. Let H_α be the Hamiltonian $\bar{H}$ with parameters $(\lambda, a) = (\lambda_\alpha, a_\alpha)$ and P_α be the probability measure associated with H_α. Then P^H in the infinite size limit is equivalent to a mixture of P_α's. The measures P_α's are called **pure states**. Moreover, in each pure state, $q(x)$ is concentrated at a_α. If there is a finite number of pure states, the free energy of the system is equal to the free energy of each pure state.*

Derivation. P^H can be viewed as a mixture of the following probability measures

$$P_a(dx) \propto e^{H(x,a)} 1_{q(x)\in da}\, \mu(dx)$$

where a range over all possible values of $q(x)$. The weight of P_a in this mixture is given by

$$w(a)da = \int e^{H(x,a)} 1_{q(x)\in da}\, \mu(dx).$$

Note that weights do not necessarily adds up to one.

Each P_a can be seen as a microcanonical ensemble $\{q(x) \in da\}$ over the measure space $(\mathcal{X}, \mu_a)$, where $\mu_a(dx) = e^{H(x,a)}\mu(dx)$. Using the result of 4.1, P_a is equivalent to the following Hamiltonian on $(\mathcal{X}, \mu)$

$$H(x, a) + \langle \lambda(a), q(x) - a \rangle$$

where

$$\lambda(a) = \arg\min_{\lambda} F(\lambda, a)$$

The weight of P_a in the mixture satisfies

$$\log w(a) = \min_{\lambda} F(\lambda, a)$$

In the infinite size limit, the mixture will concentrates on values of a that maximize $\log w(a)$. This leads to the conclusion that P^H behaves like a mixture of P_α as stated in the result. □

Remark 4.5. If the free energy function has only one dominant extremal point $(\lambda_\star, a_\star)$, then

- The macroscopic functions concentrate at the extremal value of the constraint parameters, i.e. $q(x)$ concentrates at $a_\star$.
- The free energy of the system is given by $F(\lambda_\star, a_\star)$.
- H is asymptotically equivalent to $\bar{H}(x, \lambda_\star, a_\star)$

Remark 4.6. From Remark 4.2, in order for Result 4.2 to hold, we need to make the assumption that the function

$$a \to \log \frac{\mu_c(\{q(x) \in da\})}{da}$$

is concave for any measure $\mu_c(dx)$ defined as $\mu_c(dx) = e^{H(x,c)}\mu(dx)$. For the Hamiltonians of the form $H(q(x))$, this assumption simply means that the function

$$a \to \log \frac{\mu(\{q(x) \in da\})}{da}$$

is concave. ◊

4.5.1 Application: random field Ising model

We now take a look at an application of Result 4.2. Consider $\mathcal{X} = \{1, -1\}^n$ with counting measure and the following Hamiltonian

$$H(\sigma) = \frac{\beta}{n} \sum_{i<j} \sigma_i \sigma_j + \beta \sum_i h_i \sigma_i$$

where $\beta > 0$ and the parameters h_i are i.i.d. as h. This is called the *random field Ising model*, in which (h_i) represents the random field and (σ_i) represents the spins, each of which can take the direction up ($\sigma_i = +1$) or down ($\sigma_i = -1$). The *Curie-Weiss model* corresponds the the case where h is constant. This example can be found in [56], where it is solved by the replica method. However, since the Hamiltonian can be expressed in terms of simple macroscopic functions, there is no need for replica method.

Define the following macroscopic functions $m(\sigma) = n^{-1} \sum_i \sigma_i$, called the *magnetization*, and $q(\sigma) = n^{-1} \sum_i h_i \sigma_i$. The Hamiltonian H can be written as

$$H(\sigma) \simeq \beta n (\frac{1}{2} m(\sigma)^2 + q(\sigma)),$$

in which we ignore the terms of order $o(n)$. By Result 4.2, the parametrized Hamiltonian associated with these macroscopic functions is:

$$\bar{H} = \beta n (\frac{1}{2} m^2 + q) + \hat{m} \Big(\sum_i \sigma_i - nm \Big) + \hat{q} \Big(\sum_i h_i \sigma_i - nq \Big)$$

where m, q are constraint parameters and $\hat{m}, \hat{q}$ are multipliers. The free energy of $\bar{H}$ is the sum of the free energies of $H_i(\sigma_i) := (\hat{m} + \hat{q} h_i) \sigma_i$ with the terms that does not depend on σ. We obtain $F^{\bar{H}} = n f(m, q, \hat{m}, \hat{q})$, where

$$\begin{aligned} f(m, q, \hat{m}, \hat{q}) &= \frac{1}{2} \beta m^2 + \beta q - m\hat{m} - q\hat{q} + n^{-1} \sum_i \log 2 \cosh(\hat{m} + \hat{q} h_i) \\ &\simeq \frac{1}{2} \beta m^2 + \beta q - m\hat{m} - q\hat{q} + \mathbb{E}_h \log 2 \cosh(\hat{m} + h\hat{q}), \end{aligned}$$

where the second approximation comes from the law of large numbers. Differentiating f by m, q, we obtain $\hat{m} = \beta m, \quad \hat{q} = \beta$. Plugging these into f, we get

$$-\frac{1}{2} \beta m^2 + \mathbb{E}_h \log 2 \cosh(\beta(m + h)).$$

Differentiate this function by m, we obtain

$$m = \mathbb{E} \tanh \beta(m + h). \tag{4.17}$$

If this equation has a unique equation $m_\star$, then in the infinite size limit,

- σ_i's behave like independent spins generated from $P(\sigma_i) \propto e^{\beta(m_\star + h_i)\sigma_i}$.

- $m(\sigma)$ concentrates at $m_\star$.

Consider the case $h = 0$. At high temperature ($\beta < 1$), the equation (4.17) has unique solution $m_\star = 0$. The system has zero magnetization and σ_i's behave like independent spins with no preferential direction. At low temperature ($\beta > 1$), by symmetry, (4.17) has two solutions $\pm m_\star$ and P^H behaves like a mixture of two probability measures P_+ and P_- with equal weights, where $P_\pm(\sigma_i) \propto e^{\pm\beta m_\star \sigma_i}$. Moreover, the magnetization takes one of the two values $\pm m_\star$, each with probability $1/2$. $\Diamond$

4.6 Computing asymptotic equivalents with replicas

4.6.1 Replicated Hamiltonian and replica density

Consider a disordered system with random Hamiltonian H on a configuration space $\mathcal{X}$ with underlying measure μ. Recall that the probability measure associated with H is given by

$$P^H(dx) = \frac{e^{H(x)}}{Z}\mu(dx) \tag{4.18}$$

where $Z = \int \mu(dx)e^{H(x)}$. We define the following formal object called the **replicated Hamiltonian**

$$H^{\text{rep}}(x^1, \dots, x^r) = \log \mathbb{E}e^{H(x^1)+\cdots+H(x^r)}, \quad r \to 0.$$

In other words, H^{rep} is a formal function with an infinitesimal number of arguments taking values in $\mathcal{X}$. We will not try to make sense of this crazy object. Instead, we will see that this object provides a convenient representation of the system under study and we will manipulate this object to obtain meaningful results at the end.

The corresponding probability measure $P^{H^{\text{rep}}}$, called the **replica density** and denoted by P^{rep} for short, is given by

$$P^{\text{rep}}(x^1, \dots, x^r) = \mathbb{E}e^{H(x^1)+\cdots+H(x^r)}, \quad r \to 0. \tag{4.19}$$

Remark 4.7. P^{rep} as given in the equation (4.19) is formally a probability density, since $\int dx^1, \dots, dx^r P^{\text{rep}}(x^1, \dots, x^r) = \mathbb{E}[Z^r] \to 1$ as $r \to 0$. This explains why we use the equality sign instead of the proportional sign in (4.19). This also explains why r is sent to zero instead of other values.

Remark 4.8. The replica density P^{rep} encodes information about the replicas of the system. Indeed, for any $k \in \mathbb{N}_+$, we can formally write the marginal law of k of its coordinates as

$$P_k^{\text{rep}}(x^1, \dots, x^k) = \int \mu(dx^{k+1}) \dots \mu(dx^r) P^{\text{rep}}(x^1, \dots, x^r),$$

which can be understood by first taking r larger than k then analytically continuing the formula to $r \to 0$. Next, we have

$$\begin{aligned} P_k^{\text{rep}}(x^1, \dots, x^k) &= \int \mu(dx^{k+1}) \dots \mu(dx^r) \mathbb{E}[e^{H(x^1)+\dots+H(x^r)}] \\ &= \mathbb{E} \int \mu(dx^{k+1}) \dots \mu(dx^r) e^{H(x^1)+\dots+H(x^r)} \\ &= \mathbb{E}[e^{H(x^1)+\dots+H(x^k)} Z^{r-k}] \\ &\to \mathbb{E}[e^{H(x^1)+\dots+H(x^k)} Z^{-k}], \quad r \to 0 \\ &= \mathbb{E}[P^H(x^1) \dots P^H(x^r)], \end{aligned}$$

where the last expression is the density of the replicas $X^1, \dots, X^k \overset{i.i.d}{\sim} P^H$. Since k can be any positive integer, we conclude that P^{rep} encodes information about the replicas of the system.

4.6.2 Three main steps

The replica computation of the asymptotic equivalent of a disordered system P can be summarized in the following diagram

$$P \to P^{\text{rep}} \to \bar{P}^{\text{rep}} \to \bar{P}$$

which represents a process with three steps

1. (Replication) Compute the replica density P^{rep}.
2. (Simplification) Compute the asymptotic equivalent of P^{rep}, denoted by $\bar{P}^{\text{rep}}$, which is also a replica density.
3. (Dereplication) Translate $\bar{P}^{\text{rep}}$ back to its corresponding disordered system $\bar{P}$. We conclude that the asymptotic equivalent of P^H is $\bar{P}$.

In the so-called *replica symmetric* case, we can work directly with the Hamiltonian through a similar process

$$H \to H^{\text{rep}} \to \bar{H}^{\text{rep}} \to \bar{H}$$

with the end result being $H \leftrightarrow \bar{H}$. Since $F^H \simeq F^{\bar{H}}$ and $P^H \leftrightarrow P^{\bar{H}}$, we achieve simultaneously the computation of free energy and the description of the measure P^H. For a Hamiltonian H consisting of a deterministic part H_0 and a stochastic part H_1, we only need to compute the asymptotic equivalent of H_1 since $\bar{H} = H_0 + \bar{H}_1$.

We now discuss in more details the three steps. The calculation in the replication step do not need to be exact, as we only need to keep the dominant part and throw away the rest. If this step produces a replicated Hamiltonian that depends on a few simple macroscopic functions, then the problem is likely to be solved with replicas. In the

simplification step, Result 4.2 will be applied formally. This leads to the extremization of a 'formal' free energy function F with a real number of variables. We then assume that the dominating extremal point of F has a certain form called the *replica ansatz*, which encodes in a generic and compact way the asymptotic behavior of the system. Moreover, with the replica ansatz, the function F now becomes a 'usual' function with a natural number of variables, allowing us to perform the extremization. The end result of this step is a simpler replica density $\bar{P}^{\text{rep}}$ that is asymptotically equivalent to P^{rep}. The dereplication step is the reverse of the replication step. If P^{rep} (H^{rep}) is the replication of P (H) then the dereplication of P^{rep} (H^{rep}) is P (H). The following examples will be useful for the replication and dereplication steps.

Example 4.1. If H does not contains any random parameters, then its replication is simply

$$H^{\text{rep}} = H(x^1) + \cdots + H(x^r)$$

In particular, the dereplication of rA, where A is a constant, is A.

Example 4.2. If $H^{\text{rep}} = f(r)$ for some differentiable function f such that $f(0) = 0$, then $H = f'(0)$. This follows from the fact that $H \simeq rf'(0)$ as $r \to 0$.

Example 4.3. Let X be a random variable with value in $\mathcal{X}$ and corresponding probability measure P_X. Consider the measure space $(\mathcal{X}, P_X)$. Let ϕ be a function from $\mathcal{X}$ to $\mathbb{R}^k$. The following replicated Hamiltonian

$$H^{\text{rep}} = \sum_{1 \le a < b \le r} \langle \phi(x^a), \phi(x^b) \rangle$$

corresponds to the following Hamiltonian with random parameters $Z \sim \mathcal{N}(0, I_k)$

$$H = \langle Z, \phi(x) \rangle - \frac{1}{2} \|\phi(x)\|^2$$

Example 4.4. This example will be used both for replication and dereplication in the replica analysis of inference problems in Bayes optimal setting (Chapter 7). With the same measure space considered in the previous example, consider

$$H^{\text{rep}} = \sum_{0 \le a < b \le r} \langle \phi(x^a), \phi(x^b) \rangle$$

where x^0 is a fixed vector. Using the previous example, this replicated Hamiltonian corresponds to the following Hamiltonian

$$H = \langle \phi(x^0) + Z, \phi(x) \rangle - \frac{1}{2} \|\phi(x)\|^2$$

P^H is exactly the posterior of X given Y in the following Gaussian channel

$$Y = \phi(x^0) + Z, \tag{4.20}$$

where x^0 is an unknown signal generated from the probability law P_X and Z is a standard Gaussian vector independent of X.

Remark 4.9. The dereplication step implicitly requires that each replica density corresponds to a unique disordered system. This can be 'proved' by the fact that the replica density encodes information about the replicas of the system (Remark 4.8), and that a disordered system is uniquely determined by its infinite sequence of replicas, by de Finetti theorem (Theorem 3.1).

The replica method can be seen as consisting of two parts: the qualitative part, which involves selecting a replica ansatz that describes the generic asymptotic behavior of the system, and the quantitative part, which involves the remaining computations. The qualitative part is straightforward for some problems in random matrix theory, random convex optimization, and Bayesian inference, where the replica symmetric ansatz gives the correct results. However, for the SK model, it presents the main challenge of the problem. In this case, the replica symmetric ansatz yields incorrect results at low temperature. The correct answer is given by the intriguing Parisi ansatz. Replica method will produce the correct answer if the assumptions encoded in the replica ansatz is correct.

The Parisi ansatz also describes the behaviors many other disordered systems such as generalized random energy model [32], p-spin spherical model [29] and notably the sphere packing problem [20]. Models that can be described by the Parisi ansatz are said to be in the same universal class. In some cases, we need to modify the Parisi ansatz to obtain the correct result [33]. In other cases, there is no known replica ansatz that corresponds to the qualitative behavior of the system [18]. One may wonder if it is possible to classify all universal behaviors for disordered systems, and whether each of these behavior can be encoded in some replica ansatz.

Chapter 5

Random matrix theory

Random matrix theory (RMT) has been successfully applied to wireless communication [24] and machine learning [27]. One of the main technical tools behind these applications is the method of deterministic equivalents, which goes beyond the classic Stieltjes transform. The crucial idea behind this tool is that, in order to study a random symmetric matrix M, instead of computing the limit of $\mathrm{Tr}(M - zI)^{-1}$, which is the Stieltjes transform of the spectral measure of M, the method computes a deterministic matrix that behaves like the resolvent $(M - zI)^{-1}$. From this we can study not only the limiting behaviors of the eigenvalues of M but also the eigenvectors, which often contains information about the signal hidden behind the data of matrix M.

In this chapter, we will show that deterministic equivalents can be computed by the replica method. While the standard techniques such as Gaussian method and Bai-Silverstein heuristic tend to probe the intricate dependencies between matrix entries, the replica method takes a global view as it deals with macroscopic variables. The two kind of methods thus complement each other. Moreover, the vast literature on random matrix theory could serve as an excellent testing ground for practicing replica computation, and replica method could be a useful tool to try when the standard computations is too complicated.

In section 5.1, we will present basic tools and concepts in the resolvent approach to random matrix theory. In section 5.2, we will show how the replica method can be used to study resolvents of large random matrices. Section 5.3 shows several applications of the replica method to different random matrix models.

5.1 Basic tools and concepts

The *resolvent* of a symmetric matrix $M \in \mathbb{R}^{n\times n}$ is defined as

$$Q_M(z) = (M - zI)^{-1}$$

for $z \in \mathbb{C}_+$. This matrix contains both information about the eigenvalues and the eigenvectors of M and will be the central object of our study. It turns out that the resolvent of

a large random matrices often behaves like a deterministic matrix, called the *deterministic equivalent*. Rigorously speaking, the deterministic equivalent of a symmetric random matrix $Q \in \mathbb{R}^{n\times n}$ is a deterministic matrix $\bar{Q} \in \mathbb{R}^{n\times n}$ such that for any deterministic matrix A with unit operator norm,

$$\frac{1}{n}\operatorname{Tr} A(Q-\bar{Q}) \to 0,$$

and for any vectors a, b with unit Euclidean norm

$$a^\top (Q-\bar{Q})b \to 0 \tag{5.1}$$

as $n \to \infty$, where the convergence is almost sure or in probability.

Given a matrix M, the first thing we are interested in is its eigenvalues $(\lambda_i)_{i=1}^n$, which uniquely correspond to the *spectral measure* of M, defined as

$$\mu_M(dx) = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}(dx) \tag{5.2}$$

For a random matrix $M \in \mathbb{R}^{n\times n}$, we are interested in its limiting spectral density (LSD), i.e. the limiting shape that emerges from plotting its eigenvalues as $n \to \infty$. This can be defined as the weak limit of the sequence of spectral measures μ_M as $n \to \infty$. In general, a sequence of measure (μ_n) *converges weakly* to the measure μ if

$$\lim_{n\to\infty} \int f(x)\mu_n(dx) = \int f(x)\mu(dx) \tag{5.3}$$

for any continuous f with compact support. The LSD μ of a random matrix can be cleverly computed by first computing a function $g_\mu(z)$ called *Stieltjes transform* of μ and then recover μ from this function. In general, the Stieltjes transform of a measure μ with compact support $\operatorname{supp}(\mu)$ is defined as

$$g_\mu(z) = \int_{\mathbb{R}} \frac{\mu(d\lambda)}{\lambda - z}, \quad z \in \mathbb{C}\backslash \operatorname{supp}(\mu). \tag{5.4}$$

For a matrix M with real eigenvalues, we define Stieltjes transform of M as $g_M(z) = g_{\mu_M}(z)$. In particular, if the eigenvalues of M are $\lambda_1, \dots, \lambda_n \in \mathbb{R}$, then

$$g_M(z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda_i - z} = \frac{1}{n}\operatorname{Tr} Q_M(z)$$

A measure can be recovered from its Stieltjes transform by the following result

Theorem 5.1. *Let $g_\mu(z)$ be the Stieltjes transform of a measure μ with compact support. If μ admits a density at x, then*

$$\mu(x) = \frac{1}{\pi}\lim_{y\downarrow 0} \operatorname{Im} g_\mu(x+iy) \tag{5.5}$$

Otherwise, if μ has an isolated mass at x, then

$$\mu(\{x\}) = \lim_{y\downarrow 0} -iy g_\mu(x+iy) \tag{5.6}$$

Now the question is how to compute the Stieltjes transform of the LSD. This is where the resolvent and deterministic equivalent come into the scene. The properties of deterministic equivalent implies that

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Tr}(Q(z) - \bar{Q}(z)) = 0,$$

from which we obtain $g_\mu(z) = \lim_{n\to\infty} \frac{1}{n} \operatorname{Tr} \bar{Q}(z)$. This limit can be computed easily since $\bar{Q}(z)$ is deterministic.

For many random matrix models, $g(z)$ is characterized as the solution of some fixed point equations. To compute the LSD, we need to compute $g(z)$ for complex values of z near the real axis. The fixed point equations might have several solutions. To determine which solution corresponds to the correct $g(z)$, we use the fact that $g(z) \simeq -\frac{1}{z}$ as $z \to \infty$. The details of how to recover a measure μ from the fixed point equations that define its Stieltjes transform can be found in any textbook on RMT. Most of our discussions will reach their final point when the fixed point equations is obtained.

The deterministic equivalent is also useful for studying spiked models. Suppose M is a random matrix with spectrum consisting of continuous bulks and isolated eigenvalues. When z is outside the bulks, $Q(z)$ and $\bar{Q}(z)$ are asymptotically equivalent, so we expect that they have asymptotically the same singular points[1] in this region. These singular points are no other than the isolated eigenvalues of M. On the other hand, for any deterministic vector u, exploiting the fact that $u^\top Q(z) u \simeq u^\top \bar{Q}(z) u$, we can compute the correlation between u and the eigenvectors corresponding to the isolated eigenvalues.

The remaining question is how to compute the deterministic equivalent of resolvents. This can be done by standard techniques such as Gaussian method and Bai-Silverstein heuristic (Chapter 2 of [27]). In the next section we will show how to do this with the replica method.

5.2 Deterministic equivalent and replicas

To compute the deterministic equivalent of a random matrix, the general strategy is to consider a quadratic Hamiltonian that is related in some way to this random matrix. Let us consider the problem of computing the deterministic equivalent of $Q = (M - zI)^{-1}$ for some symmetric random matrix $M \in \mathbb{R}^{n\times n}$. We can consider the Hamiltonian

$$H = \frac{1}{2} z \|x\|^2 - \frac{1}{2} x^\top M x,$$

on the measure space $(\mathbb{R}^n, \mu)$, where $\mu(dx) = (2\pi)^{-n/2} dx$. The measure μ is for removing the constant term in the calculations of free energy (Example 1.8). Suppose by replica computations we obtain $H \leftrightarrow \bar{H} = -\frac{1}{2} x^\top \bar{Q} x$. Since $P^H \sim \mathcal{N}(0, Q)$ and $P^{\bar{H}} \sim \mathcal{N}(0, \bar{Q})$, we conclude that Q asymptotically equivalent to $\bar{Q}$.

[1] Singular points of a meromorphic function are values at which the function blows up.

Let us consider another problem where we want to compute the deterministic equivalents of $(MM^\top - zI)^{-1}$ and $(M^\top M - zI)^{-1}$ for some large random matrix $M \in \mathbb{R}^{n\times n}$. This can be obtained by studying the Hamiltonian

$$H(x,y) = -\frac{1}{2}t(\|x\|^2 + \|y\|^2) - x^\top My$$

on the measure space $(\mathbb{R}^n, \mu)$, where $\mu(dx) = (2\pi)^{-n/2}dx$. The marginal laws of x and y of P^H is given by

$$P_x^H \sim \mathcal{N}(0, (tI - t^{-1}MM^\top)^{-1})$$
$$P_y^H \sim \mathcal{N}(0, (tI - t^{-1}M^\top M)^{-1})$$

Therefore, if we can show that H is asymptotic equivalent to a certain deterministic $\bar{H}$, then by computing x- and y-margins of $P^{\bar{H}}$ we obtain the deterministic equivalents of P_x^H and P_y^H, which implies the deterministic equivalents of $(tI - t^{-1}MM^\top)^{-1}$ and $(tI - t^{-1}M^\top M)^{-1}$. The equivalents for the resolvents can be easily deduced from this by a simple change of variable.

The behavior of a large random matrix is generally insensitive to the specific law governing its entries. For example, the results for the GOE matrix hold more generally for entries with mean zero and variance n^{-1}. To simplify the calculations, all the examples we provide in the next section will involve Gaussian matrices, although the results hold for more general assumptions.

5.3 Applications

5.3.1 GOE

Let A be a GOE matrix. We are interested in computing the deterministic equivalent of the resolvent matrix $Q(z) = (A - zI)^{-1}$. To do this we will study the system with the following Hamiltonian

$$H = \frac{1}{2}z\|x\|^2 - \frac{1}{2}x^\top Ax$$

on the measure space $(\mathbb{R}^n, \mu)$, where $\mu(dx) = (2\pi)^{-n/2}dx$. This Hamiltonian is related to the resolvent $Q(z)$ by the fact that $P^H \sim \mathcal{N}(0, Q(z))$. The Hamiltonian H can be written as the sum of the deterministic part $H_0 = \frac{1}{2}z\|x\|^2$ and the stochastic part $H_1 = -\frac{1}{2}x^T Ax$. The replication of H_1 is

$$H_1^{\text{rep}} = \frac{1}{4n}\sum_{a,b}\langle x^a, x^b\rangle^2$$

Next, we will compute the asymptotic equivalent of H_1^{rep}. By formally applying Result 4.2 for the configuration space $(\mathbb{R}^n)^r$ with macroscopic functions $\langle x^a, x^b\rangle/n$ with $1 \leq$

$a \le b \le r$, also called the *overlaps*, the parametrized Hamiltonian H_1^{rep} has the following form

$$\bar{H}_1^{\text{rep}} = \frac{n}{4}\sum_{a,b}(Q^{ab})^2 + \sum_{a\le b}\hat{Q}^{ab}(\langle x^a, x^b\rangle - nQ^{ab})$$

where Q^{ab}'s are the constraint parameters and $\hat{Q}^{ab}$'s are the multipliers. Assuming that the free energy function has a unique dominant extremal point $(Q_\star, \hat{Q}_\star)$. This uniqueness implies that the extremal point must be invariant by permutations of replica indexes, otherwise the points obtained from $(Q_\star, \hat{Q}_\star)$ by these permutations are also dominant extremal points (note that the free energy function is invariant by replica permutations). Therefore, the $Q_\star^{ab}$ has the following form, called *replica symmetric ansatz*

$$\begin{aligned} Q_\star^{aa} &= q, \quad Q_\star^{ab} = s, \\ \hat{Q}_\star^{aa} &= \hat{q}, \quad \hat{Q}_\star^{ab} = \hat{s}, \end{aligned}$$

On the other hand, under the replica density P^{rep}, for any $a, b \in [r]$, the overlap $\langle x^a, x^b\rangle/n$ concentrates at $Q_\star^{ab}$ (Remark 4.5). With the replica symmetric ansatz, $\langle x^a, x^b\rangle/n$ concentrates at q if $a = b$ and at s if $a \neq b$. This implies that for any two replicas X^1, X^2 of the system, $\|X^1\|^2/n \simeq q$ and $\langle X^1, X^2\rangle/n \simeq s$ (recall that the replica density encodes information about the replicas of the system, see Remark 4.8). Since $\langle X^1, X^2\rangle/n \simeq \mathbb{E}[\langle X^1, X^2\rangle/n] = 0$, we have $s = 0$ and the ansatz for Q can be narrowed down to

$$Q^{ab} = \delta_{ab} q$$

Moreover, for $a \neq b$, $\hat{Q}^{ab}$ must be 0, otherwise the overlaps between the two different replicas is nonzero. So the ansatz for $\hat{Q}$ is narrowed down to

$$\hat{Q}^{ab} = \frac{1}{2}\delta_{ab}\hat{q}$$

Here we add the factor of $1/2$ to make later computations more convenient. With this ansatz, we have

$$\bar{H}^{\text{rep}} = \sum_a \frac{1}{4}nq^2 + \frac{1}{2}\hat{q}(\|x^a\|^2 - nq)$$

Dereplicating $\bar{H}_1^{\text{rep}}$ (Example 4.1), we obtain

$$\bar{H}_1 = \frac{1}{4}nq^2 + \frac{\hat{q}}{2}(\|x\|^2 - nq)$$

Since $\bar{H} = H_0 + \bar{H}_1$, we have

$$\bar{H} = \frac{1}{2}(\hat{q} + z)\|x\|^2 + \frac{1}{4}nq^2 - \frac{1}{2}nq\hat{q}$$

We thus obtain the free energy of $\bar{H}$ as $nf(q,\hat{q})$, where

$$f(q,\hat{q}) = -\frac{1}{2}\log(-\hat{q}-z) + \frac{1}{4}q^2 - \frac{1}{2}q\hat{q}$$

The saddle point of $f(q,\hat{q})$ satisfies

$$\begin{aligned}\hat{q} &= q \\ q &= \frac{1}{-\hat{q}-z},\end{aligned}$$

from which we have

$$q^2 + zq + 1 = 0$$

From the expression for $\bar{H}$, the deterministic equivalent of $Q(z)$ is

$$\bar{Q}(z) = -(\hat{q}+z)^{-1}I = qI$$

from which we obtain the limiting Stieltjes transform $g(z) = q(z)$. From this we can obtain the limiting spectral density

$$\mu(dx) = \frac{1}{2\pi}\sqrt{4-x^2}1_{x\in[-2,2]}\,dx.$$

Remark 5.1. In the general replica symmetric computations, the following assumptions are equivalent:

- The uniqueness of the dominant extremal point
- The replica symmetric ansatz
- The concentration of the overlaps

5.3.2 A model in wireless communication

We give here a more complicated example that is studied in the paper [25] and presented in a more pedagogical way in Chapter 6 of the book [24]. Consider the following random matrix

$$B = \sum_{k=1}^{K} M_k M_k^\top$$

where

$$M_k = R_k^{\frac{1}{2}} X_k T_k^{\frac{1}{2}}$$

with the following hypothesis for $k = 1,\dots,K$:

1. $X_k \in \mathbb{R}^{n\times n_k}$ are matrices with independent entries following $\mathcal{N}(0, n_k^{-1})$. Note that in the original paper, the entries of X_k are complex Gaussian. To have a clean presentation we consider the real Gaussian entries instead.

2. $n, n_k \to \infty$, $\lim_{n\to\infty} n/n_k = c_k$. This hypothesis is weaker than that in the original paper, so that the presentation is clearer.

3. $R_k \in \mathbb{R}^{n\times n}, T_k \in \mathbb{R}^{n_k\times n_k}$ are deterministic positive definite matrices, satisfying some tightness condition as $n \to \infty$.

Our purpose is to find the deterministic equivalent of $Q(z) = (zI - B)^{-1}$. For $z \in \mathbb{R}$ such that $zI - B$ is positive definite, consider the following Hamiltonian

$$H(x, y_1, \ldots, y_k) = -\frac{1}{2}z\|x\|^2 - \frac{1}{2}\sum_{k=1}^{K}\|y_k\|^2 + \sum_{k=1}^{K} x^\top M_k y_k, \quad x \in \mathbb{R}^n, y \in \mathbb{R}^{n_k}$$

which is related to the studied model by the fact that $P_x^H \sim \mathcal{N}(0, Q(z)$. H can be divided into the deterministic part H_0 and stochastic part H_1 as

$$H_0 = -\frac{1}{2}z\|x\|^2 - \frac{1}{2}\sum_{k=1}^{K}\|y_k\|^2$$
$$H_1 = \sum_{k=1}^{K} x^\top M_k y_k$$

The replication of H_1 can be easily computed as

$$H_1^{\text{rep}} = \sum_{a,b;k} \frac{1}{n_k}\left(x^{a\top} R_k x^b\right)\left(y_k^{a\top} T_k y_k^b\right)$$

Define the following macroscopic functions

$$Q_{xk}^{ab} = \frac{x^{a\top} R_k x^b}{n} \qquad Q_{yk}^{ab} = \frac{y_k^{a\top} T_k y_k^b}{n_k}$$

The Hamiltonian associated with this change of variable is

$$\bar{H}_1^{\text{rep}} = \sum_{a,b,k} nQ_{xk}^{ab}Q_{yk}^{ab} + \sum_{a\le b;k} \hat{Q}_{xk}^{ab}(x^{a\top} R_k x^b - nQ_x^{ab}) + \hat{Q}_{yk}^{ab}(y_k^{a\top} T_k y_k^b - n_k Q_{yk}^{ab})$$

With the following replica symmetric ansatz

$$Q_{xk}^{ab} = \delta_{ab} q_{xk} \qquad \hat{Q}_{xk}^{ab} = \frac{1}{2}\delta_{ab}\hat{q}_{xk}$$
$$Q_{yk}^{ab} = \delta_{ab} q_{yk} \qquad \hat{Q}_{yk}^{ab} = \frac{1}{2}\delta_{ab}\hat{q}_{yk}$$

we have

$$\bar{H}_1^{\text{rep}} = \sum_{a;k} nq_{xk}q_{yk} + \frac{1}{2}\hat{q}_{xk}(x^{a\top}R_k x^a - nq_{xk}) + \frac{1}{2}\hat{q}_{yk}(y_k^{a\top}T_k y_k^a - n_k q_{yk})$$

Dereplicate $\bar{H}_1^{\text{rep}}$, we obtain

$$\bar{H}_1 = \sum_{k} nq_{xk}q_{yk} + \frac{1}{2}\hat{q}_{xk}(x^{\top}R_k x - nq_{xk}) + \frac{1}{2}\hat{q}_{yk}(y_k^{\top}T_k y_k - n_k q_{yk})$$

Since $\bar{H} = H_0 + \bar{H}_1$, we have

$$\bar{H} = -\frac{1}{2}x^{\top}\Big(zI - \sum_k \hat{q}_{xk}R_k\Big)x - \frac{1}{2}\sum_k y_k^{\top}(I - \hat{q}_{yk}T_k)y_k + \frac{n}{2}\sum_k q_{xk}q_{yk} - q_{xk}\hat{q}_{xk} - q_{yk}\hat{q}_{yk}$$

The free energy of $\bar{H}$ is

$$-\frac{1}{2}\log\det\Big(zI - \sum_k \hat{q}_{xk}R_k\Big) - \frac{1}{2}\sum_k \log\det(I - \hat{q}_{yk}T_k) + \frac{n}{2}\sum_k q_{xk}q_{yk} - q_{xk}\hat{q}_{xk} - q_{yk}\hat{q}_{yk}$$

Differentiate by q_{xk}, q_{yk}, we obtain $\hat{q}_{xk} = q_{yk}$, $\hat{q}_{yk} = c_k q_{xk}$, from which we have the fixed point equations

$$q_{xk} = \frac{1}{n}\operatorname{Tr} R_k\Big(\sum_{k'} q_{yk'}R_{k'} - zI\Big)^{-1}$$
$$q_{yk} = \frac{1}{n_k}\operatorname{Tr} T_k(c_k q_{xk}T_k - I)^{-1}$$

Also from the form of $\bar{H}$, we obtain the following deterministic equivalent for the resolvent

$$(zI - B)^{-1} \leftrightarrow \Big(zI - \sum_k q_{yk}R_k\Big)^{-1}$$

It is easy to check that this is the same result given in [25].

5.3.3 A model with variance profile

In this section, we will use the replica method to analyze a random matrix model with a variance profile introduced in [49]. In this case, there are two natural choices of macroscopic functions, leading to different systems of fixed point equations. One choice recovers known results, while the other leads to a system of equations that is easier to solve for low-rank variance profiles, especially those generated by a profile function of class $\mathcal{C}^1$. In hindsight, these two solutions can be proven to be equivalent through straightforward algebraic manipulations. However, in the literature, the latter is only established for a variance profile of rank one.

Let $A \in \mathbb{R}^{m\times n}$ be a deterministic matrix whose columns and rows are uniformly bounded in the Euclidean norm. Let Y be a $m \times n$ matrix with independent entries

$Y_{ij} \sim \mathcal{N}(0, V_{ij}/n)$, where the variance profile $V = (V_{ij})$ is a bounded sequence of positive numbers. Let $\Sigma = A + Y$. We want to find deterministic equivalent of the following matrices

$$Q(z) = (\Sigma\Sigma^\top - zI)^{-1}, \qquad \tilde{Q}(z) = (\Sigma^\top\Sigma - zI)^{-1}$$

in the limit where $m/n \to c \in (0, \infty)$.

For $t > 0$, consider the following Hamiltonian

$$H(x, y) = -\frac{1}{2}t(\|x\|^2 + \|y\|^2) - x^\top \Sigma y$$

As shown in Section 5.2, the deterministic equivalents for the resolvents of $\Sigma\Sigma^\top$ and $\Sigma^\top\Sigma$ can be computed from H. At first step, we write H as the sum of the deterministic part H_0 and the stochastic random part H_1,

$$\begin{aligned} H_0 &= -\frac{1}{2}t(\|x\|^2 + \|y\|^2) - x^\top A y \\ H_1 &= -x^\top Y y. \end{aligned}$$

The replication of H_1 is

$$H_1^{\text{rep}} = \frac{1}{2}\sum_{i,j}\sigma_{ij}^2 \sum_{a,b} x_i^a y_j^a x_i^b y_j^b$$

for which there are two natural choices of macroscopic functions.

macroscopic functions based directly on the entries of the variance profile

We consider here a choice of parameters for the replicated Hamiltonian H_1^{rep} that will lead to the results given in [49]. We have

$$\begin{aligned} H_1^{\text{rep}} &= \frac{1}{2}\sum_{a,b}\sum_j y_j^a y_j^b \sum_i \sigma_{ij}^2 x_i^a x_i^b \\ &= \frac{1}{2}\sum_{a,b}\sum_j y_j^a y_j^b x^{a\top} D_j x^b \end{aligned}$$

where $D_j = \text{diag}(\sigma_{ij}^2, i = 1, \dots, m)$. Similarly,

$$H_1^{\text{rep}} = \frac{1}{2}\sum_{a,b}\sum_i x_i^a x_i^b y^{a\top} \tilde{D}_i y^b$$

where $\tilde{D}_i = \text{diag}(\sigma_{ij}^2, j = 1, \dots, n)$. Therefore H_1^{rep} can be written in this symmetrical form

$$\begin{aligned} H_1^{\text{rep}} &= \frac{1}{4}\sum_{a,b}\sum_i x_i^a x_i^b y^{a\top} \tilde{D}_i y^b \\ &+ \frac{1}{4}\sum_{a,b}\sum_j y_j^a y_j^b x^{a\top} D_j x^b \end{aligned}$$

Define the following macroscopic functions

$$Q_{xj}^{ab} = \frac{1}{n} x^{a\top} D_j x^b \qquad Q_{yi}^{ab} = \frac{1}{n} y^{a\top} \tilde{D}_i y^b$$

With this change of variables, the equivalent of H_1^{rep} has the following form

$$\begin{aligned}\bar{H}_1^{\text{rep}} = &\sum_{i,a,b} \frac{n}{4} Q_{yi}^{ab} x_i^a x_i^b + \sum_{j;a\le b} \hat{Q}_{xj}^{ab} \left(x^{a\top} D_j x^b - n Q_{xj}^{ab} \right) \\ &+ \sum_{j;a,b} \frac{n}{4} Q_{xj}^{ab} y_j^a y_j^b + \sum_{j;a\le b} \hat{Q}_{yi}^{ab} \left(y^{a\top} \tilde{D}_i y^b - n Q_{yi}^{ab} \right)\end{aligned}$$

With the replica symmetric ansatz

$$\begin{aligned} Q_{xj}^{ab} &= \delta_{ab} q_{xj} & \hat{Q}_{xj}^{ab} &= \frac{1}{2} \delta_{ab} \hat{q}_{xj} \\ Q_{yi}^{ab} &= \delta_{ab} q_{yi} & \hat{Q}_{yi}^{ab} &= \frac{1}{2} \delta_{ab} \hat{q}_{yi}, \end{aligned}$$

$\bar{H}_1^{\text{rep}}$ can be simplified to

$$\begin{aligned}\bar{H}_1^{\text{rep}} = &\sum_{a,i} \frac{n}{4} q_{yi} (x_i^a)^2 + \sum_{a,j} \frac{1}{2} \hat{q}_{xj} (x^{a\top} D_j x^a - n q_{xj}) \\ &+ \sum_{a,j} \frac{n}{4} q_{xj} (y_j^a)^2 + \sum_{a,i} \frac{1}{2} \hat{q}_{yi} (y^{a\top} \tilde{D}_i y^a - n q_{yi}).\end{aligned}$$

Dereplicate $\bar{H}_1^{\text{rep}}$, we obtain

$$\begin{aligned}\bar{H}_1 = &\sum_{i} \frac{1}{4} q_{yi} x_i^2 + \sum_{j} \frac{1}{2} \hat{q}_{xj} (x^{\top} D_j x - n q_{xj}) \\ &+ \sum_{j} \frac{1}{4} q_{xj} y_j^2 + \sum_{i} \frac{1}{2} \hat{q}_{yi} (y^{\top} \tilde{D}_i y - n q_{yi}).\end{aligned}$$

Since $\bar{H} = H_0 + \bar{H}_1$, we have

$$\bar{H} = -\frac{1}{2} x^{\top} D_{\psi_x} x - \frac{1}{2} y^{\top} D_{\psi_y} y - x^{\top} A y - \frac{n}{2} \langle q_x, \hat{q}_x \rangle - \frac{n}{2} \langle q_y, \hat{q}_y \rangle$$

where ψ_x, ψ_y are such that

$$\begin{aligned} D_{\psi_x} &= tI - \frac{1}{2} D_{q_y} - \sum_j \hat{q}_{xj} D_j \\ D_{\psi_y} &= tI - \frac{1}{2} D_{q_x} - \sum_i \hat{q}_{yi} \tilde{D}_i \end{aligned}$$

The free energy density of $\bar{H}$ is

$$f = -\frac{1}{2n} \log \det \begin{pmatrix} D_{\psi_x} & A \\ A^\top & D_{\psi_y} \end{pmatrix} - \frac{1}{2}\langle q_x, \hat{q}_x \rangle - \frac{1}{2}\langle q_y, \hat{q}_y \rangle$$

We will use the following identities for computing the derivatives of f:

$$\begin{aligned} \det \begin{pmatrix} D_{\psi_x} & A \\ A^\top & D_{\psi_y} \end{pmatrix} &= \det(D_{\psi_x} - A D_{\psi_y}^{-1} A^\top) \det D_{\psi_y} \\ &= \det(D_{\psi_y} - A^\top D_{\psi_x}^{-1} A) \det D_{\psi_x} \end{aligned}$$

Differentiating f by $q_{xj}, \hat{q}_{xj}, q_{yi}, \hat{q}_{yi}$, we obtain the fixed point equations

$$\hat{q}_{xj} = (1/2n) \operatorname{Tr}(D_{\psi_y} - A^\top D_{\psi_x}^{-1} A)^{-1} E_j \tag{5.7}$$

$$q_{xj} = (1/n) \operatorname{Tr}(D_{\psi_x} - A D_{\psi_y}^{-1} A^\top)^{-1} D_j \tag{5.8}$$

$$\hat{q}_{yi} = (1/2n) \operatorname{Tr}(D_{\psi_x} - A D_{\psi_y}^{-1} A^\top)^{-1} E_i \tag{5.9}$$

$$q_{yi} = (1/n) \operatorname{Tr}(D_{\psi_y} - A^\top D_{\psi_x}^{-1} A)^{-1} \tilde{D}_i \tag{5.10}$$

where the matrix E_i is defined as having a value of one at (i, i) and zero elsewhere. Compare (5.7) with (5.10), (5.8) with (5.9), we have

$$D_{q_y} = 2 \sum_j \hat{q}_{xj} D_j \qquad\qquad D_{q_x} = 2 \sum_i \hat{q}_{yi} \tilde{D}_i$$

Plugging these equations into the definition of D_{ψ_x}, D_{ψ_y}, we have

$$D_{\psi_x} = tI - D_{q_y} \qquad\qquad D_{\psi_y} = tI - D_{q_x}$$

From these and (5.8), (5.10), we obtain

$$\begin{aligned} t - \psi_{yj} &= (1/n) \operatorname{Tr}(D_{\psi_x} - A D_{\psi_y}^{-1} A^\top)^{-1} D_j \\ t - \psi_{xi} &= (1/n) \operatorname{Tr}(D_{\psi_y} - A^\top D_{\psi_x}^{-1} A)^{-1} \tilde{D}_i \end{aligned}$$

By computing the marginals of x and y in $P^{\bar{H}}$, we have

$$\begin{aligned} (tI - t^{-1}\Sigma^\top \Sigma)^{-1} &\leftrightarrow (D_{\psi_x} - A D_{\psi_y}^{-1} A^\top)^{-1} \\ (tI - t^{-1}\Sigma \Sigma^\top)^{-1} &\leftrightarrow (D_{\psi_y} - A^\top D_{\psi_x}^{-1} A)^{-1} \end{aligned}$$

From this we can obtain the result of [49] by simple changes of variables.

Ordered parameters based on SVD of the variance profile

Suppose that variance profile $V = (\sigma_{ij}^2)_{i,j}$ can be written as $V = \sum_{k=1}^K u_k v_k^\top$ where u_k, u_k are vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$. This can be obtained from the singular value decomposition

of $V = \sum_k \lambda_k \tilde{u}_k \tilde{v}_k^\top$ with unit vectors $\tilde{u}_k, \tilde{v}_k$ by setting $u_k = \sqrt{\lambda_k} u_k$, $v_k = \sqrt{\lambda_k} v_k$. The Hamiltonian H_1^{rep} can be written as

$$H_1^{\text{rep}} = \frac{1}{2n} \sum_{k=1}^{K} (x^{a\top} D_{u_k} x^b)(y^{a\top} D_{v_k} y^b)$$

With the following choice of macroscopic functions

$$Q_{xk}^{ab} = \frac{1}{n} x^{a\top} D_{u_k} x^b \qquad Q_{yk}^{ab} = \frac{1}{n} y^{a\top} D_{v_k} y^b$$

the replica computation (whose details are very similar to the previous computation) gives

$$\begin{aligned} (tI - t^{-1}\Sigma\Sigma^\top)^{-1} \quad &\leftrightarrow \quad (D_{\psi_x} - A D_{\psi_y}^{-1} A^\top)^{-1} \\ (tI - t^{-1}\Sigma^\top\Sigma)^{-1} \quad &\leftrightarrow \quad (D_{\psi_y} - A^\top D_{\psi_x}^{-1} A)^{-1} \end{aligned}$$

where ψ_x, ψ_y are vectors such that

$$\begin{aligned} D_{\psi_x} &= tI - \sum_k q_{yk} D_{u_k} \\ D_{\psi_y} &= tI - \sum_k q_{xk} D_{v_k} \end{aligned}$$

and q_{xk}, q_{yk} for $k \in [K]$ is the solution of the following system of $2K$ equations

$$\begin{aligned} q_{xk} &= (1/n) \operatorname{Tr} D_{u_k} \left(D_{\psi_x} - A D_{\psi_y}^{-1} A^\top \right)^{-1} \\ q_{yk} &= (1/n) \operatorname{Tr} D_{v_k} \left(D_{\psi_y} - A^\top D_{\psi_x}^{-1} A \right)^{-1} \end{aligned}$$

If the variance profile is given by $\sigma_{ij}^2 = f(i/m, j/n)$ for some function $f \in \mathcal{C}^1([0,1]^2)$, then it can be well approximated by a matrix with rank $K = o(n)$. In this case, the fixed point equations based on the SVD of the variance profile is much easier to solve.

5.3.4 A spiked model

Having computed the deterministic equivalents for various random matrix models, now we show how spiked models can be studied from such results. We consider here a spiked model presented in Theorem 2.13 and 2.14 of [27] and originally studied in [6]. Our presentation is quite different in the way that it is solely based on deterministic equivalents with no further tricks. In this model, we consider a $p \times n$ matrix Z with independent entries following $\mathcal{N}(0, 1/n)$. Let

$$M = (I + P)^{\frac{1}{2}} Z Z^\top (I + P)^{\frac{1}{2}},$$

where the matrix P is symmetric, low-rank with the spectral decomposition

$$P = \sum_{i=1}^{k} \ell_i u_i u_i^\top$$

with unit vectors $u_1, \dots, u_k$ and $\ell_1 \geq \cdots \geq \ell_k > \sqrt{c}$. The matrix M is a low-rank perturbed version of the Marchenko-Pastur model. We are interested in the isolated eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$ of M as well as how the corresponding unit eigenvectors $v_1, \dots, v_k$ are related to $u_1, \dots, u_k$. We consider the setting where $p, n \to \infty$ and $p/n \to c \in (0, \infty)$ while k is fixed.

Define $Q(z) = (M - zI)^{-1}$. Let $C = I + P$. It can be shown that (by the replica method or any other tools)

$$Q(z) \leftrightarrow \bar{Q}(z) = -\frac{1}{z}(I + m(z)C)^{-1},$$

where m is the solution of

$$z + \frac{1}{m} = \frac{1}{n} \operatorname{Tr}(I + mC)^{-1} C \tag{5.11}$$

We will not provide the derivation of this result, as it can be done similarly to previous examples.

The spectrum of C consists mainly of eigenvalue 1 with no more than k eigenvalues that are different from 1. As $n \to \infty$, since the trace only depends on the eigenvalues of a matrix, we can replace C by I in (5.11) and get

$$z + \frac{1}{m(z)} = \frac{c}{m(z) + 1} \tag{5.12}$$

Recall that the isolated eigenvalues are also the singular points of $\bar{Q}(z)$, which are are values z_i such that $1 + m(z_i) + m(z_i)\ell_i = 0$. From this we obtain $m(z_i) = -1/(\ell_i + 1)$. From (5.12), we have $z_i = (\ell_i + c)(\ell_i + 1)/\ell_i$. Since $\ell_i > \sqrt{c}$, z_i form a decreasing sequence, so $z_i = \lambda_i$, In summary,

$$\lambda_i = \frac{(\ell_i + c)(\ell_i + 1)}{\ell_i}. \tag{5.13}$$

As $z \to \lambda_i$, we have

$$Q(z) \sim \frac{v_i v_i^\top}{\lambda_i - z}, \quad \bar{Q}(z) \sim \frac{u_i u_i^\top}{-z(1 + m(z) + m(z)\ell_i)}.$$

From $u_i^\top Q u_i \simeq u_i^\top \bar{Q} u_i$, we obtain

$$\frac{|u_i^\top v_i|^2}{\lambda_i - z} \sim \frac{1}{-z(1 + m(z) + m(z)\ell_i)}, \quad z \to \lambda_i.$$

Therefore

$$|u_i^\top v_i|^{-2} = \lim_{z\to\lambda_i} \frac{z(1+m(z)+m(z)\ell_i)}{z-\lambda_i} = \lambda_i(1+\ell_i)m'(\lambda_i) \tag{5.14}$$

where the second equality follows from $1+m(\lambda_i)+m(\lambda_i)\ell_i = 0$. Differentiating (5.12) by z, we obtain

$$m'(z) = \Big(\frac{1}{m(z)^2} - \frac{c}{(m(z)+1)^2}\Big)^{-1}$$

From this and $m(\lambda_i) = -1/(\ell_i+1)$, we can compute $m'(\lambda_i)$ in (5.14) and obtain

$$|u_i^\top v_i|^2 = \frac{\ell_i^2 - c}{\ell_i^2 + c\ell_i}.$$

Bibliographical notes

We refer readers to [66] and [101] for textbooks on RMT with a theoretical flavor, focusing on the joint density and microstructure of eigenvalues. More applied textbooks include [110] and [24] for wireless communication, [4] for statistics, and [27] for machine learning.

The idea of computing the deterministic equivalent from the replica method is not new and can be found in [19] and [89], where $\mathbb{E}[M]$ is defined as the asymptotic equivalent of M. The replica method also works for Haar matrices; in [89], many results from free probability are derived using the replica method. Asymptotic equivalents for models with Haar matrices can also be computed using tools from free probability [98] [111].

Spiked models have found various applications in classification in machine learning. Recent literature includes [57] in the context of online learning, and [112] and [23] for kernel puncturing. These papers aim to reduce the cost of classification with a small loss in performance. In a different type of spiked model, [11] obtained a very general and elegant result, using only elementary linear algebra, with some additional assumptions on the low-rank perturbation.

It is interesting that non-linear models of random matrices can be written as a weighted sum of a GOE matrix and a Marchenko-Pastur matrix, independent of each other, as shown in [94], which generalizes the results of [84] and [12].

In this manuscript, we only consider the expansion up to $O(r)$, where r is the number of replicas. Higher-order terms can be used to compute the fluctuations of free energy, as shown in [74] and [75] for the case of random matrices. It would be great if we could integrate the fluctuation calculations into the framework of the deterministic equivalent. On the other hand, using techniques from RMT, it is possible to compute free energy fluctuations [48] [50], as well as fluctuations of the spikes [26].

Chapter 6

Random convex optimization

Section 6.1 will introduce the Convex Gaussian min-max theorem (CGMT), a powerful tool to study random convex optimization problems in high dimension. We will see that although the theorem is stated and proved for min-max optimization problems, it holds for much relaxed condition. Section 6.2 will give some applications of the CGMT. Section 6.3 we will derive the CGMT from the replica method. From this it is possible to derive other CGMT-like results where the Gaussian matrix is replaced by other types of random matrices.

6.1 Convex Gaussian min-max theorem

Convex Gaussian min-max theorem states that the following optimization problem

$$\min_x \max_y x^\top W y + f(x, y) \tag{6.1}$$

where $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ and $W \in \mathbb{R}^{m \times n}$ with i.i.d. standard normal variables and f is a real function that satisfies certain convexity conditions, is asymptotically equivalent as $m, n \to \infty$ to the following problem

$$\min_x \max_y \|x\| \langle \xi_y, y \rangle + \|y\| \langle \xi_x, x \rangle + f(x, y) \tag{6.2}$$

where $\xi_x \sim \mathcal{N}(0, I_m)$ and $\xi_y \sim \mathcal{N}(0, I_n)$ are independent. The independence is in the sense that the optimized value and the statistical properties of the optimizers in the two problems are asymptotically the same.

It turns out that the CGMT can be applied with much relaxed condition and still gives correct result. Firstly it applies to not just min-max problems but also any extremization problems that are not too crazy, i.e. having exponentially many extremal points achieving roughly the same extremal values. Secondly, the domain over x, y does not need to be a product of two compact space, it can also be any compact set in $\mathbb{R}^m \times \mathbb{R}^n$.

6.2 Some applications

6.2.1 Operator norm of a random matrix

Let A be a $m \times n$ matrix with independent standard Gaussian entries. The operator norm of A is defined as

$$\|A\| = \max_{x:\|x\|=1} \|Ax\|,$$

or equivalently

$$\|A\| = \max_{\|x\|=1,\|y\|=1} x^\top Ay \tag{6.3}$$

By CGMT, this problem is equivalent to

$$\max_{\|x\|=1,\|y\|=1} \langle \xi_x, x\rangle + \langle \xi_y, y\rangle \tag{6.4}$$

which has the maximum $\|\xi_x\| + \|\xi_y\| \simeq \sqrt{m} + \sqrt{n}$. We thus obtain $\|A\| \simeq \sqrt{m} + \sqrt{n}$.

6.2.2 Spiked GOE

We want to study the maximum eigenvalue and the corresponding eigenvector of

$$W + \lambda uu^\top \tag{6.5}$$

where W is a GOE matrix of size n and u is a unit vector in $\mathbb{R}^n$. These correspond to the solution of the following optimization problem

$$\max_{\|x\|=1} x^\top Wx + \lambda \langle u, x\rangle^2 \tag{6.6}$$

Result 6.1. *Let G be a Gaussian matrix with independent standard Gaussian entries. If the following optimization problem has a unique extremal point*

$$\operatorname*{extr}_x x^\top Gx + f(x), \tag{6.7}$$

then it is equivalent to the following problem

$$\operatorname*{extr}_x \sqrt{2}\langle g, x\rangle + f(x)$$

where $g \sim \mathcal{N}(0, I_n)$, in the limit $n \to \infty$.

Proof. The problem can be written as

$$\operatorname*{extr}_{x=y} x^\top Gy + f(x)$$

and the result is obtained by applying the CGMT. □

As a consequence of this result and the fact that $W \overset{d}{=} \frac{G+G^\top}{\sqrt{2n}}$, the problem (6.6) is equivalent to

$$\max_{\|x\|=1} 2\langle \xi, x\rangle + \lambda \langle u, x\rangle^2 \tag{6.8}$$

where $\xi \sim \mathcal{N}(0, n^{-1} I_n)$. By introducing Lagrange multipliers, we need to extremize

$$\mathcal{L}(x, q, \alpha, \beta) = 2\langle \xi, x\rangle + \lambda q^2 - \alpha(\|x\|^2 - 1) + \beta(\langle u, x\rangle - q)$$

Differentiating by x, we obtain

$$\hat{x} = \frac{\beta u + 2g}{2\alpha} \tag{6.9}$$

and

$$\begin{aligned} \operatorname*{extr}_x \mathcal{L} &= \lambda q^2 + \alpha + \beta q + \frac{\|\beta u + 2g\|^2}{4\alpha} \\ &\simeq \lambda q^2 + \alpha + \beta q + \frac{4 + \beta^2}{4\alpha} \end{aligned}$$

in which the last expression follows from $\langle g, u\rangle \simeq 0, \|g\| \simeq 1$ and $\|u\| = 1$. Differentiating the previous equation by q, α, we obtain

$$\begin{aligned} 4\alpha^2 &= 4 + \beta^2 \\ q &= \frac{\beta}{2\lambda} \end{aligned}$$

and we obtain

$$\operatorname*{extr}_{x,q,\alpha} \mathcal{L} = \sqrt{\beta^2 + 4} - \frac{\beta^2}{4\lambda} \tag{6.10}$$

so we need to extremize (in fact maximize) this function in β.

If $\lambda < 1$, then (6.10) has a unique minimum at $\beta = 0$. Equation (6.9) now becomes $\hat{x} = \xi$, so the eigenvector $\hat{x}$ behaves like a Gaussian noise uncorrelated to the signal u.

If $\lambda > 1$, the maximum is $\lambda + \frac{1}{\lambda}$, achieved at $\beta = 2\sqrt{\lambda^2 - 1}$. We thus obtain $\alpha = \lambda$ and the eigenvector $\hat{x}$ is related to the signal u with the relation

$$\hat{x} = \sqrt{1 - \frac{1}{\lambda^2}}\, u + \frac{1}{\lambda}\xi$$

As $\lambda \to \infty$ the signal part in this equation dominates and $\hat{x}$ get closer to u. In particular,

$$\lim_{n\to\infty} \langle \hat{x}, u\rangle = \sqrt{1 - \frac{1}{\lambda^2}}$$

6.2.3 Regression on Gaussian mixture

In this example we will recover the main result in [63]. Contrary to the statement that CGMT is not powerful enough to deal with the model in this paper, the main result can be derived by a straightforward application of CGMT. In the paper's setting, we want to classify the data that consists of N points in $\mathbb{R}^D$ that comes from K Gaussian clusters. The cluster k has N_k data points, centered at an unknown vector $\mu_k \in \mathbb{R}^D$, with covariance matrix C_k. In other words, if x_{ki} is the i-th data point in the cluster k, then

$$x_{ki} = \mu_k + C_k^{1/2} Z_{ki} \tag{6.11}$$

where $Z_{ki} \sim \mathcal{N}(0, I_D)$ for all k, i. We consider the setting where $N_k/N \to \rho_k$ and $N/D \to \alpha$. The classification is done by solving the following convex optimization problem

$$\min_{w,b} \sum_{k\in[K],\, i\in[N_k]} \ell\left(w^\top x_{ki} + b, e_k\right) + R(w) \tag{6.12}$$

where $w \in \mathbb{R}^D$, $b \in \mathbb{R}$, ℓ is a loss function and R is for regularization. For simplicity we assume $e_1, \dots, e_K$ are real numbers instead of being multi-dimensional vectors as in the original paper.

Analysis of the model with CGMT Plugging (6.11) into (6.12), we have

$$\min_{w,b} \sum_{k,i} \ell\Big(w^\top(\mu_k + C_k^{1/2} Z_{ik}) + b, e_k\Big) + R(w)$$

It is more convenient to consider a more general problem of extremizing

$$\min_{w} L(\mu_1^\top w, \dots, \mu_k^\top w, G_1 C_1^{1/2} w, \dots, G_k C_k^{1/2} w) + R(w)$$

where $G_1, \dots, G_k$ independent Gaussian matrices. Let $y_k = G_k C_k^{\frac{1}{2}} w$ and $m_k = \mu_k^\top w$. By introducing the Lagrange multipliers, we need to extremize the following function

$$\begin{aligned} L(m_1, \dots, m_k, y_1, \dots, y_k) + \sum_k \hat{m}_k(\mu_k^\top w - m_k) \\ + \sum_k v_k^\top (G_k C_k^{\frac{1}{2}} w - y_k) \end{aligned}$$

By the CGMT, this problem is equivalent to extremizing

$$\begin{aligned} L(m_1, \dots, m_k, y_1, \dots, y_k) + \sum_k \hat{m}_k(\mu_k^\top w - m_k) - v_k^\top y_k \\ + \sum_k \|v_k\| g_k^\top C_k^{\frac{1}{2}} w + \|C_k^{\frac{1}{2}} w\| h_k^\top v_k \end{aligned}$$

where g_k, h_k are standard Gaussian vectors. Let $\hat{Q}_k = \|v_k\|^2$, $Q_k = \|C_k^{\frac{1}{2}} w\|^2 = w^\top C_k w$. By introducing the Lagrange multipliers, we need to extremize the following function

$$\begin{aligned} L(m_1, \ldots, m_k, y_1, \ldots, y_k) + \sum_k \hat{m}_k(\mu_k^\top w - m_k) - v_k^\top y_k \\ + \sum_k \sqrt{\hat{Q}_k}\, g_k^\top C_k^{\frac{1}{2}} w + \sqrt{Q_k} h_k^\top v_k \\ + \sum_k \frac{1}{2}\hat{V}_k(w^\top C_k w - Q_k) - \frac{1}{2} V_k(\|v_k\|^2 - \hat{Q}_k) \end{aligned}$$

This is a quadratic function in each v_k. Extremize over all v_k and rearrange the terms, we obtain the equivalent form

$$\psi_L(m, Q, V) + \psi_R(\hat{m}, \hat{Q}, \hat{V}) + \sum_k \frac{1}{2} V_k \hat{Q}_k - \frac{1}{2}\hat{V}_k Q_k - m_k \hat{m}_k \tag{6.13}$$

where

$$\begin{aligned} \psi_L &= \operatorname*{extr}_y L(m_1, \ldots, m_K, y_1, \ldots, y_K) + \sum_k \frac{\left\|y_k - \sqrt{Q_k} h_k\right\|^2}{2V_k} \\ \psi_R &= \operatorname*{extr}_w R(w) + b^\top w + \frac{1}{2} w^\top A w \end{aligned}$$

and

$$\begin{aligned} b &= \sum_k \hat{m}_k \mu_k + \sqrt{\hat{Q}_k} g_k^\top C_k^{\frac{1}{2}} \\ A &= \sum_k \hat{V}_k C_k \end{aligned}$$

The equation (6.13) is the replica symmetric free energy obtain in equation the (168) of [63]. In summary, we need to find the extremal point of (6.13), then substitute the value of $\hat{m}, \hat{Q}, \hat{V}$ into ψ_R. Then the minimizer of the original problem behaves like the minimizer of ψ_R.

6.3 CGMT and replicas

In this section we will derive the CGMT for the case where $\operatorname{extr}_{x,y} = \max_{x,y}$ from the replica method. Consider the following problem

$$\max_{(x,y)\in K} x^\top W y + f(x, y)$$

where W is a m by n matrix with independent standard Gaussian entries. Consider the following Hamiltonian

$$H(x, y) = \beta x^\top W y + \beta f(x, y), \quad (x, y) \in K, \beta > 0.$$

H can be written as the sum of the deterministic part $H_0 = \beta f(x, y)$ and $H_1 = \beta x^T W y$. We have

$$H_1^{\text{rep}} = \frac{\beta^2}{2} \sum_{a,b} \langle x^a, x^b \rangle \langle y^a, y^b \rangle$$

We make the following change of variables

$$Q_x^{ab} = \langle x^a, x^b \rangle, \quad Q_y^{ab} = \langle y^a, y^b \rangle$$

The Hamiltonian that corresponds to this change of variables is

$$\bar{H}_1^{\text{rep}} = \frac{\beta^2}{2} \sum_{a,b} Q_x^{ab} Q_y^{ab} + \sum_{1 \le a \le b \le r} \hat{Q}_x^{ab} (\langle x^a, x^b \rangle - Q_{ab}) + \hat{Q}_y^{ab} (\langle y^a, y^b \rangle - Q_y^{ab})$$

With the following replica symmetry ansatz

$$Q_x^{ab} = \begin{cases} m_x, & a = b \\ q_x, & a \neq b \end{cases}, \quad \hat{Q}_x^{ab} = \begin{cases} \hat{m}_x, & a = b \\ \hat{q}_x, & a \neq b \end{cases}$$

and a similar y-ansatz, we have

$$\begin{aligned} \bar{H}_1^{\text{rep}} = \frac{\beta^2}{2}(r m_x m_y + r(r-1) q_x q_y) + \sum_a \hat{m}_x (\|x^a\|^2 - m_x) + \hat{m}_y (\|y^a\|^2 - m_y) \\ + \sum_{a<b} \hat{q}_x (\langle x^a, x^b \rangle - q_x) + \hat{q}_y (\langle y^a, y^b \rangle - q_y) \end{aligned}$$

Dereplicate $\bar{H}_1^{\text{rep}}$, we obtain

$$\begin{aligned} \bar{H}_1 = \frac{\beta^2}{2}(m_x m_y - q_x q_y) - m_x \hat{m}_x - m_y \hat{m}_y + \frac{1}{2} q_x \hat{q}_x + \frac{1}{2} q_y \hat{q}_y \\ + \sqrt{\hat{q}_x} \langle \xi_x, x \rangle + \Big(\hat{m}_x - \frac{\hat{q}_x}{2} \Big) \|x\|^2 + \sqrt{\hat{q}_y} \langle \xi_y, y \rangle + \Big(\hat{m}_y - \frac{\hat{q}_y}{2} \Big) \|y\|^2 \end{aligned}$$

where ξ_x, ξ_y are standard Gaussian vectors. Next we make the following rescaling

$$\begin{aligned} (q_x, m_x) &\to (q_x, q_x + 2\delta_x/\beta) \\ (\hat{q}_x, \hat{m}_x) &\to (\beta^2 \hat{q}_x, \frac{1}{2} \beta^2 \hat{q}_x + \beta \hat{\delta}_x) \end{aligned}$$

and similarly for the y-parameters. The reason for this rescaling is that, as β approaches infinity is that for $x^1, x^2 \overset{i.i.d}{\sim} P^H$, we expect that the Euclidean distance $\|x^1 - x^2\|$ is of order $\beta^{-1/2}$, leading to $m_x - q_x = O(\beta^{-1})$. On the other hand, the rescaling of $(\hat{q}_x, \hat{m}_x)$ ensures that the coefficients for the variable x in $\bar{H}_1$ are of order $O(\beta)$.

Making these substitutions and using the fact that $\bar{H} = H_0 + \bar{H}_1$, we have

$$\begin{aligned} \beta^{-1} H \simeq f(x, y) + q_x \delta_y + q_y \delta_x - \hat{q}_x \delta_x - q_x \hat{\delta}_x - \hat{q}_y \delta_y - q_y \hat{\delta}_y \\ + \sqrt{\hat{q}_x} \langle \xi_x, x \rangle + \hat{\delta}_x \|x\|^2 + \sqrt{\hat{q}_y} \langle \xi_y, y \rangle + \hat{\delta}_y \|y\|^2 \end{aligned} \tag{6.14}$$

as $\beta \to \infty$. So the original problem is equivalent to extremizing (6.14). Differentiating (6.14) by δ_x, δ_y, we obtain $q_x = \hat{q}_y$ and $q_y = \hat{q}_x$, so we need to extremize

$$f(x,y) + \sqrt{q_y}\langle \xi_x, x\rangle + \sqrt{q_x}\langle \xi_y, y\rangle + \hat{\delta}_x(\|x\|^2 - q_x) + \hat{\delta}_y(\|y\|^2 - q_y)$$

which is the same problem of extremizing

$$f(x,y) + \|x\|\,\langle \xi_y, y\rangle + \|y\|\,\langle \xi_x, x\rangle.$$

Remark 6.1. Although the CGMT can be derived from the replica method, in the literature, for the same problem, the two methods obtain the results in drastically different ways. The reason is that there are several ways of preparing a problem before doing the calculations, which has a great impact on how the calculations is done afterwards. To better illustrate this, let us take for example the following optimization problem

$$\min_{x\in\mathbb{R}^n} L(Wx)$$

where W is a $m \times n$ Gaussian matrix ($m < n$) and L is a convex function. With the way the CGMT and the replica method are normally used in the literature, one would have rewritten the problem as

$$\min_x \max_y y^\top W x - \hat{L}(y)$$

where $\hat{L}$ is the convex dual of L, before applying the CGMT, and would have considered the distribution with density

$$P(x) \propto e^{-\beta L(Wx)},$$

before doing the replica calculations.

If we prepare the problem differently, by rewriting it as

$$\operatorname*{extr}_{x,y,\lambda} L(y) + \lambda^\top(Wx - y)$$

before applying the CGMT, and by considering the joint density

$$P(x,y) \propto e^{-\beta L(y)}\delta(y - Wx)$$

before doing the replica calculations, the two methods become very similar, leading the to realization that the replica method can derive the CGMT as has been shown.

6.4 Random optimization with IRO matrices

With the replica method, we obtain a result similar to the CGMT for isotropically orthogonal random (IRO) matrices. In the same way that the CGMT is used to analyze optimization problems involving Gaussian matrices, this result will allow us to study

numerous optimization problems that involves IRO matrices, especially signal recovering problems using IRO matrices for measurements. Certain random optimization problems with IRO matrices can also be analyzed by the CGMT [103] with some trick that only works for quadratic loss.

In this section, we will first state the main result and test it against a spiked model involving square IRO matrices. Then we will develop some spherical integral identity which allow us to derive the main result using replicas.

6.4.1 Result and consequences

Definition 6.1. The matrix $O \in \mathbb{R}^{m\times n}$ with $m \le n$ is **isotropically random orthogonal** (IRO) if it is sampled uniformly from the manifold $OO^\top = I_m$.

Remark 6.2. The IRO matrix O can be sampled by $(GG^\top)^{-1/2}G$ where G is a random $m \times n$ Gaussian matrix with $m \le n$. When $m = n$, O is said to be sampled from the Haar measure of the orthogonal group.

Remark 6.3. The IRO matrix is invariant in law by left and right multiplications by orthogonal matrices.

Result 6.2. *Let $O \in \mathbb{R}^{m\times n}$ with $m \le n$ be an IRO matrix. Then the following optimization problem*

$$\operatorname*{extr}_{x,y} x^\top Oy + f(x,y)$$

where (x, y) is in some domain of $\mathbb{R}^{m+n}$, is equivalent to the following problem when $n \to \infty$

$$\operatorname{extr} f(x,y) + \sqrt{\hat\delta_x}\langle \xi_x, x\rangle + \sqrt{\hat\delta_y}\langle \xi_y, y\rangle + \frac{\delta_x\|y\|^2 + \delta_y\|x\|^2}{1+\sqrt{1+4\delta_x\delta_y}} - \frac{n}{2}(\delta_x\hat\delta_x + \delta_y\hat\delta_y) \tag{6.15}$$

where ξ_x, ξ_y are standard Gaussian vectors and the saddle point is computed for all variables in the expression.

Remark 6.4. If the term $4\delta_x\delta_y$ colored in red is removed, we recover the CGMT.

Remark 6.5. If the function $f(x, y)$ is separable, meaning it can be expressed as the sum of terms, each depending on only one coordinate, then the problem (6.15) is easy to analyze, since it can be reduced to optimization problems in one variable. If $f(x, y)$ can be expressed in terms of separable functions, then we can reduce (6.15) to a separable problem by using Lagrange multipliers.

When O is a square matrix, by setting the domain in Result 6.2 to $\{x = y\}$, we have

Result 6.3. *Let $O \in \mathbb{R}^{n\times n}$ an IRO matrix. The following problem*

$$\operatorname*{extr}_{x} x^\top Ox + f(x)$$

is asymptotically equivalent to

$$\operatorname*{extr}_{x,\delta,\hat{\delta}} f(x) + \sqrt{\hat{\delta}}\langle \xi, x\rangle + \frac{\delta\|x\|^2}{1+\sqrt{1+\delta^2}} - \frac{n}{4}\delta\hat{\delta} \tag{6.16}$$

Next, we will test Result 6.2, in particular Result 6.3, on the following spiked model. Let $O \in \mathbb{R}^{n\times n}$ be an IRO matrix. We consider the random matrix $M = O + O^\top + \lambda uu^\top$ where $\lambda > 0$ and u is a unit vector in $\mathbb{R}^n$. Let λ_{max} and u_{max} be the maximum eigenvalue of M and the corresponding unit eigenvalue. We will derive the following result

$$\lim_{n\to\infty} \lambda_{max} = \sqrt{\lambda^2+4}$$
$$\lim_{n\to\infty} (u_{max}^\top u)^2 = \frac{\lambda}{\sqrt{\lambda^2+4}}$$

First, note that λ_{max} and u_{max} are the maximum and maximizer of the following optimization problem

$$\max_{\|x\|=1} \lambda\langle u, x\rangle^2 + 2x^\top Ox.$$

We have

$$\begin{aligned}
&\max_{\|x\|=1} \lambda\langle u, x\rangle^2 + 2x^\top Ox \\
\leftrightarrow &\operatorname*{extr}_{\delta,\hat{\delta},\|x\|=1} \lambda\langle u, x\rangle^2 + 2\sqrt{\hat{\delta}}\langle \xi, x\rangle + \frac{2\delta}{1+\sqrt{1+\delta^2}} - \frac{n}{2}\delta\hat{\delta}, \quad \text{by Result 6.3} \\
\leftrightarrow &\operatorname*{extr}_{\delta,\hat{\delta},\|x\|=1} \lambda\langle u, x\rangle^2 + 2\sqrt{\hat{\delta}}\langle \xi, x\rangle + \frac{2\delta}{1+\sqrt{1+\delta^2}} - \frac{1}{2}\delta\hat{\delta} && \text{(a)} \\
\leftrightarrow &\operatorname{extr} \lambda q^2 + 2\sqrt{\hat{\delta}}\langle \xi, x\rangle + \frac{2\delta}{1+\sqrt{1+\delta^2}} - \frac{1}{2}\delta\hat{\delta} + \hat{q}(\langle u, x\rangle - q) - \mu(\|x\|^2 - 1) && \text{(b)} \\
\leftrightarrow &\operatorname{extr} \lambda q^2 + \frac{2\delta}{1+\sqrt{1+\delta^2}} - \frac{1}{2}\delta\hat{\delta} - q\hat{q} + \mu + \frac{4\hat{\delta}+\hat{q}^2}{4\mu} && \text{(c)} \\
\leftrightarrow &\operatorname{extr} \lambda q^2 + \frac{2\delta}{1+\sqrt{1+\delta^2}} - q\hat{q} + \frac{2}{\delta} + \frac{1}{8}\delta\hat{q}^2 && \text{(d)} \\
\leftrightarrow &\operatorname{extr} \lambda q^2 + \frac{2\sqrt{\delta^2+1}}{\delta} - \frac{2q^2}{\delta} && \text{(e)} \\
= &\sqrt{\lambda^2+4}
\end{aligned}$$

Explanations:

(a) By rescaling $\xi \leftarrow \xi/\sqrt{n}$ and $\hat{\delta} \leftarrow n\hat{\delta}$. Note that ξ now follows the law $\mathcal{N}(0, I/n)$.

(b) By Lagrange multipliers. Here we decouple the problem at the expense of adding the variables $\mu, q, \hat{q}$ to the arguments of extremization.

(c) By extremizing over x, we have $x = \frac{2\sqrt{\hat{\delta}}\xi+\hat{q}u}{2\mu}$. The terms involving x becomes $\frac{\|2\sqrt{\hat{\delta}}\xi+\hat{q}u\|^2}{4\mu} \simeq \frac{4\hat{\delta}+\hat{q}^2}{4\mu}$ since $\|\xi\|^2 \simeq 1, \langle \xi, u\rangle \simeq 0$.

(d) By extremizing over $\hat{\delta}$, we obtain the equation $\delta\mu = 2$. The terms involving $\hat{\delta}$ becomes 0.

(e) By simplifying the expression involving δ and by extremizing over q, we obtain the equation $\hat{q} = \frac{4q}{\delta}$. The terms involving q becomes $-\frac{2q^2}{\delta}$.

In the last step, by differentiating (e) by δ we obtain $q^2 = \frac{1}{\sqrt{1+\delta^2}}$. From (d) we have $\delta = 2/\lambda$, so $q^2 = \lambda/\sqrt{\lambda^2+4}$. Since q is the constraint parameter for $u_{max}^\top u$, we conclude that

$$\lim_{n\to\infty} (u_{\max}^\top u)^2 = \frac{\lambda}{\sqrt{\lambda^2+4}}.$$

Remark 6.6. The limiting spectral density of $O + O^\top$ is given by

$$\mu(dx) = \frac{dx}{\sqrt{4-x^2}}, \quad |x| < 2.$$

This can be proved from the fact that the limiting spectral density of O is uniform on the unit circle and $O^\top = O^{-1}$. In contrast to the spiked GOE model, in this case there is no phase transition: the top eigenvalue of the matrix $O + O^\top + \lambda uu^\top$ is isolated as soon as $\lambda > 0$.

6.4.2 Derivation of the result

A Spherical integrals

We give here some results on spherical integrals that will be useful when dealing with IRO matrices.

Lemma 6.1. *Let u, v be independent uniform vector in $\sqrt{n}S^{n-1}$. Define*

$$\begin{aligned} S(t) := \lim_{n\to\infty} \frac{1}{n}\mathbb{E}e^{tu^\top v} &= \operatorname*{extr}_\lambda \lambda - 1 - \frac{1}{2}\log(\lambda^2 - t^2) \\ &= \frac{-1+\sqrt{1+4t^2}}{2} - \frac{1}{2}\log\left(\frac{1+\sqrt{1+4t^2}}{2}\right) \end{aligned}$$

Proof. We have $S(t) = Z(t)/Z(0)$, where

$$Z(t) = \int_{\mathbb{R}^n\times\mathbb{R}^n} dudv\, e^{tu^\top v}\delta(u^\top u - n)\delta(v^\top v - n)$$

Using the Fourier representation of Dirac delta function and stationary point method, we obtain

$$Z(t) \sim C(n,t)e^{nI(t)}$$

where $C(n,t)$ is bounded by a polynomial in n and

$$I(t) = \frac{1}{2} \operatorname*{extr}_{\lambda,\mu} \lambda + \mu - \log(\lambda\mu - t^2)$$

From the asymptotic formula of $Z(t)$, we have $S(t) = I(t) - I(0)$, leading to the result stated in the lemma. □

Corollary 6.1. *Let $O \in \mathbb{R}^{m\times n}$ with $m \leq n$ be an IRO matrix and A is a matrix with rank $r \leq m$ and singular values $\sigma_1, \ldots, \sigma_r > 0$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{n \operatorname{Tr} AO} = \sum_{a=1}^{r} S(\sigma_a) \tag{6.17}$$

where S is defined in Lemma 6.1. Note that the result does not depend on m.

Proof. Since the law of O is invariant in law by left and right multiplications by orthogonal matrices, the left hand side of (6.17) only depends on the singular values of A. Without loss of generality, we can assume that A is a $n \times m$ matrix with entries $\sigma_1, \ldots, \sigma_r$ on the first r entries of the diagonal and zero elsewhere. The left hand side of (6.17) is therefore

$$\lim_{n\to\infty} \frac{1}{n} \log e^{\sigma_1 O_{11} + \cdots + \sigma_r O_{rr}} = \lim_{n\to\infty} \frac{1}{n} \log e^{\sigma_1 e_1^\top u_1 + \ldots \sigma_r e_r^\top u_r}$$

where e_a is the vector with 1 and the a-th position and zero elsewhere and $u_1, \ldots, u_r$ is the first r rows of the matrix O. From this we can see why the left hand side of (6.17) does not depend on m. Indeed, the law of the first r rows of O does not depends on m as they can be generated from r independent standard Gaussian vectors in $\mathbb{R}^n$.

The right hand side of the previous equation is unchanged if the vectors $e_1, \ldots, e_r$ are replaced by any set of r orthonormal vectors. Therefore it is unchanged if $(e_1, \ldots, e_r)$ is replaced by uniformly random orthonormal vectors $v_1, \ldots, v_r$ independent of $u_1, \ldots, u_r$. We thus conclude that the left hand side of (6.17) is equal to

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{\sigma_1 u_1^\top v_1 + \cdots + \sigma_r u_r^\top v_r}$$

where $(u_1, \ldots, u_r)$ and $(v_1, \ldots, v_r)$ are independent uniformly random sets of orthonormal vectors in $\mathbb{R}^n$. Using the same method of Lemma 6.1, we can obtained the result given by the lemma. □

B A useful result on the singular values

We will need the following result for our replica computations in the next section.

Lemma 6.2. *Let $x^1, \dots, x^r \in \mathbb{R}^m$, $y^1, \dots, y^r \in \mathbb{R}^n$ and $r \le \min(m, n)$. Then the positive singular values of*

$$\sum_{a=1}^{r} y^a x^{a\top}$$

are the square roots of the eigenvalues of the matrix $Q_x Q_y$, where $Q_x = \left\{\langle x^a, x^b\rangle\right\}_{a,b=1}^r$ and $Q_y = \left\{\langle y^a, y^b\rangle\right\}_{a,b=1}^r$

Proof. Consider the matrices X, Y with columns $(x_1, \dots, x_r)$ and $(y_1, \dots, y_r)$. Then $y_1 x_1^\top + \cdots + y_r x_r^\top = YX^\top$. The singular values of $YX^\top$ are square root of the the eigenvalues of $YX^\top XY^\top$, which has the same eigenvalues as $X^\top XY^\top Y = Q_x Q_y$. □

C Replica computation

Now we are ready to perform replica computations that derive Result 6.2. Let $O \in \mathbb{R}^{m\times n}$ with $m \le n$ and IRO matrix. Consider the following optimization problem

$$\max_{x,y} x^\top O y + f(x, y)$$

where f is a function such that this problem is not too crazy, avoiding scenarios in which exponentially many local maxima achieve roughly the same local maximum. In such case the replica symmetric ansatz can give the correct result. Here we provide the calculations for the maximization problem but the result turns out to be valid for general extremization problem. To study this problem, we consider the following Hamiltonian

$$H = \beta x^\top O y + \beta f(x, y)$$

which can be separated into the deterministic part H_0 and stochastic part H_1 as

$$H_0 = \beta f(x, y) \qquad H_1 = \beta x^\top O y$$

The replication of H_1 is given by

$$H_1^{\text{rep}} = \log \mathbb{E} e^{\beta n \operatorname{Tr}(AO)}$$

where

$$A = \sum_a y^a x^{a\top}/n$$

Since the rank of A cannot be larger than r and $S(0) = 0$, by Lemma 6.1, we have

$$H_1^{\text{rep}} \simeq n \sum_a S(\beta\sigma_a) \tag{6.18}$$

where $\sigma^1, \dots, \sigma^r$ are the r largest singular values of A. By Lemma 6.2, these singular values are square roots of the eigenvalues of the matrix $Q_x Q_y$, where

$$Q_x^{ab} = \langle x^a, x^b \rangle / n, \qquad Q_y^{ab} = \langle y^a, y^b \rangle / n \tag{6.19}$$

We can write $H_1^{\text{rep}} = h(\{Q_x^{ab}, Q_y^{ab}\})$ for some function h. The Hamiltonian associated with the macroscopic functions Q_x, Q_y is $\bar{H}_1^{\text{rep}} = (\star) + (\star\star)$, where

$$\begin{aligned}(\star) &= h(\{Q_x^{ab}, Q_y^{ab}\}) \\ (\star\star) &= \sum_{a \le b} \hat{Q}_x^{ab}(\langle x^a, x^b \rangle - n Q_x^{ab}) + \hat{Q}_y^{ab}(\langle y^a, y^b \rangle - n Q_y^{ab})\end{aligned}$$

with constraint parameters Q_x, Q_y and multipliers $\hat{Q}_x, \hat{Q}_y$. Here we abuse the notations by using the same symbol for a macroscopic function and its corresponding constraint parameters. We consider the following replica symmetric ansatz for the dominating extremal point

$$Q_x^{ab} = \begin{cases} m_x, & a = b \\ q_x, & a \neq b \end{cases}, \quad \hat{Q}_x^{ab} = \begin{cases} \hat{m}_x, & a = b \\ \hat{q}_x, & a \neq b \end{cases}$$

and a similar y-ansatz. With this ansatz, the term $(\star\star)$ becomes

$$\begin{aligned}&\sum_a \hat{m}_x(\|x^a\|^2 - n m_x) + \sum_{a<b} \hat{q}_x(\langle x^a, x^b \rangle - n q_x) \\ +&\sum_a \hat{m}_y(\|y^a\|^2 - n m_y) + \sum_{a<b} \hat{q}_y(\langle y^a, y^b \rangle - n q_y)\end{aligned}$$

The dereplication of $(\star\star)$ is

$$\begin{aligned}\text{dereplicate}(\star\star) =& \sqrt{\hat{q}_x}\langle \xi_x, x \rangle + \Big(\hat{m}_x - \frac{\hat{q}_x}{2}\Big)\|x\|^2 + \sqrt{\hat{q}_y}\langle \xi_y, y \rangle + \Big(\hat{m}_y - \frac{\hat{q}_y}{2}\Big)\|y\|^2 \\ &+ n\Big(\frac{1}{2} q_x \hat{q}_x + \frac{1}{2} q_y \hat{q}_y - m_x \hat{m}_x - m_y \hat{m}_y\Big)\end{aligned}$$

We now consider the term $(\star)$ under the replica symmetric ansatz. We have

$$Q_x Q_y = \Big(q_x \mathbb{1}\mathbb{1}^\top + (m_x - q_x) I\Big)\Big(q_y \mathbb{1}\mathbb{1}^\top + (m_y - q_y) I\Big),$$

which is a polynomial of the matrix $\mathbb{1}\mathbb{1}^\top$. Since $\mathbb{1}\mathbb{1}^\top$ has eigenvalues $r, 0^{(r-1)}$ (here $0^{(r-1)}$ means 0 repeated r times), the singular values σ_a's are

$$\sqrt{(m_x + (r-1)q_x)(m_y + (r-1)q_y)}, \sqrt{(m_x - q_x)(m_y - q_y)}^{(r-1)}$$

From this and (6.18), we have

$$(\star) = nS\left(\beta\sqrt{(m_x + (r-1)q_x)(m_y + (r-1)q_y)}\right) + n(r-1)S\left(\beta\sqrt{(m_x - q_x)(m_y - q_y)}\right)$$

By dereplicating $(\star)$, in this case taking $\partial_{r=0}$, we have

$$\begin{aligned}\text{dereplicate}(\star) = n\beta S'\left(\beta\sqrt{(m_x-q_x)(m_y-q_y)}\right)\frac{q_x(m_y-q_y)+q_y(m_x-q_x)}{2\sqrt{(m_x-q_x)(m_y-q_y)}}\\ + nS\left(\beta\sqrt{(m_x-q_x)(m_y-q_y)}\right)\end{aligned}$$

Next we make the following rescaling

$$\begin{aligned}(q_x, m_x) &\to (q_x, q_x + 2\delta_x/\beta)\\ (\hat{q}_x, \hat{m}_x) &\to (\beta^2\hat{q}_x, \frac{1}{2}\beta^2\hat{q}_x + \beta\hat{\delta}_x)\end{aligned}$$

and similarly for the y-parameters. The reason for this rescaling is that, as β approaches infinity is that for $x^1, x^2 \overset{i.i.d}{\sim} P^H$, we expect that the Euclidean distance $\|x^1 - x^2\|$ is of order $\beta^{-1/2}$, leading to $m_x - q_x = O(\beta^{-1})$. On the other hand, the rescaling of $(\hat{q}_x, \hat{m}_x)$ ensures that the coefficients for the variable x in $\bar{H}_1$ are of order $O(\beta)$. From this rescaling, as $\beta \to \infty$, we have

$$\begin{aligned}\beta^{-1}\text{dereplicate}(\star) &\simeq \frac{2n(q_x\delta_y + q_y\delta_x)}{1+\sqrt{1+16\delta_x\delta_y}}\\ \beta^{-1}\text{dereplicate}(\star\star) &= \sqrt{\hat{q}_x}\langle\xi_x, x\rangle + \hat{\delta}_x\|x\|^2 + \sqrt{\hat{q}_y}\langle\xi_y, y\rangle + \hat{\delta}_y\|y\|^2\\ &\quad - n(\hat{q}_x\delta_x + q_x\hat{\delta}_x + \hat{q}_y\delta_y + q_y\hat{\delta}_y)\end{aligned}$$

From this and the fact that $\bar{H} = H_0 + \bar{H}_1$, we have

$$\begin{aligned}\beta^{-1}\bar{H} \simeq f(x,y) + \frac{2n(q_x\delta_y + q_y\delta_x)}{1+\sqrt{1+16\delta_x\delta_y}} + \sqrt{\hat{q}_x}\langle\xi_x, x\rangle + \hat{\delta}_x\|x\|^2 + \sqrt{\hat{q}_y}\langle\xi_y, y\rangle + \hat{\delta}_y\|y\|^2\\ - n(\hat{q}_x\delta_x + q_x\hat{\delta}_x + \hat{q}_y\delta_y + q_y\hat{\delta}_y)\end{aligned} \tag{I}$$

As $\beta \to \infty$, $\bar{H}$ is the Hamiltonian for extremizing the expression given by (I), which is equivalent to the problem of extremizing

$$f(x,y) + \sqrt{\hat{q}_x}\langle\xi_x, x\rangle + \sqrt{\hat{q}_y}\langle\xi_y, y\rangle + \frac{2(\delta_x\|y\|^2 + \delta_y\|x\|^2)}{1+\sqrt{1+16\delta_x\delta_y}} - n(\delta_x\hat{q}_x + \delta_y\hat{q}_y) \tag{II}$$

Indeed, by introducing the constraint parameters q_x, q_y and multipliers $\hat{q}_x, \hat{q}_y$ for the functions $\|x\|^2, \|y\|^2$ in (II) we obtain (I). By the following rescaling $\delta_x, \delta_y \leftarrow 2\delta_x, 2\delta_y$ and change of notation $\hat{\delta}_x, \hat{\delta}_y \leftarrow \hat{q}_x, \hat{q}_y$, we obtain the Result 6.2.

Bibliographical notes

CGMT has been used in hundreds of papers that analyze optimization-based algorithms, such as PhaseMax [35] [93] and compressed sensing [104], just to name a few. It is also

used to study theoretical models in machine learning with Gaussian data [71] [106] [54]. Models with slightly more realistic data [34] can be transformed into an equivalent form ready to be analyzed by CGMT, thanks to the Gaussian equivalence principle [42] [41].

Techniques from RMT can be also applied to analyze random convex optimization problems. For instance, this approach has been used to study problems in semi-supervised learning [64] [65] and multitask learning [108], [107].

Chapter 7

Bayes-optimal inference

In this chapter, we will study inference problems in Bayes-optimal setting. We will first derive some elementary results on Gaussian channels which will be important for studying inference problems in high dimension. One of the simplest such problem is the factorization of a rank-1 symmetric matrix corrupted by noise. This problem is then solved by the replica and cavity method, offering two complementing points of view. The ideas involved in this simple problem can be applied to a more complicated model of multitask learning on Gaussian mixtures, which is studied in details with simulations.

7.1 Bayes-optimal setting

In this chapter, we will be interested in inference problems in the so-called **Bayes optimal setting**. In this setting, we want to estimate a signal X from an observation Y, given the probability law P_X that is used to generate X and the conditional law $P_{Y|X}$. It makes sense in this setting to ask what the best estimator is, according to certain measure of performance. Some commonly used performance metrics include the probability of exact recovery and the (normalized) mean squared error. The conditional law of X given Y is called the **posterior law**, given by

$$P(x|Y) = \frac{P_X(x)P(Y|x)}{P(Y)} \propto P_X(x)P(Y|x)$$

In the Bayes-optimal setting, the posterior law contains all the information that we can extract from X given Y. If the posterior depends only on the variable x via some function $S(Y)$, meaning that knowing $S(Y)$ provides full information about the posterior, then we can estimate X from $S(Y)$ without any loss of information. For this reason, $S(Y)$ is called the **sufficient statistics** for estimating X from Y.

7.2 Gaussian channels

Due to special properties of the Gaussian distribution, Gaussian channels are at the meeting point of estimation theory, statistical physics, and information theory. More-

over, they will be important to the study of some high-dimensional Bayesian inference problems.

7.2.1 Overlaps, free energy and mutual information

Consider the following Gaussian channels

$$Y_i = \sqrt{\lambda_i} X_i + Z_i, \quad i = 1, \dots, n \tag{7.1}$$

with inputs X_i, outputs Y_i and SNRs λ_i. The vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is generated from a known distribution $P_{\boldsymbol{X}}$ and Z_i are independent standard Gaussian noises. Let $\hat{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$. The **overlap** of the signal X_i is defined as

$$\mathcal{O}_{\boldsymbol{X},i}(\boldsymbol{\lambda}) := \mathbb{E}[\hat{X}_i X_i] = \mathbb{E}[\hat{X}_i^2] \tag{7.2}$$

On the other hand, the posterior density of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is given by

$$P(\boldsymbol{x}|\boldsymbol{Y}) \propto P_{\boldsymbol{X}}(\boldsymbol{x}) e^{H(\boldsymbol{x},\boldsymbol{Y})} \tag{7.3}$$

where

$$H(\boldsymbol{x}, \boldsymbol{Y}) = \sum_i \sqrt{\lambda_i} Y_i x_i - \frac{1}{2} \lambda_i x_i^2$$

In other words, $P(d\boldsymbol{x}|\boldsymbol{Y})$ is the probability measure associated with the Hamiltonian $H(\boldsymbol{x}, \boldsymbol{Y})$ in which $\boldsymbol{Y}$ plays the role of random parameters, over the space $\mathbb{R}^n$ with underlying measure $P_{\boldsymbol{X}}$. Let $F_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \mathbb{E}[F^H]$ where F^H is the free energy of the Hamiltonian H. Abusing the terminology, we call $F_{\boldsymbol{X}}(\boldsymbol{\lambda})$ the **free energy** of the Gaussian channels (7.1). The free energy $F_{\boldsymbol{X}}(\boldsymbol{\lambda})$ is related to the mutual information $I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = I(\boldsymbol{X}; \boldsymbol{Y})$ by the following formula (Remark 7.1)

$$I_{\boldsymbol{X}}(\boldsymbol{\lambda}) + F_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \frac{1}{2} \sum_{i=1}^n \lambda_i \mathbb{E}[X_i^2] \tag{7.4}$$

Recall that the I-MMSE formula states that

$$\partial_{\lambda_i} I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \frac{1}{2} \mathbb{E}[(X_i - \hat{X}_i)^2] \tag{7.5}$$

From (7.4) and (7.5), we obtain the following relation between the overlaps and the free energy

$$\partial_{\lambda_i} F_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \frac{1}{2} \mathcal{O}_{\boldsymbol{X},i}(\boldsymbol{\lambda}) \tag{7.6}$$

Remark 7.1. The formula (7.4) can be proved as follows. From (7.3), we have $P(\boldsymbol{x}|\boldsymbol{Y}) = P_{\boldsymbol{X}}(x) e^{H(\boldsymbol{x},\boldsymbol{Y}) - F^H}$. From the definition $I(\boldsymbol{X}; \boldsymbol{Y}) = \mathbb{E} \log \frac{P(\boldsymbol{X}|\boldsymbol{Y})}{P_{\boldsymbol{X}}(\boldsymbol{X})}$, we obtain

$$I(\boldsymbol{X}; \boldsymbol{Y}) = \mathbb{E}[H(\boldsymbol{X}, \boldsymbol{Y})] - \mathbb{E}[F^H].$$

The term $\mathbb{E}[H(\boldsymbol{X}, \boldsymbol{Y})]$, in which H denotes the Hamiltonian, not entropy, is exactly the right hand side of (7.4). This proves the result.

Next we will consider some particular examples of Gaussian channels.

7.2.2 Rademacher signal

Consider the Gaussian channel given by

$$Y = \sqrt{\lambda}X + Z, \tag{7.7}$$

where the Rademacher signal X takes values of 1 and -1 with equal probabilities and the standard Gaussian noise Z is independent of X. We have

$$P(x|Y) \propto e^{\sqrt{\lambda}Yx}, \tag{7.8}$$

from which we obtain the posterior distribution as

$$P(x|Y) = \frac{e^{\sqrt{\lambda}Yx}}{2\cosh(\sqrt{\lambda}Y)} \tag{7.9}$$

and the MMSE estimator $\hat{X}_{\text{MMSE}} = \mathbb{E}[X|Y]$ as

$$\hat{X} = \sum_{x=\pm 1} xP(x|Y) = \tanh(\sqrt{\lambda}Y). \tag{7.10}$$

The overlap between the MMSE estimator and the signal is therefore

$$\begin{aligned}
\mathbb{E}[X\hat{X}_{\text{MMSE}}] &= \mathbb{E}[X\tanh(\sqrt{\lambda}(\sqrt{\lambda}X + Z)] \\
&= \frac{1}{2}\mathbb{E}[\tanh(\lambda + \sqrt{\lambda}Z)] - \frac{1}{2}\mathbb{E}[\tanh(-\lambda + \sqrt{\lambda}Z)] \\
&= \frac{1}{2}\mathbb{E}[\tanh(\lambda + \sqrt{\lambda}Z)] - \frac{1}{2}\mathbb{E}[\tanh(-\lambda - \sqrt{\lambda}Z)] \\
&= \mathbb{E}[\tanh(\sqrt{\lambda}Z + \lambda)]
\end{aligned} \tag{7.11}$$

Next, the error $\mathbb{P}(\hat{X} \neq X)$ for any estimator $\hat{X}$ of X is minimized by the maximum-likelihood estimator:

$$\begin{aligned}
\hat{X}_{\text{ML}} &= \operatorname*{argmax}_{x=\pm 1} P(x, Y) \\
&= \operatorname*{argmax}_{x=\pm 1} e^{\sqrt{\lambda}Yx}
\end{aligned} \tag{7.12}$$

This gives us the maximum-likelihood estimator as:

$$\hat{X}_{\text{ML}} = \text{sgn}(Y). \tag{7.13}$$

The Bayes risk is therefore

$$\begin{aligned}
\mathbb{P}(X \neq \hat{X}_{ML}) &= \frac{1}{2}\mathbb{P}(X = 1, \hat{X}_{ML} = -1) + \frac{1}{2}\mathbb{P}(X = -1, \hat{X}_{ML} = 1) \\
&= \frac{1}{2}\mathbb{P}(X = 1, Y < 0) + \frac{1}{2}\mathbb{P}(X = -1, Y > 0) \\
&= \mathbb{P}(X = -1, Y > 0) \\
&= \mathbb{P}(Z > \sqrt{\lambda}).
\end{aligned}$$

7.2.3 Correlated Gaussian signals

Consider T Gaussian channels, where the signals $X_1, \dots, X_T$ have a joint distribution of $\mathcal{N}(0, \boldsymbol{M})$ and are independent of Gaussian noises $Z_1, \dots, Z_T$ that are independently distributed as $\mathcal{N}(0,1)$. Specifically, we have:

$$Y_t = \sqrt{\lambda_t} X_t + Z_t, \quad t = 1, \dots, T.$$

Let $\hat{X}_t = \mathbb{E}[X|\boldsymbol{Y}]$ be the MMSE estimator for X_t. Since $(X_t, Y_1, \dots, Y_T)$ is a Gaussian vector, $\hat{X}_t$ is a linear combination of $Y_1, \dots, Y_T$. Therefore

$$\begin{aligned} \mathrm{MMSE}_t &:= \mathbb{E}[(X_t - \hat{X}_t)^2] \\ &= \min_{\boldsymbol{\beta}_t \in \mathbb{R}^T} \mathbb{E}\left[(X_t - \langle \boldsymbol{\beta}_t, \boldsymbol{Y} \rangle)^2\right]. \end{aligned}$$

This can be written as a quadratic optimization problem

$$\mathrm{MMSE}_t = \min_{\boldsymbol{\beta}_t \in \mathbb{R}^T} \left\{ M_{tt} - 2\boldsymbol{a}_t^T \boldsymbol{\beta}_t + \boldsymbol{\beta}_t^T \boldsymbol{A} \boldsymbol{\beta}_t \right\}$$

with

$$\begin{aligned} \boldsymbol{a}_t &= (\mathbb{E}[X_t Y_s])_{s=1}^T = \left(\sqrt{\lambda_t} M_{ts}\right)_{s=1}^T = \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{e}_t \\ \boldsymbol{A} &= (\mathbb{E}[Y_s Y_{s'}])_{s,s'=1}^T = \left(\sqrt{\lambda_s \lambda_{s'}} M_{ss'} + \delta_{ss'}\right)_{s,s'=1}^T = \boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2}. \end{aligned}$$

This optimization problem admits a unique minimizer $\boldsymbol{\beta}_t = \boldsymbol{A}^{-1} \boldsymbol{a}_t$, from which we obtain

$$\hat{\boldsymbol{X}} = \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2})^{-1} \boldsymbol{Y} \tag{7.14}$$

$$\mathrm{MMSE}_t = [\boldsymbol{M} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{M})^{-1}]_{tt} \tag{7.15}$$

$$\mathbb{E}[X_t \hat{X}_t] = [\boldsymbol{M} - \boldsymbol{M} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{M})^{-1}]_{tt}. \tag{7.16}$$

7.3 Rank-1 matrix factorization

In this section we will study with replicas the problem of factorizing a rank-1 symmetrical matrix corrupted by noise. We consider here the following model

$$Y = \sqrt{\frac{\lambda}{n}} x^0 x^{0\top} + Z \tag{7.17}$$

where $x^0 \in \mathbb{R}^n$ is an unknown signal with i.i.d entries generated from the a known probability distribution P_X with finite second moment and the noise Z is standard Gaussian, independent of the signal x^0. We want to infer the vector x^0 from the observation Y. The fundamental object of our study is the posterior $P(x|Y)$ which contains all the information about x^0 that can be extracted from Y.

Consider the measure space $(\mathbb{R}^n, P_X^{\otimes n})$. Using the result from Example 4.4 with $\phi(x) = \sqrt{\frac{\lambda}{n}} x x^\top$, since $\langle \phi(x), \phi(y) \rangle = \frac{\lambda}{n} \langle x, y \rangle^2$ the posterior $P(x|Y)$ has the following replicated Hamiltonian

$$H^{\text{rep}} = \frac{\lambda}{n} \sum_{0 \le a < b \le r} \langle x^a, x^b \rangle^2$$

Note that H^{rep} is a Hamiltonian in variables $x^1, \dots, x^r$, not including x^0. The Hamiltonian corresponding to the macroscopic functions $\langle x^a, x^b \rangle / n, 0 \le a < b \le r$ is

$$\bar{H}^{\text{rep}} = \lambda n \sum_{0 \le a < b \le r} Q_{ab}^2 + \sum_{0 \le a < b \le r} \hat{Q}_{ab} (\langle x^a, x^b \rangle - n Q_{ab})$$

with the constraint parameters Q^{ab}'s and multipliers $\hat{Q}^{ab}$'s. In the replica symmetric ansatz, $Q^{ab} = q$ and $\hat{Q}^{ab} = \hat{q}$ for all $0 \le a < b \le r$. This ansatz encodes the assumption that if x^1, x^2 are replicas of the system, then $\langle x^0, x^1 \rangle \simeq \langle x^1, x^2 \rangle$, which basically says that these inner products concentrates, as the equality of their concentrated values follows from the fact that $\mathbb{E}[\langle x^0, x^1 \rangle] = \mathbb{E}[\langle x^1, x^2 \rangle]$. With the replica symmetric ansatz, we have

$$\bar{H}^{\text{rep}} = \lambda n \frac{r(r+1)}{2} q^2 + \sum_{0 \le a < b \le r} \hat{q} (\langle x^a, x^b \rangle - n q)$$

Again, with Example 4.4, we can dereplicate $\bar{H}^{\text{rep}}$ and obtain

$$\bar{H} = \frac{n \lambda q^2}{2} - \frac{n q \hat{q}}{2} + \langle \hat{q} x^0 + \sqrt{\hat{q}} \xi, x \rangle - \frac{1}{2} \hat{q} \|x\|^2 \tag{7.18}$$

where $\xi \sim \mathcal{N}(0, I_n)$. The free energy of $\bar{H}$ is given by $n f(q, \hat{q})$, where

$$f(q, \hat{q}) \simeq \frac{1}{2} \lambda q^2 - \frac{1}{2} q \hat{q} + F_X(\hat{q})$$

where $F_X(\hat{q})$ is the free energy of the Gaussian channel with signal $X \sim P_X$ and SNR $\hat{q}$. The stationary point $(q_\star, \hat{q}_\star)$ of F satisfies the following equations

$$\begin{aligned} \hat{q} &= 2\lambda q \\ q &= 2 F_X'(\hat{q}) = \mathcal{O}_X(\hat{q}) \end{aligned}$$

where $\mathcal{O}_X(\hat{q})$ is the overlap of the Gaussian channel with signal $X \sim P_X$ and SNR $\hat{q}$.

We conclude,

- from (7.18), that the posterior law $P(x|Y)$ is asymptotically equivalent to the posterior law of the following Gaussian channel

  $$Y = \sqrt{2\lambda q_\star} x^0 + \xi$$

  where $\xi \sim \mathcal{N}(0, I_n)$ is independent of x^0. In particular, if $q_\star = 0$, the problem is equivalent to estimating the signal x^0 from pure noise, so the signal is unrecoverable.

- If x^1 is drawn from from the posterior $P(x|Y)$, then

  $$\langle x^0, x^1 \rangle / n \simeq q_\star.$$

7.4 Cavity argument

We present here a useful lemma inspired by the cavity method in statistical physics. This result will be used to analyze the matrix factorization problem in Section 7.3 from another perspective. It will also be used to analyze the problem of multitask learning on Gaussian mixture model in Section 7.5. First, we need some definitions:

Definition 7.1. The inference of $\boldsymbol{X} \in \mathbb{R}^D$ from the data $\boldsymbol{Y}$ satisfies the **replica symmetric property** with **overlap** q if in the limit $D \to \infty$,

$$\langle \boldsymbol{X}, \boldsymbol{X}^1 \rangle, \langle \boldsymbol{X}^1, \boldsymbol{X}^2 \rangle, \langle \boldsymbol{X}, \hat{\boldsymbol{X}} \rangle, \|\hat{\boldsymbol{X}}\|^2 \tag{7.19}$$

all converge to the same limit q, where $\boldsymbol{X}^1, \boldsymbol{X}^2$ are sampled independently from the posterior of $\boldsymbol{X}$ given $\boldsymbol{Y}$, and $\hat{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$.

Remark 7.2. The replica symmetric property basically says that the quantities in (7.19) concentrate around their means. The fact that their concentration values are the same follows from

$$\mathbb{E}\langle \boldsymbol{X}, \boldsymbol{X}^1 \rangle = \mathbb{E}\langle \boldsymbol{X}^1, \boldsymbol{X}^2 \rangle = \mathbb{E}\langle \boldsymbol{X}, \hat{\boldsymbol{X}} \rangle = \mathbb{E}\|\hat{\boldsymbol{X}}\|^2.$$

The main result of this section is the following.

Lemma 7.1. *Suppose we want to estimate the signal $X \in \mathbb{R}$ generated by P_X from the data $\boldsymbol{Y}$ that can be split into two parts as follows. The first part, denoted by $\boldsymbol{Y}^x$, consists of the following observation on X,*

$$\boldsymbol{Y}^x = X\boldsymbol{U} + \boldsymbol{Z}, \tag{7.20}$$

where

- *$\boldsymbol{U} \in R^D$ is unknown with prior P_U,*
- *$\boldsymbol{Z} \sim \mathcal{N}(0, I_D)$,*
- *X, $\boldsymbol{U}$ and $\boldsymbol{Z}$ are independent.*

The second dataset, denoted by $\boldsymbol{Y}^u$, is independent of X. Suppose that the law $\boldsymbol{U}|\boldsymbol{Y}^u$ has the replica symmetric property with overlap q. Then in the limit $D \to \infty$,

i) The posterior of X given $\boldsymbol{Y}$ is asymptotically equivalent to the law $\bar{P}$ defined as

$$\frac{\bar{P}(dx|\boldsymbol{Y})}{P_X(dx)} \propto \exp\left(x\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}} \rangle - \frac{1}{2}qx^2\right) \tag{7.21}$$

where $\hat{\boldsymbol{U}} = \mathbb{E}[\boldsymbol{U}|\boldsymbol{Y}]$. As a consequence, the statistics $S = \langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}} \rangle$ is asymptotically sufficient for estimating X from $\boldsymbol{Y}$.

ii) $S/\sqrt{q}$ converges in law to $\sqrt{q}X + \xi$, where ξ follows standard normal distribution and is independent of X. As a result, estimating X from $\boldsymbol{Y}$ is asymptotically equivalent to estimating X from the output of a Gaussian channel with SNR q.

Proof. Since X is independent of $\boldsymbol{U}$ and $\boldsymbol{Y}^u$, we have

$$\begin{aligned}\frac{P(dx|\boldsymbol{Y})}{P_X(dx)} &= \int P(d\boldsymbol{u}|\boldsymbol{Y}^u)P(x|\boldsymbol{u},\boldsymbol{Y}^x) \\ &\propto \int P(d\boldsymbol{u}|\boldsymbol{Y}^u)\exp\Big(x\langle\boldsymbol{Y}^x,\boldsymbol{u}\rangle - \frac{1}{2}x^2\|\boldsymbol{u}\|^2\Big) \\ &:= \mathcal{A}\end{aligned}$$

Define

$$\mathcal{B} = \exp\Big(x\langle\boldsymbol{Y}^x,\hat{\boldsymbol{U}}\rangle - \frac{1}{2}qx^2\Big) \tag{7.22}$$

To prove (i), we will show that $\mathbb{E}[(\mathcal{A}-\mathcal{B})^2] \to 0$ in the high-dimensional limit $D \to \infty$ for any value of x. To do this, it is sufficient to show that $\mathbb{E}[\mathcal{A}^2]$, $\mathbb{E}[\mathcal{B}^2]$ and $\mathbb{E}[\mathcal{A}\mathcal{B}]$ converge to the same limit, using the replica symmetric property of $\boldsymbol{U}|\boldsymbol{Y}^u$. Indeed, $\mathbb{E}[\mathcal{A}^2]$ can be written as

$$\mathbb{E}\exp\Big(\sum_{a=1}^{2} x\langle\boldsymbol{Y}^x,\boldsymbol{U}^a\rangle - \frac{1}{2}x^2\|\boldsymbol{U}^a\|^2\Big)$$

where $\boldsymbol{U}^1,\boldsymbol{U}^2$ are sampled independently from $\boldsymbol{U}|\boldsymbol{Y}^u$. Substituting $\boldsymbol{Y}^x = X\boldsymbol{U} + \boldsymbol{Z}$ into the previous expression, we obtain

$$\mathbb{E}\exp\Big(\sum_{a=1}^{2} xX\langle\boldsymbol{U},\boldsymbol{U}^a\rangle + x\langle\boldsymbol{Z},\boldsymbol{U}^a\rangle - \frac{1}{2}x^2\|\boldsymbol{U}^a\|^2\Big)$$

Taking the expectation over $\boldsymbol{Z}$ and using the fact that $\mathbb{E}[e^{\langle\boldsymbol{a},\boldsymbol{Z}\rangle}] = e^{\frac{1}{2}\|\boldsymbol{a}\|^2}$, we have

$$\mathbb{E}[\mathcal{A}^2] = \mathbb{E}\exp\Big(\sum_{a=1}^{2} xX\langle\boldsymbol{U},\boldsymbol{U}^a\rangle + x^2\langle\boldsymbol{U}^1,\boldsymbol{U}^2\rangle\Big)$$

It follows from replica symmetric property of $\boldsymbol{U}|\boldsymbol{Y}^u$ that

$$\lim_{D\to\infty}\mathbb{E}[\mathcal{A}^2] = \mathbb{E}\exp\big(2qXx + qx^2\big) \tag{7.23}$$

To calculate the limits of $\mathbb{E}[\mathcal{A}\mathcal{B}]$ and $\mathbb{E}[\mathcal{B}^2]$, we follow exactly the same procedure, which involves substituting the definition of $\boldsymbol{Y}^x$, taking the expectation over $\boldsymbol{Z}$, and using the replica symmetric property. This leads us to the same limit as (7.23), thereby proving (i).

It follows immediately from the asymptotic equivalence between $P(x|\boldsymbol{Y})$ and $\bar{P}(x|\boldsymbol{Y})$ that the statistics $\langle\boldsymbol{Y}^x,\hat{\boldsymbol{U}}\rangle$ is asymptotically sufficient for estimating X from $\boldsymbol{Y}$. This means that all of the relevant information about X can be extracted from $\langle\boldsymbol{Y}^x,\hat{\boldsymbol{U}}\rangle$ instead of from $\boldsymbol{Y}$, without any loss of information in high dimensional limit.

Now we have

$$\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}} \rangle = \langle X\boldsymbol{U} + \boldsymbol{Z}, \hat{\boldsymbol{U}} \rangle = X\langle \boldsymbol{U}, \hat{\boldsymbol{U}} \rangle + \langle \boldsymbol{Z}, \hat{\boldsymbol{U}} \rangle.$$

Given that $\langle \boldsymbol{Z}, \hat{\boldsymbol{U}} \rangle \sim \mathcal{N}(0, \|\hat{\boldsymbol{U}}\|^2)$ and $\boldsymbol{Z}$ is independent of X, in the limit $D \to \infty$, this inner product converges in distribution to $\sqrt{q}\xi$, where ξ is a standard normal random variable independent of $\boldsymbol{X}$. Therefore

$$\frac{\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}} \rangle}{\sqrt{q}} \xrightarrow{d} \sqrt{q}X + \xi, \quad D \to \infty,$$

which proves (ii) since the left hand side of the last expression is also a sufficient statistics of X given $\boldsymbol{Y}$. □

Next we will use Lemma 7.1 to give another analysis of the matrix factorization problem in Section 7.3. We rewrite the model here as

$$\boldsymbol{Y} = \sqrt{\frac{\lambda}{n}}\boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{Z} \tag{7.24}$$

where $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ replace x^0, Y, Z in the previous notation. The reason for this change of notations is that, with the cavity method we will need to look into the coordinates, so it becomes necessary to distinguish between vectors and scalars.

We assume that the estimation of $n^{-1/2}\boldsymbol{X}$ given $\boldsymbol{Y}$ satisfies the replica symmetry property with overlap q. This is the same assumption made by the replica symmetric ansatz. We will use Lemma 7.1 to derive the fixed point equation for q

Let $i \in [n]$ be fixed. The cavity method involves dividing the data $\boldsymbol{Y}$ into two parts. The first part, denoted as $\boldsymbol{Y}^1$, includes the observations related to X_i, which can be compressed into

$$\sqrt{\frac{2\lambda}{n}}X_i\boldsymbol{X}_{-i} + \tilde{\boldsymbol{Z}}_i \tag{7.25}$$

where $\boldsymbol{X}_{-i} = (X_j)_{j\neq i}$ and $\tilde{\boldsymbol{Z}}_i$ is a standard Gaussian vector. The SNR λ is multiplied by 2 because each X_iX_j with $j \neq i$ appear in two channels (Proposition 2.2). We ignore the channel with X_i^2 since it contains an insignificant amount of information related to X_i. The second part consists of the remaining data $\boldsymbol{Y}^2 = \boldsymbol{Y}\backslash\boldsymbol{Y}^1$. Since the dataset $\boldsymbol{Y}^1$ only contains an insignificant amount of information relevant to $\boldsymbol{X}_{-i}$, estimating $\boldsymbol{X}_{-i}$ from $\boldsymbol{Y}$ is essentially the same as estimating it from $\boldsymbol{Y}^2$. Therefore, $n^{-1/2}\boldsymbol{X}_{-i}|\boldsymbol{Y}^2$ also satisfies the replica symmetric property with overlap q. It is easy to check that the Lemma 7.1 is applicable for this model, with X_i and $\sqrt{2\lambda/n}\boldsymbol{X}_{-i}$ respectively playing the role of X and $\boldsymbol{U}$ in the lemma. As a result, estimating X_i from $\boldsymbol{Y}$ is asymptotically equivalent to estimating it from the output of a Gaussian channel with SNR $2\lambda q$.

For distinct $i, k \in [n]$, it can be shown that $\tilde{\boldsymbol{Z}}_i$ and $\tilde{\boldsymbol{Z}}_k$ are independent. Therefore, the noises ξ_i, ξ_k of the equivalent Gaussian channels associated with X_i, X_k are independent (see the proof of Lemma 7.1-ii). Therefore $\hat{\boldsymbol{X}}_i$, which depends on ξ_i and X_i, are

asymptotically independent for all i. By the law of large number

$$q = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i^2 = \mathcal{O}_X(2\lambda q) \tag{7.26}$$

which is exactly the same fixed point equation obtained by the replica method. Note that fixed point equations may not uniquely determine overlaps, as they can have multiple solutions. However, rigorous methods ([8]) demonstrate that overlaps can be uniquely determined as the minimax point of a certain function.

7.5 Multitask learning on Gaussian mixtures

In this section based on our paper [78], we will study in details a simple model of multitask learning. Although the main results can be derived by replica computations similar to the example given in Section 7.3, we will give a more rigorous derivation of these results based on the cavity method.

Multitask learning (MTL) is a machine learning method in which multiple tasks are learned simultaneously. It can facilitate knowledge transfer between tasks and can lead to more informative data representation [92]. Although learning from related tasks can help disseminate useful information learned from one task to other tasks, the presence of unrelated tasks can also be beneficial. With the prior knowledge that two given tasks are unrelated, the algorithm can learn to ignore irrelevant features of the data distribution, resulting in better data representation [80].

We will propose a simple model of MTL based on Gaussian mixtures that focuses on capturing the transfer of knowledge between tasks, leaving out the data representation aspect. This model extends the semi-supervised learning model studied in [58], which examines the added value of unlabeled data in a one-task classification. We consider here instead multiple classification tasks, for which the data in each task are partially labeled and come from two classes. Thanks to the simplicity of our model, we can define the correlation between two tasks as a number in $[-1, 1]$. We are interested in the performance gain when correlated tasks are learned together versus when they are learned separately, assuming the best algorithm is used. This leads to the concept of *Bayes risk*, defined as the smallest possible probability of misclassifying a new data point not from the training dataset. Despite the randomness of data, in the limit where both the quantity and the dimensionality of the data are large with a fixed ratio, the Bayes risk converges towards a deterministic value.

We will derive an exact formula for the asymptotic Bayesian risk, from which we will analyze the role of task correlations and how they interact with other elements of the model, such as the proportion of labeled data in each task. It is well known that unsupervised learning on a single task with Gaussian mixture data leads to a phase transition that separates the high and low noise regimes. We demonstrate that phase transition persists to the case of multitask and study how it is affected by task correlations. In the context of source task - target task, we identify the conditions in which the source task is

most beneficial to the target task. Finally, we will derive a simple algorithm that achieves the optimal performance for supervised multitask learning on Gaussian mixtures.

7.5.1 Model

We consider T classification tasks, where task t consists of N_t data points in $\mathbb{R}^D$. The i-th data point in task t, denoted by $\boldsymbol{Y}_{ti}$, is given by

$$\boldsymbol{Y}_{ti} = V_{ti}\boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{ti} \tag{7.27}$$

where $\sigma_t > 0$. The random variables $\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{Z}$ are independent, with

$$V_{ti} \overset{i.i.d}{\sim} \mathcal{U}(\{-1, 1\}),$$
$$Z_{ti} \overset{i.i.d}{\sim} \mathcal{N}(0, I_D),$$

and $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T$ are chosen uniformly randomly on the unit sphere S^{D-1}, conditioned on the event

$$\langle \boldsymbol{U}_t, \boldsymbol{U}_{t'} \rangle = C_{tt'}, t \neq t'.$$

The matrix $\boldsymbol{C} = (\langle \boldsymbol{U}_t, \boldsymbol{U}_{t'} \rangle)_{t,t'=1}^T$ is called the *task-correlation matrix*. It follows from the definition that $\boldsymbol{C}$ is a positive definite matrix with diagonal entries all equal to 1.

In other words, the data in task t comes from two classes corresponding to two Gaussian distributions centered at $\pm \boldsymbol{U}_t$ with the same covariance $\sigma_t^2 I_D$. The positions of the centers are not known and can only be estimated from the data. The class of a data point $\boldsymbol{Y}_{ti}$ is indicated by V_{ti}, so each data point has probability $1/2$ of belonging to each class. A data point is said to be *labeled* if we know which class it belongs to, otherwise it is *unlabeled*. Independently of all other random variables, each data point in task t is labeled with probability η_t. The cases $\eta_t = 1$ and $\eta_t = 0$ correspond to supervised and unsupervised learning. $C_{tt'}$ measures the correlation between tasks t and t'. The parameters $\lambda_t = 1/\sigma_t^2$ are called the *signal to noise ratio* (SNR). As the SNR increases, the two classes separate and classification is easier. We study the model in the setting where the dimension and the amount of data in each task tends to infinity at a fixed rate $\alpha_t = \lim_{D\to\infty} N_t/D$, called the *sampling ratio*. Note that the model for semi-supervised learning studied in [58] corresponds to the case $T = 1$.

We assume to have access to the dataset $\boldsymbol{Y} = (\boldsymbol{Y}_{ti})$, the labels as well as model parameters $(\sigma_t), (\eta_t), (\alpha_t)$ and $\boldsymbol{C}$[1]. Our job is to use that available information to classify a new data point $\boldsymbol{Y}_{\text{new}}$ in any given task t

$$\boldsymbol{Y}_{\text{new}} = V_{\text{new}}\boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{\text{new}} \tag{7.28}$$

We are interested in the minimal classification error, i.e. the Bayes risk

$$\inf_{\hat{V}} \mathbb{P}(\hat{V} \neq V_{\text{new}}) \tag{7.29}$$

where the infimum is taken over all estimators of V_{new}.

[1] In fact, $\boldsymbol{\sigma}$ and $\boldsymbol{C}$ can be estimated with vanishing errors as $D \to \infty$, given that a positive fraction of labeled data is available in each task (Section F) .

7.5.2 Main result

We make the following assumption for our model:

Assumption. $\sigma_t^{-1}\boldsymbol{U}_t|\boldsymbol{Y}$ and $N_t^{-1/2}\boldsymbol{V}_t|\boldsymbol{Y}$ satisfies the replica symmetric property for all $t \in [T]$ with the overlaps denoted by q_{ut} and q_{vt} respectively.

Our main result is as follows.

Result 7.1. *i) Under the setting of the model, as $D \to \infty$, the Bayes risk converges to*

$$1 - \Phi\left(\sqrt{q_{ut}}\right),$$

where $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2} dx$

ii) The overlaps q_{ut}, q_{vt} satisfies the following equations

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{7.30a}$$

$$q_{vt} = \eta_t + (1 - \eta_t)\mathbb{E}[\tanh(\sqrt{q_{ut}}Z + q_{ut})] \tag{7.30b}$$

where $Z \sim \mathcal{N}(0, 1)$ and

$$\boldsymbol{M} = \{C_{tt'}/\sigma_t\sigma_{t'}\}_{t,t'=1}^{T}$$
$$\boldsymbol{D} = \mathrm{diag}\{\alpha_t q_{vt}\}_{t=1}^{T}$$

Remark 7.3. When $q_{ut} = 0$, the Bayes risk of task t is equal to 0.5, which corresponds to the level of classification error of a random guess. In this case, we say that the classification of task t is *impossible*. On the other hand, if q_{ut} is positive, the classification of task t is said to be *possible*.

Remark 7.4. The fixed point equations (7.30a) and (7.30b) may not uniquely determine the overlaps. Specifically, for unsupervised learning with high SNR, two solutions exist: the zero solution is unstable while the non-zero solution is stable, and the stable solution is naturally chosen as overlaps. In other cases, there is only one solution.

We can perform a sanity check of the result by considering the following special cases: if the similarity between any two tasks is zero, the result implies that MTL has the same asymptotic Bayes risks as learning task separately, which is obvious since the data from different tasks are independent, while if $\sigma_t = \sigma$ and $C_{tt'} = 1$ for all t, t', i.e. the data distributions are identical for all tasks, the asymptotic Bayes risks of all tasks are equal to that of a single task with parameters $\alpha = \sum_t \alpha_t$ and $\alpha\eta = \sum_t \alpha_t\eta_t$.

7.5.3 Consequences

We present in this section some implications of the main result.

A Supervised learning.

For supervised learning with only one task, the minimal classification error of a new data point $\boldsymbol{Y}_{\text{new}}$ is achieved by the estimator $\hat{V}_{\text{new}} = \text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \bar{\boldsymbol{Y}} \rangle)$, where $\bar{\boldsymbol{Y}} = N^{-1} \sum_i V_i \boldsymbol{Y}_i$ [58]. In the multitask case, if $\boldsymbol{Y}_{\text{new}}$ is a new data point in task t, the following algorithm achieves the optimal performance:

1. Compute

$$\bar{\boldsymbol{Y}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Y}_{ti}$$

2. Compute

$$\tilde{\boldsymbol{Y}}_t = \sum_{s=1}^{T} a_{ts} \bar{\boldsymbol{Y}}_s$$

   where $\boldsymbol{A} = (a_{ts})_{t,s=1}^{T} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}})^{-1}$.

3. The asymptotic Bayes risk is achieved by

$$\hat{V}_{\text{new}} = \text{sgn}(\langle \boldsymbol{Y}, \tilde{\boldsymbol{Y}}_t \rangle). \tag{7.31}$$

We can see that the optimal estimator for multiple tasks modifies the optimal estimators for separated tasks $\bar{\boldsymbol{Y}}_t$ by taking into account the correlations between tasks as well as their levels of difficulty and the relative sizes, measured by $\boldsymbol{C}, (\sigma_t)$ and (α_t) respectively. Interestingly, this optimal algorithm coincides with the method proposed in [**?**] using a different approach.

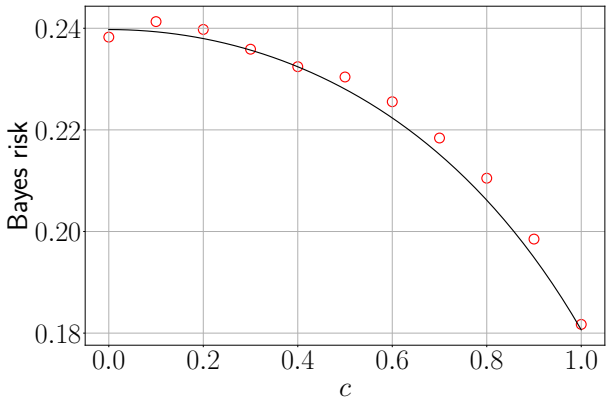


Figure 7.1: Bayes risk vs performance of the asymptotic optimal algorithm. $\alpha_1 = \alpha_2 = 1$, $\sigma_1 = 1$, $\sigma_2 = 0.5$, $D = 1000$.

B Unsupervised learning and phase transition.

A particularly interesting behavior that only occurs in the case of unsupervised learning is phase transition. One of the most well-known example of this phenomenon is *BBP phase transition* [5] which concerns a single learning task with $\lim_{D\to\infty} N/D = 1$. When $\lambda = 1/\sigma^2 \leq 1$, no estimator can achieve a smaller classification error than 0.5. In other words, the classification is objectively impossible since the two classes are statistically identical. On the other hand, we say that a task is *possible* if one can obtain a classification error smaller than 0.5. It turns out that phase transition persists to the case of multitask. Fig. 7.2 shows the performance of task 1 in terms of SNRs in the case of two tasks with $N_1 = N_2 = D$ and correlation $c = 0.7$. The classification is impossible in the region delimited by the black curve.

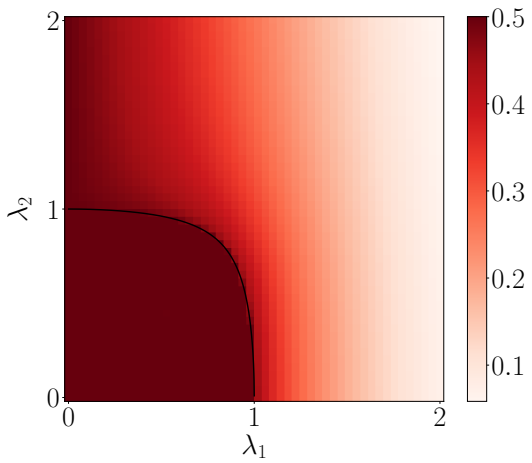


Figure 7.2: Bayes risk of Task 1 in terms of SNR of each task. Two tasks are unsupervised, with $N_1 = N_2 = D$ and correlation $c = 0.7$. The classification is impossible in the region delimited by the black curve. The impossible region is identical for two tasks.

The simulation also shows that the impossible regions are identical for both tasks. In other words, two correlated tasks are either possible or impossible. This observation can be explained by the following result:

Result 7.2. *If the tasks are* ***connected****, meaning that for any two tasks t and t', there is a sequence of tasks $t_1, \ldots, t_k$ with $k \geq 0$ such that $C_{tt_1}, C_{t_1t_2} \ldots C_{t_kt'} \neq 0$, then they are either all possible or all impossible.*

Note that phase transition disappears as soon as a positive proportion of labeled data is available, since supervised learning restricted on labeled data already produces a non-trivial performance.

In the case of two tasks with $N_1 = N_2 = D$, the region of impossible classification is

given by

$$\left\{(\lambda_1, \lambda_2) \in [0,1]^2 : (1-\lambda_1^2)(1-\lambda_2^2) \geq c^4 \lambda_1^2 \lambda_2^2\right\} \tag{7.32}$$

as shown in Figure 7.3. As the task correlation c increases from 0 to 1, this region shrinks from the unit square $[0,1]^2$ to a quarter of a disk.

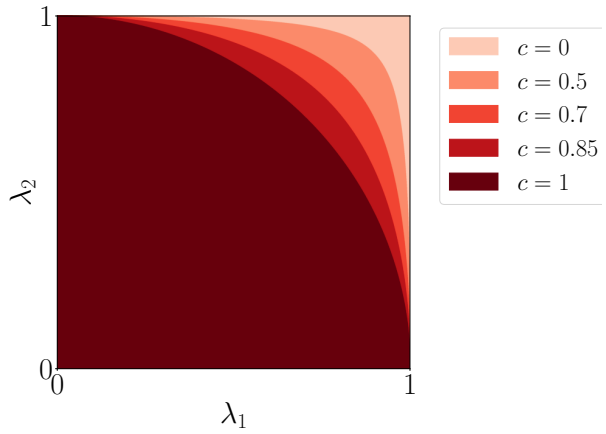


Figure 7.3: The region of impossible classification shrinks as the task correlation increases. When two tasks are uncorrelated ($c = 0$), the region of impossible classification is the whole square $[0,1]^2$. As c increases from 0 to 1, the impossible region shrinks from the unit square $[0,1]^2$ to a quarter of a disk.

Another special case where an explicit formula for the impossible region can be obtained is when there are T tasks with $N_1 = \cdots = N_T = D$, with correlation $c > 0$ between any two of them, and $\lambda_t = \lambda$ for all t. It can be shown that the classification is impossible whenever

$$\lambda \leq \frac{1}{\sqrt{1 + (T-1)c^2}}. \tag{7.33}$$

C Semi-supervised learning.

To reduce the number of model parameters in the simulation, we here focus on a specific setting consisting of one *source task* and one *target task*. The source task is comparatively easy: it can be fully labeled, have a high SNR, or have a larger dataset. We want to see how the target task benefits from the source task.

Figure 7.4 illustrates the effect of task correlation. The task correlation c ranges from 0 to 1. Note that the correlations c and $-c$ are essentially the same, since one can be transformed to another by switching labels in one task. The first task (target task) is composed of a small dataset ($\alpha_1 = 0.1$) without label ($\eta_1 = 0$), while the second task (source task) consists of a fully labeled dataset ($\eta_2 = 1$) with twice as much data ($\alpha_2 = 0.2$). If two tasks are highly correlated ($c \gtrsim 0.5$), the performance of the target

task can be significantly improved. When c is near zero, the decrease in Bayes risk is slow, in order of $O(c^2)$. Note that two tasks have the same SNR ($\lambda_1 = \lambda_2 = 4$), so when $c = 1$ they have the same data distribution and can be combined into a single task, yielding a identical performance.

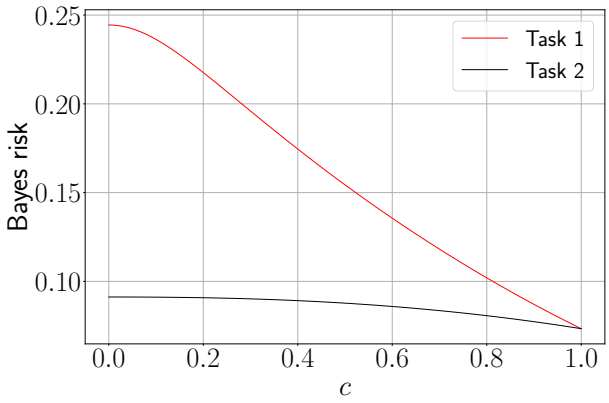


Figure 7.4: Two-task setting: Bayes risks as a function of the task correlation c, with proportions of labeled data $\eta_1 = 0$, $\eta_2 = 1$, oversampling ratios $\alpha_1 = 0.1$, $\alpha_2 = 0.2$ and SNRs $\lambda_1 = \lambda_2 = 4$. When two tasks are highly correlated ($c \gtrsim 0.5$), the performance of task 1 is significantly improved.

In Figure 7.5, we compute the rate of error reduction in the target task as a result of transferring information from the source task. We found that MTL is most effective when the SNR of the target task is near the phase transition and is smaller than that of the source task, while the proportion of labeled data is low.

Intuitively, there are three reasons for this. Firstly, the labeled data from the target task is more valuable than that of source task, even in this case where two tasks are highly correlated ($c = 0.8$). This leads to lower gain when the proportion of labeled data in the target task is high. Secondly, if the source task is more difficult than the target task, i.e. the SNR is higher in the target task, then the source task is not very useful. Finally, near the phase transition where the target task struggles, labeled data from the source task can offer valuable help.

7.5.4 Derivation of the results

A Reformulation as a tensor model

Let $\tilde{\boldsymbol{U}}_t = \sqrt{D}\boldsymbol{U}_t$, it can be shown (Remark 7.5) that in the limit $D \to \infty$, $\tilde{U}_{tj}$ are asymptotically Gaussian with covariance

$$\mathbb{E}[\tilde{U}_{tj}\tilde{U}_{t'j'}] = C_{tt'}\delta_{jj'} \tag{7.34}$$

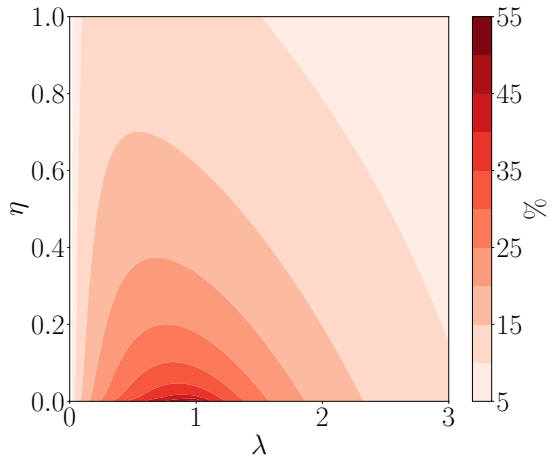


Figure 7.5: Percentage of reduction of Bayes risk in term of SNR and proportion of labeled data of the target task, with parameters $c = 0.8$, $N_1 = N_2 = D$, $\lambda_1 = 2$, $0 \leq \lambda_2 \leq 3$, $\eta_1 = 1$, $0 \leq \eta_2 \leq 1$.

Let $\boldsymbol{W}_t = \sqrt{D}\boldsymbol{U}_t/\sigma_t$, the original model can be written as a collection of one-dimensional Gaussian channels

$$Y_{ijt} = \frac{1}{\sqrt{D}} V_{ti} W_{tj} + Z_{tij} \tag{7.35}$$

for $1 \leq t \leq T, 1 \leq i \leq N_t, 1 \leq j \leq D$. As $D \to \infty$, the random variables W_{tj} are asymptotically Gaussian with covariance

$$\mathbb{E}[W_{tj} W_{t'j'}] = M_{tt'} \delta_{jj'} \tag{7.36}$$

where $M_{tt'} = C_{tt'}/(\sigma_t \sigma_{t'})$.

Next, the information conveyed by the labels can be absorbed into the prior distribution of $\boldsymbol{V}$. Specifically, if the value of V_{ti} is unknown, then its prior remains uniform over $\{-1, 1\}$. Otherwise, if it is known that $V_{ti} = 1$, then the prior of V_{ti} is given by the density $\delta(v - 1)$. Note that in this case, the posterior coincides with the prior.

The replica symmetric property of $\sigma_t^{-2}\boldsymbol{U}_t|\boldsymbol{Y}$ implies that $D^{-1/2}\boldsymbol{W}_t|\boldsymbol{Y}$ also has the replica symmetric property with overlap $\boldsymbol{q}_{ut}$.

In summary, the problem can be cast as a tensor model, whereby the objective is to estimate the signals $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ based on prior information regarding these vectors and noisy observations of the tensor products $\boldsymbol{V}_t \otimes \boldsymbol{W}_t$.

Remark 7.5. The claim that $\tilde{\boldsymbol{U}}_t$'s are asymptotically Gaussian vectors with correlation (7.34) can be seen as a direct consequence of Gibbs principle presented in Chapter 4. Another way to see this is from how the vectors $\boldsymbol{U}_t$'s are generated. To generate these vectors according to the prior distribution specified in the model, we follow these steps:

1. Generate $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{0}, I_D)$.

2. Orthonormalize $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ using Gram-Schmidt process, we obtain orthonormal vectors $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_T$

3. $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T) = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_T)\boldsymbol{C}^{1/2}$, where $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T)$ denotes the $D \times T$ matrix with columns $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T$.

In the high dimensional limit, the vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ are asymptotically orthogonal, each with norm $\simeq \sqrt{D}$. Therefore, the orthonormalizing step produces approximately $D^{-1/2}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T)$. The claim follows from this and step 3.

B Fixed point equations

To obtain the fixed point equations, we follow the same approach as the example presented in Section 7.4. We assume that the proportion of unlabeled data is positive in any task. By taking the limit of these proportions to zero, we can derive the result for the supervised case. Fix $t \in [T]$ and $i \in [N_t]$ such that V_{ti} is unknown. We divide the data $\boldsymbol{Y}$ into two parts: $\boldsymbol{Y}^1$ consisting of the observations concerning V_{ti}, namely

$$\boldsymbol{Y}_{ti} = \frac{1}{\sqrt{D}} V_{ti} \boldsymbol{W}_t + \boldsymbol{Z}_{ti}$$

and the remaining data $\boldsymbol{Y}^2 = \boldsymbol{Y} \backslash \boldsymbol{Y}^1$. Since the dataset $\boldsymbol{Y}^1$ only contains an insignificant amount of information relevant to $\boldsymbol{W}_t$, estimating $\boldsymbol{W}_t$ from $\boldsymbol{Y}$ is essentially the same as estimating $\boldsymbol{W}_t$ from $\boldsymbol{Y}^2$. Therefore, $D^{-1/2}\boldsymbol{W}_t | \boldsymbol{Y}^2$ also satisfies the replica symmetric property with overlap q_u. It is easy to check that the Lemma 7.1 is applicable, with V_{ti} and $D^{-1/2}\boldsymbol{W}_t$ respectively playing the role of X and $\boldsymbol{U}$ in the lemma. As a result, estimating V_{ti} from $\boldsymbol{Y}$ is asymptotically equivalent to estimating the signal V_{ti} from the output of the Gaussian channel with SNR q_{ut}. For distinct $i, k \in [N_t]$, since $\boldsymbol{Z}_{ti}$ and $\boldsymbol{Z}_{tk}$ are independent, it can be seen from the proof of Lemma 7.1-ii that the noises ξ_i and ξ_k of the equivalent Gaussian channels associated with V_{ti}, V_{tk} are also independent. Therefore V_{ti}, which depends on ξ_i and V_{ti}, are asymptotically independent for all i such that V_{ti} is unlabeled. By the law of large number,

$$\begin{aligned} r_{vt} &:= \lim_{N_t \to \infty} \frac{1}{(1-\eta_t)N_t} \sum_i \hat{V}_{ti}^2 \\ &= \mathcal{O}_v(q_{ut}) \end{aligned} \tag{7.37}$$

where the sum is over all $i \in [N_t]$ such that V_{ti} is unlabeled and $\mathcal{O}_v$ is the overlap function of the Gaussian channel with Rademacher signal. From 7.2.2,

$$\mathcal{O}_v(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \quad Z \sim \mathcal{N}(0,1). \tag{7.38}$$

On the other hand, from the definition of r_{vt}, we have

$$q_{vt} = \eta_t + (1-\eta_t) r_{vt} \tag{7.39}$$

The fixed point equation (7.30b) follows from (7.37), (7.38) and (7.39).

Following exactly the same cavity argument, the estimation of W_{tj} given $\boldsymbol{Y}$ is asymptotically equivalent the the estimation of the signal W_{tj} from the output of the Gaussian channel with SNR $\alpha_t q_{vt}$. Moreover, the noises corresponding to the signals W_{tj} and $W_{t'j'}$ are asymptotically independent for $(t,j) \neq (t',j')$. When $j \neq j'$, the signals W_{tj} and $W_{t'j'}$ are independent. As a result, the inference on the equivalent Gaussian channels can be performed independently on groups of T scalar Gaussian channels $(W_{tj})_{t=1}^T$. By the law of large number,

$$q_{ut} = \lim_{D\to\infty} \frac{1}{D} \sum_{j=1}^{D} \hat{W}_{tj}^2 = \mathcal{O}_{w,t}(\{\alpha_t q_{vt}\}_{t=1}^T) \tag{7.40}$$

where $\mathcal{O}_{w,t}$ are overlap functions of the Gaussian channel with signal $\mathcal{N}(0, \boldsymbol{M})$. The explicit formula for $\mathcal{O}_{w,t}$ are computed in Section 7.2.3, which gives the fixed point equation (7.30a).

C Bayes risk and optimal algorithm

Suppose we want to classify a new data point $\boldsymbol{Y}_{\text{new}}$ in task t

$$\boldsymbol{Y}_{\text{new}} = V_{\text{new}} \boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{\text{new}} \tag{7.41}$$

It is easy to check that Lemma 7.1 can be applied to this problem, with $V_{\text{new}}, \boldsymbol{U}_t$ playing the role of $X, \boldsymbol{U}$ in the lemma, as the posterior $\sigma_t^{-1}\boldsymbol{U}_t|\boldsymbol{Y}$ satisfies the replica symmetric property with overlap q_{ut}. As a result, in high dimensional limit, estimating V_{new} given $\boldsymbol{Y}, \boldsymbol{Y}_{\text{new}}$ is essentially the same as estimating the signal V_{new} from the output of the Gaussian channel with SNR q_{ut}. This implies that the minimal classification error of V_{new} is given by that of the Gaussian channel with Rademacher signal and SNR q_{ut}, which is (Section 7.2.2)

$$1 - \Phi\left(\sqrt{q_{ut}}\right),$$

According to Lemma 7.1, $S = \langle \boldsymbol{Y}_{\text{new}}, \hat{\boldsymbol{U}}_t \rangle / \sqrt{q_{ut}}$ is sufficient for estimating V_{new}. Moreover, S converges in law to the output of the Gaussian channel with signal V_{new} and SNR q_{ut}. The estimator that minimizes the Bayes risk for this channel is simply $\text{sgn}(S)$, which leads to the optimal estimator of V_{new} as $\text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \hat{\boldsymbol{U}}_t \rangle)$. The next step is to determine the value of $\hat{\boldsymbol{U}}_t$. We will take advantage of the fact that the vectors $\boldsymbol{U}_t$ are asymptotically Gaussian, so our subsequent argument will rely on the reformulation (7.34) of the model. We will need the following result

Lemma 7.2. *The following collection of Gaussian channels*

$$Y_i = c_i X + Z_i, \quad i = 1, \dots, n \tag{7.42}$$

with input X, outputs Y_i, SNR c_i^2 and independent standard Gaussian noises Z_i, is equivalent to a single Gaussian channel with signal X, output $\langle \boldsymbol{c}, \boldsymbol{Y} \rangle / \|\boldsymbol{c}\|$ and SNR $\sum_{i=1}^n c_i^2$. Moreover,

Proof. It is straightforward to verify that the statistics $S := \langle \boldsymbol{c}, \boldsymbol{Y} \rangle / \|\boldsymbol{c}\|$ is sufficient for estimating X from $\boldsymbol{Y}$. Moreover, $S = \|\boldsymbol{c}\| X + \xi$ where $\xi = \|\boldsymbol{c}\|^{-1} \langle \boldsymbol{c}, \boldsymbol{Z} \rangle$ is standard Gaussian and independent of X. This proves the claim of the lemma. □

Remark 7.6. From the proof of Lemma 7.2 we can also see that the noise ξ of the simplified channel comes from the noises of the original channels.

The Lemma 7.2 implies that, for each (t, j) fixed, the following Gaussian channels

$$Y_{tij} = \frac{1}{\sqrt{D}} V_{ti} W_{tj} + Z_{tij}, \quad i = 1, \dots, N_t$$

which share the same signal W_{tj}, can be simplified into a single Gaussian channel with output $\sqrt{N_t} \bar{Y}_{tj}$ and SNR $N_t / D \simeq \alpha_t$, where $\bar{Y}_{tj}$ is the j-th coordinate of the vector $\bar{\boldsymbol{Y}}_t$ in the algorithm.

For $(t, j) \neq (t', j')$, the noises of the simplified Gaussian channels associated with W_{tj} and $W_{t'j'}$ are independent, as a consequence of Remark 7.6. Additionally, the signals W_{tj} and $W_{t'j'}$ are independent if $j \neq j'$. Therefore, the inference on the simplified Gaussian channels can be carried out independently on each group of T channels with signals $(W_{tj})_{t=1}^T$. The MMSE estimator on each of these groups can be computed explicitly as

$$(\hat{W}_{tj})_{t=1}^T = \boldsymbol{B} (\sqrt{N_t} \bar{Y}_{tj})_{t=1}^T$$

where

$$\boldsymbol{B} = \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2} \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2})^{-1}$$

(7.2.3). Equivalently,

$$\hat{\boldsymbol{W}}_t = \sum_s B_{ts} \sqrt{N_s} \bar{\boldsymbol{Y}}_s$$

Dividing both sides by $\sqrt{D}$ and using $N_t / D \simeq \alpha_t$, we have

$$\tilde{\boldsymbol{Y}}_t := \sigma_t^{-1} \hat{\boldsymbol{U}}_t \simeq \sum_s A_{ts} \bar{\boldsymbol{Y}}_s \tag{7.43}$$

where $A_{ts} = B_{ts} \sqrt{\alpha_s}$. Therefore,

$$\boldsymbol{A} = \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\alpha}} (\boldsymbol{I} + \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\alpha}})^{-1}$$

as given in the optimal algorithm. The optimal estimator for V_{new} is $\text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \hat{\boldsymbol{U}}_t \rangle) = \text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \tilde{\boldsymbol{Y}}_t \rangle)$.

D Region of impossible recovery

In the unsupervised case, the fixed point equations are

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{7.44a}$$
$$q_{vt} = F(q_{ut}) \tag{7.44b}$$

where

$$F(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)],$$

These equations always admits $(\boldsymbol{q}_u, \boldsymbol{q}_v) = (\boldsymbol{0}, \boldsymbol{0})$ as solution. The classification is impossible if and only if this solution is stable (Remark 7.4). To analyze the stability of (7.44) around zero, let $q_{ut}, q_{vt} = O(h)$ where $h \to 0$. For vectors A and B of the same dimension, we denote $A \simeq B$ if $|A - B| \simeq O(h^2)$, where $|\,.\,|$ denotes the Euclidean norm. Using the Taylor expansion $\tanh(x) = x - x^3/3 + o(x^3)$, we get

$$q_{vt} = F(q_{ut}) \simeq q_{ut}$$

On the other hand,

$$\begin{aligned} q_{ut} &= [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \\ &\simeq [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} - \boldsymbol{D}\boldsymbol{M})]_{tt} \\ &= [\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}]_{tt} \\ &= \sum_{s=1}^{T} M_{ts}^2 \alpha_s q_{vs} \end{aligned}$$

Let

$$\boldsymbol{P} = (M_{ts}^2 \alpha_s)_{s,t=1}^T = \Big(\frac{C_{ts}^2}{\sigma_t^2 \sigma_s^2}\alpha_s\Big)_{s,t=1}^T = (\lambda_s \lambda_t C_{st}^2 \alpha_s)_{s,t=1}^T \tag{7.45}$$

In a small neighborhood of $(\boldsymbol{0}, \boldsymbol{0})$, the system of equations can be approximated up to an error of $O(h^2)$ by

$$\boldsymbol{q}_v = \boldsymbol{q}_u \tag{7.46}$$
$$\boldsymbol{q}_u = \boldsymbol{P}\boldsymbol{q}_v \tag{7.47}$$

Therefore the fixed point $(\boldsymbol{0}, \boldsymbol{0})$ is stable if and only if the module of each eigenvalue of $\boldsymbol{P}$ is not larger than 1. Using the property that AB and BA has the same eigenvalues for general square matrices A, B, the matrix $\boldsymbol{P}$ has the same eigenvalues as the following symmetric matrix

$$\boldsymbol{R} = (\sqrt{\alpha_s \alpha_t}\lambda_s \lambda_t C_{st}^2)_{s,t=1}^T \tag{7.48}$$

Note that $\boldsymbol{R}$ is a positive semidefinite (p.s.d) matrix, since it can be written as Hadamard product of p.s.d. matrices. Therefore, the classification is impossible if and only if all eigenvalues of $\boldsymbol{R}$ are not greater than 1.

When $C_{tt'} = c$ for all $t \neq t'$ and $\lambda_t = \lambda, \alpha_t = 1$ for all t, we have

$$\boldsymbol{R} = \lambda^2(c^2 \mathbb{1}\mathbb{1}^T + (1-c^2)I) \tag{7.49}$$

Note that the matrix $\mathbb{1}\mathbb{1}^T$ has eigenvalues $0, \ldots, 0, T$, so the largest eigenvalue of $\boldsymbol{R}$ is $\lambda^2(1 + (T-1)c^2)$, from with we obtain the condition for impossible classification

$$\lambda^2(1 + (T-1)c^2) \leq 1 \tag{7.50}$$

which becomes $\lambda \leq 1$ for the special case $T = 1$.

When $T = 2$ with task correlation c and $\alpha_1 = \alpha_2 = 1$, we have

$$\boldsymbol{R} = \begin{pmatrix} \lambda_1^2 & c^2\lambda_1\lambda_2 \\ c^2\lambda_1\lambda_2 & \lambda_2^2 \end{pmatrix} \tag{7.51}$$

It is clear that the (λ_1, λ_2)-domain of impossible classification is a subset of $[0,1]^2$, otherwise at least one task is achievable. All eigenvalues of $\boldsymbol{R}$ are less than 1 if and only if $\mathrm{Tr}(\boldsymbol{I} - \boldsymbol{R}) \geq 0$ and $\det(\boldsymbol{I} - \boldsymbol{R}) \geq 0$. The first condition is already satisfied for $(\lambda_1, \lambda_2) \in [0,1]^2$ while the second condition is equivalent to

$$(1-\lambda_1^2)(1-\lambda_2^2) \geq c^4\lambda_1^2\lambda_2^2 \tag{7.52}$$

E Connected tasks are either all possible or all impossible

We will prove that if tasks are connected, then they are either all possible or all impossible, as stated in Result 7.2. As a reminder, for any task t, the value of q_{ut} is always non-negative. If $q_{ut} = 0$, then the task t is impossible; otherwise, it is possible.

Consider T Gaussian channels with outputs $(Y_t)_{t=1}^T$, signals $(X_t)_{t=1}^T$ having joint distribution $\mathcal{N}(0, \boldsymbol{M})$ and independent standard Gaussian noises. The SNRs for each channel are $(\alpha_t q_{vt})_{t=1}^T$. Then the right-hand side of (7.44a) corresponds to the overlap between the signal X_t and its MMSE estimator (Section 7.2.3).

Suppose by contradiction that the tasks can be split into non-empty sets such S and S' such that $q_{ut} = 0$ for all $t \in S$ while $q_{ut} > 0$ for all $t \in S'$. Since the tasks are connected, there exists correlated tasks t, t' such that $t \in S, t' \in S'$. Therefore, there exists t, t' such that $q_{ut} = 0, q_{ut'} > 0$ and $C_{tt'} \neq 0$.

Since $\mathbb{E}[X_t X_{t'}] = M_{tt'} = C_{tt'}/(\sigma_t\sigma_{t'}) \neq 0$, X_t is correlated with $X_{t'}$. Moreover, as $q_{vt'} = F(q_{ut'})$ and $q_{ut'} > 0$, we have $q_{vt'} > 0$. This implies that X_t is not independent of $\boldsymbol{Y} = \left\{\sqrt{\alpha_s q_{vs}} X_s + Z_s\right\}_{s=1}^T$, leading to $q_{ut} = \mathbb{E}[X_t \mathbb{E}[X_t|\boldsymbol{Y}]] > 0$, a contradiction.

F Estimating model parameters from data

Although it is assumed that the model parameters $\boldsymbol{C}$ and (σ_t) are available for the analysis, we show here that they can indeed be estimated with vanishing errors as $D \to$

∞, given that a positive fraction of labeled data is available in each task, i.e. $\eta_t > 0$ for all t. First consider the supervised learning case. Let

$$\bar{\boldsymbol{Y}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Y}_{ti} \tag{7.53}$$

Then we have

$$\bar{\boldsymbol{Y}}_t = \boldsymbol{U}_t + \sqrt{\frac{\sigma_t^2}{N_t}} \bar{\boldsymbol{Z}}_t \tag{7.54}$$

where

$$\bar{\boldsymbol{Z}}_t = \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Z}_{ti} \tag{7.55}$$

It is clear that $\bar{\boldsymbol{Z}}_t \overset{i.i.d}{\sim} \mathcal{N}(0, I_D)$ for $t = 1, \ldots, T$. We consider the following estimator of $C_{tt'}$ for $t \neq t'$:

$$\hat{C}_{tt'} = \langle \bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Y}}_{t'} \rangle \tag{7.56}$$

Insert (7.54) into the definition of $\hat{C}_{tt'}$ and use the fact that $\langle \bar{\boldsymbol{Z}}_t, \bar{\boldsymbol{Z}}_{t'} \rangle = O(\sqrt{D})$, $\langle \bar{\boldsymbol{U}}_t, \bar{\boldsymbol{Z}}_{t'} \rangle = O(1)$, which are direct consequences of Central Limit Theorem, we obtain $\hat{C}_{tt'} = C_{tt'} + O(D^{-1/2})$. Moreover

$$\|\bar{\boldsymbol{Y}}_t\|^2 = 1 + \frac{\sigma_t^2}{\alpha_t} + O(D^{-1/2}), \tag{7.57}$$

from which σ_t can also be estimated.

In the case where the proportion of labeled data is positive for all tasks, we can restrict the above estimators on the labeled data and obtain the approximate values of $\boldsymbol{C}$ and (σ_t) with errors converging to zero when $D \to \infty$.

Bibliographical notes

We refer readers to [60], [61], [9], [53], [59], and [72] for various tensor models. Another important model in Bayes-optimal setting is the generalized linear model [7] which generalizes many preexisting models in compressed sensing, statistical learning and communication. A review of statistical methods for inference problems can be found in [113]. It is interesting that in some cases, the optimization-based methods can nearly reach or achieve the optimal performance [65, 106, 71, 63, 3].

Chapter 8

SK model

In this chapter, we consider the Sherrington-Kirkpatrick (SK), a disordered system defined by the Hamiltonian

$$H(\sigma) = \frac{\beta}{\sqrt{n}} \sum_{i<j} J_{ij}\sigma_i\sigma_j \tag{8.1}$$

in which $\beta > 0$, $\sigma = (\sigma_i) \in \{-1, 1\}^n$ and the parameters J_{ij} are independent standard Gaussian random variables. In contrast with previously considered models, the SK model has much more complex and interesting behavior.

Let us study the SK model with replicas. The replicated Hamiltonian is

$$H^{\text{rep}}(\sigma^1, \ldots, \sigma^r) = \frac{\beta^2}{2n} \sum_{a<b} \langle \sigma^a, \sigma^b \rangle^2 + \frac{n\beta^2 r}{4} \tag{8.2}$$

Applying Result 4.2 with the macroscopic functions $\langle \sigma^a, \sigma^b \rangle / n$ for $1 \leq a < b \leq r$, the parametrized Hamiltonian is

$$\bar{H}^{\text{rep}} = \frac{n\beta^2}{2} \sum_{a<b} Q_{ab}^2 + \sum_{a<b} \hat{Q}^{ab}(\langle \sigma^a, \sigma^b \rangle - nQ^{ab}) + \frac{n\beta^2 r}{4}$$

where Q^{ab}'s are constraint parameters and $\hat{Q}^{ab}$ are multipliers. The free energy of $\bar{H}^{\text{rep}}$ is

$$\frac{n\beta^2}{2} \sum_{a<b} Q_{ab}^2 + \Psi(\{\hat{Q}^{ab}\}) - n \sum_{a<b} Q^{ab}\hat{Q}^{ab} + \frac{n\beta^2 r}{4}$$

where $\Psi(\{\hat{Q}^{ab}\})$ is the free energy of the Hamiltonian $\sum_{a<b} \hat{Q}^{ab}\langle \sigma^a, \sigma^b \rangle$. Differentiate with respect to Q^{ab}, we have

$$\hat{Q}_{ab} = \beta^2 Q_{ab} \tag{8.3}$$

and we obtain

$$\bar{H}^{\text{rep}} = -\frac{n\beta^2}{2} \sum_{a<b} Q_{ab}^2 + \sum_{a<b} \beta^2 Q^{ab}\langle \sigma^a, \sigma^b \rangle + \frac{n\beta^2 r}{4} \tag{8.4}$$

8.1 Replica symmetric ansatz

In the replica symmetric ansatz, we assume that the dominant extremal point of the free energy function is achieved at

$$Q^{ab} = q$$

for all $a < b$. With this ansatz, we have

$$\bar{H}^{\text{rep}} = -\frac{\beta^2}{2}\frac{r(r-1)}{2}nq^2 + \beta^2 q \sum_{a<b}\langle \sigma^a, \sigma^b \rangle + \frac{n\beta^2 r}{4}$$

Dereplicating $\bar{H}^{\text{rep}}$, we obtain

$$\bar{H} = \frac{n\beta^2 q^2}{4} + \langle \beta\sqrt{q}\xi, \sigma \rangle - \frac{n\beta^2 q}{2} + \frac{n\beta^2}{4} \tag{8.5}$$

where ξ is a standard Gaussian vector. The free energy of $\bar{H}$ is given by $nf(q)$, where

$$f(q) = \frac{\beta^2(1-q)^2}{4} + \mathbb{E}\log 2\cosh(\beta\sqrt{q}Z), \quad Z \sim \mathcal{N}(0,1).$$

The stationary point of $f(q)$ satisfies

$$q = \mathbb{E}[\tanh^2(\beta\sqrt{q}Z)]. \tag{8.6}$$

From (8.5), we conclude that the replica symmetric solution corresponds the description that σ_i are asymptotically independent under P^H, with marginal law

$$P(\sigma_i) \propto e^{\beta h_i \sigma_i}$$

where h_i are independent random variables drawn from $\mathcal{N}(0, q)$ and q solves the fixed point equation (8.6).

It turns out that the replica symmetric solution is only correct when β is below a certain threshold β_c. When $\beta > \beta_c$, the correct form of the dominant extremal point is given by Parisi ansatz. The idea is to divide $[r]$ into groups of equal size, and impose that the value of Q^{ab} depends on whether a, b belong to the same or different group. This is called 1-step replica symmetry breaking (1RSB). This procedure is then applied repeatedly, with each group being divided into subgroups of equal size, and so on. This results in k-step RSB for $k = 1, 2, \ldots$. The k-step RSB ansatz provides increasingly accurate descriptions of the SK model as k increases. However, to obtain the correct result for any value of $\beta > \beta_c$, no finite k-step RSB suffices, and we need full RSB where $k \to \infty$.

In the following, we will present the Parisi ansatz and clarify the probabilistic structures they correspond to, drawing from the discussion in Chapter 3.

8.2 Replica symmetry breaking

Consider the following sequences

$$0 < x_1 < \cdots < x_k < 1$$
$$0 < q_0 < q_1 < \cdots < q_k < 1$$

Denote $x = (x_1, \ldots, x_k)$ and $q = (q_0, \ldots, q_k)$. We need to always keep in mind that $r \to 0$. Let Π_x be the collection of all nested partitions of the set $[r]$ into r/x_1 groups of size x_1, each of them divided into x_1/x_2 groups of size x_2, and so on. We assign the level 0 to $[r]$ and levels $1, \ldots, k$ to the groups that appear in the subsequent divisions. For any $a, b \in [r]$, consider the group with highest level that contains both of them, and define their similarity $s(a, b)$ as the level of this group. For any $\pi \in \Pi_x$, define the following $r \times r$ matrix:

$$Q_\pi^{ab} = q_\ell, \text{ if } s(a, b) = \ell.$$

for different a, b in $[r]$.

The k-RSB ansatz proposes that a dominant extremal point of the free energy function has the form Q_π for some $\pi \in \Pi_x$. Since the free energy function is symmetric by permutation of replicas, Q_π is a dominant extremal point for any $\pi \in \Pi_x$. Also by symmetry and Result 4.2, the equivalent measure $\bar{P}^{\text{rep}}$ is a mixture of $\bar{P}_\pi^{\text{rep}}$ for all $\pi \in \Pi_x$, each with an equal weight, where $\bar{P}_\pi^{\text{rep}}$ is the probability measure associated with the Hamiltonian $\bar{H}^{\text{rep}}(\cdot, Q_\pi)$ given by (8.4). $\bar{P}_\pi^{\text{rep}}$ can be written more simply as

$$\bar{P}_\pi^{\text{rep}}(\sigma^1, \ldots, \sigma^r) \propto \exp\Big(\sum_{a<b} \beta^2 Q_\pi^{ab} \langle \sigma^a, \sigma^b \rangle\Big)$$

Under $\bar{P}_\pi^{\text{rep}}$, $\sigma^1, \ldots, \sigma^r$ are conditionally independent given $\xi = (\xi^a)$, with the conditional law

$$\mathbb{P}(\sigma^a | \xi) \propto \exp(\beta \langle \xi^a, \sigma^a \rangle) \tag{8.7}$$

where $\xi^1, \ldots, \xi^r$ are centered Gaussian vectors in $\mathbb{R}^n$ such that $\xi^a \overset{d}{=} h_a^{\otimes n}$, where

$$\mathbb{E}[h_a^2] = q_k,$$
$$\mathbb{E}[h_a h_b] = Q_\pi^{ab}, \quad a \neq b. \tag{8.8}$$

The replica density $\bar{P}^{\text{rep}}$ uniquely encodes a disordered system $\bar{P}$ that is asymptotically equivalent to P^H. We will describe $\bar{P}$ through the infinite exchangeable sequence of replicas derived from it. This is the same sequence that is derived from $\bar{P}^{\text{rep}}$ by marginalization. Since $\bar{P}^{\text{rep}}$ is a uniform mixture of $\bar{P}_\pi^{\text{rep}}$ for all $\pi \in \Pi_x$, this sequence is the same as $(\sigma^{a_1}, \sigma^{a_k}, \ldots)$, where $a_1, a_2, \ldots$ is a random sequence drawn from $[r]$ and $(\sigma^1, \ldots, \sigma^r) \sim \bar{P}_\pi^{\text{rep}}$ for some fixed $\pi \in \Pi_x$. Since $\sigma^1, \ldots, \sigma^r$ is conditionally independent given $\xi^1, \ldots, \xi^r$, this sequence can be generated by first generate the Gaussian vectors

$\xi^{a_1}, \xi^{a_2} \dots$ then generate $\sigma^{a_1}, \sigma^{a_2} \dots$ independently according to (8.7). We call that these Gaussian vectors are attached to the sequence $a_1, a_2, \dots$

The infinite sequence $a_1, a_2, \dots$ generates a nested partition of $\mathbb{N}$, where i, j inherit the relation between a_i, a_j. This nested partition of $\mathbb{N}$ can be represented by an infinite tree with $k+1$ levels (not including the root): the root is $\mathbb{N}$, each node different from the leaves has an infinite number of children representing the blocks of next level. Each leaf is labeled by a sequence $(n_1, \dots, n_{k+1}) \in \mathbb{N}^{k+1}$. We say that the root is at level 0, its children are at level 1, and so on. The leaves are at level $(k+1)$. For two leaves α and γ, their similarity $s(\alpha, \gamma)$ is defined as the level of their youngest common ancestor.

The correlations between $\xi^1, \dots, \xi^r$ given by (8.8) imply that the Gaussian vector attached to the leaf α has law $G_\alpha^{\otimes n}$, where

$$\begin{aligned} \mathbb{E}[G_\alpha^2] &= q_k, \\ \mathbb{E}[G_\alpha G_\gamma] &= q_{s(\alpha,\gamma)}. \end{aligned}$$

These correlations between the G_α's can be implemented by generating independent standard Gaussian random variables $g_\emptyset, (g_n), (g_{n_1,n_2}), \dots$ and let

$$G_\alpha = \sqrt{q_0} g_\emptyset + \sqrt{q_1 - q_0} g_{n_1} + \dots + \sqrt{q_k - q_{k-1}} g_{n_1,\dots,n_k} \tag{8.9}$$

for $\alpha = (n_1, \dots, n_{k+1})$. Note that G_α are the same for all leaves coming from the same parent.

We have thus translated the k-step RSB assumption into a description of the infinite exchangeable sequence derived from the asymptotic equivalent $\bar{P}$ of the SK model. In summary, this sequence can be generated by the following steps

- Generate a random nested partition of $\mathbb{N}$ with parameters $(\alpha_0, \dots, \alpha_k) = (0, x_1, \dots, x_k)$ as described in Section 3.5.
- Construct the tree of $k+1$ levels associated with this nested partition.
- Assign each leaf $\alpha = (n_1, \dots, n_{k+1})$ with a Gaussian vector $\xi^\alpha \overset{d}{=} G_\alpha^{\otimes n}$, where G_α are described by (8.9).
- For each leaf α, generate the random variables σ^α independently from

$$P(\sigma^\alpha) \propto \exp\left(\beta \langle \xi^\alpha, \sigma^\alpha \rangle\right).$$

The sequence (σ^α) for all leaves α corresponds to the exchangeable sequence derived from $\bar{P}$.

Note that for all leaves α's from the same parent, σ^α's are generated from the same law, under which the coordinates are independent. This law is referred to as a pure state. Consequently, the system behaves like a mixture of pure states, each with asymptotically independent coordinates. We can check that this description is consistent with the one given by [68].

It is difficult to compute the free energy function $F(x, q)$ from this description. The best way to do this is going back to the replica formalism. The calculation can be found in [82]. It turns out that we need to minimizing this function to obtain the values of the parameters x, q.

The full RSB is obtained by taking the limit where $k \to \infty$, in which the sequence x and q can be encoded in an increasing function $q(x)$ from $[0, 1]$ to $[0, 1]$.

8.3 Bibliographical notes

The SK model was initially studied using the replica method in [95]. However, the replica symmetry ansatz used in the paper made incorrect predictions at low temperatures. In 1979, Parisi proposed a replica ansatz that led to a 'less wrong' solution [81], known today as the one-step replica symmetry breaking (RSB) ansatz. In a subsequent paper, he proposed the full RSB scheme that provided the correct solution [82], a fact that we know today. It took several years for Parisi and his collaborators to fully understand the implications of the RSB ansatz, as presented in [68] [70], and later in the book [69], which compiled the developments on the subject. A reference on the SK model for non-physicists is provided by [73].

Parisi's formula for the free energy of the SK model was rigorously proven by Talagrand [99], building upon a breakthrough by Guerra [45]. However, the proof only confirmed the formula for the free energy and did not prove Parisi's predictions regarding the structure of the Gibbs measure, especially the ultrametricity, i.e. the fact that the pure states can be organized into a tree-like structure. The ultrametricity conjecture was finally proved by Panchenko in [79].

Perspectives

Combinatorial optimization. Although the subject of combinatorial optimization is mentioned in the introduction, it is not presented in this manuscript. I am still unable to connect the replica method with Aldous' objective method [1]. This seems to be a difficult question.

Formal construction of other exchangeable structures. Exchangeable structures such that Polya urn model and Chinese restaurant process have remarkably succinct descriptions by formal sets. The advantage of these constructions is that they greatly simplify calculations and make the exchangeability appear in a natural way. Examples of random exchangeable structures are abundant in the literature, especially in Bayesian topic model. It is interesting to know if, besides the examples in Chapter 3, there are other formal constructions of exchangeable structures.

Hermitian matrices. We have only considered real symmetric matrices. However, the replica method can also be applied to Hermitian matrices. A random Hermitian matrix A can be studied with the following Hamiltonian

$$H = -w^\dagger A w, \quad w \in \mathbb{C}^n. \tag{8.10}$$

For convenience, we choose the underlying measure of $\mathbb{C}^n$ as

$$\mu(dw) = \prod_i \frac{d\,\mathrm{Re}(w_i) d\,\mathrm{Im}(w_i)}{\pi}$$

With this underlying measure, the free energy of the Hamiltonian H is $-\log\det A$.

Similarly to the Gaussian measures on $\mathbb{R}^n$, in this case we have $\mathbb{E}[ww^\dagger] = A^{-1}$ if $w \sim P^H$. If we can find a deterministic matrix $\bar{A}$ such that the following probability measure

$$\bar{P}(dw) \propto \exp\Big(- w^\dagger \bar{A} w \Big)$$

is asymptotically equivalent to A, then we can conclude that $\bar{A}^{-1} \leftrightarrow A^{-1}$.

In this case the integration over the disorders also has an explicit form, so the replicated Hamiltonian can be computed quite easily. However, the macroscopic functions

have complex values, so if we want to reduce this to the case of real values, we have to double the number of macroscopic functions by splitting each of them into the real and imaginary part. The calculations are still doable but ugly. I wonder if there is a more elegant way to handle this.

Free probability. It will be beneficial to see the free probability in the point of view of the replica method. With the replica method, R- and S-transform as well as many other transforms can be seen as the extremizer of some potential functions. From this point of view, it is likely that some surprising connections in free probability, such as the ones described in [10], can be explained in simple way. Recovering results in free probability with the replica method could bring new insights and ideas. Another questions is whether the free cumulant can arise from replica computations, in the same way that classical cumulant-moment formula can be derived by replicas.

Gaussian equivalence in random matrix theory. Many non-linear random matrix models behave like the weighted sum of a GOE matrix and a Marchenko Pastur matrix that are independent of each other. For example, let W, Z be large random matrices with independent standard Gaussian entries, B is a Bernoulli mask with density ρ. Then the following three models

$$(ZZ^\top) \odot B, \quad f(ZZ^\top)_0, \quad f(WX)f(WX)^\top$$

all have this behavior. Here f is a non-linear function that is applied pointwise. For the matrix in the middle, the subindex zero means the diagonal is set to zero. For the first matrix, its spectrum remains the same if B is replaced by a matrix with independent Gaussian entries with mean p and variance $p(1-p)$. The rigorous proofs of these facts often rely on complicated arguments, deriving the Stieltjes transform for the limiting spectral density and realizing their behaviors mentioned above. The question is whether there is a simple explanation for this behavior. This might be connected to the fact that in the replica computation, we only care about the leading term of a logarithm, which is insensitive to the details of the random variables.

Dynamics of disordered system. This thesis only concerns about the Gibbs measure of disordered system, which describes the system at equilibrium. The dynamics aspect of disordered system is an interesting and challenging topic to learn.

Fluctuation of the free energy. This thesis only deals with the leading term of the free energy. It would be interesting to learn how to compute the fluctuations around the leading term.

Appendix A

Gaussian integrals

The following results are needed for Gaussian random matrix models.

Lemma A.1. *For $A \in \mathbb{R}^{n\times n}$ symmetric positive definite and $b \in \mathbb{C}^n$,*

$$\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}x^\top Ax + b^\top x\right) dx = \sqrt{\frac{(2\pi)^n}{\det A}} \exp\left(\frac{1}{2}b^\top A^{-1}b\right)$$

Proof. For the case $n = 1$, we have

$$\begin{aligned}\int_{\mathbb{R}} \exp\left(-\frac{1}{2}ax^2 + bx\right) dx &= \int_{\mathbb{R}} \exp\left(-\frac{1}{2}a\Big(x - \frac{b}{a}\Big)^2 + \frac{b^2}{2a}\right) \\ &= \sqrt{\frac{2\pi}{a}} \exp\left(\frac{b^2}{2a}\right)\end{aligned}$$

If A is diagonal, the integral is a product of one-dimensional integrals and the result follows immediately. For the general case, suppose $A = O^\top \tilde{A} O$ where $\tilde{A}$ is diagonal and $OO^\top = I$. Let $\tilde{x} = Ox$ and $\tilde{b} = Ob$, then $d\tilde{x} = dx$ and

$$\begin{aligned}\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}x^\top Ax + b^\top x\right) dx &= \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\tilde{x}^\top \tilde{A}\tilde{x} + \tilde{b}^\top \tilde{x}\right) d\tilde{x} \\ &= \sqrt{\frac{(2\pi)^n}{\det \tilde{A}}} \exp\left(\frac{1}{2}\tilde{b}^T \tilde{A}^{-1}\tilde{b}\right) \\ &= \sqrt{\frac{(2\pi)^n}{\det A}} \exp\left(\frac{1}{2}b^T A^{-1}b\right)\end{aligned}$$

□

Lemma A.2. *Let A be a positive definite Hermitian matrix and $u, v \in \mathbb{C}^n$, then*

$$\int_{\mathbb{C}^n} dz\, e^{-z^\dagger Az + u^\dagger z + z^\dagger v} = \frac{\pi^n}{\det A} e^{u^\dagger A^{-1} v} \tag{A.1}$$

where $dz = \prod_{i=1}^n d(\operatorname{Re} z_i) d(\operatorname{Im} z_i)$

Proof. With the same argument in the proof of Lemma A.1, it is sufficient to prove the result for $n = 1$. For $a > 0$ and $u, v \in \mathbb{Z}$, we have

$$\begin{aligned}\int_{\mathbb{C}} \exp\left(-a|z|^2 + \bar{u}z + \bar{z}v\right) dz &= \int_{\mathbb{R}^2} \exp\left(-a(x^2+y^2) + x(\bar{u}+v) + iy(\bar{u}-v)\right) dxdy \\ &= \frac{\pi}{a} \exp\left(\frac{\bar{u}v}{a}\right)\end{aligned}$$

in which the second equality follows from Lemma A.1. □

Bibliography

[1] D. Aldous and J. M. Steele, *The objective method: probabilistic combinatorial optimization and local weak convergence*, in Probability on discrete structures, Springer, 2004, pp. 1–72.

[2] D. J. Aldous, *The $\zeta(2)$ limit in the random assignment problem*, Random Structures & Algorithms, 18 (2001), pp. 381–418.

[3] B. Aubin, F. Krzakala, Y. Lu, and L. Zdeborová, *Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization*, Advances in Neural Information Processing Systems, 33 (2020), pp. 12199–12210.

[4] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, vol. 20, Springer, 2010.

[5] J. Baik, G. B. Arous, and S. Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, The Annals of Probability, 33 (2005), pp. 1643–1697.

[6] J. Baik and J. W. Silverstein, *Eigenvalues of large sample covariance matrices of spiked population models*, Journal of multivariate analysis, 97 (2006), pp. 1382–1408.

[7] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, *Optimal errors and phase transitions in high-dimensional generalized linear models*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 5451–5460.

[8] J. Barbier and N. Macris, *The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference*, Probability theory and related fields, 174 (2019), pp. 1133–1185.

[9] J. Barbier, N. Macris, and L. Miolane, *The layered structure of tensor estimation and its mutual information*, in 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2017, pp. 1056–1063.

[10] F. Benaych-Georges, *On a surprising relation between the Marchenko-Pastur law, rectangular and square free convolutions*, in Annales de l'IHP Probabilités et statistiques, vol. 46, 2010, pp. 644–652.

[11] F. Benaych-Georges and R. R. Nadakuditi, *The singular values and vectors of low rank perturbations of large rectangular random matrices*, Journal of Multivariate Analysis, 111 (2012), pp. 120–135.

[12] L. Benigni and S. Péché, *Eigenvalue distribution of some nonlinear models of random matrices*, Electronic Journal of Probability, 26 (2021), pp. 1–37.

[13] N. Berestycki, *Recent progress in coalescent theory*, arXiv preprint arXiv:0909.3985, (2009).

[14] J. Bertoin, *Random fragmentation and coagulation processes*, vol. 102, Cambridge University Press, 2006.

[15] P. Billingsley, *On the distribution of large prime divisors*, Periodica Mathematica Hungarica, 2 (1972), pp. 283–289.

[16] D. M. Blei, T. L. Griffiths, and M. I. Jordan, *The nested Chinese Restaurant Process and Bayesian nonparametric inference of topic hierarchies*, Journal of the ACM (JACM), 57 (2010), pp. 1–30.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent Dirichlet allocation*, Journal of machine Learning research, 3 (2003), pp. 993–1022.

[18] J.-P. Bouchaud and M. Mézard, *Universality classes for extreme-value statistics*, Journal of Physics A: Mathematical and General, 30 (1997), p. 7997.

[19] J. Bun, J.-P. Bouchaud, and M. Potters, *Cleaning large correlation matrices: tools from random matrix theory*, Physics Reports, 666 (2017), pp. 1–109.

[20] P. Charbonneau, J. Kurchan, G. Parisi, P. Urbani, and F. Zamponi, *Fractal free energy landscapes in structural glasses*, Nature communications, 5 (2014), p. 3725.

[21] S. Chatterjee, *A note about the uniform distribution on the intersection of a simplex and a sphere*, Journal of Topology and Analysis, 9 (2017), pp. 717–738.

[22] P. Comon, *Independent component analysis, a new concept?*, Signal processing, 36 (1994), pp. 287–314.

[23] R. Couillet, F. Chatelain, and N. Le Bihan, *Two-way kernel matrix puncturing: towards resource-efficient PCA and spectral clustering*, in International Conference on Machine Learning, PMLR, 2021, pp. 2156–2165.

[24] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*, Cambridge University Press, 2011.

[25] R. Couillet, M. Debbah, and J. W. Silverstein, *A deterministic equivalent for the analysis of correlated MIMO multiple access channels*, IEEE Transactions on Information Theory, 57 (2011), pp. 3493–3514.

[26] R. Couillet and W. Hachem, *Fluctuations of spiked random matrix models and failure diagnosis in sensor networks*, IEEE Transactions on Information Theory, 59 (2012), pp. 509–525.

[27] R. Couillet and Z. Liao, *Random matrix methods for machine learning*, Cambridge University Press, 2022.

[28] H. Crane, *The ubiquitous Ewens sampling formula*, (2016).

[29] A. Crisanti and H.-J. Sommers, *The spherical p-spin interaction spin glass model: the statics*, Zeitschrift für Physik B Condensed Matter, 87 (1992), pp. 341–354.

[30] B. Derrida, *From random walks to spin glasses*, Physica D: Nonlinear Phenomena, 107 (1997), pp. 186–198.

[31] B. Derrida and H. Flyvbjerg, *Statistical properties of randomly broken objects and of multivalley structures in disordered systems*, Journal of Physics A: Mathematical and General, 20 (1987), p. 5273.

[32] B. Derrida and E. Gardner, *Solution of the generalised random energy model*, Journal of Physics C: Solid State Physics, 19 (1986), p. 2253.

[33] B. Derrida and P. Mottishaw, *Finite size corrections in the random energy model and the replica approach*, Journal of Statistical Mechanics: Theory and Experiment, 2015 (2015), p. P01021.

[34] O. Dhifallah and Y. M. Lu, *A precise performance analysis of learning with random features*, arXiv preprint arXiv:2008.11904, (2020).

[35] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, *Phase retrieval via linear programming: Fundamental limits and algorithmic improvements*, in 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2017, pp. 1071–1077.

[36] P. Diaconis and D. Freedman, *A dozen de Finetti-style results in search of a theory*, in Annales de l'IHP Probabilités et statistiques, vol. 23, 1987, pp. 397–423.

[37] P. Donnelly and G. Grimmett, *On the asymptotic distribution of large prime factors*, Journal of the London Mathematical Society, 2 (1993), pp. 395–404.

[38] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, *On robust regression with high-dimensional predictors*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 14557–14562.

[39] R. S. Ellis, H. Touchette, and B. Turkington, *Thermodynamic versus statistical nonequivalence of ensembles for the mean-field blume–emery–griffiths model*, Physica A: Statistical Mechanics and its Applications, 335 (2004), pp. 518–538.

[40] B. Fristedt, *The structure of random partitions of large integers*, Transactions of the American Mathematical Society, 337 (1993), pp. 703–735.

[41] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, *The gaussian equivalence of generative models for learning with shallow neural networks*, in Mathematical and Scientific Machine Learning, PMLR, 2022, pp. 426–471.

[42] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, *Modeling the influence of data structure on learning in neural networks: The hidden manifold model*, Physical Review X, 10 (2020), p. 041044.

[43] T. Griffiths, M. Jordan, J. Tenenbaum, and D. Blei, *Hierarchical topic models and the nested Chinese restaurant process*, Advances in neural information processing systems, 16 (2003).

[44] T. L. Griffiths and Z. Ghahramani, *The Indian Buffet Process: An introduction and review.*, Journal of Machine Learning Research, 12 (2011).

[45] F. Guerra and F. L. Toninelli, *The thermodynamic limit in mean field spin glass models*, Communications in Mathematical Physics, 230 (2002), pp. 71–79.

[46] D. Guo, S. Shamai, and S. Verdú, *Mutual information and minimum mean-square error in Gaussian channels*, IEEE transactions on information theory, 51 (2005), pp. 1261–1282.

[47] D. Guo, Y. Wu, S. S. Shitz, and S. Verdú, *Estimation in Gaussian noise: Properties of the minimum mean-square error*, IEEE Transactions on Information Theory, 57 (2011), pp. 2371–2385.

[48] W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, and L. Pastur, *A new approach for mutual information analysis of large dimensional multi-antenna channels*, IEEE Transactions on Information Theory, 54 (2008), pp. 3987–4004.

[49] W. Hachem, P. Loubaton, and J. Najim, *Deterministic equivalents for certain functionals of large random matrices*, (2007).

[50] ——, *A CLT for information-theoretic statistics of gram random matrices with a given variance profile*, (2008).

[51] K. Handa, *The two-parameter Poisson–Dirichlet point process*, (2009).

[52] A. Hyvärinen and E. Oja, *Independent component analysis: algorithms and applications*, Neural networks, 13 (2000), pp. 411–430.

[53] A. Jagannath, P. Lopatto, and L. Miolane, *Statistical thresholds for tensor PCA*, (2020).

[54] A. Kammoun and M.-S. AlouiniFellow, *On the precise error analysis of support vector machines*, IEEE Open Journal of Signal Processing, 2 (2021), pp. 99–118.

[55] N. E. Karoui, *Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results*, arXiv preprint arXiv:1311.2445, (2013).

[56] F. Krzakala and L. Zdeborová, *Statistical physics methods in optimization and machine learning*, 2021.

[57] H. Lebeau, R. Couillet, and F. Chatelain, *A random matrix analysis of data stream clustering: coping with limited memory resources*, in International Conference on Machine Learning, PMLR, 2022, pp. 12253–12281.

[58] M. Lelarge and L. Miolane, *Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting*, in 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), IEEE, 2019, pp. 639–643.

[59] ———, *Fundamental limits of symmetric low-rank matrix estimation*, Probability Theory and Related Fields, 173 (2019), pp. 859–929.

[60] T. Lesieur, F. Krzakala, and L. Zdeborová, *Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications*, Journal of Statistical Mechanics: Theory and Experiment, 2017 (2017), p. 073403.

[61] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová, *Statistical and computational phase transitions in spiked tensor estimation*, in 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 511–515.

[62] J. Lin, *On the Dirichlet distribution*, Department of Mathematics and Statistics, Queens University, (2016), pp. 10–11.

[63] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, *Learning Gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions*, Advances in Neural Information Processing Systems, 34 (2021), pp. 10144–10157.

[64] X. Mai and R. Couillet, *A random matrix analysis and improvement of semi-supervised learning for large dimensional data*, The Journal of Machine Learning Research, 19 (2018), pp. 3074–3100.

[65] ———, *Consistent semi-supervised graph regularization for high dimensional data*, The Journal of Machine Learning Research, 22 (2021), pp. 4181–4228.

[66] M. L. Mehta, *Random matrices*, Elsevier, 2004.

[67] M. Mézard and G. Parisi, *Replicas and optimization*, Journal de Physique Lettres, 46 (1985), pp. 771–778.

[68] M. Mézard, G. Parisi, N. Sourlas, G. Toulouse, and M. Virasoro, *Replica symmetry breaking and the nature of the spin glass phase*, Journal de Physique, 45 (1984), pp. 843–854.

[69] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, vol. 9, World Scientific Publishing Company, 1987.

[70] M. Mézard and M. A. Virasoro, *The microstructure of ultrametricity*, Journal de Physique, 46 (1985), pp. 1293–1307.

[71] F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborova, *The role of regularization in classification of high-dimensional noisy Gaussian mixture*, in International conference on machine learning, PMLR, 2020, pp. 6874–6883.

[72] L. Miolane, *Fundamental limits of low-rank matrix estimation: the non-symmetric case*, arXiv preprint arXiv:1702.00473, (2017).

[73] A. Montanari and S. Sen, *A short tutorial on mean-field spin glass techniques for non-physicists*, arXiv preprint arXiv:2204.02909, (2022).

[74] A. L. Moustakas and S. H. Simon, *On the outage capacity of correlated multiple-path MIMO channels*, IEEE Transactions on Information Theory, 53 (2007), pp. 3887–3903.

[75] A. L. Moustakas, S. H. Simon, and A. M. Sengupta, *MIMO capacity through correlated channels in the presence of correlated interferers and noise: A (not so) large N analysis*, IEEE Transactions on Information Theory, 49 (2003), pp. 2545–2561.

[76] M.-T. Nguyen, *Derivatives of mutual information in Gaussian channels*, arXiv preprint arXiv:2303.02500, (2023).

[77] ———, *Formal construction of some exchangeable structures*, arXiv preprint arXiv:2311.05002, (2023).

[78] M.-T. Nguyen and R. Couillet, *Asymptotic Bayes risk of semi-supervised multitask learning on Gaussian mixture*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 5063–5078.

[79] D. Panchenko, *The Parisi ultrametricity conjecture*, Annals of Mathematics, (2013), pp. 383–393.

[80] B. R. Paredes, A. Argyriou, N. Berthouze, and M. Pontil, *Exploiting unrelated tasks in multi-task learning*, in Artificial intelligence and statistics, PMLR, 2012, pp. 951–959.

[81] G. Parisi, *Toward a mean field theory for spin glasses*, Physics Letters A, 73 (1979), pp. 203–205.

[82] ———, *A sequence of approximated solutions to the SK model for spin glasses*, Journal of Physics A: Mathematical and General, 13 (1980), p. L115.

[83] ———, *Nobel lecture: Multiple equilibria*, Reviews of Modern Physics, 95 (2023), p. 030501.

[84] J. Pennington and P. Worah, *Nonlinear random matrix theory for deep learning*, Advances in neural information processing systems, 30 (2017).

[85] M. Perman, J. Pitman, and M. Yor, *Size-biased sampling of Poisson point processes and excursions*, Probability Theory and Related Fields, 92 (1992), pp. 21–39.

[86] J. Pitman, *Coalescents with multiple collisions*, Annals of Probability, (1999), pp. 1870–1902.

[87] ———, *Combinatorial stochastic processes: Ecole d'eté de probabilités de Saint-Flour XXXII-2002*, Springer, 2006.

[88] J. Pitman and M. Yor, *The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator*, The Annals of Probability, (1997), pp. 855–900.

[89] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*, Cambridge University Press, 2020.

[90] J. K. Pritchard, M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data*, Genetics, 155 (2000), pp. 945–959.

[91] O. Rioul, W. Cheng, and S. Guilley, *Cumulant expansion of mutual information for quantifying leakage of a protected secret*, in 2021 IEEE International Symposium on Information Theory (ISIT), IEEE, 2021, pp. 2596–2601.

[92] S. Ruder, *An overview of multi-task learning in deep neural networks*, arXiv preprint arXiv:1706.05098, (2017).

[93] F. Salehi, E. Abbasi, and B. Hassibi, *A precise analysis of phasemax in phase retrieval*, in 2018 IEEE International Symposium on Information Theory (ISIT), IEEE, 2018, pp. 976–980.

[94] D. Schröder, H. Cui, D. Dmitriev, and B. Loureiro, *Deterministic equivalent and error universality of deep random features learning*, arXiv preprint arXiv:2302.00401, (2023).

[95] D. Sherrington and S. Kirkpatrick, *Solvable model of a spin-glass*, Physical review letters, 35 (1975), p. 1792.

[96] D. Song and A. Gupta, *l^p-norm uniform distribution*, Proceedings of the American Mathematical Society, 125 (1997), pp. 595–601.

[97] T. Speed, *Cumulants and partition lattices 1*, Australian Journal of Statistics, 25 (1983), pp. 378–388.

[98] R. Speicher and C. Vargas, *Free deterministic equivalents, rectangular random matrix models, and operator-valued free probability theory*, Random Matrices: Theory and Applications, 1 (2012), p. 1150008.

[99] M. Talagrand, *The Parisi formula*, Annals of mathematics, (2006), pp. 221–263.

[100] ———, *Large deviations, Guerra's and ASS schemes, and the Parisi hypothesis.*, Journal of Statistical Physics, 126 (2007).

[101] T. Tao, *Topics in random matrix theory*, vol. 132, American Mathematical Society, 2023.

[102] Y. Teh, M. Jordan, M. Beal, and D. Blei, *Sharing clusters among related groups: Hierarchical Dirichlet Processes*, Advances in neural information processing systems, 17 (2004).

[103] C. Thrampoulidis and B. Hassibi, *Isotropically random orthogonal matrices: Performance of LASSO and minimum conic singular values*, in 2015 IEEE International Symposium on Information Theory (ISIT), IEEE, 2015, pp. 556–560.

[104] C. Thrampoulidis, S. Oymak, and B. Hassibi, *Recovering structured signals in noise: Least-squares meets compressed sensing*, in Compressed Sensing and its Applications: MATHEON Workshop 2013, Springer, 2015, pp. 97–141.

[105] ——, *Regularized linear regression: A precise analysis of the estimation error*, in Conference on Learning Theory, PMLR, 2015, pp. 1683–1709.

[106] C. Thrampoulidis, S. Oymak, and M. Soltanolkotabi, *Theoretical insights into multiclass classification: A high-dimensional asymptotic view*, Advances in Neural Information Processing Systems, 33 (2020), pp. 8907–8920.

[107] M. Tiomoko, R. Couillet, and F. Pascal, *PCA-based multi-task learning: a random matrix approach*, in International Conference on Machine Learning, PMLR, 2023, pp. 34280–34300.

[108] M. Tiomoko, H. Tiomoko, and R. Couillet, *Deciphering and optimizing multi-task learning: a random matrix approach*, in ICLR 2021-9th International Conference on Learning Representations, 2021.

[109] H. Touchette, *Equivalence and nonequivalence of ensembles: Thermodynamic, macrostate, and measure levels*, Journal of Statistical Physics, 159 (2015), pp. 987–1016.

[110] A. M. Tulino, S. Verdú, et al., *Random matrix theory and wireless communications*, Foundations and Trends® in Communications and Information Theory, 1 (2004), pp. 1–182.

[111] C. Vargas, *A general solution to (free) deterministic equivalents*, Contributions of Mexican Mathematicians Abroad in Pure and Applied Mathematics, 709 (2018), p. 131.

[112] T. Zarrouk, R. Couillet, F. Chatelain, and N. Le Bihan, *Performance-complexity trade-off in large dimensional statistics*, in 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2020, pp. 1–6.

[113] L. Zdeborová and F. Krzakala, *Statistical physics of inference: Thresholds and algorithms*, Advances in Physics, 65 (2016), pp. 453–552.