Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics
Department of Electrical Engineering

Master Thesis Winter 2024-2025

Victor Kawasaki-Borruat

Sampling for Statistical Physics with Diffusion Processes

Leveraging Disorder-Dependent Diffusions to Access Low-Temperature Gibbs Measures

Submission Date: 3 March 2025

Advisor D-ITET: Prof. Dr. Hans-Andrea Loeliger
Advisor D-MATH: Prof. Dr. Yuansi Chen

To my past self who failed first semester at EPFL, and doubted himself more than ever.
Look at us now, we graduated.

Preface

Jack of all trades, master of none. A feeling which has often confused, questioned and even haunted my identity has now ironically, become my strength. This thesis has somehow perfectly captured this representation of my academic journey; a lot of mathematics, some computer science, a dash of theoretical physics, all put elegantly together in a tastefully beautiful problem.

Thank you to Prof. Yuansi Chen for giving me the just the right amount of supervision, and finding a topic which combined not only my strengths, but also my weaknesses. With your guidance, I was not only able to learn how to complete a Master thesis, but also something about myself. You pushed me to be more independent, and I cannot thank you enough for the time I got to spend working with you. I would also like to thank Prof. Amos Lapidoth, for advising me throughout my Master studies, perhaps in a more indirect way. Your advice, along with overly strong coffee, were and always will be welcome.

Thank you to Prof. Loeliger, who, perhaps unbeknownst to him, inspired me to study Statistical Mechanics and its ties to Information Theory. Attending the AECC course in my Bachelor's sparked my initial interest, and now here I am – full circle. Back to graphical models, albeit with a twist.

Finally, thank you to my family for supporting me from the (literal) beginning and throughout what only seems to be the start of an exciting journey. Thank you to my friends who were there before, and those I've met along the way. Thank you to my wonderful partner, for putting up with my odd work habits and believing in me.
I seriously could not have done it without any of you. You are all a source of inspiration, and I am proud to say that you are a part of my life.

Victor Kawasaki-Borruat

Abstract

Sampling from low-temperature spin glass Gibbs measures is computationally hard beyond the tri-critical point β_{tri}, but is expected to be feasible up to the dynamical threshold $\beta_{\text{d}} \geq \beta_{\text{tri}}$. Standard diffusion-based methods fail due to the emergence of metastable states in the free energy Φ_{RS} along the Approximate Message Passing (AMP) algorithm trajectory, leading to faulty estimation of the Gibbs mean. In this thesis, we propose disorder-dependent diffusions (DDDs), a natural anisotropic extension of stochastic localization that incorporates the disorder structure $\mathbf{A} \in \mathbb{R}^{n \times n}$ of the spin glass. We prove that DDDs allow AMP to have a controllable trajectory in phase space, offering the potential to circumvent both metastable and impossible regions. We analyse the theoretical guarantees of DDDs through a rigorous proof of the AMP state evolution and convergence for the planted Sherrington-Kirkpatrick model, generalizing the AMP convergence proofs of Alaoui, Montanari, and Sellke (2024) and Montanari and Wu (2024) to account for anisotropic side information.

We also supplement this result with numerical applications of DDDs to solve sparse rank-one matrix factorization problems, implying the capacity to sample from sparse SK models. We provide strong numerical evidence demonstrating the advantages of DDDs in both estimation and sampling for the sparse case.

Contents

List of Figures

Notation

Note 1
This thesis will presume an undergraduate-level understanding of Mathematics from the reader (mostly in Probability Theory, Statistics, Linear Algebra and Stochastic Processes). While we try to be as comprehensive as possible in introducing concepts, some definitions will be assumed to be known; in which case we will provide a blue link to a definition.

Random Variables, Vectors and Matrices

- □ Capital letters $X \in \mathbb{R}$ denote a one-dimensional random variable (RV)
- □ Bold lowercase letters $\mathbf{x}$ denote (random) vectors of suitable dimensions
- □ Bold uppercase letters $\mathbf{X}$ denote (random) matrices of suitable dimensions
- □ As the unique exception to the above, we will denote vector-valued Brownian Motion with bold uppercase letters $\mathbf{B}_t$, along with a time index t
- □ For a vector $\mathbf{x} \in \mathbb{R}^n$ or matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, we denote its transpose by $\mathbf{x}^T \in \mathbb{R}^{1 \times n}$ resp. $\mathbf{X}^T \in \mathbb{R}^{m \times n}$
- □ For a random variable X, and a probability distribution $\mu : \mathbb{R} \to [0,1]$, $X \sim \mu$ denotes that X is distributed according to μ
- □ We use usual notation for expected value $\mathbb{E}[\cdot]$, variance $\text{Var}(\cdot)$ and covariance $\text{Cov}(\cdot)$
- □ We will sometimes subscript $\mathbb{E}$ with a random variable or distribution, if not clear from context
- □ We say a random vector $\mathbf{x}$ is independent and identically distributed (i.i.d.) if all elements are independent and have the same distribution
- □ We denote the n-dimensional null vector by $\mathbf{0}_n$, the all-ones vector by $\mathbf{1}_n$, and the identity matrix by $\mathbf{I}_n$
- □ Letters like X, Y, R, S will typically denote signal values, whereas Z, W, G will typically denote (Gaussian) noise values. The same goes for vectors and matrices.
- □ The notation $\mathcal{N}(\mu, \sigma^2)$ will be reserved for Gaussian RVs, of mean μ and variance $\sigma^2 > 0$. This also extends to vectors, with mean vectors and covariance matrices.
- □ $\bar{\nu}$ will denote the uniform distribution on $\{+1, -1\}$. We will interchangeably call it the *symmetric Bernoulli* or *Rademacher* distribution

Iterations, Indexing

- A subscript on an RV X_n denote an indexed random variable, typically from a sequence $\{X_n\}_{n \in \mathcal{I}}$, for an index set $\mathcal{I} \subseteq \mathbb{N}$
- We index vectors with superscripts or subscripts, depending on context. For example, superscripts in parentheses, i.e. $\mathbf{x}^{(k)}, \mathbf{x}^{(k+1)}$, will be used to denote vector-valued iterations of an algorithm, whereas $\mathbf{x}_t$ will typically denote a time-indexed vector-valued stochastic process
- We denote the i-th component of a vector $\mathbf{x}$ by dropping the boldface and subscripting: $\mathbf{x}[i] = x_i$
- We denote the (i, j)-th entry of a matrix $\mathbf{X}$ by dropping the boldface and subscripting: $\mathbf{X}[i, j] = X_{i,j}$

Stochastic Convergence

- $X_n \overset{d}{=} X$ denotes convergence in distribution for any kind of random object X
- $X_n \overset{a.s.}{=} X$ denotes almost sure convergence for any kind of random object X. This is analogous to
- We use the notation $\text{p-lim}_{n \to \infty} X_n = X$ to denote convergence in probability

Linear Algebra

- $\mathbb{S}_+^n$ is the cone of $n \times n$ positive semi-definite matrices
- $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product
- $\| \cdot \|_p$ denotes the ℓ_p-norm
- For $\mathbf{Q} \in \mathbb{S}_+^n$ and $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_{\mathbf{Q}}^2 = \langle \mathbf{x}, \mathbf{Q}\mathbf{x} \rangle$ is a skewed norm

Chapter 1

Introduction

How does one generate synthetic data? This fundamental question lies at the heart of generative modeling, a rapidly advancing field in machine learning and artificial intelligence (AI). Much like how people “learn” social cues from experience, generative models aim to learn complex probability distributions from data, enabling the generation of new samples that mimic the original dataset. For instance, AI-generated images of faces are not those of real people, yet they feel plausible. They accurately emulate the “distribution” of all human faces with high fidelity.

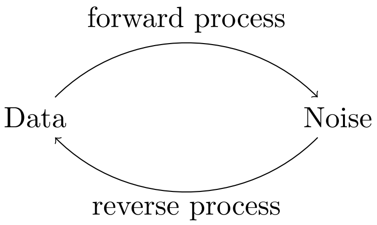


Figure 1.1: The simplistic idea of diffusion-based generative modeling.

The use of *diffusion processes* for sampling gained significant attention in the context of generative modeling with the work of Sohl-Dickstein, Weiss, Maheswaranathan, and Ganguli (2015). In their pioneering paper, they introduced a framework known as diffusion denoising probabilistic models (DDPMs), which leverages the reverse-time dynamics of diffusion processes to generate samples from complex probability distributions. The approach begins with a forward diffusion process, progressively adding noise to the data until it becomes indistinguishable from a Gaussian distribution. By learning the reverse diffusion process via neural networks, they demonstrated how one could iteratively remove the noise from the system to recover samples from the original data distribution (see Figure 1.1). DDPMs were later popularised by Ho, Jain, and Abbeel (2020), and the framework was further extended to fit the mathematical setting of stochastic differential equations (SDEs) in Song, Sohl-Dickstein, Kingma, Kumar, Ermon, and Poole (2021). These models, while empirically successful, are primarily understood through an applied lens, leaving many theoretical aspects unexplored. In particular, the theoretical limitations of diffusion-based sampling methods remain unclear. Understanding these limitations is crucial, not only for improving existing methods but also for guiding the development of principled alternatives. Recent advances by Chen, Chewi, Li, Li, Salim, and Zhang (2023) show that learning this time-reversed process simply reduces to estimating the log-likelihood of the forward law (or score), a well-known problem to statisticians.

This thesis will mostly consist of an in-depth review of this relatively new topic and present a theoretical analysis of diffusion-based sampling methods, focusing on their performance when restricted to a specific class of parametric probability measures: Gibbs measures. Gibbs measures are central objects in the statistical mechanics of disordered systems, representing the equilibrium distribution of all possible configurations. In particular, we will study Gibbs measures of spin glasses, which were popularized by Mezard, Parisi, and Virasoro (1987) to physicists, and to mathematicians by Talagrand (2011)[1].

Spin glasses are a class of disordered systems where the interactions between components (spins) are governed by a random energy landscape. To each possible configuration $\mathbf{x} = (x_1, \ldots, x_n)$, with each spin $x_i \in \{-1, +1\}$, we associate the probability of such a state. This is the Gibbs measure. It is parameterized by an inverse temperature $\beta > 0$, which controls the influence of the system's energy $H(\mathbf{x})$ on the probability of observing a given configuration $\mathbf{x}$:

$$\mu_\beta(\mathbf{x}) = \frac{1}{Z_\beta} \exp\left(-\beta H(\mathbf{x})\right),$$

where Z_β is a normalization constant. As the inverse temperature β increases, spin glasses undergo phase transitions that fundamentally alter their energy landscapes, transitioning from a simple homogeneous phase to a "glassy phase" with many local minima.

These phase transitions have profound implications for sampling methods. Beyond the dynamical threshold β_d, the energy landscape becomes highly non-convex, making sampling computationally hard for any algorithm (Alaoui, Montanari, and Sellke (2023)). Classical sampling techniques, such as Markov Chain Monte Carlo (MCMC), have been extensively studied in this context. Physicists have predicted fast mixing for MCMC algorithms (Sompolinsky and Zippelius (1981); Mezard et al. (1987)), but rigorous mathematical guarantees have proven elusive, with polynomial convergence established only for vanishing temperature intervals (Aizenman and Holley (1987)), or less than half of the expected regime (Bauerschmidt and Bodineau (2019)). These challenges motivate the exploration of new approaches, such as diffusion-based sampling, which may offer novel insights into tackling or bypassing this computationally challenging regime.

[1] Incidentally, Parisi won the Physics Nobel Prize in 2021 and Talagrand the Abel Prize in 2024!

Contributions

In this thesis we will focus on the limitations of diffusion-based sampling methods, and offer a potential improvement by incorporating the disorder in the diffusion process' driving Brownian motion. While we only treat the case of the Sherrington-Kirkpatrick (SK) model (which is already solved), we lay the theoretical foundation of disorder-dependent diffusions, as well as offer promising numerical simulations for the problem of sparse rank-one matrix factorization (Chapter 5).

The main technical challenge lies in a subroutine of the sampling algorithm, which is done via Approximate Message Passing (AMP). Due to the disorder dependency in the driving noise, existing AMP results from Alaoui et al. (2024); Montanari and Wu (2024) fail to accurately predict our case, as they only rely on isotropic diffusions. To remedy this, we first propose a non-rigorous but intuitive characterization of the disorder-dependent State Evolution using tools from Statistical Physics (Section 4.4), and follow up with a rigorous statement on the State Evolution of our Disorder-Dependent AMP in Theorem 4.4.1.

Thesis Outline

This thesis aims to provide a rigorous mathematical framework for solving the $\mathbb{Z}_2$-synchronization problem (optimizing the SK model) at low signal-to-noise-ratio (SNR) with anisotropic side-information, thus implying a diffusion-based sampling algorithm. We do so by presenting the current state-of-the-art research in the topic, while building upon it by incorporating the anisotropic diffusion. Due to the multiple sources of the problem (High-Dimensional Statistics, Compressed Sensing, Statistical Physics, etc..), the problem itself is a challenge to explain in linear fashion.

This work is divided into five chapters which are meant to be read sequentially, but with frequent cross-references. Each to reinforce pre-established concepts, foreshadow important ideas, and draw parallels between topics. Chapter 2 introduces the idea of sampling general probability measures using diffusion processes. Chapter 3 first discusses the key objects of interest – spin glasses – and then ties back into Chapter 2 by explaining how to sample from Gibbs measures with diffusions. We then present Approximate Message Passing as the main subroutine of the sampling algorithm, and highlight its limitations. Finally, we propose disorder-dependent diffusions as a solution to such limitations in Section 3.4, as well as suggest the associated AMP algorithm. The main result for disorder-dependent AMP lies in Theorem 3.4.1, which states that AMP accurately computes the required Gibbs mean. Chapter 4 derives the AMP algorithm from Variational Inference principles (supported by Statistical Physics derivations from Chapter 3) and presents the analysis of both AMP algorithms (isotropic and anisotropic) simultaneously. This culminates in the characterization of both AMP algorithms' State Evolutions in Theorem 4.4.1, allowing the proof of Theorem 3.4.1. Finally, Chapter 5 extends the work on the SK model to its sparse planted counterpart: $\mathbb{Z}_2$-synchronization, along with many promising numerical results.

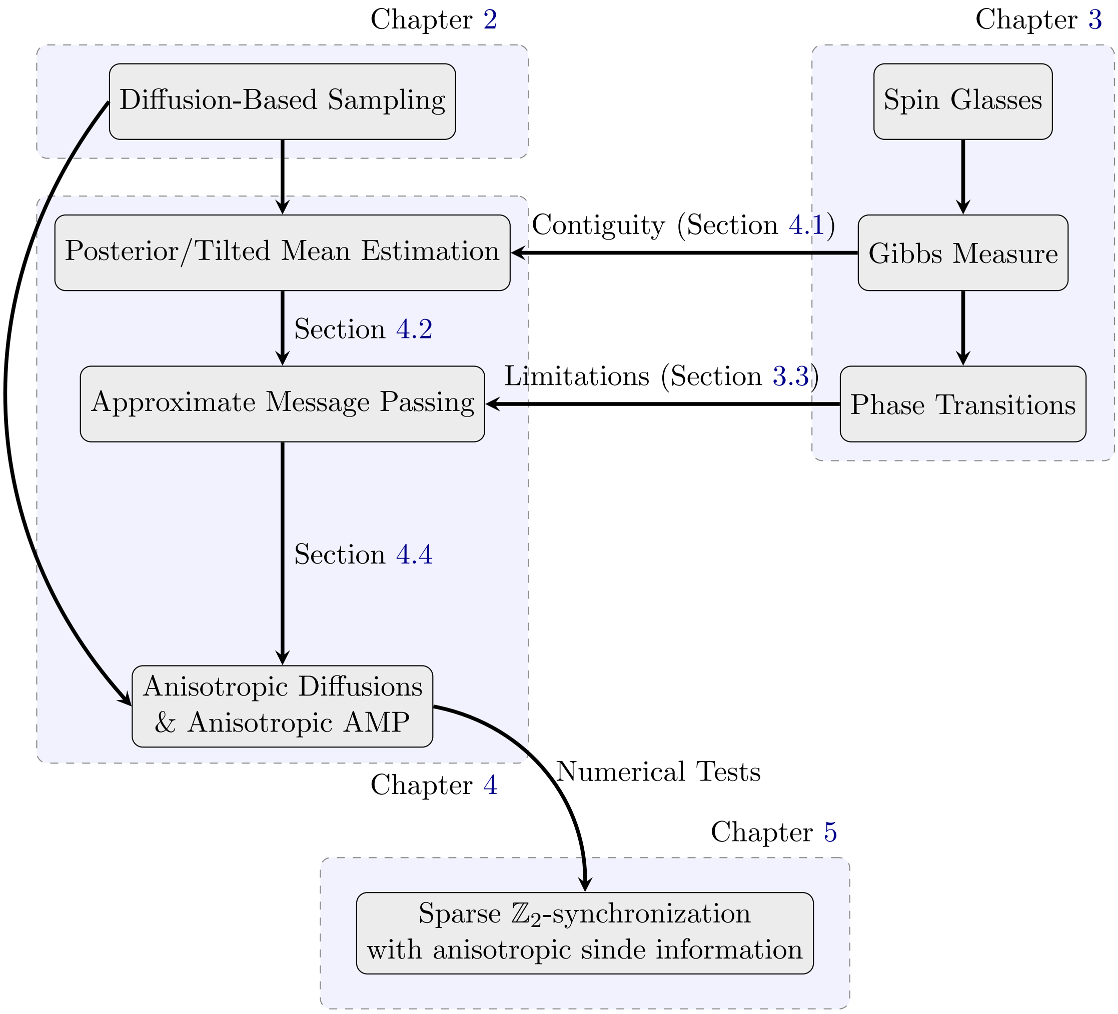


Figure 1.2: Dependencies and ties between topics presented in the thesis.

Chapter 2

Sampling with Diffusion Processes

This chapter will present the overarching topic of the thesis. Namely, we explore how diffusion processes can be reversed to create sampling algorithms.

Section 2.1 will first present the basics of diffusion processes, as well as how their time-reversals. In Section 2.2, we explain using Stochastic Localization (pioneered by Eldan (2013)) how the reverse stochastic differential equation (SDE) successfully produces a sample (recall Figure 1.1 for motivation). We then present a basic sampling algorithm using diffusion processes in Section 2.3.

2.1 Diffusion Processes

Diffusion processes are a class of stochastic processes that model the random motion of particles in a medium, governed by a particular differential equation. Mathematically, they are described by Itô or Stratonovich (but we will only discuss Itô) SDEs, where the dynamics are driven by a deterministic drift term and random volatility term, modeled as Brownian Motion. These correspond respectively to f and g in Eq. (2.1.1) below.

We begin by laying some fundamentals. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, $(\mathcal{F}_t)_{0\leq t\leq T}$ be a non-decreasing family of sub-σ algebras, and $\mathbf{B} = (\mathbf{B}_t, \mathcal{F}_t)$ be a Brownian Motion.

Definition 2.1.1 (Diffusion Process)
A (vector-valued) stochastic process $(\mathbf{x}_t)_{t\geq 0}$ is called a diffusion process *or* process of the diffusion type[1] *relative to a Brownian Motion $(\mathbf{B}_t)_{t\geq 0}$ if it expressible as*

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t f(s,\omega)\mathrm{d}s + \int_0^t g(s,\omega)\mathrm{d}\mathbf{B}_s, \quad \mathbf{x}_0 \sim \mu. \tag{2.1.1}$$

As per definition of Itô processes, we may also express the above in SDE form

$$\mathrm{d}\mathbf{x}_t = f(t,\omega) + g(t,\omega)\mathrm{d}\mathbf{B}_t. \tag{2.1.2}$$

Finally, for an initial disitrbution $\mathbf{x}_0 \sim \mu$*, we will denote the* law of the process at time *t as $\mu_t := \mathcal{L}(\mathbf{x}_t)$. It is the distribution of the random variable associated to the current timestep t.*

Forward Diffusion Process

Let μ be a probability distribution on $\mathbb{R}^n$, which we wish to sample from. The role of the forward process will be to turn our initial sample $\mathbf{x}_0 \sim \mu$ to Gaussian noise, by injecting more noise at each timestep. Namely, we will consider the Ornstein-Uhlenbeck (OU) process, which is famously known to accomplish this task exponentially fast (Bakry, Gentil, and Ledoux (2013)) under various divergence metrics:

$$\mathrm{d}\mathbf{x}_t = -\mathbf{x}_t\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \tag{2.1.3}$$

where $\mathbf{B}_t$ denotes standard Brownian Motion.

It may seem counterintuitive to describe a process taking us *away* from the desired distribution μ, but we recall that

i.) We presumably do not know μ or have access to an oracle producing its samples (hence we do not have $\mathbf{x}_0$).

ii.) We will reverse this process, allowing for the sampling procedure.

Backward Diffusion Process

A standard result from the theory of stochastic processes by Anderson (1982) shows that such a process can be *time-reversed*. Setting a stopping time $T > 0$ allows us to define a process with reversed time-indices:

$$\mathbf{x}^{\leftarrow}_t := \mathbf{x}_{T-t}, \quad \text{for } t \in [0,T]. \tag{2.1.4}$$

This process also admits an SDE description, given by

$$\mathrm{d}\mathbf{x}^{\leftarrow}_t := \left(\mathbf{x}^{\leftarrow}_t + 2\nabla \ln\left(\mu_{T-t}(\mathbf{x}^{\leftarrow}_t)\right)\right)\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \quad \mathbf{x}^{\leftarrow}_0 \sim \mu_T,$$

where $\mathbf{B}_t$ is the reversed Brownian motion. Since we will only focus on the reverse, we keep the notation simple to avoid overloading. As we can see, the reverse process only depends on the score of the forward law $\nabla \ln(\mu_{T-t})$. The following result is key to computing this score analytically.

[1] We use Definition 7 of Liptser and Shiryayev (1977), which also requires that both drift and volatility be $\mathcal{F}^{\mathbf{x}}_t$-measurable. Since this lies on the more measure-theoretic side, and is not relevant to the rest of the thesis, we do not go into the details of such definitions.

Proposition 2.1.1 (Tweedie's Formula, Efron (2011))
Let X be a random variable, and $Y = X + Z$ be a noisy observation, with $Z \sim \mathcal{N}(0, \gamma^2)$. Then,

$$\mathbb{E}[X|Y = y] = y + \gamma^2 \nabla_y \ln(p_Y(y)), \tag{2.1.5}$$

where p_Y denotes the marginal law of Y.

Proof. Let $X \sim \mu$, and $Y = X + Z$ as above. Then the marginal law of Y is

$$p_Y(y) = \int \mu(x) \frac{1}{\sqrt{2\pi}\gamma} \exp\left(-\frac{(y-x)^2}{2\gamma^2}\right) \mathrm{d}x. \tag{2.1.6}$$

Thus,

$$\begin{aligned} \frac{1}{\gamma^2}(\mathbb{E}[X|Y = y] - y) &= \mathbb{E}\left[\frac{X - Y}{\gamma^2}\middle| Y = y\right] \\ &= \int \frac{x-y}{\gamma^2} p_{X|Y=y}(x) \mathrm{d}x \\ &= \int \frac{x-y}{\gamma^2} \frac{\mu(x) p_Z(y-x)}{p_Y(y)} \mathrm{d}x \\ &= \int \frac{\partial_y p_Y(y)}{p_Y(y)} \mathrm{d}x = \nabla_y \ln(p_Y(y)) \end{aligned} \tag{2.1.7}$$

□

We integrate the OU process from Eq. (2.1.3) using Itô's formula,

$$\mathbf{x}_t = e^{-t}\mathbf{x}_0 + \sqrt{1 - e^{-2t}}\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_n). \tag{2.1.8}$$

Using Proposition 2.1.1 in Eq. (2.1.8), we obtain

$$\nabla \ln(\mu_t(\mathbf{x}_t)) = \frac{1}{1 - e^{-2t}} \left(\mathbb{E}[e^{-t}\mathbf{x}_0 | \mathbf{x}_t = x_t] - x_t\right). \tag{2.1.9}$$

With an adequate change of variables, we finally obtain the reverse diffusion process, as in Montanari (2023).

$$\mathrm{d}\mathbf{x}_t^{\leftarrow} = \left(-\frac{1+t}{t(1+t)}\mathbf{x}_t^{\leftarrow} + \frac{\mathbf{m}(\sqrt{(1+t)}\mathbf{x}_t^{\leftarrow}; t)}{\sqrt{t(1+t)}}\right)\mathrm{d}t + \frac{\mathrm{d}\mathbf{B}_t}{\sqrt{t(1+t)}}, \tag{2.1.10}$$

where

$$\mathbf{m}(\mathbf{y}; \mathbf{t}) = \mathbb{E}[\mathbf{x} | t\mathbf{x} + \sqrt{t}\mathbf{z} = \mathbf{y}], \quad (\mathbf{x}, \mathbf{z}) \sim \mu \otimes \mathcal{N}(0, \mathbf{I}_n) \tag{2.1.11}$$

is shorthand notation for a Gaussian observation of the initial distribution. After the intuition provided at the beginning of this chapter, simulating the reverse SDE in Eq. (2.1.10) should allow us to produce a sample from μ. However, it seems strange that in the process of trying to create samples, we must compute a mean involving said sample. We explore this further in the following Section.

2.2 Stochastic Localization

Stochastic Localization was initially developed as a proof technique by Eldan (2013), (2019). The main concept is that it allows us to study the properties of probability measure-valued stochastic processes which "localize" as $t \to \infty$, i.e. the random measures will contract to an atom measure $\delta_{\mathbf{x}_*}$. We will later see to what atom measure it will localize, and how to leverage this.

Consider a probability measure μ on $\mathbb{R}^n$, a positive time index $t > 0$ and a vector $\mathbf{y} \in \mathbb{R}^n$ such that it is a noisy observation of a sample $\mathbf{x}_* \sim \mu$:

$$\mathbf{y}(t) := t\mathbf{x}_* + \mathbf{B}_t. \tag{2.2.1}$$

With this, we can write the posterior distribution of $\mathbf{x}_*$ given $\mathbf{y}$ by characterization of Brownian Motion:

$$\begin{aligned}\mathbb{P}(\mathrm{d}\mathbf{x}_* | t\mathbf{x}_* + \mathbf{B}_t = \mathbf{y}) &= \frac{1}{Z'(\mathbf{y},t)}\mu(\mathrm{d}\mathbf{x}_*)\exp\left(-\frac{1}{2t}\|\mathbf{y} - t\mathbf{x}_*\|_2^2\right)\\ &= \frac{1}{Z(\mathbf{y},t)}\mu(\mathrm{d}\mathbf{x}_*)\exp\left(\langle \mathbf{y}, \mathbf{x}_*\rangle - \frac{t}{2}\|\mathbf{x}_*\|_2^2\right),\end{aligned} \tag{2.2.2}$$

where Z is a normalizing constant. We notice that this is a random tilt of μ, and that by Theorem 7.12 Liptser and Shiryayev (1977), we have that the process $(\mathbf{y}(t))_{t \geq 0}$ is the unique solution to the following SDE

$$\mathrm{d}\mathbf{y}(t) = \mathbf{m}(\mathbf{y}(t);t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \quad \mathbf{y}(0) = \mathbf{0}_n. \tag{2.2.3}$$

It is clear by a simple change of variable

$$\mathbf{y}(t) = \sqrt{t(1+t)}\mathbf{x}_t^{\leftarrow}$$

that the processes (2.1.4) and (2.2.1) satisfy the same SDEs. Let $\mu_{\mathbf{y},t}$ denote the the probability distribution in Eq. (2.2.2). We have by properties of stochastic localization that the measure valued process $(\mu_{\mathbf{y}(t),t})_{t \geq 0}$

i.) is a martingale,

ii.) $\mu_{\mathbf{y}(t),t} \xrightarrow{t\to\infty} \delta_{\mathbf{x}_*}$, almost surely, for some random sample $\mathbf{x}_*$ of μ.

Thus, by simulating the SDE in Eq. (2.2.3), the resulting law of the process will "localize" to an atom over a random sample of μ. To access this sample, we notice that as $t \to \infty$,

$$\frac{\mathbf{y}(t)}{t} \approx \mathbf{x}_*,$$

as the variance of the noise will be of order $\mathcal{O}\left(\frac{1}{t}\right)$, making it negligible.

2.3 A General Sampling Algorithm

Now, we may finally put all the above elements together to form a simple sampling algorithm. Such an algorithm was first proposed in Alaoui et al. (2024), and introduced the idea of "Algorithmic Stochastic Localization". This was quickly followed by the same authors in (2023), where the connection to DDPMs was made more apparent. Since then, diffusion-based sampling has been attracting more attention, with a more general framework presented in Montanari and Wu (2024) for posterior sampling in high dimensions.

In order to simulate the reverse SDE (2.2.3), Alaoui et al. (2024) propose two main ideas:

- Discretizing $\mathbf{y}(t)$ to $\widehat{\mathbf{y}}(t)$ with a standard Euler method.
- A subroutine to compute $\mathbf{m}(\widehat{\mathbf{y}}(t); t)$ at each timestep.

We summarize this in the following Algorithm 1 (Algorithm 1, Montanari and Wu (2024)).

Algorithm 1: General Diffusion-Based Sampling Scheme

```
  Input: Parameters $\delta, L > 0$
1 Set $\widehat{\mathbf{y}}(0) = 0$,
2 for $\ell = 0, \dots, L-1$ do
3     Estimate $\mathbf{m}(\mathbf{y}(t); t)$ with $\widehat{\mathbf{m}}(\mathbf{y}(t); t)$
4     Update $\widehat{\mathbf{y}}_{\ell+1} = \widehat{\mathbf{y}}_\ell + \delta \widehat{\mathbf{m}}(\widehat{\mathbf{y}}_\ell; \ell) + \sqrt{\delta}\mathbf{w}_{\ell+1}$
5 end
6 Draw $\{x_i^{\mathrm{alg}}\}_{i \le n}$ conditionally independent with $\mathbb{E}[x_i^{\mathrm{alg}} | \mathbf{y}, \{\mathbf{w}_\ell\}] = \widehat{m}_i(\mathbf{A}, \widehat{\mathbf{y}}_L)$
7 return $\mathbf{x}^{\mathrm{alg}}$
```

While the framework is general, its implementation depends critically on the posterior mean estimation step (Line 3). For example, in the case of spin glasses, this step involves Approximate Message Passing (AMP). In Section 3.3, we address the specific challenges of this step in the context of spin glasses and will develop how AMP provides a principled solution to the estimation problem in Chapter 4.

Chapter 3

Statistical Physics & Sampling

Statistical Physics (SP) was originally conceived as a probabilistic method to infer macroscopic properties of systems involving enormous amounts of interacting microscopic parts; the most common examples being gases or models of magnetism. To avoid computing every single interaction, physicists came up with ingenious methods to show that for large enough amounts of interacting agents, the general behaviour of the overall system can typically be approximated by the mean. This is the "self-averaging" property, and is none other than the concentration of measure phenomenon in mathematics. As computing power has increased, so has the size of datasets, and we are now approaching the large scale regime of statistical mechanics. This has prompted scientists to begin applying concepts of statistical mechanics to data-related problems, such as high-dimensional statistics or statistical learning. It is now our turn to understand the key underlying concepts of SP, to hopefully provide mathematically strong foundations for sampling algorithms.

Many frameworks from SP have provided excellent methods in modeling complex, high-dimensional systems in physics of information (Mezard and Montanari (2009)), statistics (Zdeborová and Krzakala (2016), Maillard (2021)) and communication theory (Tanaka (2002), Montanari and Tse (2006)). As discussed in the introduction, the problem of sampling from such objects' Gibbs measures with diffusions has only recently become of interest. While literature on the topic is still at an early stage, we base ourselves on the analysis done by Ghio et al. (2023) on the limitations of such methods.

We will first introduce the fundamental concepts of spin glasses, and explain the current challenges in the context of sampling. Finally, we introduce a new concept called *disorder-dependent diffusions*, which will be the main contribution of the thesis. We exhibit their potential in Section 3.4 for the sampling of spin glasses. In particular, we will treat the case of sampling from the SK Gibbs measure (see 3.1.1).

3.1 Spin Glasses

The study of spin systems in statistical mechanics began with the introduction of the celebrated Ising model in 1925. Originally conceived to model ferromagnetism, it considered spins $x_i \in \{\pm 1\}^n$ to only interact with their nearest neighbours on a lattice of given geometry. It has since then become the canonical example of statistical physics, as being the foundational model to study phase transitions, critical phenomena, and renormalization.

The Sherrington Kirkpatrick Model

In the 1970s, physicists developed a generalisation of the Ising model, introducing spin glass theory. This theory allows the modeling of more complex behaviour, such as random or frustrated interactions, unconstrained to a lattice. Such assumptions allow the spin glasses to exhibit much more complicated energy landscapes, which characterizes their 'disordered' nature. Below we will present the models of interest: the Sherrington-Kirkpatrick (Sherrington and Kirkpatrick (1975)) model and the p-spin glass (Mezard et al. (1987)). Both models have symmetrical i.i.d. random interactions, and still allow for an external field h to influence the spins in a deterministic manner.

Definition 3.1.1 (SK Model)
For $n \geq 1$, let $\mathbf{x}$ be a point on the n-dimensional Boolean hypercube $\mathcal{C}^n := \{+1, -1\}^n$. We define the ***SK Hamiltonian*** *as*

$$H(\mathbf{x}) := \frac{1}{\sqrt{n}} \sum_{i<j} J_{ij} x_i x_j + h \sum_i x_i, \tag{3.1.1}$$

where the interaction terms $(J_{ij})_{i,j\leq n}$ are i.i.d. distributed as follows:

$$J_{ij} \sim \begin{cases} \mathcal{N}(0,1), & \text{for } i \neq j, \\ \mathcal{N}(0,2), & \text{for } i = j, \end{cases} \tag{3.1.2}$$

Remark 3.1
We can alternatively define the SK Hamiltonian by means of a GOE matrix (see Definition (A.1.2)) $\mathbf{A}$ and an external field vector $\mathbf{h} := h \cdot \mathbf{1}$,

$$H(\mathbf{x}) = \frac{\beta}{2} \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle + \langle \mathbf{h}, \mathbf{x} \rangle. \tag{3.1.3}$$

This is the notation we will mainly use from now on.

Associated to each Hamiltonian is a corresponding probability distribution, allowing us the characterize how likely to be in a given state $\mathbf{x} \in \mathcal{C}^n$ upon observing the energy of the spin glass. A crucial parameter of this distribution is the *inverse temperature* β, which will dictate the phase transitions associated to the complexity of this probability measure.

Definition 3.1.2 (Gibbs Measure)
For a fixed parameter $\beta > 0$ and a given Hamiltonian $H(\mathbf{x})$, we associate to this Hamiltonian a probability measure, called the ***Gibbs measure****, defined as*

$$\mu_{\beta,\mathbf{A}}(\mathrm{d}\mathbf{x}) := \frac{1}{Z(\beta,\mathbf{A})} e^{\beta H_n(\mathbf{x})} \bar{\nu}(\mathrm{d}\mathbf{x}), \tag{3.1.4}$$

where $\bar{\nu}$ is the uniform distribution on $\mathcal{C}^n$, and $Z(\beta,\mathbf{A})$ is a normalizing constant; known in the physics literature as the ***partition function***

$$Z(\beta,\mathbf{A}) := 2^{-n} \sum_{\mathbf{x}\in\mathcal{C}^n} \exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle + \langle \mathbf{h}, \mathbf{x}\rangle\right). \tag{3.1.5}$$

Note that we have absorbed the β-term in the external field h. This can be done w.l.o.g., and avoids notational overload.

The Spase SK Model

Another example of less interest to physicists but of higher statistical significance is the addition of *sparsity constraints.* We present the canonical example, which will serve as the main case of interest in disorder-dependent diffusions.

Definition 3.1.3 (Sparse SK Model)
Exactly as in Definition 3.1.1, the ***sparse SK model*** *with external field $h \in \mathbb{R}$ has the Hamiltonian*

$$H(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i<j}^{n} J_{ij} x_i x_j + h \sum_{i=1}^{n} x_i, \tag{3.1.6}$$

*but $\mathbf{x}$ is constrained to the ρ-**sparse Boolean hypercube** $\mathcal{C}_\rho^n := \left\{\mathbf{x} \in \{+1, 0-1\}^n : \|\mathbf{x}\|_2^2 = \rho n\right\}$ for $\rho \in (0,1)$.*

Higher-Order Spin Glasses

We will also consider further generalisations of spin glasses, where the interactions are not restricted to being twofold. We define these objects in the following.

Definition 3.1.4 (p-spin Glass)
For $n \geq 1$ and $P \geq 2$, let $\mathbf{x} \in \Sigma^n$. The P-spin glass' Hamiltonian is given by

$$H_n^{(P)}(\mathbf{x}) := \sum_{p=2}^{P} \frac{c_p}{n^{(p-1)/2}} \sum_{i_1,\dots,i_p} J_{i_1,\dots i_p} x_{i_1} \dots x_{i_p} + h \sum_i x_i, \tag{3.1.7}$$

where $(c_p)_{p\geq 2}$ has fast decay to zero, e.g. $\sum_p c_p 2^p < \infty$, and $J_{i_1,\ldots i_p}$ are i.i.d. standard Gaussian random variables. Note that c_p controls whether the Hamiltonian considers mixtures of interactions or not. In the case of $c_p = \begin{cases}\Delta, & \text{for a unique } p_*, \\ 0 \text{ otherwise}, \end{cases}$ *we call this a* ***pure*** p_****-spin model****. Otherwise, it is said to be* ***mixed****. The P-spin Gibbs measure and partition function are given analogously to those of the SK model.*

A final important function what we will call the ***centered covariance****:*

$$\xi(t) := \beta^2 \sum_{p=2}^{P} c_p^2 t^p, \tag{3.1.8}$$

as it offers compact notation for the covariance of two Hamiltonian with no external field ($h = 0$):

$$\mathbb{E}\left[H_n^{(P)}(\mathbf{x})H_n^{(P)}(\mathbf{x}')\right] = n\xi\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right). \tag{3.1.9}$$

Brief Introduction to Replica Symmetry

Replica symmetry (RS) was introduced in the 1970s by physicists to study disordered systems; it is a simplifying assumption where all replicas of the system are assumed to have identical statistical properties. The main implication is that we may assume that for any two replicas of the system $\mathbf{x}^{(a)}, \mathbf{x}^{(b)}$, where the disorder $\mathbf{A}$ is assumed to be the same, their overlap is also the same.

Proposition 3.1.1 (Replica Ansatz)
For any two replicas $\mathbf{x}^{(a)}, \mathbf{x}^{(b)}$, it holds that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i^{(a)} x_i^{(b)} = q_{a,b}, \tag{3.1.10}$$

and $q_{a,b}$ is the same value for all $a \neq b$. This quantity is known as the ***overlap*** *or* ***Edwards-Anderson (EA) order parameter****.*

It is the main phase exhibited in high temperatures (low β) which coincides with the parameter range of interest for this work. This corresponds to being below the Almeida-Thouless (AT) line (de Almeida and Thouless (1978)). A detailed discussion of RS and the AT line is beyond the scope of this thesis, but its applicability to our regime is essential to the methods we develop.

The RS free energy provides a tractable and accurate approximation for the system's thermodynamic properties in this phase. In this RS phase, and in the thermodynamic limit[1] of $n \to \infty$, the RS free energy Φ_{RS} of a spin system with Hamiltonian $H_n(\mathbf{x})$ and inverse

[1] We will also refer to this as the *large system* limit.

temperature β is well approximated by the expectation of the logarithm of the partition function.

$$\Phi_{RS}(q) \approx \lim_{n\to\infty} -\frac{1}{\beta}\mathbb{E}[\log Z_n], \tag{3.1.11}$$

where the expectation is over the randomness of the interaction terms.

The Replica Trick

However, due to the intractability of computing this expectation of the logarithm, physicists came up with the famous *replica trick* (Mezard et al. (1987)). Despite being non-rigorous, it has served its physical purpose of providing a simpler solution via the clever use of

$$\mathbb{E}[\log Z] = \lim_{m\to 0} \frac{\mathbb{E}[Z_n^m] - 1}{m}, \tag{3.1.12}$$

which allows simpler computations to be made using the assumption of identical statistical properties.

Replica Symmetric Free Energy

We will state that the **RS free energy** of the SK model is given by the formula

$$\Phi_{RS}(q) := \ln(2) + \frac{\beta^2(1-q)^2}{4} + \mathbb{E}\left[\log\cosh\left(\beta\sqrt{q}G + h\right)\right], \tag{3.1.13}$$

where $G \sim \mathcal{N}(0,1)$, and $q \in (0,1)$ is the overlap from Proposition 3.1.1, and is given by a constant determined by the fundamental relation

$$q := \mathbb{E}\left[\tanh^2(\beta\sqrt{q}G + h)\right]. \tag{3.1.14}$$

Clearly, q given as in Eq. (3.1.14) maximizes Φ_{RS}, as is clear by a simple differentiation. Moreover, it is the unique maximizer of Φ_{RS}, which is the main characterization of the RS phase. For rigorous mathematical results, see Chapter 1 of Talagrand (2011).

Mean Field Theory

Mean field theory (MFT) is a very efficient tool in statistical mechanics, as it offers a powerful way to understand the steady state of many interacting particles. The leading principle is to approximate the effective field acting on a site x_i by a global field. This in turn cancels correlations between spins. A bold assumption, mathematically presented below. We denote the *thermal average* or *magnetization* of a spin x_i by the common physicists' notation

$$m_i := \langle x_i \rangle,$$

where $\langle\cdot\rangle$ denotes averaging under the Gibbs measure (3.1.4). The zero-correlation assumption equates to

$$(x_i - m_i)(x_j - m_j) = 0. \tag{3.1.15}$$

While it has been proven to work very well for ferromagnetic models, MFT quite drastically fails to capture the equilibrium behaviour of the SK model. This is in part due to the randomness introduced in the coupling A_{ij}, as well as their $\mathcal{O}(1/\sqrt{n})$ scaling, instead of[2] $\mathcal{O}(1/n)$.

Using (3.1.15), we rewrite the Hamiltonian from Eq. (3.1.1) by expanding the quadratic term:

$$\begin{aligned} H(\mathbf{x}) &= \sum_{i<j} A_{ij} x_i x_j + h \sum_{i=1}^{n} x_i \\ &\overset{a)}{=} \sum_{i<j} A_{ij} \left(x_i m_j + x_j m_i - m_i m_j \right) + h \sum_{i=1}^{n} x_i \\ &\overset{b)}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} m_j x_i + \sum_{i<j}^{n} A_{ij} m_i m_j + h \sum_{i=1}^{n} x_i \end{aligned} \tag{3.1.16}$$

where $a)$ uses the MFT Ansatz and $b)$ accounts for a change of variables. We can now regroup the first and last term as the *effective field acting on site x_i*. Finally, the middle term may be dropped, as it does not act on *any* site. We thus have a mean-field Hamiltonian

$$H^{(\mathrm{MF})}(\mathbf{x}) = \sum_{i=1}^{n} h_{\mathrm{eff}}^{(i)} x_i, \tag{3.1.17}$$

which is essentially a decoupled Hamiltonian, where

$$h_{\mathrm{eff}}^{(i)} := \sum_{j=1}^{n} A_{ij} m_j + h. \tag{3.1.18}$$

Due to the independence of the sites, the Gibbs measure can now be given per site, and we may finally give the accurate formulation for m_i:

$$\begin{aligned} m_i &:= \langle x_i \rangle \\ &= \sum_{s \in \{1,-1\}} s \cdot \mathbb{P}_{\mathrm{Gibbs}}(x_i = s) \\ &= \mathbb{P}_{\mathrm{Gibbs}}(x_i = 1) + \mathbb{P}_{\mathrm{Gibbs}}(x_i = -1) \\ &= \frac{\exp(\beta h_{\mathrm{eff}}^{(i)}) - \exp(-\beta h_{\mathrm{eff}}^{(i)})}{\exp(\beta h_{\mathrm{eff}}^{(i)}) + \exp(-\beta h_{\mathrm{eff}}^{(i)})} \\ &= \tanh(\beta h_{\mathrm{eff}}^{(i)}) \\ &= \tanh\left(\beta \sum_{j=1}^{n} A_{ij} m_j + h \right), \end{aligned} \tag{3.1.19}$$

where we recall that β gets "absorbed" into the constant external field for notational convenience. The above equation works quite well in Ising-type models, where the interactions

[2] Recall the definition of the Hamiltonian (3.1.1), compared to Ising models with deterministic couplings.

do not depend on x_i, and the magnetizations m_i thus do not depend on x_i either. The complexity of spin glasses lies in the fact that Eq. (3.1.19) does *not* represent the thermal average at all. Indeed, x_i provides a contribution to the global field via the random A_{ij}'s, but in turn influences itself via "double-disorder". This must be accounted for and will be done so in the following Section.

The TAP Equations

The Thouless-Anderson-Palmer (TAP) equations are a set of self-consistent equations describing the behaviour of the local average spin $\langle x_i \rangle$, where $\langle \cdot \rangle$ denotes the mean with respect to the Gibbs measure (3.1.4). They were first introduced by Thouless, Anderson, and Palmer (1977) as an improvement on mean-field theory, presented above. We begin with the aforementioned notion of correcting the double-disorder.

The concept of *self-susceptibility* χ_i corresponds to the reaction of the thermal average m_i under a change in the effective mean field $h_{\text{eff}}^{(i)}$. Mathematically, we have

$$\chi_i := \frac{\partial m_i}{\partial h_{\text{eff}}^{(i)}}. \tag{3.1.20}$$

Still assuming the mean field equation has the form

$$m_i = \tanh(\beta h_{\text{eff}}^{(i)}),$$

the TAP equations can be computed by removing the action of m_i onto the effective field through the couplings A_{ij}. This is justified as otherwise, x_i influences the field, which in turn influences x_i, creating a feedback loop, over-inflating the effective field. We thus adjust $h_{\text{eff}}^{(i)}$ by subtracting the self-susceptibility, though the coupling A_{ij}.

$$\begin{aligned} h_{\text{eff}}^{(i)} &= h + \sum_{j=1}^{n} A_{ij} \left(m_j - m_i A_{ij} \chi_j \right) \\ &= h + \sum_{j=1}^{n} A_{ij} \left(m_j - m_i \cdot \beta \left(1 - \tanh^2(\beta h_{\text{eff}}^{(j)}) \right) \right) \\ &= h + \sum_{j=1}^{n} A_{ij} m_j - \beta \sum_{j=1}^{n} A_{ij}^2 \left(1 - m_j^2 \right) \cdot m_i. \end{aligned} \tag{3.1.21}$$

Here, we apply a concentration result on the ℓ_2-norm of Gaussian vectors (e.g. Bernstein, see Vershynin (2018)), asserting that in the large system limit, $\sum_{j=1}^{n} A_{ij}^2 \approx 1$. Moreover, the sum $\sum_{j=1}^{n} A_{ij}^2 m_j^2$ must be treated a bit more carefully. Due to the assumed independence of the A_{ij} along the same row or column, we may treat this as a sum of independent Gaussian RVs. We consider the random variable $M_j := A_{ij} m_j \sim \mathcal{N}\left(0, \frac{m_j^2}{n}\right)$. Then, again by concentration of the norm of Gaussian vectors, we have

$$\sum_{j=1}^{n} M_j^2 \approx \frac{1}{n} \sum_{j=1}^{n} m_j^2, \tag{3.1.22}$$

which corresponds to the definition of the EA order parameter from the Replica Ansatz (Proposition 3.1.1). Thus, plugging Eq. (3.1.22) back into the mean field of Eq. (3.1.21), and that back into Eq. (3.1.19), our final TAP equation can be expressed as

$$m_i = \tanh\left(h + \beta \sum_{j=1}^{n} A_{ij} m_j - \beta(1-q) m_i\right). \tag{3.1.23}$$

Remark 3.2
We note that this derivation is more of a heuristic one, as presented in Chapter II of the book by Mezard et al. (1987). A more thorough derivation can be done by means of the cavity method (from the same reference), and a full rigorous proof of the TAP equations can be found in the book by Talagrand (2011). Full treatment of this topic, while incredibly interesting, remains outside of the scope of this thesis, and was considered to be besides the point.

The TAP Free Energy

Turning a steady-state condition (such as eq. (3.1.23)) into a variational problem can make it easier to solve, which is how the *TAP free energy* $\mathcal{F}_{\text{TAP}}$ was introduced.

$$\mathcal{F}_{\text{TAP}}(\mathbf{m}, q) := -\frac{\beta}{2}\langle \mathbf{m}, \mathbf{A}\mathbf{m}\rangle - \langle h\mathbf{1}, \mathbf{m}\rangle - \sum_{i=1}^{n} h(m_i) - \frac{n\beta^2(1-q)^2}{4}, \tag{3.1.24}$$

where $h(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right)$ is the binary entropy function. It is quite simple to see that the stationary points of the above functional correspond to solutions of Eq. (3.1.23).

A long-standing conjecture by statistical physicists claims that the minima of the TAP free energy correspond to pure states of the Gibbs measure (3.1.4). This variational relationship between Gibbs states $\mathbf{m}$ and stationary points of the TAP free energy has been explored in physics for the past 40 years (Bray, Moore, and Young (1984); Cavagna, Giardina, Parisi, and Mézard (2003)), and more recently by mathematicians (Talagrand (2011); Chen and Panchenko (2018)). A recent paper by Fan, Mei, and Montanari (2020) rigorously presented a long-lasting result conjectured by physicists on the amount of critical points of $\mathcal{F}_{\text{TAP}}$. This was motivated by the statistical estimation problem of $\mathbb{Z}_2$-synchronization, which we will discuss later in Section 4.2.

3.2 Sampling Gibbs Measures with Diffusions

Sampling from the SK Model

We now concentrate on the measure on interest; the SK Gibbs measure (3.1.4) with no external field ($h = 0$). Plugging (3.1.4) into the posterior (2.2.2) yields the following tilted measure:

$$\mu_{\mathbf{y}(t),t}(\mathrm{d}\mathbf{x}) = \frac{1}{Z(\mathbf{y}(t),t)}\frac{1}{Z(\mathbf{A},\beta)}\exp\left(\frac{\beta}{2}\langle \mathbf{x},\mathbf{A}\mathbf{x}\rangle + \langle \mathbf{y},\mathbf{x}\rangle - \frac{t}{2}\|\mathbf{x}\|_2^2\right)\bar{\nu}(\mathrm{d}\mathbf{x}). \tag{3.2.1}$$

Since this is the main object of interest, we slow down here to introduce some more convenient notation which we will continually reuse. We first change the notation of the tilted measure

$$\mu_{\mathbf{y}(t),t} \to \mu_{\mathbf{A},\mathbf{y}(t)}, \tag{3.2.2}$$

to better highlight the dependence of the tilted measure on *both* the disorder $\mathbf{A}$ as well as the observation process $\mathbf{y}(t)$ (where t is implicit in $\mathbf{y}(t)$). Next, we accordingly adjust the notation of the associated tilted mean:

$$\mathbf{m}(\mathbf{y}(t);t) \to \mathbf{m}(\mathbf{A},\mathbf{y}(t)). \tag{3.2.3}$$

Tilted Mean Computation

In order to simulate the reverse SDE from Eq. (2.2.3) by means of Algorithm 1, we still require access to an accurate and efficient way to compute $\mathbf{m}(\mathbf{A},\mathbf{y}(t))$. Due to the discrete nature of the tilted measure and its exponential scaling with n, direct computation of the tilted mean is intractable. To address this, Alaoui et al. (2024) propose an efficient polynomial-time Approximate Message Passing algorithm, which we outline below.

Algorithm 2: SK Tilted Mean Estimation

```
Input: $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathbf{y} \in \mathbb{R}^n$, parameters $\beta, q \in (0,1), \eta > 0$, iteration numbers $K_{\mathrm{AMP}}, K_{\mathrm{NGD}}$
$\widehat{\mathbf{m}}^{-1} = \mathbf{0}_n$,
for $k = 0, \dots, K_{\mathrm{AMP}} - 1$ do
    $\widehat{\mathbf{m}}^k = \tanh(\mathbf{z}^k)$,    $\mathsf{b}_k = \frac{\beta^2}{n}\sum_{i=1}^n \left(1 - \tanh(z_i^k)^2\right)$
    $\mathbf{z}^{k+1} = \beta\mathbf{A}\widehat{\mathbf{m}}^k + \mathbf{y} - \mathsf{b}_k\widehat{\mathbf{m}}^{k-1}$,
end
$\mathbf{u}^0 = \mathbf{z}^{K_{\mathrm{AMP}}}$,
for $k = 0, \dots, K_{\mathrm{NGD}}$ do
    $\mathbf{u}^{k+1} = \mathbf{u}^k - \eta \cdot \nabla\mathcal{F}_{\mathrm{TAP}}(\widehat{\mathbf{m}}^{+,k}; \mathbf{y}, q)$,
    $\widehat{\mathbf{m}}^{+,k+1} \tanh\left(\mathbf{u}^{k+1}\right)$,
end
Output: $\widehat{\mathbf{m}}^{+,K_{\mathrm{NGD}}}$
```

The algorithm makes use of a constant q, which corresponds to the EA order parameter of the tilted spin glass at time t. This constant is given by

$$\begin{cases} q_0 = 0, \\ q_{k+1}(\beta,t) = \mathbb{E}_Z\left[\tanh\left(\beta^2 q_k(\beta,t) + t + \sqrt{\beta^2 q_k(\beta,t) + t}G\right)^2\right], \\ q = \lim_{k\to\infty} q_k, \end{cases} \tag{3.2.4}$$

with $G \sim \mathcal{N}(0,1)$.

Lines 1-5 of Algorithm 2 outline the AMP section, which provides a reliable way to estimate the posterior mean. Lines 6-10 are a Natural Gradient Descent (NGD) section, used to refine to initial AMP estimate $\widehat{\mathbf{m}}^{K_{\text{AMP}}}$. This second step is well-known to work, as it relies on the local convexity of the TAP free energy (explored in Celentano (2022); Celentano, Fan, and Mei (2023)), and addresses the previously mentioned connection between Gibbs states and the global minimum of $\mathcal{F}_{\text{TAP}}$. We provide a full derivation in Section 4.2, but will not elaborate on convexity or the NGD part.

3.3 Limitations of AMP for Tilted Mean Estimation

The AMP section of Algorithm 2 works remarkably well in the RS phase of the SK model (and general p-spin models), as this algorithm is designed to converge at the order parameter q, which maximizes Φ_{RS} (see Alaoui, Montanari, and Sellke (2021); Bolthausen (2012)). Due to the implicit goal of maximizing Φ_{RS}, AMP struggles if the spin glass finds itself in a metastable state. The performance of AMP is thus tightly related to the phase of the spin glass of interest. Indeed, one key component in the study of spin glasses is their phase transitions (see Panchenko (2012); Talagrand (2011) for rigorous results). Two main quantities are the *dynamical threshold* β_d and the *critical threshold* β_c, illustrated in Figure 3.1.

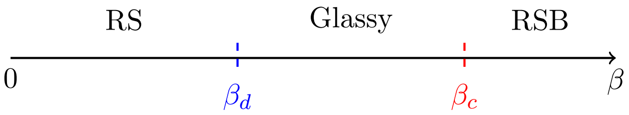


Figure 3.1: The different phases of spin glasses

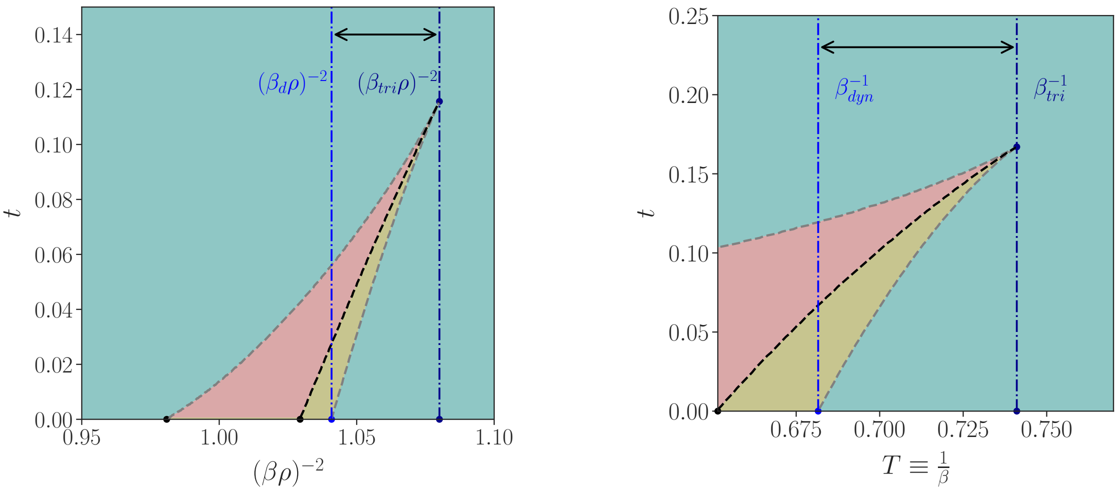
In the RS phase, the formulation (3.1.13) of Φ_{RS} is valid, and has a single maximizer. Moreover, the associated Gibbs measure concentrates around a dominant configuration. In the glassy phase $\beta \in (\beta_d, \beta_c)$, metastable states appear, i.e. Φ_{RS} is still valid but no longer has a single maximizer. For $\beta \geq \beta_c$, we enter a phase called *Replica Symmetry Breaking* (RSB), which is beyond the scope of this thesis.

AMP Regions of Convergence

It is clear from the discussion above that AMP has no guaranteed success for $\beta \geq \beta_d$, and the computational hardness in the glassy phase was predicted by Zdeborová and Krzakala (2016); Gamarnik, Moore, and Zdeborová (2022). This was also the main negative result from Alaoui et al. (2024, 2023), presenting the impossibility of efficiently sampling past β_d. Conversely, they have shown that for the SK model (where $\beta_d = \beta_c = 1$), sampling is achievable in polynomial time. This was first shown for $\beta < \frac{1}{2}$ but improved to $\beta < 1$ due to Celentano (2022), covering the entire replica symmetric regime.

In the case of p-spin glasses, Alaoui et al. (2023) demonstrate that the problem of sampling from the full RS regime still remains open. This is due to the algorithm trajectory of AMP in the phase space over $(\beta, t) \in (0, \beta_d) \times \mathbb{R}_+$. Indeed, due to the tilting of the measure, we also accordingly modify the landscape of Φ_{RS}, potentially trapping the AMP iterates in newly created metastable regions, preventing convergence.

The investigation by Ghio et al. (2023) showed by inspection of phase diagrams the emergence of such regions for multiple problems exhibiting a random first-order phase transition. Figure 3.2 illustrates such a transition for two different cases – the sparse (Fig. 3.2a) case and the pure 3-spin (Fig. 3.2b) – where the number of metastable states progressively increases as t increases, effectively shortening the range of AMP's performance. This culminates at a tri-point β_{tri}, representing the computational barrier to isotropic diffusion-based sampling of Gibbs measures.



(a) RS free energy phases for a planted sparse SK model (Definition 3.1.3) with sparsity $\rho = 0.08$. The metastable region begins around $\beta_{\mathrm{d}} \approx 12.2$ for $t = 0$, and the tri-point prohibits isotropic diffusion-based sampling beyond $\beta_{\mathrm{tri}} \approx 12.02$.

(b) RS free energy phases for a pure 3-spin model (Definition 3.1.4). We do not treat this spin glass in this work.

Figure 3.2: Phase diagrams highlighting two cases where the dynamical-to-computational gap occurs in the $((\beta\rho)^{-2}, t)$ resp. (β^{-1}, t)-plane. The choice of the x-axis is to better highlight the low-temperature gap between β_{d} and β_{tri}, and is spanned by the black arrow. The blue region is where Φ_{RS} has a single maximizer q, and thus AMP converges. The yellow regions denote metastable regions, where Φ_{RS} begins to develop a second maximizer $q' > q$. The red regions denote algorithmically hard regions, by the fact that the *second* (new) stationary point q' is the global maximizer of Φ_{RS}, trapping AMP in the first local maximum q. The code is courtesy of Ghio et al., which we have adapted for expository purposes.

Thus, we see that the AMP trajectory of Algorithm 2 is a vertical line in the (β, t) (resp. (β^{-1}, t)) phase space, with a fixed β and increasing t at every iteration. Figure 3.3 highlights the statistical-to-computational gap in p-spin models, where β_d and the β_{tri} are no longer coincident. In this scenario, the Gibbs measure remains statistically accessible up to β_d, but AMP fails to converge beyond β_{tri}, warranting novel methods in order to sample from the full RS regime.

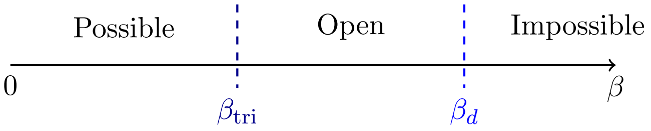


Figure 3.3: Illustration of the statistical-to-computational gap in p-spin or sparse models, as well as the feasibility of sampling.

3.4 Anisotropic Disorder-Dependent Diffusion

Here we introduce the anisotropic diffusion process, and show how it can help us in bridging the gap of sampling in $[\beta_{\text{tri}}, \beta_d)$, explained previously in Section 3.3. We consider here a natural extension to the observation process given in Eq. (2.2.1), as proposed in Section 4.2 of Montanari (2023). For a positive semi-definite (PSD) matrix $\mathbf{Q} \in \mathbb{S}_+^n$, we define the anisotropic process

$$\mathbf{y}(t) = \mathbf{Q}t\mathbf{x}_* + \mathbf{Q}^{1/2}\mathbf{B}_t. \tag{3.4.1}$$

It consequently satisfies a very similar SDE to Eq. (2.2.3),

$$\mathrm{d}\mathbf{y}(t) = \mathbf{Q}\mathbf{m}(\mathbf{A}, \mathbf{y}(t), \beta)\mathrm{d}t + \mathbf{Q}^{1/2}\mathrm{d}\mathbf{B}_t, \quad \mathbf{y}(0) = \mathbf{0}_n, \tag{3.4.2}$$

where $\mathbf{m}(\mathbf{A}, \mathbf{y}(t), \beta)$ is now the posterior mean upon linear (not scalar) observation at inverse temperature β:

$$\mathbf{m}(\mathbf{A}, \mathbf{y}(t), \beta) = \mathbb{E}\left[\mathbf{x}_* | \mathbf{Q}t\mathbf{x}_* + \sqrt{\mathbf{Q}t}\mathbf{z}\right]. \tag{3.4.3}$$

We again write the tilted measure:

$$\mu_{\mathbf{A},\mathbf{y}^{\text{anis}}(t)}(\mathbf{x}) = \frac{1}{Z(\mathbf{A}, \mathbf{y}^{\text{anis}}(t), \beta)} \mu_{\mathbf{A},\beta}(\mathbf{x}) \exp\left(-\frac{1}{2t}\|\mathbf{y}^{\text{anis}}(t) - t\mathbf{Q}\mathbf{x}\|^2_{\mathbf{Q}^{-1}}\right). \tag{3.4.4}$$

But how will this help in the issue that arose in Section 3.3?

We observe that as in the isotropic case, the tilt yields a term of the type

$$\exp\left(\langle \mathbf{y}(t), \mathbf{x}\rangle - \frac{t}{2}\|\mathbf{Q}\mathbf{x}\|^2_{\mathbf{Q}^{-1}}\right).$$

The intuition here is to also observe that by carefully choosing $\mathbf{Q}$ to depend on the disorder $\mathbf{A}$, we can allow for an effective temperature change, reflected in the following proposition.

Proposition 3.4.1 (Leveraging the Disorder)
Let $\mathbf{Q} = \lambda \mathbf{I}_n + c\mathbf{A}$ *for some* $\lambda > 2c \cdot \lambda_{\max}(\mathbf{A}) > 0$. *Then, the tilted measure in* (3.4.4) *becomes*

$$\mu_{\mathbf{A},\mathbf{y}^{\text{anis}}(t)}(\mathbf{x}) = \frac{1}{Z'(\mathbf{A},\mathbf{y}(t))} \exp\left(\frac{\beta - ct}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle + \langle \mathbf{y}(t), \mathbf{x}\rangle - \frac{\lambda t}{2}\|\mathbf{x}\|_2^2\right) \tag{3.4.5}$$

Proof. We defer the proof to Appendix B.1.1 □

By the nature of Eq. (3.4.4), we now see that the effective inverse temperature β has been lowered to $\beta - ct$. This is the main advantage of disorder-dependent diffusion sampling, as it offers a non-linear algorithm trajectory through phase space, potentially allowing circumvention the red regions of Figures 3.2a and 3.2b. Following Proposition 3.4.1, the new SDE we with to compute is thus

$$\mathrm{d}\mathbf{y}(t) = \mathbf{Q}\mathbf{m}(\mathbf{A}, \mathbf{y}(t), \beta - ct)\mathrm{d}t + \mathbf{Q}^{1/2}\mathrm{d}\mathbf{B}_t, \tag{3.4.6}$$

which would allow us to *increase the temperature* as we progress through the phase diagram, allowing for a trajectory akin to that depicted in Figure 3.4.

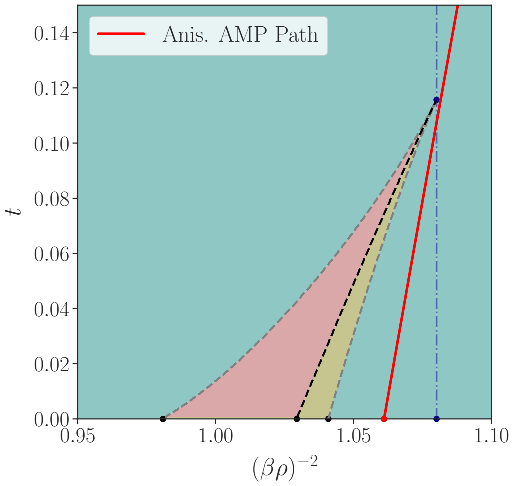


Figure 3.4: Visual depiction of the AMP trajectory (red) with decreasing β through the $((\beta\rho)^{-2}, t)$ phase space of the sparse SK model, $\rho = 0.08$. We highlight that this is *not* an actual implementation. As we can see, we begin our sampling procedure at $\beta = 12.135$, which is clearly out of bounds for isotropic AMP, and we breach the tri-point before encoutering any metastable states.

Remark 3.3
From here on out, we will use the parameters $(\lambda, c) = (3, 1)$ *or* $(\lambda, c) = (2.01, 1)$. *This was chosen for simplicity and to mitigate numerical issues with the eigenvalues of* $\mathbf{A}$ *later on.*

Tilted Mean Estimation

With all this in mind, we turn to the problem of estimating the anisotropic tilted mean associated to $\mu_{\mathbf{A},\mathbf{y}^{\text{anis}}(t)}$. We claim that Algorithm 3 successfully asymptotically estimates the tilted mean, which will be the main result of this thesis.

Theorem 3.4.1
Let $\mathbf{x}, \mathbf{A}, \mathbf{y}$ be the spin glass configuration, disorder matrix and anisotropic reverse diffusion process as described at the beginning of the Section. Let $\widehat{\mathbf{m}}^k$ be the output of Algorithm 3 with parameters $(\mathbf{A}, \mathbf{y}, \beta, K_{\text{AMP}})$, and let $\mathbf{m}(\mathbf{A}, \mathbf{y})$ denote the true tilted mean associated to Eq. (3.4.4). Then, for all $\epsilon > 0$, there exists a $K \in \mathbb{N}$ such that for all $k > K$,

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\|\mathbf{m}(\mathbf{A},\mathbf{y}) - \widehat{\mathbf{m}}^k(\mathbf{A},\mathbf{y})\|_2^2\right] < \epsilon. \tag{3.4.7}$$

In particular, we have

$$\lim_{k\to\infty} \operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\|\mathbf{m}(\mathbf{A},\mathbf{y}) - \widehat{\mathbf{m}}^k(\mathbf{A},\mathbf{y})\|_2^2\right] = 0. \tag{3.4.8}$$

Proof. We will devote the next chapter to lay down all the heuristics and derivations required for the proof of this Theorem. □

Algorithm 3: Anisotropic Tilted Mean Estimation via AMP

```
Input: $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathbf{y} \in \mathbb{R}^n$, parameters $\lambda, t, \beta > 0$, iteration numbers $K_{\text{AMP}}$
$\widehat{\mathbf{m}}^{(-1)} = \mathbf{0}_n$,
for $k = 0, \dots, K_{\text{AMP}} - 1$ do
    $\gamma_k = \frac{1}{n}\|\widehat{\mathbf{m}}^{(k-1)}\|_2^2$,
    $\eta_k = \frac{(\gamma_k + (\lambda+\beta)t)^2}{\sqrt{\gamma_k + 2t\beta^{-1}\gamma_k + t^2 + \lambda t}}$,
    $\widehat{\mathbf{m}}^{(k)} = \tanh(\eta_k \mathbf{z}^{(k)})$,    $\mathsf{b}_k = \frac{\beta^2}{n}\sum_{i=1}^n \eta_k \left(1 - \tanh(z_i^{(k)})^2\right)$
    $\mathbf{z}^{(k+1)} = \beta \mathbf{A}\widehat{\mathbf{m}}^k + \mathbf{y} - \mathsf{b}_k \widehat{\mathbf{m}}^{(k-1)}$,
end
Output: $\tanh(\mathbf{z}^{(K_{\text{AMP}})})$
```

Remark 3.4
Notice that Algorithm 3 is very similar to Lines 1-5 of Algorithm 2. We are thus essentially claiming that the same algorithm solves both cases. This is be fully developed and justified in Chapter 4.

Anisotropic Diffusion-Based Sampling Algorithm

We now have the (albeit unrealistic) result that as $k \to \infty$, Algorithm 3 correctly estimates the anisotropically tilted Gibbs mean from Eq. (3.4.3). We can now adapt Algorithm 1 to fit with our AMP iteration in Algorithm 4, accounting for the decrease in inverse temperature at each time step:

Algorithm 4: Disorder-Dependent Diffusion-Based Sampling Scheme

```
Input: Parameters $\mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{y} \in \mathbb{R}^n, \beta \in (0,1), \delta, L > 0, K_{\text{AMP}} \in \mathbb{N}_+$.
Set $\widehat{\mathbf{y}}(0) = 0$,
for $\ell = 0, \ldots, L-1$ do
    Compute $\beta_t = \beta - \ell\delta$,
    Set $\widehat{\mathbf{m}}(\mathbf{y}(t); t)$ as the output of Algorithm 3 with parameters $(\mathbf{A}, \mathbf{y}, \beta_t, K_{\text{AMP}})$
    Update $\widehat{\mathbf{y}}_{\ell+1} = \widehat{\mathbf{y}}_\ell + \delta\widehat{\mathbf{m}}(\widehat{\mathbf{y}}_\ell; \ell) + \sqrt{\delta}\mathbf{w}_{\ell+1}$
end
Draw $\{x_i^{\text{alg}}\}_{i \le n}$ conditionally independent with $\mathbb{E}[x_i^{\text{alg}} | \mathbf{y}, \{\mathbf{w}_\ell\}] = \widehat{m}_i(\mathbf{A}, \widehat{\mathbf{y}}_L)$
return $\mathbf{x}^{\text{alg}}$
```

We do not have any proof of convergence for this algorithm, but we think that mentioning it is of value to the reader, as it provides good insight on the simplicity of DDD-based sampling.

Chapter 4

Estimation of the Tilted Mean

Recalling the simple sampling algorithm from Algorithm 1, estimating the tilted mean $\mathbf{m}(\mathbf{A}, \mathbf{y}(t))$ plays a crucial role in the procedure. Theorem 1 of Montanari and Wu (2024) establishes that the error in this estimation propagates to the sampling scheme, affecting its accuracy in Wasserstein-2 distance.

In our case of interest the task is however not a simple one, as it requires estimation of a value with “no data”. Indeed, the SK model (3.1.1) is mostly noise. To remedy this, we will instead turn this tilted mean problem into a Bayesian estimation problem, by means of a *planted model* $\mathbb{P}$ whose joint probability distribution $\mathbb{P}(\mathbf{A}, \mathbf{y})$ is mutually contiguous to that of our SK model's disorder $\mathbf{A}$ and reverse SDE $\mathbf{y}$ (2.2.3). This is analogous to the methods used by Alaoui et al. (2024) in the context of sampling from the SK model, which we will report below.

We will solve this signal-recovery problem with Mean Field Variational Inference in Section 4.2, all while leveraging the mutual contiguity to allow usage of tools developed in Section 3.1. This will lead us to Section 4.3 where we will introduce Approximate Message Passing (AMP), whose analysis will be the main technical part of the thesis. We will lay out all the necessary tools to analyse Algorithm 2, and provide an extension to the analysis in Section 4.4, which will treat our disorder-dependent observation process $\mathbf{y}(t)$ introduced in Section 3.4.

4.1 The Planted Model and Contiguity

We will outline a more general contiguity proof in Section 4.1, which encapsulates that of Alaoui et al. (2024), allowing for simultaneous analysis of their model and our disorder-dependent model from Section 3.4, leading us to the popular problem of $\mathbb{Z}_2$-synchronization (Singer (2011)).

To reformulate our problem in terms of this promised simpler $\mathbb{Z}_2$-synchronization model, we employ the notion of contiguity. Originally introduced by LeCam, it is the asymptotic equivalent of absolute continuity for sequences of probability measures. Intuitively, as $n \to \infty$, two contiguous probability measures become increasingly indistinguishable – as if overlapping – and assign zero probability to the same events. Formally, we have the following.

Definition 4.1.1 (Contiguity)
A sequence of probability measures $\{Q_n\}_{\geq 1}$ is said ***contiguous*** *with respect to the sequence $\{P_n\}_{n\geq 1}$ if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ for every sequence of measurable sets A_n. We denote this $Q_n \lhd P_n$. If the reverse is also true, we say that Q_n and P_n are* ***mutually contiguous****, and denote it $Q_n \lhd \ \rhd P_n$.*

A key consequence of mutual contiguity is that all statistics under both probability measures have the same probability-zero events. We summarize LeCam's celebrated first lemma in the following statement.

Lemma 4.1.1 (Theorem 6.4, Vaart (1998))
Let P_n and Q_n be sequences of probability measures. Then the following statements are equivalent:

i.) $Q_n \rhd P_n$.

ii.) If $\frac{\mathrm{d}P_n}{\mathrm{d}Q_n} \overset{Q_n}{\to} U$ along a subsequence, then $P(U > 0) = 1$.

iii.) If $\frac{\mathrm{d}Q_n}{\mathrm{d}P_n} \overset{P_n}{\to} V$ along a subsequence, then $\mathbb{E}[V] = 1$.

iv.) For any statistics T_n, if $T_n \overset{P_n}{\to} 0$, then $T_n \overset{Q_n}{\to} 0$.

An immediate and very useful consequence thereof is stated below.

Corollary 4.1.1
If P_n and Q_n are probability measures such that

$$\frac{\mathrm{d}P_n}{\mathrm{d}Q_n} \overset{Q_n}{\to} \exp\left(\mathcal{N}(\mu, \sigma^2)\right), \tag{4.1.1}$$

then $Q_n \lhd P_n$. Moreover, $Q_n \lhd \ \rhd P_n$ if $\mu = -\frac{\sigma^2}{2}$.

Proof. Clearly, the log-normal random variable is always positive, implying $Q_n \rhd P_n$ by point ii) of Lemma 4.1.1. For the converse statement, notice that by

$$\mathbb{E}\left[\exp\left(\mathcal{N}(\mu, \sigma^2)\right)\right] = e^{\mu + \frac{\sigma^2}{2}} \tag{4.1.2}$$

equals to one if and only if $\mu = -\frac{\sigma^2}{2}$, implying $P_n \rhd Q_n$ by point iii) of Lemma 4.1.1. □

Random Model

We present here the joint probability distribution for the data generated by the SK model $(\mathbf{A}, \mathbf{x})$ from Def. 3.1.1 & the disorder-dependent reverse diffusion process $\mathbf{y}(t)$ from Eq. (3.4.2). Namely, the random objects of interest are $\mathbf{A}, \mathbf{x}, \mathbf{y}(t), \mathbf{Q}$, and we will denote their distribution under the random model as $\mathbb{Q}$:

$$\mathbb{Q} : \begin{cases} \mathbf{A} \sim \mu_{\mathrm{GOE}}, \\ \mathbf{x}|\mathbf{A} \sim \mu_{\mathbf{A},\beta}, \\ \mathbf{Q} := \lambda \mathbf{I}_n + \mathbf{A}, \\ \mathbf{y}(t) := \int_0^t \mathbf{Qm}(\mathbf{A}, \mathbf{y}(s)) \mathrm{d}s + \int_0^t \mathbf{Q}^{1/2} \mathrm{d}\mathbf{B}_s, \end{cases} \tag{4.1.3}$$

where $\mathbf{B}_s$ is a standard Brownian Motion. Due to the covariance matrix $\mathbf{Q}$, the situation is slightly more complicated than Alaoui et al. (2024). We must take into account the dependency of $\mathbf{y}$ on $\mathbf{Q}$ (i.e. on $\mathbf{A}$). We now explicitly derive the various distributions under $\mathbb{Q}$.

$$\mathbb{Q}(\mathrm{d}\mathbf{A}) = \frac{1}{Z_{\mathrm{GOE}}} \exp\left(-\frac{n}{4}\|\mathbf{A}\|_F^2\right) \mathrm{d}\mathbf{A}, \quad Z_{\mathrm{GOE}} = \int e^{-\frac{n}{4}\|\mathbf{A}\|_F^2} \mathrm{d}\mathbf{A} \tag{4.1.4}$$

$$\mathbb{Q}(\mathrm{d}\mathbf{x}|\mathbf{A}) = \frac{1}{Z_\beta(\mathbf{A})} \exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{Ax}\rangle\right) \bar{\nu}(\mathrm{d}\mathbf{x}), \tag{4.1.5}$$

where $\mathrm{d}\mathbf{A}$ denotes the Lebesgue measure over $\mathbb{R}^{n\times n}_{\mathrm{sym}}$. The normalizer of the SK measure is given by

$$\begin{aligned} Z_\beta(\mathbf{A}) &= \int \exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{Ax}\rangle\right) \bar{\nu}(\mathrm{d}\mathbf{x}) \\ &= 2^{-n} \sum_{\mathbf{x}\in\mathcal{C}^n} \exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{Ax}\rangle\right) \end{aligned} \tag{4.1.6}$$

To derive a law for the observation process, we will first define $\mathbb{W}_\mathbf{Q} : \mathsf{C}[0,T] \to \mathbb{R}$ as the Wiener measure of an anisotropic Brownian Motion $\sqrt{\mathbf{Q}}\mathbf{B}_t$. Recall that this is valid by assumed positive semi-definiteness of $\mathbf{Q}$. An important characterization of processes of the form of $\mathbf{y}(t)$ is given in the following Theorem.

Theorem 4.1.1 (Theorem 7.12; i), Liptser and Shiryayev (1977))
Let $\xi = (\xi_t)_{t\le T}$ be a diffusion process with a differential as in Eq. (2.1.1), where

$$\int_0^T \mathbb{E}[|f(\xi, s)|]\mathrm{d}s < \infty.$$

Let $\alpha = (\alpha_t(x))_{t\le T}$ be a functional such that for almost all $t \in [0,T]$,

$$\alpha_t(\xi) = \mathbb{E}[f(\xi, t)].$$

It follows that the random process $\bar{W} = (\bar{W}_t)_{t\leq T}$ defined as

$$\bar{W}_t = \xi_t - \int_0^t \alpha_s(\xi)\mathrm{d}s$$

is a Wiener process, and ξ is also a diffusion process with respect to $\bar{W}$, in the sense that

$$\mathrm{d}\xi_t = \alpha_t(\xi) + \mathrm{d}\bar{W}_t. \tag{4.1.7}$$

It follows from Theorem 4.1.1 that there exists another Brownian motion $\bar{\mathbf{B}}_t$, such that $\mathbf{y}(t)$ admits an expression like the following:

$$\mathbf{y}(t) = t\mathbf{x}_* + \sqrt{\mathbf{Q}}\bar{\mathbf{B}}_t, \tag{4.1.8}$$

where $\mathbf{x}_* \sim \mu_{\beta,\mathbf{A}}$ is independent of $\bar{\mathbf{B}}_t$. By characterization of Brownian motion, this allows the use of Girsanov's Theorem to write the Radon-Nikodym derivative:

$$\begin{aligned}\frac{\mathrm{d}\mathbb{Q}(\cdot|\mathbf{A},\mathbf{x})}{\mathrm{d}\mathbb{W}_{\mathbf{Q}}}(\mathbf{y}) &= \exp\left(\int_0^T \langle \mathbf{Q}^{1/2}\mathbf{x}, \mathrm{d}\mathbf{B}(t)\rangle - \frac{1}{2}\int_0^T \|\mathbf{Q}^{1/2}\mathbf{x}\|_2^2\right)\\ &= \exp\left(\langle \mathbf{Q}^{1/2}\mathbf{x}, \mathbf{B}(T)\rangle - \frac{T}{2}\|\mathbf{Q}^{1/2}\mathbf{x}\|_2^2\right)\\ &= \exp\left(\langle \mathbf{Q}^{1/2}\mathbf{x}, \mathbf{Q}^{-1/2}(\mathbf{y}(T) - T\mathbf{Q}\mathbf{x})\rangle - \frac{T}{2}\|\mathbf{Q}^{1/2}\mathbf{x}\|_2^2\right)\\ &= \exp\left(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3T}{2}\|\mathbf{Q}^{1/2}\mathbf{x}\|_2^2\right). \end{aligned} \tag{4.1.9}$$

Hence

$$\begin{aligned}\mathbb{Q}(\mathrm{d}\mathbf{A}, \mathrm{d}\mathbf{y}) &= \int_{\mathbf{x}} \mathbb{Q}(\mathrm{d}\mathbf{y}|\mathbf{A},\mathbf{x})\mathbb{Q}(\mathrm{d}\mathbf{A},\mathrm{d}\mathbf{x}) \qquad (4.1.10)\\ &= \int_{\mathbf{x}} \mathbb{Q}(\mathbf{y}|\mathbf{A},\mathbf{x})\mathbb{Q}(\mathrm{d}\mathbf{x}|\mathbf{A})\mathbb{Q}(\mathrm{d}\mathbf{A})\\ &= \int_{\mathbf{x}}\Bigg[\exp\left(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3\lambda T}{2}\|\mathbf{x}\|_2^2 - \frac{3T}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle\right) Z_{\text{GOE}}^{-1}\exp\left(-\frac{n}{4}\|\mathbf{A}\|_F^2\right) \times\\ &\quad Z_\beta^{-1}(\mathbf{A})\exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle\right)\bar{\nu}(\mathrm{d}\mathbf{x})\Bigg]\mathrm{d}\mathbf{A}\mathbb{W}_{\mathbf{Q}}(\mathrm{d}\mathbf{y})\\ &= \frac{Z_\beta^{-1}(\mathbf{A})}{Z_{\text{GOE}}}\int_{\mathbf{x}}\Bigg[\exp\left(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3\lambda T}{2}\|\mathbf{x}\|_2^2 + \frac{\beta_*}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \frac{n}{4}\|\mathbf{A}\|_F^2\right)\bar{\nu}(\mathrm{d}\mathbf{x})\Bigg]\mathrm{d}\mathbf{A}\mathbb{W}_{\mathbf{Q}}(\mathrm{d}\mathbf{y}). \end{aligned}$$

with $\beta_* := \beta - 3T$. Furthermore, we can multiply both numerator and denominator of $\mathbb{Q}(\mathrm{d}\mathbf{A}, \mathrm{d}\mathbf{y})$ by $e^{-\frac{\beta^2 n}{4}}$:

$$\begin{aligned}\mathbb{Q}(\mathrm{d}\mathbf{A}, \mathrm{d}\mathbf{y}) = \frac{Z_{\text{SK}}^{-1}(\mathbf{A},\beta)}{Z_{\text{GOE}}}\int_{\mathbf{x}}\Bigg[\exp\Bigg(&\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3\lambda T}{2}\|\mathbf{x}\|_2^2 - \frac{\beta_*}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle\\ &-\frac{\beta^2 n}{4} - \frac{n}{4}\|\mathbf{A}\|_F^2\Bigg)\bar{\nu}(\mathrm{d}\mathbf{x})\Bigg]\mathrm{d}\mathbf{A}\mathbb{W}_{\mathbf{Q}}(\mathrm{d}\mathbf{y}), \end{aligned} \tag{4.1.11}$$

with

$$Z_{\mathrm{SK}}(\mathbf{A},\beta) := 2^{-n} \sum_{\mathbf{x}\in\{\pm 1\}^n} \exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \frac{\beta^2 n}{4}\right) \tag{4.1.12}$$

being the normalizer of a rescaled SK measure at inverse temperature β.

Planted Model

The planted model – whose distribution will be denoted $\mathbb{P}$ – first generates the configuration $\mathbf{x}$ from a uniform distribution on $\mathcal{C}^n$, and creates a spiked GOE matrix in the direction of $\mathbf{x}$; $\mathbf{A} = \frac{\beta}{n}\mathbf{x}\mathbf{x}^T + \mathbf{W}$ as the planted disorder. Then, $\mathbf{y}(t)$ is generated as a linear observation of $\mathbf{x}$ with sensing matrix and noise variance equal to $\mathbf{Q}t$. In particular, we have

$$\mathbb{P} : \begin{cases} \mathbf{x} \sim \bar{\nu} \\ \mathbf{A}|\mathbf{x} = \frac{\beta_*}{n}\mathbf{x}\mathbf{x}^T + \mathbf{W}, \quad \mathbf{W} \sim \mu_{\mathrm{GOE}}, \\ \mathbf{Q} := \lambda \mathbf{I}_n + \mathbf{A}, \\ \mathbf{y}(t) = \mathbf{Q}t\mathbf{x} + \sqrt{\mathbf{Q}}\mathbf{B}_t. \end{cases} \tag{4.1.13}$$

The marginal, conditional, and joint distributions of $\mathbf{x}$ and $\mathbf{A}$ are straightforwardly given by

$$\mathbb{P}(\mathrm{d}\mathbf{x}) = \bar{\nu}(\mathrm{d}\mathbf{x}), \tag{4.1.14}$$

$$\mathbb{P}(\mathrm{d}\mathbf{A}|\mathbf{x}) = \frac{1}{Z_{\mathrm{pl}}} \exp\left(-\frac{n}{4}\|\mathbf{A} - \frac{\beta}{n}\mathbf{x}\mathbf{x}^T\|_F^2\right) \mathrm{d}\mathbf{A}, \tag{4.1.15}$$

$$\mathbb{P}(\mathrm{d}\mathbf{A},\mathrm{d}\mathbf{x}) = \frac{1}{Z_{\mathrm{pl}}} \exp\left(-\frac{n}{4}\|\mathbf{A} - \frac{\beta}{n}\mathbf{x}\mathbf{x}^T\|_F^2\right) \bar{\nu}(\mathrm{d}\mathbf{x})\mathrm{d}\mathbf{A}. \tag{4.1.16}$$

$$\tag{4.1.17}$$

To determine the law of the observation process, we will again use Girsanov's Theorem, which is directly applicable by the form of $\mathbf{y}(t)$ in Eq. (4.1.13). Let $\mathbb{W}_{\mathbf{Q}}$ be as in the previous section (with $\mathbf{Q}$ as in Eq. (4.1.13) this time), then

$$\frac{\mathrm{d}\mathbb{P}(\cdot|\mathbf{A},\mathbf{x})}{\mathrm{d}\mathbb{W}_{\mathbf{Q}}}(\mathbf{y}) = \exp\left(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3T}{2}\|\mathbf{Q}\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\right). \tag{4.1.18}$$

Now we will write the joint distribution for the anisotropic planted model with the matrix $\mathbf{Q} = \lambda\mathbf{I} + \mathbf{A}$.

$$
\begin{aligned}
\mathbb{P}(\mathrm{d}\mathbf{A}, \mathrm{d}\mathbf{y}) &= \int_{\mathbf{x}} \mathbb{P}(\mathrm{d}\mathbf{y}|\mathbf{A},\mathbf{x})\mathbb{P}(\mathrm{d}\mathbf{A}, \mathrm{d}\mathbf{x}) \qquad (4.1.19)\\
&= \int_{\mathbf{x}} \Bigg[\exp\left(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3\lambda T}{2}\|\mathbf{x}\|_2^2 - \frac{3T}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle \right) \times \\
&\quad \frac{1}{Z_{\mathrm{pl}}} \exp\left(-\frac{n}{4}\|\mathbf{A} - \frac{\beta}{n}\mathbf{x}\mathbf{x}^T\|_F^2 \right) \bar{\nu}(\mathrm{d}\mathbf{x}) \Bigg] \mathrm{d}\mathbf{A}\mathbb{W}_{\mathbf{Q}}(\mathrm{d}\mathbf{y}) \\
&= \frac{1}{Z_{\mathrm{pl}}} \int_{\mathbf{x}} \Bigg[\exp\left(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3\lambda T}{2}\|\mathbf{x}\|_2^2 - \frac{3T}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \frac{n}{4}\|\mathbf{A} - \frac{\beta}{n}\mathbf{x}\mathbf{x}^T\|_F^2 \right) \bar{\nu}(\mathrm{d}\mathbf{x}) \Bigg] \mathrm{d}\mathbf{A}\mathbb{W}_{\mathbf{Q}}(\mathrm{d}\mathbf{y}) \\
&= \frac{1}{Z_{\mathrm{pl}}} \int_{\mathbf{x}} \Bigg[\exp\Bigg(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3\lambda T}{2}\|\mathbf{x}\|_2^2 - \frac{3T}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \frac{n}{4}\|\mathbf{A}\|_F^2 \\
&\quad + \frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \frac{\beta^2 n}{4} \Bigg) \bar{\nu}(\mathrm{d}\mathbf{x}) \Bigg] \mathrm{d}\mathbf{A}\mathbb{W}_{\mathbf{Q}}(\mathrm{d}\mathbf{y}) \\
&= \frac{1}{Z_{\mathrm{pl}}} \int_{\mathbf{x}} \Bigg[\exp\Bigg(\langle \mathbf{x}, \mathbf{y}(T)\rangle - \frac{3\lambda T}{2}\|\mathbf{x}\|_2^2 - \frac{\beta_*}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \frac{n}{4}\|\mathbf{A}\|_F^2 - \frac{\beta^2 n}{4} \Bigg) \bar{\nu}(\mathrm{d}\mathbf{x}) \Bigg] \mathrm{d}\mathbf{A}\mathbb{W}_{\mathbf{Q}}(\mathrm{d}\mathbf{y}).
\end{aligned}
$$

As we can see, we are eerily close to the joint random distribution of (4.1.11), up to $Z_{\mathrm{SK}}(\mathbf{A}, \beta)$. We must still take a closer look at Z_{pl} and Z_{GOE}.

Lemma 4.1.2
Consider a GOE matrix $\mathbf{A}$ *and its planted version as above, with their distributions given by* (4.1.4) *and* (4.1.15) *respectively. Then,* $Z_{\mathrm{pl}} = Z_{\mathrm{GOE}}$.

Proof. See Appendix B.2.1. □

With this final result, we may state the contiguity result.

Proposition 4.1.1
The two Borel probability distributions $\mathbb{Q}$ *and* $\mathbb{P}$ *from Eqs.* (4.1.3) *resp.* (4.1.13) *on* $\mathsf{C}([0,T], \mathbb{R}^n) \times \mathbb{R}^{n\times n}_{\mathrm{sym}}$ *are mutually contiguous for* $\beta < 1$.

Proof. Combining our newfound Eq. (4.1.19) with the random model in (4.1.11), we can see that

$$
\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\mathbf{A}, \mathbf{y}) = Z_{\mathrm{SK}}(\mathbf{A}, \beta). \qquad (4.1.20)
$$

We know from Aizenman, Lebowitz, and Ruelle (1987) that Z_{SK} from (4.1.12) has log-normal fluctuations for $\beta < 1$ (see Appendix B.2.2 for a derivation of the correspondence between our and their model). By Corollary 4.1.1, we have that $\mathbb{Q}$ and $\mathbb{P}$ are mutually contiguous. □

We will now restrict our analysis to the planted model from Eq. (4.1.13).

4.2 TAP Variational Inference

$\mathbb{Z}_2$-synchronization

The data generated in the planted model is none other than the $\mathbb{Z}_2$-*synchronization* problem in the high-dimensional statistics world. The aim is to recover the i.i.d. Rademacher signal $\mathbf{x} \sim \mathrm{Unif}\{1,-1\}^{\otimes n}$ from a spiked matrix observation with signal-to-noise ratio (SNR) β^2

$$\mathbf{A} = \frac{\beta}{n}\mathbf{x}\mathbf{x}^T + \mathbf{W}, \quad \mathbf{W} \sim \mu_{\mathrm{GOE}}.$$

Remark 4.1
This rank-one matrix factorization problem is known to exhibit a phase transition at $\beta = 1$. *Under this threshold, no estimator can produce significant correlation with the signal* $\mathbf{x}$, *whereas over this threshold, strictly positive correlation in achievable (Deshpande, Abbe, and Montanari (2015)). This transition in estimation capability arises from the inability to extract a leading eigenvector of* $\mathbf{A}$ *to initialize the estimation procedure (Baik, Arous, and Péché (2005), Benaych-Georges and Nadakuditi (2011), Knowles and Yin (2013)).*

While we are indeed necessarily restricted to the subcritical regime, since contiguity holds only for $\beta < 1$ *by Proposition 4.1.1, the estimation problem remains feasible thanks to the additional side information* $\mathbf{y}(t)$

$$\mathbf{y}(t) = \mathbf{Q}t\mathbf{x} + \sqrt{\mathbf{Q}t}\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_n),$$

which will morally act as a leading eigenvector.

In the random model (Eq. (4.1.3)), the goal was to estimate the tilted mean, defined in Eq. (3.4.4)). In the planted model, this tilted mean corresponds to the posterior mean of $\mathbf{x}$ given observations $\mathbf{A}, \mathbf{y}(t)$. In particular, it is the Bayes estimator of $\mathbf{x}$ with respect to quadratic loss, i.e.

$$\mathbf{m}(\mathbf{A}, \mathbf{y}(t)) := \int \mathrm{x}\mathbb{P}(\mathrm{d}\mathbf{x}|\mathbf{A}, \mathbf{y}(t))$$

is the minimizer of the mean squared error (MSE)

$$\mathrm{MSE} = \mathbb{E}[\|\mathbf{x} - \widehat{\mathbf{m}}\|_2^2]. \tag{4.2.1}$$

The posterior distribution is quite straightforward to compute from the model in Eq. (4.1.13), but we will provide it here for further reference:

$$\begin{aligned}\mathbb{P}(\mathrm{d}\mathbf{x}|\mathbf{A}, \mathbf{y}(t)) &\propto \exp\left(-\frac{n}{4}\|\mathbf{A} - \frac{\beta}{n}\mathbf{x}\mathbf{x}^T\|_F^2 - \frac{1}{2t}\|\mathbf{y}(t) - \mathbf{Q}t\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\right)\bar{\nu}(\mathrm{d}\mathbf{x}) \\ &\propto \exp\left(\frac{\beta - t}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle + \langle \mathbf{y}(t), \mathbf{x}\rangle - \frac{(\lambda + \beta)t}{2}\|\mathbf{x}\|_2^2\right)\bar{\nu}(\mathrm{d}\mathbf{x}).\end{aligned} \tag{4.2.2}$$

The challenge is now how to compute this posterior mean.

Mean Field Variational Inference

Due to the intractability of computing the posterior mean of $\mathbb{P}(\mathrm{d}\mathbf{x}|\mathbf{A},\mathbf{y}(t))$, we will follow a very popular road in Bayesian inference: *variational inference* (VI). Before we get into details, we introduce three concepts: entropy, the Kullback-Leibler (KL) divergence, and product measures.

Definition 4.2.1 (Entropy)
Let Q be a probability distribution with density $q(\mathbf{x})$ with respect to the Lebesgue measure $\mathrm{d}\mathbf{x}$. *The* ***entropy*** *of Q is defined as*

$$\mathsf{h}(q) := \int_{\mathbf{x}} q(\mathbf{x}) \log\left(\frac{1}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \tag{4.2.3}$$

The entropy provides a measure of uncertainty of a system described by a probability distribution.

Definition 4.2.2 (KL Divergence)
Let P and Q be two probability measures such that the support of P is contained in the support of Q, i.e.

$$\operatorname{supp}(P) \subseteq \operatorname{supp}(Q),$$

and let $p(\mathbf{x}), q(\mathbf{x})$ represent their densities respectively. Then, we define the ***KL divergence of P with respect to Q*** *as*

$$\mathscr{D}(P\|Q) := \int_{\mathbf{x}\in\operatorname{supp}(P)} p(\mathbf{x}) \log\left(\frac{p(\mathbf{x}}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x}. \tag{4.2.4}$$

The KL divergence (also known in Information Theory as relative entropy), is a measure of the inefficiency of assuming Q when the actual distribution is P. In other words, it quantifies in bits (or nats) how "bad" replacing P with Q is (Cover and Thomas (2006)).

Definition 4.2.3 (Product Measure)
Let $(\Omega_1,\mathcal{F}_1),(\Omega_2,\mathcal{F}_2)$ be measurable spaces, and let μ_1 and μ_2 be measures on these spaces. The measure $\mu_1\times\mu_2$ on $(\Omega_1,\mathcal{F}_1)\times(\Omega_2,\mathcal{F}_2)$ is called a ***product measure*** *if for all $A_1\in\Omega_1, A_2\in\Omega_2$, it satisfies the property*

$$(\mu_1\times\mu_2)(A_1\times A_2) = \mu_1(A_1)\mu_2(A_2). \tag{4.2.5}$$

It is clear from Definition 4.2.3 that when two probability measures satisfy the product measure property, then the random variables associated to the spaces are independent.

The particular path of VI we will follow will be Mean Field Variational Inference (MFVI), which attempts to best approximate the posterior by a product measure $\hat{q}(\mathrm{d}\mathbf{x}) = \prod_{i=1}^{n}\hat{q}(\mathrm{d}x_i)$, such that $\hat{q}$ minimizes the KL-divergence w.r.t. $p(\mathbf{x}|\mathbf{A},y)$ over the class of product measures

$\mathcal{Q}$, i.e.

$$\hat{q}(\mathbf{x}) = \arg\min_{q\in\mathcal{Q}} \mathscr{D}\left(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{A},\mathbf{y}(t))\right). \tag{4.2.6}$$

In other words, we minimize the cost of replacing the posterior distribution $p(\mathbf{x}|\mathbf{A},\mathbf{y}(t))$ with a probability distribution $q(\mathbf{x})$ with independent random variables, much akin to the assumption in MFT made in Section 3.1, which negates the correlations between spins. We begin by first explicitly calculating the objective from Eq. (4.2.6) using Definition 4.2.2:

$$\begin{aligned}
\mathscr{D}\left(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{A},\mathbf{y}(t)))\right) &:= \int q(\mathbf{x})\log\left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{A},\mathbf{y}(t))}\right)\mathrm{d}\mathbf{x} \\
&\overset{a)}{=} \int q(\mathbf{x})\log(q(\mathbf{x}))\mathrm{d}\mathbf{x} \\
&\quad - \int q(\mathbf{x})\log\left(\frac{1}{Z}\exp\left(\frac{\beta-t}{2}\langle\mathbf{x},\mathbf{A}\mathbf{x}\rangle + \langle\mathbf{y},\mathbf{x}\rangle\right)\bar{\nu}(\mathbf{x})\right)\mathrm{d}\mathbf{x} \\
&\overset{b)}{\propto} -\mathsf{h}(q) - \mathbb{E}_q\left[\frac{\beta-t}{2}\langle\mathbf{x},\mathbf{A}\mathbf{x}\rangle + \langle\mathbf{y},\mathbf{x}\rangle\right],
\end{aligned} \tag{4.2.7}$$

where $a)$ holds by plugging in Eq. (4.2.2) into $p(\mathbf{x}|\mathbf{A},\mathbf{y}(t))$, and $b)$ holds by removing all irrelevant constant factors.

Approximating via the Bayes Mean

The above quantity is still not trivial to minimize. To simplify this task, we parametrize q by its mean $m_i := \mathbb{E}_{q_i}[x_i]$ and notice that if we suppose that $q_i(x_i = 1) = p, q_i(x_i = -1) = 1-p$, then we have

$$m_i = q_i(x_i = 1) - q_i(x_i = -1) = 2p - 1 \tag{4.2.8}$$

i.e. $p = \frac{1+m_i}{2}$. The differential entropy term thus turns into a sum which can solely be described in term of $\mathbf{m} = \mathbb{E}_q[\mathbf{x}]$:

$$\mathsf{h}(\mathbf{m}) := \sum_{i=1}^{n}\left(\frac{1+m_i}{2}\log\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2}\log\left(\frac{1-m_i}{2}\right)\right). \tag{4.2.9}$$

Secondly, we notice that the quadratic form admits a decomposition

$$\mathbb{E}_{\mathbf{x}\sim q}\left[\langle\mathbf{x},\mathbf{A}\mathbf{x}\rangle\right] = \sum_{i,j} A_{ij}\mathbb{E}_q[x_i x_j] = \sum_{i,j} A_{ij}\left(\mathbb{E}_q[x_i]\mathbb{E}[x_j] + \mathrm{Cov}(x_i, x_j)\right),$$

where the covariance term is zero since q is a product measure. Note that this covariance is *not* zero under the posterior, but by a well-founded conjecture by physicists that the Gibbs measure concentrates around one pure state, it is safe to assume for the sake of this MFVI scheme that correlations decay. This is also part of an accuracy-efficiency trade-off. Finally, we obtain the mean field following objective function:

$$\mathscr{D}\left(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{A},\mathbf{y}(t)))\right) \propto \mathcal{F}_{\mathrm{MF}} := -\mathsf{h}(\mathbf{m}) - \frac{\beta-t}{2}\langle\mathbf{m},\mathbf{A}\mathbf{m}\rangle - \langle\mathbf{y},\mathbf{m}\rangle. \tag{4.2.10}$$

Unfortunately, it has been shown that similar to mean field approximation of the SK model, the MFVI objective of Eq. (4.2.10) does not provide an adequate optimization landscape, and leads to incorrect estimates of the Bayes posterior mean.

Theorem 4.2.1 (Theorem 1.3, Fan et al. (2020))
Denote $\mathcal{S}_{\beta,n} = \{\widehat{\mathbf{m}} \in (-1,1)^n : \nabla \mathcal{F}_{\text{MF}} = \mathbf{0}\}$. *There exists a constant* $\beta_0 > 0$ *and a constant* $\varepsilon(\beta) > 0$ *for every* $\beta > \beta_0$, *such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\inf_{\widehat{\mathbf{m}}\in\mathcal{S}_{\beta,n}} \frac{1}{n^2} \|\mathbf{m}\mathbf{m}^T - \widehat{\mathbf{m}}\widehat{\mathbf{m}}^T\|_F^2 > \varepsilon(\beta) \right) = 1. \qquad (4.2.11)$$

The search for an adequate objective now continues.

TAP Free Energy Objective

Looking at the MFVI objective function in eq. (4.2.10), we directly notice that its stationary points satisfy the equation

$$\mathbf{m} = \tanh\left(\beta\mathbf{A}\mathbf{m} + \mathbf{y}\right),$$

a tantalizingly similar equation to the mean field assumption on the SK model in Eq. (3.1.19). In the same way that MFT does not accurately describe the Gibbs mean of the SK model, the MFVI objective's stationary points do *not* provide an accurate characterization of the Bayes estimator. To rectify this, we now turn to the tools developed in the Spin Glass Theory section, and will proceed to "TAP-ify" our objective. Adding the Onsager term as in Eq. (3.1.23), we obtain a TAP objective function (free energy), which is given by

$$\mathscr{F}_{\text{TAP}}(\mathbf{m};\mathbf{y}) = -\mathsf{h}(\mathbf{m}) - \frac{\beta(t)}{2}\langle \mathbf{m}, \mathbf{A}\mathbf{m}\rangle - \langle \mathbf{y}, \mathbf{m}\rangle - \frac{\beta(t)^2 n}{4}(1 - Q(\mathbf{m}))^2, \qquad (4.2.12)$$

where $Q(\mathbf{m}) := \frac{\|\mathbf{m}\|_2^2}{n}$.

Theorem 2.1 of Celentano et al. (2023) rigorously establishes that this TAP free energy correctly describes the asymptotic landscape of the problem. In particular, it states that the Bayes estimator is arbitrarily close to a local minimum of $\mathcal{F}_{\text{TAP}}$. Our focus now shifts to minimizing this TAP free energy. This corresponds to solving the TAP equations derived in Eq. (3.1.23), and we will now explore methods for obtaining their solutions.

4.3 An Iterative solution to the TAP Equations

Such a solution to the TAP equations originated from Bolthausen (2012)'s seminal work on constructing an iterative solution to the TAP equations of the SK model with parameters (β, h) (see Eqs. (3.1.23)). It was shown that in the RS phase, the following iterative scheme

$$\mathbf{m}^{(k+1)} = \tanh\left(h + \beta\mathbf{A}\mathbf{m}^{(k)} - \beta^2(1-q)\mathbf{m}^{(k-1)}\right), \quad \mathbf{m}^{(0)} = \mathbf{0} \qquad (4.3.1)$$

with q as in Eq. (3.1.14) approaches a fixed point, yielding a solution to Eq. (3.1.23). The main novelty introduced in this paper was the use of *Gaussian conditioning*, which we will present formally in the subsequent section.

Approximate Message Passing

Rather than following Bolthausen's derivation, we will follow that of a more general approach presented in Bayati and Montanari (2011).

Let $\mathbf{A} = \mathbf{G} + \mathbf{G}^T$, where $\mathbf{G} \in \mathbb{R}^{n\times n}$ is a Gaussian matrix (see Definition A.1.1). Now we consider the iteration

$$\mathbf{h}^{(k+1)} = \beta \mathbf{A}\mathbf{m}^{(k)} - \mathsf{b}_k \mathbf{m}^{(k-1)}, \tag{4.3.2}$$
$$\mathbf{m}^{(k)} = f(\mathbf{h}^{(k)}), \tag{4.3.3}$$

for a non-linear function $f : \mathbb{R} \to \mathbb{R}$ applied component-wise to the input vector, and $\mathbf{m}^{-1} = \mathbf{0}$. Clearly, past the second iteration, the correlation between the $\mathbf{m}^{(k)}$'s and the matrix $\mathbf{A}$ is non-zero and will continue to increase. This makes the task of "simply" expressing the iterates $\mathbf{h}^{(k)}, \mathbf{m}^{(k)}$ seemingly impossible. The main innovation by Bolthausen was to consider the σ-algebra of past iterations, which we will denote

$$\mathscr{F}_k = \{\mathbf{m}^0, \mathbf{m}^{(1)}, \dots, \mathbf{m}^{(k)}\}.$$

Noticing that conditioning on the knowledge of $\mathscr{F}_k$ is equivalent to conditioning on knowledge of

$$\mathscr{E}_k = \left\{\mathbf{h}^{(1)} + \mathsf{b}_0 \mathbf{m}^{(-1)} = \mathbf{A}\mathbf{m}^{(0)}, \dots, \mathbf{h}^{(k)} + \mathsf{b}_{k-1}\mathbf{m}(k-1) = \mathbf{A}\mathbf{m}^{(k-1)}\right\},$$

which is a linear (or affine, depending on the point of view) observation of the columns of $\mathbf{A}$. Conditioning a Gaussian matrix on linear observations of itself results in a new Gaussian vector plus some non-Gaussian terms. The key idea is to choose $\{\mathsf{b}_k\}_{k\geq 0}$ in such a way that these non-Gaussian terms vanish, allowing for a simpler Gaussian analysis of the iterates $\mathbf{h}^k$ and $\mathbf{m}^k$. Before getting into technicalities and formal statements, we will lay out a few important quantities and notions.

Definition 4.3.1 (Pseudo-Lipschitz Function)
For $k \geq 1$, we say a function $\psi : \mathbb{R}^m \to \mathbb{R}$ is ***pseudo-Lipschitz*** *of order k and denote it $\psi \in \mathrm{PL}(k)$ if there exists a constant $L > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$,*

$$|\psi(\mathbf{x}) - \psi(\mathbf{y})| \leq L\left(1 + \|\mathbf{x}\|_2^{k-1} + \|\mathbf{y}\|_2^{k-1}\right)\|\mathbf{x} - \mathbf{y}\|_2. \tag{4.3.4}$$

For the iterates $\{\mathbf{m}^{(0)}, \mathbf{m}^{(1)}, \dots, \mathbf{m}^{(k-1)}\}$, we define the matrix

$$\mathbf{M}_k := \left[\mathbf{m}^{(0)}|\mathbf{m}^{(1)}|\dots|\mathbf{m}^{(k-1)}\right]. \tag{4.3.5}$$

Associated to this matrix, we will define the projections of $\mathbf{m}^{(k)}$ along and perpendicular to its column space as $\mathbf{m}_{\parallel}^{(k)}$ and $\mathbf{m}_{\perp}^{(k)}$, respectively. We will denote $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)$ the coefficient of the first projection, namely

$$\mathbf{m}_{\parallel}^{(k)} := \sum_{i=1}^{k-1} \alpha_i \mathbf{m}^{(i)}, \tag{4.3.6}$$

$$\mathbf{m}_{\perp}^{(k)} := \mathbf{m}^{(k)} - \mathbf{m}_{\parallel}^{(k)}. \tag{4.3.7}$$

Finally, we define a sequence $\{\tau_k\}_{k\geq 1}$ recursively

$$\begin{cases} \tau_1^2 := \frac{1}{n}\|\mathbf{m}^{(1)}\|_2^2, \\ \tau_{k+1}^2 := \mathbb{E}\left[(f(\tau_k Z))^2\right], \end{cases} \tag{4.3.8}$$

with $Z \sim \mathcal{N}(0,1)$, and f the same non-linear function as above. This recursion may seem quite arbitrary, but we will show that it is *absolutely fundamental* to the analysis of AMP algorithms. With these definitions in hand, we are now ready to summarize the results of Bayati and Montanari (2011).

Gaussian Conditioning & State Evolution

Lemma 4.3.1 (Lemma 15, Bayati and Montanari)
Let $\{\mathbf{G}(n)\}_n$ be a sequence of Gaussian matrices $\mathbf{G} \in \mathbb{R}^{n\times n}$ indexed by n. Then the following hold for all $k \in \mathbb{N}$:

i.)

$$\mathbf{h}^{(k)}|_{\mathscr{E}_k} \overset{d}{=} \sum_{i=1}^{k-1} \alpha_i \mathbf{h}^{(i)} + \tilde{\mathbf{G}}\mathbf{m}_{\perp}^{(k)} + \tilde{\mathbf{M}}_{k-1}\vec{o}(1), \tag{4.3.9}$$

where $\tilde{\mathbf{G}}$ is an independent copy of $\mathbf{G}$ and the coefficients α_i satisfy $\mathbf{m}_{\parallel}^{(k)} = \sum_{i=1}^{k-1} \alpha_i \mathbf{m}^{(i)}$. The matrix $\tilde{\mathbf{M}}_k$ is s.t. the columns form an orthogonal basis of the column space of $\mathbf{M}_k$, and $\tilde{\mathbf{M}}_k^T \tilde{\mathbf{M}}_k = n\mathbf{I}_k$.

ii.) For any $\phi : \mathbb{R}^k \to \mathbb{R}$ with $\phi \in \mathrm{PL}(2)$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \phi(h_i^{(2)}, \dots, h_i^{(k)}) \overset{\text{a.s.}}{=} \mathbb{E}\left[\psi\left(\tau_1 Z_1, \dots, \tau_k Z_k\right)\right], \tag{4.3.10}$$

where $Z_1, \dots Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

iii.) For all $1 \leq r, s \leq k$, the following holds

$$\lim_{n\to\infty} \frac{1}{n}\langle \mathbf{h}^{(r+1)}, \mathbf{h}^{(s+1)} \rangle \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n}\langle \mathbf{m}^{(r)}, \mathbf{m}^{(s)} \rangle, \tag{4.3.11}$$

and all limits exist, are bounded, and are constant random variables.

iv.) For all $1 \le r, s \le k$, and for any Lipschitz function ψ, the following holds

$$\lim_{n\to\infty} \frac{1}{n} \langle \mathbf{h}^{(r+1)}, \psi(\mathbf{h}^{(s+1)}) \rangle \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n} \langle \mathbf{h}^{(r+1)}, \mathbf{h}^{(s+1)} \rangle \langle \psi'(\mathbf{h}^{(s+1)}) \rangle, \tag{4.3.12}$$

and all limits exist, are bounded and are constant random variables.

v.) For all $0 \le r \le k$, the following limit exists and there are positive constants ρ_r such that

$$\lim_{n\to\infty} \frac{1}{n} \langle \mathbf{m}_\perp^{(r)}, \mathbf{m}_\perp^{(r)} \rangle \overset{\text{a.s.}}{>} \rho_r. \tag{4.3.13}$$

For a proof of this Lemma, we refer to Section 3 of Bayati and Montanari (2011). The particular instance of Lemma 4.3.1 is a simpler version of their main result, presented in their Section 4. With this characterization, we can now give the main result of AMP: the state evolution (SE).

Theorem 4.3.1 (Theorem 4, Bayati and Montanari)
Let $\{\mathbf{G}(n)\}_n$ be a sequence of Gaussian matrices in the previous Lemma. Then, for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ and all $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(m_i^{(k)}\right) \overset{\text{a.s.}}{=} \mathbb{E}\left[\psi\left(f(\tau_k Z)\right)\right], \tag{4.3.14}$$

with τ_k as in Eq. (4.3.8).

This characterization of AMP iterates is paramount, as thanks to the Gaussian conditioning trick, we are able to show that the AMP iterates are asymptotically Gaussian, allowing for much simpler analysis. This phenomenon is closely related to density evolution in message-passing algorithms (Richardson and Urbanke (2008)), giving AMP its name and making it a powerful tool for high-dimensional inference.

Remark 4.2 (Application to Signal Recovery)
Message-passing algorithms have historically been widely used compressed sensing, making a significant appearance in a paper by Donoho, Maleki, and Montanari (2009). Later, a rigorous analysis was proposed by Bayati and Montanari (2011), inspired by Bolthausen's conditioning trick. In the latter, the conditioning trick was extended to signal recovery schemes involving Gaussian matrices, i.e. inverse problems of the kind

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n), \tag{4.3.15}$$

where $\mathbf{G}$ is Definition A.1.1. The application of AMP for compressed sensing problems is quite fascinating, but unfortunately does not fit the current topic. We defer the interested reader to Appendix C.1 for a 2-step tilted mean estimation, which was unfortunately unfruitful in our case, but could be interesting in different scenarios.

Remark 4.3
The conditioning trick was later shown to be applicable to a much wider class of random matrices. Validity for rotationally invariant matrices (see Definition A.2.3) was shown by Takeuchi (2017), and the same concept was used to prove tight asymptotics of finite sample-size AMP iterates by Cademartori and Rush (2024).

4.4 Low-Rank Matrix Factorization

We are now finally ready to address the $\mathbb{Z}_2$-synchronization problem introduced at the beginning of the Chapter (see the data generated in Section 4.2). Our approach draws intuition and results from multiple areas, and we first provide a proof for the isotropic case before extending it to the anisotropic setting. Given the interwoven nature of the arguments, a linear exposition would be difficult. To give a high-level understanding of the structure of this section, we provide the reader with a diagram in Figure 4.1.

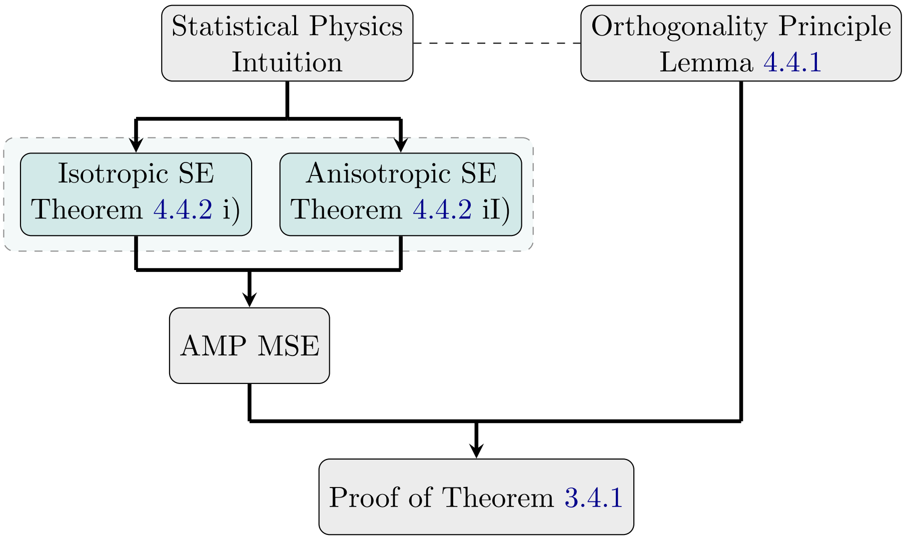


Figure 4.1: Structure of the proof of Theorem 3.4.1, pulling from various sections to come. The dashed line represents a shared intuition and heuristic, which will lead to rigorous proofs. The flow of the proof is vertical, but some notion of "parallelism" will be required, as we will develop the proofs for isotropic and non-isotropic side-information simultaneously.

Statistical Physics Intuition & the Orthogonality Principle

Before presenting the rigorous results, we first offer a heuristic derivation. We note that this derivation is *not intended as a rigorous proof. It is highly hand-wavey, but offers good insight into the problem.* Readers solely interested in rigorous proofs may skip this Section. We begin with a well-known result from Bayesian Statistics.

Lemma 4.4.1 (Orthogonality Principle)
Let $\mathbf{x} \in \mathbb{R}^n$ be a random signal and let $\mathbf{y} \in \mathbb{R}^m$ be the corresponding observations, with the posterior Bayes estimator w.r.t. the MSE (4.2.1) given by

$$\mathbf{m} := \mathbb{E}[\mathbf{x}|\mathbf{y}].$$

Then, the estimation error $\mathbf{x} - \mathbf{m}$ is orthogonal to the estimate $\mathbf{m}$ in expectation, implying the variance decomposition:

$$\mathbb{E}[\|\mathbf{x} - \mathbf{m}\|_2^2] = \mathbb{E}\left[\|\mathbf{x}\|_2^2\right] - \mathbb{E}\left[\|\mathbf{m}\|_2^2\right]. \tag{4.4.1}$$

Proof. See Appendix B.3. □

Corollary 4.4.1
For the data generated in $\mathbb{Z}_2$-synchronization, we have by Lemma 4.4.1 that

$$\frac{1}{n}\mathbb{E}[\|\mathbf{x} - \mathbf{m}\|_2^2] = 1 - q, \tag{4.4.2}$$

with

$$q := \frac{1}{n}\mathbb{E}[\|\mathbf{m}\|_2^2]$$

the order parameter of the planted model.

From this useful decomposition, we will aim to use tools from Statistical Physics (in particular, Section 3.1) to compute the order parameter q of the planted model in both the isotropic and anisotropic cases. Recall that in the planted model, our disorder is given by $\mathbf{A} = \mathbf{W} + \frac{\beta}{n}\bar{\mathbf{x}}\bar{\mathbf{x}}^T$, where we will now differentiate the planted signal $\bar{\mathbf{x}}$ from the signal $\mathbf{x}$ we test against. The Hamiltonian is now given by

$$H(\mathbf{x}) := \sum_{i<j}\left(W_{ij} + \frac{\beta}{n}\bar{x}_i\bar{x}_j\right)x_i x_j + \sum_{i=1}^{n} y_i x_i. \tag{4.4.3}$$

Using the same MFT tricks from Section 3.1, we obtain the effective field on the i-th spin[1]

$$\beta h_{\text{eff}}^{(i)} := \frac{\beta^2}{n}\sum_{j=1}^{n}\bar{x}_j x_j \bar{x}_i + \beta\sum_{i=1}^{n} W_{ij} m_j + y_i. \tag{4.4.4}$$

We now use Proposition 3.1.1 (the Replica Ansatz) to claim that the overlap between the planted and test signals is equal to q, i.e. $\frac{1}{n}\langle\bar{\mathbf{x}}, \mathbf{x}\rangle = q$.

Next, we claim that the sum

$$\beta\sum_{j=1}^{n} W_{ij} m_j \approx \beta\sqrt{q}W$$

[1] Recall that we absorb β into the external field, here $\mathbf{y}$.

is approximately Gaussian $W \sim \mathcal{N}(0,1)$ with variance $\frac{1}{n}\sum_{i=1}^{n}\beta^2 m_i^2 = \beta^2 q$, due to row-wise independence of the W_{ij} terms and the Replica Ansatz again. We thus have that

$$\beta h_{\text{eff}}^{(i)} = \beta^2 q\bar{x}_i + \beta\sqrt{q}W + y_i. \tag{4.4.5}$$

Turning to our Bayesian estimation problem, we are still interested in recovering $\bar{\mathbf{x}}$. The way we will use our mean field derivation is to consider that the mean field is our *observation* of the signal, from which we will derive our posterior / Bayes mean. We first derive a general formulation for the posterior mean before treating the isotropic and non-isotropic cases separately.

Proposition 4.4.1 (Gaussian Mean Field)
Let $H = a\bar{X} + \sqrt{b}Z$ be an observation of a Rademacher RV $\bar{X}$, where $Z \sim \mathcal{N}(0,1)$ is independent of $\bar{X}$, and $a, b > 0$. Then, the posterior mean of $\bar{X}$ given H is given by

$$\mathbb{E}[\bar{X}|H] = \tanh\left(\frac{a}{b}H\right) \tag{4.4.6}$$

Proof. See Appendix B.4. □

Isotropic y

In this case, $\mathbf{y} = t\bar{\mathbf{x}} + \sqrt{t}\mathbf{z}$, which when plugged into Eq. (4.4.5) yields

$$\beta h_{\text{eff}}^{(i)} = \beta^2 q\bar{x}_i + \beta\sqrt{q}W + t\bar{x}_i + \sqrt{t}z_i, \tag{4.4.7}$$

which by assumed independence of W and z_i yields

$$\beta h_{\text{eff}}^{(i)} = (\beta^2 q + t)\bar{x}_i + \sqrt{\beta^2 q + t}G, \quad G \sim \mathcal{N}(0,1). \tag{4.4.8}$$

Notice that this matches the Gaussian mean field assumed in Proposition 4.4.1. We thus get that the i-th coordinate of the Bayes estimator is given by

$$m_i = \tanh\left(\frac{\beta^2 q + t}{\beta^2 q + t}\beta h_{\text{eff}}^{(i)}\right) = \tanh\left((\beta^2 q + t)\bar{x}_i + \sqrt{\beta^2 q + t}G\right) \tag{4.4.9}$$

Applying a strong law of large numbers (SLLN) allows us to compute the expected norm of $\mathbf{m}$, giving us

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m_i^2 = \mathbb{E}_{\bar{X},Z}\left[\tanh^2\left((\beta^2 q + t)\bar{X} + \sqrt{\beta^2 q + t}Z\right)\right], \tag{4.4.10}$$

where $\bar{X} \sim \bar{\nu}$. A simple calculation leads to the final expression

$$q := \lim_{n\to\infty}\frac{1}{n}\mathbb{E}[\|\mathbf{m}\|_2^2] = \mathbb{E}_Z\left[\tanh^2\left((\beta^2 q + t) + \sqrt{\beta^2 q + t}Z\right)\right] \tag{4.4.11}$$

for the isotropic case.

Anisotropic $\mathbf{y}$

In the case, $\mathbf{y} = \mathbf{Q}t\bar{\mathbf{x}} + \sqrt{\mathbf{Q}t}\mathbf{z}$, with $\mathbf{Q} = \lambda \mathbf{I}_n + \mathbf{W} + \frac{\beta}{n}\bar{\mathbf{x}}\bar{\mathbf{x}}^T$. We can begin by re-writing it as

$$\mathbf{y} = (\lambda + \beta)\, t\bar{\mathbf{x}} + t\mathbf{W}\bar{\mathbf{x}} + \sqrt{\mathbf{Q}t}\mathbf{z}. \tag{4.4.12}$$

Furthermore, we will offer the following representation for the noise term.

Proposition 4.4.2
For $\mathbf{Q} = \lambda \mathbf{I}_n + \mathbf{W} + \frac{\beta}{n}\bar{\mathbf{x}}\bar{\mathbf{x}}^T$ *and* $\mathbf{z}$ *a standard* n*-dimensional Gaussian vector, then*

$$\frac{1}{n}\|\sqrt{\mathbf{Q}t}\mathbf{z} - \sqrt{\lambda t}\mathbf{z}\|_2 \in \mathcal{O}\left(\frac{1}{2\sqrt{\lambda n}}\right). \tag{4.4.13}$$

Proof. See Appendix B.5.

□

We will use Proposition 4.4.2 as a reasonable approximation, and plug Eq. (4.4.12) into Eq. (4.4.5).

$$\begin{aligned} \beta h_{\text{eff}}^{(i)} &= \beta^2 q\bar{x}_i + \beta\sum_{j=1}^{n} W_{ij}m_j + (\lambda+\beta)t\bar{x}_i + t\sum_{j=1}^{n} W_{ij}\bar{x}_j + \sqrt{\lambda t}z_i \\ &= \left(\beta^2 q + (\lambda+\beta)t\right)\bar{x}_i + \sum_{j=1}^{n} W_{ij}\left(\beta m_j + t\bar{x}_j\right) + \sqrt{\lambda t}z_i. \end{aligned} \tag{4.4.14}$$

Applying the same Gaussian approximation trick for the sum $\sum_{j=1}^{n} W_{ij}\left(\beta m_j + t\bar{x}_j\right)$, we can consider it as the sum of n independent Gaussians $W_j \sim \mathcal{N}(0, \sigma_j^2)$ with

$$\sigma_j^2 = \frac{1}{n}\left(\beta^2 m_j^2 + 2\beta t m_j\bar{x}_j + t^2\underbrace{\bar{x}_j^2}_{=1}\right).$$

Thus, we obtain the Gaussain approximation

$$\sum_{j=1}^{n} W_{ij}\left(\beta m_j + t\bar{x}_j\right) \approx \sqrt{\beta^2 q + 2\beta t\frac{1}{n}\langle \mathbf{m}, \bar{\mathbf{x}}\rangle + t^2}W, \tag{4.4.15}$$

with $W \sim \mathcal{N}(0,1)$. By Lemma 4.4.1, we have that $n^{-1}\mathbb{E}[\langle \mathbf{m}, \bar{\mathbf{x}}\rangle] = n^{-1}\mathbb{E}[\|\mathbf{m}\|_2^2] = q$, which allows for the final simplification of the mean field into

$$\beta h_{\text{eff}}^{(i)} = \left(\beta^2 q + (\lambda+\beta)t\right)\bar{x}_i + \sqrt{\beta^2 q + 2\beta t q + t^2 + \lambda t}Z. \tag{4.4.16}$$

Plugging this Gaussian mean field into Proposition 4.4.1 leads to the Bayes mean

$$m_i = \tanh\left(\frac{\left(\beta^2 q + (\lambda+\beta)t\right)}{\beta^2 q + 2\beta t q + t^2 + \lambda t}\left(\left(\beta^2 q + (\lambda+\beta)t\right)\bar{x}_i + \sqrt{\beta^2 q + 2\beta t q + t^2 + \lambda t}Z\right)\right). \tag{4.4.17}$$

We then use the same SLLN trick as in Eq (4.4.10), leading to the anisotropic order parameter

$$q = \mathbb{E}\left[\tanh^2\left(\frac{\left(\beta^2 q + (\lambda+\beta)t\right)^2}{\beta^2 q + 2\beta t q + t^2 + \lambda t} + \frac{\left(\beta^2 q + (\lambda+\beta)t\right)}{\sqrt{\beta^2 q + 2\beta t q + t^2 + \lambda t}} Z\right)\right]. \tag{4.4.18}$$

AMP State Evolution

We now take a step back from the Statistical Physics aspect, and consider the AMP algorithm. Continuing with the introduced notation of the planted signal being $\bar{\mathbf{x}} \in \mathbb{R}^n$ such that $\bar{x}_i \overset{i.i.d.}{\sim} \bar{\nu}$, we have our rank-one matrix given by

$$\mathbf{A} = \frac{\beta}{n}\bar{\mathbf{x}}\bar{\mathbf{x}}^T + \mathbf{W}, \quad \mathbf{W} \sim \mathrm{GOE}(n), \tag{4.4.19}$$

and a side information vector

$$\mathbf{y}(t) = \begin{cases} t\bar{\mathbf{x}} + \sqrt{t}\mathbf{z}, & \text{in the isotropic case,} \\ \mathbf{Q}t\bar{\mathbf{x}} + \sqrt{\mathbf{Q}t}\mathbf{z}, & \text{in the anisotropic case.} \end{cases} \tag{4.4.20}$$

In order to prove results for Algorithm 3, we will demonstrate and prove a more general case of AMP, from which results for our case will follow. Consider the iteration:

$$\mathbf{x}^{(k+1)} = \beta \mathbf{A} f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{k-1}), \tag{4.4.21}$$

with $f_k : \mathbb{R} \to \mathbb{R}$ a Lipschitz function applied element-wise to the input vector, and b_k given by

$$\mathsf{b}_k = \beta^2 \langle f_k'(\mathbf{x}^{(k)}) \rangle := \frac{\beta^2}{n}\sum_{i=1}^{n} f_k'(\mathbf{x}^{(k)}), \tag{4.4.22}$$

where g' denotes the weak derivative of a Lipschitz function $g : \mathbb{R} \to \mathbb{R}$. As hinted towards in Theorem 4.3.1, the state evolution of AMP depends on the non-linearity f_k. The same idea holds for this rank-one matrix AMP, which we synthesize in the following.

Theorem 4.4.1
Consider the spiked GOE matrix $\mathbf{A}$ *from the planted model from Eq.* (4.4.19)*, the side information* $\mathbf{y}$ *from Eq.* (4.4.20) *(either one), and the iteration from Eq.* (4.4.21)*, initialized at* $\mathbf{x}^{(-1)} = \mathbf{0}_n$ *and* $f_0(\mathbf{x}^{(-1)}) = \mathbf{0}_n$*. Assume* f_k *to be Lipschitz continuous for all* $k \in \mathbb{N}$*.*

i.) For $\mathbf{y}$ *isotropic, consider* $(\mu_k, \sigma_k)_{k\geq 0}$*, defined recursively as*

$$\mu_{k+1} := \beta^2 \mathbb{E}\left[\bar{X} f_k((\mu_k + t)\bar{X} + \sqrt{\sigma_k^2 + t}G)\right], \tag{4.4.23}$$

$$\sigma_{k+1}^2 := \beta^2 \mathbb{E}\left[f_k((\mu_k + t)\bar{X} + \sqrt{\sigma_k^2 + t}G)^2\right], \tag{4.4.24}$$

ii.) For $\mathbf{y}$ *anisotropic, consider* $(\mu_k, \sigma_k)_{k\geq 0}$*, defined recursively as*

$$\mu_{k+1} := \beta^2 \mathbb{E}\left[\bar{X} f_k\left((\mu_k + (\lambda+\beta)t)\bar{X} + \sqrt{\sigma_k^2 + \lambda t}G\right)\right], \tag{4.4.25}$$

$$\sigma_{k+1}^2 := \beta^2 \mathbb{E}\left[\left(f_k\left((\mu_k + (\lambda+\beta)t)\bar{X} + \sqrt{\sigma_k^2 + \lambda t}G\right) + t\bar{X}\right)^2\right], \tag{4.4.26}$$

initialized at $\mu_0 = \sigma_0 = 0$*, where* $G \sim \mathcal{N}(0,1)$ *is a Gaussian independent of* $\bar{X} \sim \bar{\nu}$*.*

Then, for any pseudo-Lipschitz function $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *of order 2, we have*

i.)

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i^{(k)}, \bar{x}_i\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left((\mu_k + t)\bar{X} + \sqrt{\sigma_k^2 + t}G, \bar{X}\right)\right], \tag{4.4.27}$$

and

ii.)

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i^{(k)}, \bar{x}_i\right) = \mathbb{E}\left[\psi\left((\mu_k + (\lambda+\beta)t)\bar{X} + \sqrt{\sigma_k^2 + \lambda t}G, \bar{X}\right)\right], \tag{4.4.28}$$

for the isotropic and anisotropic case respectively.

Note 2
We note that the isotropic case holds almost surely, whereas the anisotropic case holds in probability. We believe this to be a proof artifact, and that extending to almost sure convergence is possible.

Proof sketch of i). We will provide a heuristic idea of the proof as it is both interesting and insightful, but will defer the rigorous computations to Appendix B.6.

First, we decompose the iteration (4.4.21) as follows:

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \beta \mathbf{A} f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}) \\
&= \beta \left(\mathbf{W} + \frac{\beta}{n} \bar{\mathbf{x}} \bar{\mathbf{x}}^T \right) f_k(\mathbf{x}^{(k)}) + t\bar{\mathbf{x}} + \sqrt{t}\mathbf{z} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}) \\
&= \left(\frac{\beta^2}{n} \langle f_k(\mathbf{x}^{(k)}, \bar{\mathbf{x}} \rangle + t \right) \bar{\mathbf{x}} + \mathbf{W} \beta f_k(\mathbf{x}^{(k)}) + \sqrt{t}\mathbf{z} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}). \qquad (4.4.29)
\end{aligned}$$

Notice that by grouping the signal terms (factors of $\bar{\mathbf{x}}$) we obtain a term that in the large system limit would resemble

$$\lim_{n \to \infty} \left(\frac{\beta^2}{n} \langle f_k(\mathbf{x}^{(k)}, \bar{\mathbf{x}} \rangle + t \right) = \beta^2 \mathbb{E}[\bar{X} f_k(X_k)] + t, \qquad (4.4.30)$$

where X_k is the empirical distribution of the iterates $\mathbf{x}^{(k)}$. Assuming that it has the form given in i.), we see that this term is none other that $\mu_{k+1} + t$, matching the factor of the signal $\bar{X}$.

Next, we consider the terms

$$\mathbf{W} \beta f_k(\mathbf{x}^{(k)}) - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}), \qquad (4.4.31)$$

which match those of a "standard" AMP iteration from Eq. (4.3.2). This AMP's SE is given by Theorem 4.3.1, giving us the σ_{k+1}^2 term, being the variance of a Gaussian RV W. By assumed independence of $\sqrt{t}\mathbf{z}$ and $\sigma_k W$, we obtain the $\sqrt{\sigma_k^2 + t} G$ Gaussian term, acting as the noise. □

Proof sketch of ii). Again, we will provide some intuition, but the full proof is deferred to Appendix B.7.

We apply the same decomposition as for i), but by taking account for anisotropic $\mathbf{y}$ which gives us

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \beta \mathbf{A} f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}) \qquad (4.4.32) \\
&\overset{a)}{=} \frac{\beta^2}{n} \langle f_k(\mathbf{x}^{(k)}), \bar{\mathbf{x}} \rangle \bar{\mathbf{x}} + \beta \mathbf{W} f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}).
\end{aligned}$$

Then, applying Proposition 4.4.2 allows to approximate $\sqrt{\mathbf{Q}t}\mathbf{z}$ as $\sqrt{\lambda}\mathbf{z}$. We do the same regrouping as for point i), leading us to the desired state evolution. Convergence in probability holds versus almost-sure convergence due to the use of Proposition 4.4.2. □

This characterization is clearly very powerful, and we will exploit it to choose the optimal non-linearity f_k, leading us to *Bayes AMP*. Using Theorem 4.4.1, we can define a performance metric for our AMP algorithm.

Corollary 4.4.2 (Corollary 3.2, Feng, Venkataramanan, Rush, and Samworth (2021))
Considering the normalized ℓ_2-norm, and the random variable

$$X_k := \begin{cases} (\mu_k + t)\bar{X} + \sqrt{\sigma_k^2 + t}G, \\ (\mu_k + (\lambda + \beta)t)\bar{X} + \sqrt{\sigma_k^2 + \lambda t}G, \end{cases}$$

in the isotropic resp. anisotropic case, we can characterize the squared error

$$\frac{1}{n}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2^2 = \mathbb{E}\left[(f_k(X_k))^2\right], \tag{4.4.33}$$

and the empirical correlation of AMP

$$\frac{|\langle \mathbf{x}^{(k)}, \bar{\mathbf{x}}\rangle|}{\|\mathbf{x}^{(k)}\|_2 \cdot \|\bar{\mathbf{x}}\|_2} = \frac{|\mathbb{E}[\bar{X} f_k(X_k)]|}{\sqrt{\mathbb{E}[f_k(X_k)^2]}} \tag{4.4.34}$$

It follows that choosing

$$f_k(X_k) = \mathbb{E}[\bar{X}|X_k], \tag{4.4.35}$$

Indeed, by orthogonality of Bayes estimators, we have that

$$\mathbb{E}[\bar{X} f_k(X_k)] = \mathbb{E}[f_k^2(X_k)]. \tag{4.4.36}$$

i.e. the Bayes denoiser, leads to a minimization resp. maximization of Eqs. (4.4.33) resp. (4.4.34). We note in passing that in the case of $\mathbb{Z}_2$-synchronization with isotropic side information, we have

$$f_k(X_k) = \tanh(X_k), \tag{4.4.37}$$

which thus turns Eq. (4.4.21) into lines 1-5 of Algorithm 2, and we will use $\widehat{\mathbf{m}}^{(k)} = \mathbf{x}^{(k)}$ interchangeably from here on out. AMP with the non-linearities chosen as in (4.4.35) is called *Bayes AMP*.

Reduction to One-Dimensional SE

We may now adapt the result of Theorem 4.4.1 with the denoiser chosen in Eq. (4.4.37).

Theorem 4.4.2
Consider the setup of Theorem (4.4.1), but now that $f_k(X_k) = \mathbb{E}[\bar{X}|X_k]$. Then,

i.) For isotropic $\mathbf{y}$, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i^{(k)}, \bar{x}_i\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left((\gamma_k + t)\bar{X} + \sqrt{\gamma_k^2 + t}G, \bar{X}\right)\right], \tag{4.4.38}$$

for $G \sim \mathcal{N}(0,1)$ independent of $\bar{X} \sim \bar{\nu}$, and γ_k is recursively defined as

$$\gamma_{k+1} = \beta^2 \mathbb{E}\left[\tanh^2\left((\gamma_k + t) + \sqrt{\gamma_k + t}G\right)\right] \tag{4.4.39}$$

initialized at $\gamma_0 = 0$.

For anisotropic $\mathbf{y}$*, we have*

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i^{(k)}, \bar{x}_i\right) = \mathbb{E}\left[\psi\left(\frac{(\gamma_k + (\lambda+\beta)t)^2}{\gamma_k + 2t\beta^{-1}\gamma_k + t^2 + \lambda t}\bar{X} + \sqrt{\frac{(\gamma_k + (\lambda+\beta)t)^2}{\gamma_k + 2t\beta^{-1}\gamma_k + t^2 + \lambda t}}G, \bar{X}\right)\right], \tag{4.4.40}$$

for $G \sim \mathcal{N}(0,1)$ independent of $\bar{X} \sim \bar{\nu}$, and γ_k is recursively defined as

$$\gamma_{k+1} = \beta^2 \mathbb{E}\left[\tanh^2\left(\frac{(\gamma_k + (\lambda+\beta)t)^2}{\gamma_k + 2t\beta^{-1}\gamma_k + t^2 + \lambda t} + \sqrt{\frac{(\gamma_k + (\lambda+\beta)t)^2}{\gamma_k + 2t\beta^{-1}\gamma_k + t^2 + \lambda t}}G\right)\right] \tag{4.4.41}$$

initialized at $\gamma_0 = 0$.

Proof. We split the proof into two parts again.

i.) Using (4.4.36), it follows from their definitions that

$$\mu_{k+1} := \beta^2 \mathbb{E}[\bar{X} f_k(X_k)] = \beta^2 \mathbb{E}[f_k^2(X_k)] =: \sigma_{k+1}^2,$$

allowing direct substitution with γ_{k+1}. We can remove the expectation over $\bar{X}$ due to symmetry and parity of $\tanh^2(\cdot)$.

ii.) In this case, notice that by Proposition 4.4.1,

$$f_k(X_k) = \tanh\left(\frac{\mu_k + (\lambda+\beta)t}{\sigma_k^2 + \lambda t} X_k\right),$$

which correctly leads us to Algorithm 3. We simply expand the definition of σ_{k+1}^2, which allows the formulation

$$\begin{aligned}\sigma_{k+1}^2 &:= \mathbb{E}\left[\left(\beta f_k(X_k) + t\bar{X}\right)^2\right] \\ &= \underbrace{\beta^2 \mathbb{E}\left[f_k(X_k)^2\right]}_{:=\gamma_{k+1}} + 2t\underbrace{\beta \mathbb{E}[\bar{X} f_k(X_k)]}_{:=\beta^{-1}\gamma_{k+1}} + t^2 \underbrace{\mathbb{E}\left[\bar{X}^2\right]}_{=1} = \gamma_{k+1} + 2t\beta^{-1}\gamma_k + t^2.\end{aligned}$$

Plugging this into f_k leads to the desired state evolution.

□

Remark 4.4
Now that we have shown our state evolutions respectively, assuming that they both have a single fixed-point, we see that they read

i.)

$$\gamma_* = \beta^2 \mathbb{E}\left[\tanh^2\left(\gamma_* + t + \sqrt{\gamma_* + t} G\right)\right], \tag{4.4.42}$$

and

ii.)

$$\gamma_* = \beta^2 \mathbb{E}\left[\tanh^2\left(\frac{(\gamma_* + (\lambda+\beta)t)^2}{\gamma_* + 2t\beta^{-1}\gamma_* + t^2 + \lambda t} + \sqrt{\frac{(\gamma_* + (\lambda+\beta)t)^2}{\gamma_* + 2t\beta^{-1}\gamma_* + t^2 + \lambda t}} G\right)\right], \tag{4.4.43}$$

both of which match their heuristic derivations using Statistical Physics, from Eqs. (4.4.11) *and* (4.4.18) *by the change of variable* $\gamma_* = \beta^2 q$.

Using Corollary 4.4.2 we can now characterize the error made by AMP at each iteration using state evolution.

Proposition 4.4.3
Let $\widehat{\mathbf{m}}^{(k)}$ *denote the* k*-th iterate of Bayes AMP (aka Algorithm 2 resp. 3), then the average error at each timestep* k *is given by*

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\mathbb{E}[\|\bar{\mathbf{x}} - \widehat{\mathbf{m}}^{(k)}\|_2^2] = 1 - \frac{\gamma_{k+1}}{\beta^2}. \tag{4.4.44}$$

Moreover, we have

$$\lim_{k\to\infty} \operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\mathbb{E}[\|\widehat{\mathbf{m}}^{(k)} - \bar{\mathbf{x}}\|_2^2] = 1 - \frac{\gamma_*}{\beta^2}, \tag{4.4.45}$$

where γ_* *is the fixed-point of the SE.*

Proof. Considering the PL function $\psi(u,v) = u \cdot v$, we have by Corollary 4.4.2

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\|\bar{\mathbf{x}} - \widehat{\mathbf{m}}^{(k)}\|_2^2] = \mathbb{E}[(f_k(X_k) - \bar{X})^2] = \beta^{-2}\sigma_{k+1}^2 - 2\beta^{-2}\mu_{k+1} + 1 = 1 - \frac{\gamma_{k+1}}{\beta^2}.$$

□

By Lemma 4.4.1 and Remark 4.4, we have thus shown that AMP achieves the Bayes risk.

Proof of Theorem 3.4.1

Finally, we have all the necessary tools to prove the main Theorem. We note that by the bias-variance decomposition, we can write out the error that AMP achieves with respect to the actual Bayes mean $\mathbf{m}(\mathbf{A},\mathbf{y}) \equiv \mathbf{m}$.

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\|\bar{\mathbf{x}}-\widehat{\mathbf{m}}^{(k)}\|_2^2\right]=\lim_{n\to\infty}\left(\frac{1}{n}\mathbb{E}\left[\|\mathbf{m}-\widehat{\mathbf{m}}^{(k)}\|_2^2\right]+\frac{1}{n}\mathbb{E}\left[\|\bar{\mathbf{x}}-\mathbf{m}\|_2^2\right]\right). \tag{4.4.46}$$

Using Lemma 4.4.1 to substitute in the second term of the RHS, and Proposition 4.4.3 to substitute the LHS, we obtain

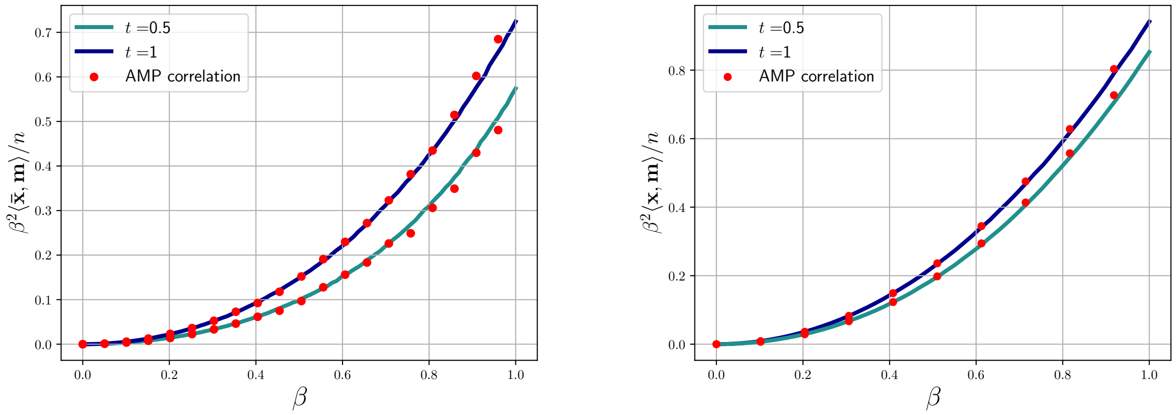
$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\|\mathbf{m}-\widehat{\mathbf{m}}^{(k)}\|_2^2\right]=\left(1-\frac{\gamma_{k+1}}{\beta^2}\right)-\left(1-\frac{\gamma_*}{\beta^2}\right)=\frac{\gamma_*}{\beta^2}-\frac{\gamma_{k+1}}{\beta^2}. \tag{4.4.47}$$

Since $\gamma_k \xrightarrow{k\to\infty} \gamma_*$, we have thus proven the Theorem.

□

4.5 Numerical Simulations

Here we present simulations to verify our hypotheses and exhibit the validity of Theorem 3.4.1. Figures 4.2a and 4.2b report the results of numerical simulations of Algorithms 2 resp. 3. We plot $\gamma_k(\beta,t)$ as a function of β, for two fixed instances of $t\in\{0.5,1\}$. Against this, we plot the empirical correlation of AMP $\rho_k := \frac{\langle\bar{\mathbf{x}},\mathbf{m}^{(k)}\rangle}{n}$, as presented in Corollary 4.4.2. For numerical purposes, we actually plot $\beta^2\rho_k$.



(a) (b)

Figure 4.2: Empirical correlation of AMP for the $\mathbb{Z}_2$-synchronization problem of dimension $n = 2000$ with isotropic (left) and anisotropic (right) side information with $\lambda = 3$, both averaged over 20 trials. The blue lines denotes the theoretical SE, described in Eqs. (4.4.39) resp. (4.4.41). The red dots are the results from simulating AMP.

We see that our AMP iterations follow the expected behaviour, pointing toward applications to open problems.

Conclusion

In this Chapter, we have explored the interplay between AMP, the statistical mechanics of spin glasses, and variational inference, to develop an understanding of the role played by TAP corrections in achieving optimal Bayesian estimation in $\mathbb{Z}_2$-synchronization. By reformulating the MFVI objective through a TAP-inspired modification, we have shown that the optimization landscape aligns with the Bayes-optimal conditions predicted by statistical physics (Section 4.2). The resulting TAP equations establish a connection between high-dimensional inference and the thermodynamic formalism of spin glasses, justifying in hindsight the phase transitions of Fig 3.2a in terms of fixed-point stability of the associated AMP algorithm's state evolution (Section 4.4). While a deeper connection to the RS free energy Φ_{RS} from Eq. (3.1.13) exists, we omitted this connection for the sake of brevity. Interested readers are invited to read Montanari and Venkataramanan (2019); Montanari and Wu (2024); Bolthausen (2019), and more works by the authors of Lesieur, Krzakala, and Zdeborová (2017).

Finally, we have extended the framework of AMP to account for anisotropic side information, resulting in Theorems 4.4.1 and 4.4.2. A very important note is that Theorem 4.4.1 holds for any prior. The assumption on the prior $\bar{\nu}$ can be reduced to mild regularity conditions and a second-moment assumption of $\mathbb{E}[\bar{X}^2] = 1$, which does not impede on the statement's generality. This in turn has allowed us to provide a proof of Theorem 3.4.1 by expressing the MSE in terms of the SE from Proposition 4.4.3. We plan to further employ this Theorem to tackle the open problem of interest – sampling from the sparse SK model – in the next Chapter.

Chapter 5

The Sparse SK Model

5.1 Setup

We introduced the sparse SK model in Definition 3.1.3, and have used it as the primary motivating problem for DDDs (see Fig. 3.2a and surrounding discussion). Formally, the sparse SK model is a type of *diluted* spin glass (see Parisi, Ricci-Tersenghi, and Rizzo (2014)) but with a sparsity constraint on the configuration space (ρ-sparse Boolean hypercube $\mathcal{C}_{\rho}^{n}$), rather than the interactions.

This model is of particular interest to theoretical physicists, as the sparsity constraint may enable modeling of more "realistic" problems. We recall that our sampling algorithm for the SK model (Alg. 2) heavily relies on AMP in the planted model, with results transferring via mutual contiguity of the random and planted measures (see Section 4.1). We expect contiguity to hold for the sparse SK model in an analogous manner, though the range of validity will depend on both β and ρ. We do not provide a contiguity proof and will directly work in the planted case, which is a statistical estimation problem. Our analysis of the AMP algorithm will imply a polynomial time sampler, as explained at the end of Section 3.4.

In this chapter we will present how results from Chapter 4 apply to the sparse planted model. We will begin by deriving the AMP algorithm and SE for a general denoiser f_k, demonstrating that the SE from Theorem 4.4.1 holds in this case too. We then move on to deriving the Bayes AMP algorithm, which will differ for the isotropic and anisotropic cases. Finally, we present a numerical study supporting the gain achieved by DDDs.

5.2 AMP and State Evolution Equations

Planted Model

We begin by recalling the planted model. Let $\bar{\mathbf{x}} \in \mathcal{C}_\rho^n$ be a random vector such that each entry is i.i.d. along $\bar{\nu}_\rho$, the Rademacher-Bernoulli distribution with sparsity $\rho \in (0,1)$. Consider $\mathbf{W} \in \mathbb{R}^{n\times n}$ a GOE matrix, and the observation

$$\mathbf{A} := \frac{\beta}{n}\bar{\mathbf{x}}\bar{\mathbf{x}}^T + \mathbf{W}, \tag{5.2.1}$$

where $\beta > 0$. To this spiked rank-one sparse matrix, we couple an observation vector $\mathbf{y}$, which is given by

$$\mathbf{y}(t) = \begin{cases} t\bar{\mathbf{x}} + \mathbf{B}_t, & \text{isotropic}, \\ \mathbf{Q}t\bar{\mathbf{x}} + \sqrt{\mathbf{Q}t}\mathbf{B}_t, & \text{anisotropic}, \end{cases} \tag{5.2.2}$$

with $\mathbf{Q} := \lambda\mathbf{I}_n + \mathbf{A}$, for some $\lambda > 2$, and $\mathbf{B}_s$ a standard Brownian Motion.

AMP Algorithm

We now notice that by construction, the planted posterior $\mathbb{P}(\mathrm{d}\mathbf{x}|\mathbf{A},\mathbf{y})$ exactly has the same form as the one for the SK model in Eq. (4.2.2). To avoid repetition, we refer the reader to Section 4.2 to understand how we derive the AMP algorithm using variational inference and TAP corrections. We now have the following iteration:

$$\mathbf{x}^{(k+1)} = \beta\mathbf{A}f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}), \tag{5.2.3}$$

where f_k is a separable Lipschitz function, and $b_k := \frac{1}{n}\sum_{i=1}^n f'(x_i^{(k)})$.

We notice that by the same decomposition as in Eqs. (B.6.1) and (B.7.2), we can write the iteration as

$$\mathbf{x}^{(k+1)} = \frac{\beta^2}{n}\langle\bar{\mathbf{x}}, f_k(\mathbf{x}^{(k)})\rangle\bar{\mathbf{x}} + \beta\mathbf{W}f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}). \tag{5.2.4}$$

By applying the same decomposition for isotropic resp. anisotropic $\mathbf{y}$, we obtain an analogous result as Theorem 4.4.1.

Corollary 5.2.1
Let X_k be a random variable such that

$$X_k = \begin{cases} (\mu_k + t)\bar{X} + \sqrt{\sigma_k^2 + t}G, \\ (\mu_k + (\lambda+\rho\beta)t)\bar{X} + \sqrt{\sigma_k^2 + \lambda t}G \end{cases} \tag{5.2.5}$$

in the isotropic resp. anisotropic case with $(\mu_k, \sigma_k)_{k\geq 0}$ *defined as*

$$\begin{cases} \mu_{k+1} = \beta^2 \mathbb{E}\left[\bar{X} f_k(X_k)\right], \\ \sigma_{k+1}^2 := \mathbb{E}\left[(\beta f_k(X_k))^2\right], \end{cases} \tag{5.2.6}$$

resp.

$$\begin{cases} \mu_{k+1} = \beta^2 \mathbb{E}\left[\bar{X} f_k(X_k)\right], \\ \sigma_{k+1}^2 := \mathbb{E}\left[\left(\beta f_k(X_k) + \bar{X} t\right)^2\right]. \end{cases} \tag{5.2.7}$$

Then by Theorem 4.4.1, we have that for any PL(2) function $\psi : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ *the following holds*

i.)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(x_i^{(k)}, \bar{x}_i\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left((\mu_k + t)\bar{X} + \sqrt{\sigma_k^2 + t} G, \bar{X}\right)\right]. \tag{5.2.8}$$

ii.)

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(x_i^{(k)}, \bar{x}_i\right) = \mathbb{E}\left[\psi\left((\mu_k + (\lambda + \rho\beta)t)\bar{X} + \sqrt{\sigma_k^2 + \lambda t} G\right)\right], \tag{5.2.9}$$

with $G \sim \mathcal{N}(0,1)$ *independent of everything else.*

Proof. The only difference from Chapter 4 lies in the fact that $\mathbb{E}[\bar{X}^2] = \rho$ and $\frac{1}{n}\|\bar{\mathbf{x}}\|_2^2 = \rho$, which slightly alters the anisotropic signal $\mathbf{y}$, whence the factor $\rho\beta$. The rest follows from Theorem 4.4.1. □

Reduction via Bayes AMP

Following Eq. (4.4.35), we can tailor our choice of f_k to best suit Bayes-optimal estimation. Let μ_k^t denote either $\mu_k + t$ or $\mu_k + (\lambda + \rho\beta)t$ for the isotropic resp. anisotropic case, and σ_k^t denote $\sigma_k^2 + t$ or $\sigma_k^t + \lambda t$ similarly. The optimal denoiser choice in both cases is given by

$$\begin{aligned} \mathbb{E}[\bar{X} | \mu_k^t \bar{X} + \sqrt{\sigma_k^t} G] &= \frac{\mathbb{P}(\bar{X} = 1 | \mu_k^t + \sqrt{\sigma_k^t} G) - \mathbb{P}(\bar{X} = 1 | -\mu_k^t + \sqrt{\sigma_k^t} G)}{\mathbb{P}(\bar{X} = 1 | \mu_k^t + \sqrt{\sigma_k^t} G) + \mathbb{P}(\bar{X} = 1 | -\mu_k^t + \sqrt{\sigma_k^t} G) + \mathbb{P}(\bar{X} = 0 | \sqrt{\sigma_k^t} G)} \\ &= \frac{\frac{\rho}{2} \exp\left(-\frac{(g-\mu_k^t)^2}{2\sigma_k^t}\right) - \frac{\rho}{2} \exp\left(-\frac{(g+\mu_k^t)^2}{2\sigma_k^t}\right)}{\frac{\rho}{2} \exp\left(-\frac{(g-\mu_k^t)^2}{2\sigma_k^t}\right) + \frac{\rho}{2} \exp\left(-\frac{(g+\mu_k^t)^2}{2\sigma_k^t}\right) + (1-\rho) \exp\left(-\frac{g^2}{2\sigma_k^t}\right)} \\ &= \frac{\frac{\rho}{2} \exp\left(g \frac{\mu_k^t}{\sigma_k^t}\right) - \frac{\rho}{2} \exp\left(-g \frac{\mu_k^t}{\sigma_k^t}\right)}{\frac{\rho}{2} \exp\left(g \frac{\mu_k^t}{\sigma_k^t}\right) + \frac{\rho}{2} \exp\left(-g \frac{\mu_k^t}{\sigma_k^t}\right) + (1-\rho) \exp\left(\frac{(\mu_k^t)^2}{2\sigma_k^t}\right)}, \end{aligned} \tag{5.2.10}$$

with $g = \mu_k^t \bar{X} + \sqrt{\sigma_k^t} G$, $G \sim \mathcal{N}(0,1)$ independent of $\bar{X}$.

Isotropic Case

Recalling that the consequence of such a choice leads to a simpler one-dimensional SE $(\gamma_k)_{k\geq 0}$ with $\gamma_k = \mu_k = \sigma_k^2$ by Eq. (4.4.36); we deduce that in the isotropic case,

$$\mu_k^t := \mu_k + t = \gamma_k + t = \sigma_k^2 + t = \sigma_k^t. \tag{5.2.11}$$

We can thus write the Bayes AMP denoiser in the isotropic case as

$$f_k(X_k) = \frac{\rho \sinh(X_k)}{\rho \cosh(X_k) + (1-\rho)\exp\left(\frac{\gamma_k + t}{2}\right)}, \tag{5.2.12}$$

with $\gamma_{k+1} := \lim_{n\to\infty} \frac{\beta^2}{n} \|\mathbf{m}^{(k)}\|_2^2$ by Corollary 4.4.2. Indeed, we recover the AMP non-linearity for sparse rank-one matrix factorization (Eq. (B13), Ghio et al.), adapted from Lesieur et al. (2017). Likewise, a direct computation of the derivative gives us the component-wise Onsager term

$$\mathsf{b}_k = \frac{\rho^2 + \rho(1-\rho)\exp\left(\frac{\gamma_k+t}{2}\right)\cosh(X_k)}{\left(\rho\cosh(X_k) + (1-\rho)\exp\left(\frac{\gamma_k+t}{2}\right)\right)^2}. \tag{5.2.13}$$

Finally, we can compute the one-dimensional state evolution $(\gamma_k)_{k\geq 0}$. Using (4.4.36) we obtain

$$\begin{aligned}\gamma_{k+1} &:= \beta^2 \mathbb{E}\left[\bar{X} f_k(X_k)\right] \\ &= \beta^2\rho^2 \mathbb{E}_G\left[\frac{\sinh(\gamma_k + t + \sqrt{\gamma_k + t}G)}{\rho\cosh(\gamma_k + t + \sqrt{\gamma_k + t}G) + (1-\rho)\exp\left(\frac{\gamma_k+t}{2}\right)}\right],\end{aligned} \tag{5.2.14}$$

again matching the state evolution computed by Ghio et al. in Eq. (B16). We summarize these derivations in Algorithm 5.

Algorithm 5: Sparse Tilted Mean Estimation: Isotropic Side Information

```
Input: $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathbf{y} \in \mathbb{R}^n$, parameters $\beta, \rho, t > 0$, iteration numbers $K_{\mathrm{AMP}}$
$\widehat{\mathbf{m}}^{-1} = \mathbf{0}_n$,
for $k = 0, \ldots, K_{\mathrm{AMP}} - 1$ do
    $\gamma_k := \frac{1}{n}\|\widehat{\mathbf{m}}^{(k-1)}\|_2^2$,
    $\widehat{\mathbf{m}}^{(k)} = \frac{\rho\sinh(\mathbf{z}^{(k)})}{\rho\cosh(\mathbf{z}^{(k)}) + (1-\rho)\exp((\gamma_k+t)/2)}$,    $\mathsf{b}_k = \frac{1}{n}\sum_{i=1}^n \frac{\rho^2+\rho(1-\rho)\exp\left(\frac{\gamma_k+t}{2}\right)\cosh(z_i^{(k)})}{\left(\rho\cosh(z_i^{(k)})+(1-\rho)\exp\left(\frac{\gamma_k+t}{2}\right)\right)^2}$
    $\mathbf{z}^{(k+1)} = \beta\mathbf{A}\widehat{\mathbf{m}}^k + \mathbf{y} - \beta^2\mathsf{b}_k\widehat{\mathbf{m}}^{k-1}$,
end
$\widehat{\mathbf{m}}^{(K_{\mathrm{AMP}})} = \frac{\rho\sinh(\mathbf{z}^{(K_{\mathrm{AMP}})})}{\rho\cosh(\mathbf{z}^{(K_{\mathrm{AMP}})}) + (1-\rho)\exp((\gamma_{K_{\mathrm{AMP}}}+t)/2)}$,
Output: $\widehat{\mathbf{m}}^{(K_{\mathrm{AMP}})}$
```

Anisotropic Case

Similarly, using Eq. (4.4.36) allows for a simplification of the SE by consider $\gamma_k = \mu_k$, we can rewrite Eq. (5.2.18) by expanding σ_k^2:

$$\begin{cases} \gamma_{k+1} \coloneqq \beta^2 \mathbb{E}[\bar{X} f_k(X_k)], \\ \sigma_{k+1}^2 \coloneqq \beta^2 \mathbb{E}[f(X_k)^2] + 2\beta t \mathbb{E}[\bar{X} f_k(X_k)] + \mathbb{E}[\bar{X}] t^2 = \gamma_{k+1} + 2t\beta^{-1}\gamma_{k+1} + \rho t^2. \end{cases} \tag{5.2.15}$$

Plugging this one-dimensional SE into Eq. (5.2.10) yields

$$f_k(X_k) = \frac{\rho \sinh(\eta_k X_k)}{\rho \cosh(\eta_k X_k) + (1-\rho)\exp\left(\frac{(\gamma_k + (\lambda+\rho\beta)t)^2}{2(\gamma_k + 2t\beta^{-1}\gamma_k + \rho t^2 + \lambda t)}\right)}, \tag{5.2.16}$$

with $\eta_k \coloneqq \frac{\gamma_k + (\lambda+\rho\beta)t}{\gamma_k + 2t\beta^{-1}\gamma_k + \rho t^2 + \lambda t}$. The Onsager term is similarly computed and given by

$$\mathsf{b}_k = \eta_k \frac{\rho^2 + \rho(1-\rho)\exp\left(\frac{\gamma_k + t}{2}\right)\cosh(\eta_k X_k)}{\left(\rho\cosh(\eta_k X_k) + (1-\rho)\exp\left(\frac{\gamma_k + t}{2}\right)\right)^2}. \tag{5.2.17}$$

Finally, we can compute the anisotropic SE

$$\begin{aligned} \gamma_{k+1} &= \beta^2 \rho^2 \mathbb{E}\left[\frac{\sinh(\eta_k X_k)}{\rho\cosh(\eta_k X_k) + (1-\rho)\exp\left(\frac{(\gamma_k + (\lambda+\beta)t)^2}{2(\gamma_k + 2t\beta^{-1}\gamma_k + \rho t^2 + \lambda t)}\right)}\right] \\ &= \beta^2 \rho^2 \mathbb{E}_G\left[\frac{\sinh\left(\eta_k \bar{G}_k\right)}{\rho\cosh(\eta_k \bar{G}_k) + (1-\rho)\exp\left(\frac{(\gamma_k + (\lambda+\beta)t)^2}{2(\gamma_k + 2t\beta^{-1}\gamma_k + \rho t^2 + \lambda t)}\right)}\right] \end{aligned} \tag{5.2.18}$$

with $\bar{G}_k \coloneqq \gamma_k + (\lambda + \rho\beta)t + \sqrt{\gamma_k + 2t\beta^{-1}\gamma_k + \rho t^2 + \lambda t}G$, $G \sim \mathcal{N}(0,1)$. We summarize these results in Algorithm 6.

Algorithm 6: Sparse Tilted Mean Estimation: Anisotropic Side Information

```
Input: $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathbf{y} \in \mathbb{R}^n$, parameters $\lambda, \beta, \rho, t > 0$, iteration numbers $K_{\text{AMP}}$
$\widehat{\mathbf{m}}^{-1} = \mathbf{0}_n$,
for $k = 0, \dots, K_{\text{AMP}} - 1$ do
    $\gamma_k \coloneqq \frac{\beta^2}{n} \|\widehat{\mathbf{m}}^{(k-1)}\|_2^2$,
    $\eta_k = \frac{\gamma_k + (\lambda+\rho\beta)t}{\gamma_k + 2t\beta^{-1}\gamma_k + \rho t^2 + \lambda t}$,    $\xi_k = \frac{(\gamma_k + (\lambda+\rho\beta)t)^2}{\gamma_k + 2t\beta^{-1}\gamma_k + \rho t^2 + \lambda t}$,
    $\widehat{\mathbf{m}}^{(k)} = \frac{\rho\sinh\left(\eta_k \mathbf{z}^{(k)}\right)}{\rho\cosh\left(\eta_k \mathbf{z}^{(k)}\right) + (1-\rho)\exp(\xi_k/2)}$,    $\mathsf{b}_k = \frac{\eta_k}{n}\sum_{i=1}^n \frac{\rho^2 + \rho(1-\rho)\exp(\xi_k/2)\cosh(\eta_k z_i^{(k)})}{\left(\rho\cosh(\eta_k z_i^{(k)}) + (1-\rho)\exp(\xi_k/2)\right)^2}$
    $\mathbf{z}^{(k+1)} = \beta \mathbf{A}\widehat{\mathbf{m}}^k + \mathbf{y} - \beta \mathsf{b}_k \widehat{\mathbf{m}}^{k-1}$,
end
$\widehat{\mathbf{m}}^{(K_{\text{AMP}})} = \frac{\rho\sinh(\mathbf{z}^{(K_{\text{AMP}})})}{\rho\cosh(\mathbf{z}^{(K_{\text{AMP}})} + (1-\rho)\exp(\xi_{K_{\text{AMP}}}/2)}$,
Output: $\widehat{\mathbf{m}}^{(K_{\text{AMP}})}$
```

5.3 Numerical Simulations

In this section we present results of numerical experiments, testing Algorithms 5 and 6. As mentioned at the beginning of the Chapter, we solely focus on the planted model, which is a statistical estimation problem. The main goal of this section is to empirically determine whether the conjectured superiority of DDDs is valid.

The first experiment was an implementation of Algorithms 5 and 6, plotted against their theoretical SEs (Eqs. 5.2.14 and 5.2.18 respectively). We chose parameters $n = 2000, \rho = 0.08, \lambda = 2.01$, and ran the AMP algorithms for 100 iterations, respectively. We report these results in Figure 5.1.

The second experiment was to test the plausibility of our anisotropic SE equation (5.2.18). We tested it for two different values of t; $t = 0$ and $t = 0.1$. We report the simulations comparing the isotropic and anisotropic SE equations in Figure 5.2.

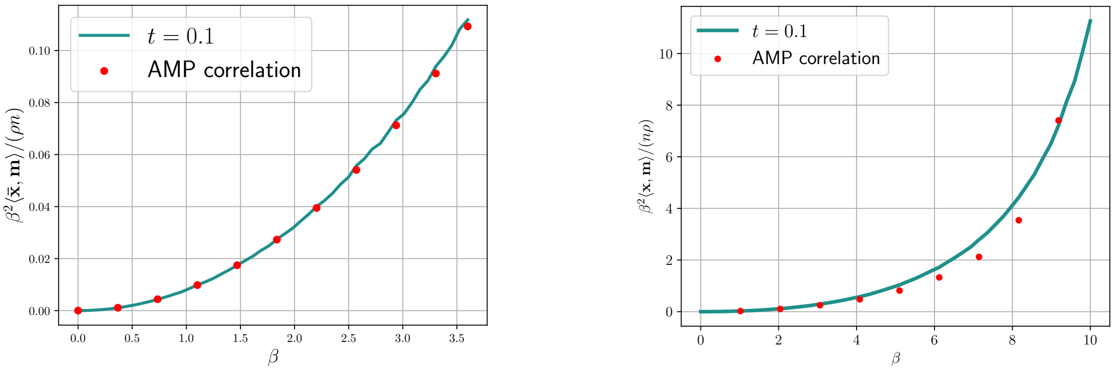
Finally, the third experiment aimed to re-create the full phase diagram, as seen previously in Figures 3.2b, 3.2a. Due to extremely long computing times, we could only produce diagrams of sizes $50 \times 50, 100 \times 100$ and 200×200 at best. To compute the full phase diagram, we aimed to find regions in phase space where the fixed point of Eqs. (5.2.14) and (5.2.18) are not unique. This is equivalent to metastable regions, and are problematic for AMP convergence. To find them, we computed the recursive SE equations, but with different starting points.

$$\begin{cases} \gamma_0^{(0)} = 0, \\ \gamma_{k+1}^{(0)} = \beta^2 \rho^2 \mathbb{E}_G \left[\dfrac{\sinh(\gamma_k^{(0)} + t + \sqrt{\gamma_k^{(0)} + t} G)}{\rho \cosh(\gamma_k^{(0)} + t + \sqrt{\gamma_k^{(0)} + t} G) + (1 - \rho) \exp\left(\frac{\gamma_k^{(0)} + t}{2} \right)} \right], \end{cases} \tag{5.3.1}$$

$$\begin{cases} \gamma_0^{(1)} = 1, \\ \gamma_{k+1}^{(1)} = \beta^2 \rho^2 \mathbb{E}_G \left[\dfrac{\sinh(\gamma_k^{(1)} + t + \sqrt{\gamma_k^{(1)} + t} G)}{\rho \cosh(\gamma_k^{(1)} + t + \sqrt{\gamma_k^{(1)} + t} G) + (1 - \rho) \exp\left(\frac{\gamma_k^{(1)} + t}{2} \right)} \right]. \end{cases} \tag{5.3.2}$$

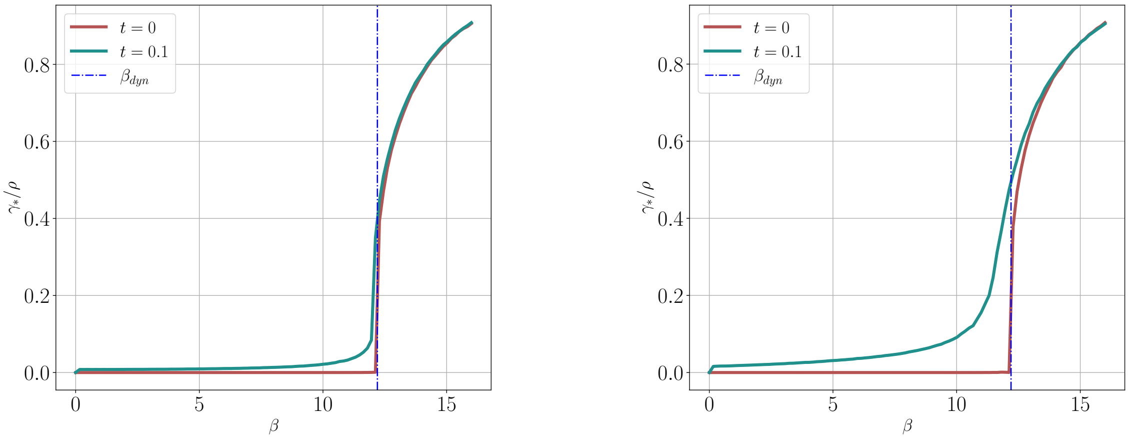
We then plot their difference $|\gamma_*^{(0)} - \gamma_*^{(1)}|$ on a heatmap to produce the phase diagram, with $\gamma_* := \lim_{k \to \infty} \gamma_k$.

We report these diagrams in Figure 5.3, and discuss all results at the end of the section.



(a) (b)

Figure 5.1: The correlation of Algorithm 5's (left) and Algorithm 6's (right) respective estimators with $\bar{\mathbf{x}}$ (red dots) against the predicted SE from Eqs. (5.2.14) resp. (5.2.18) (blue line).

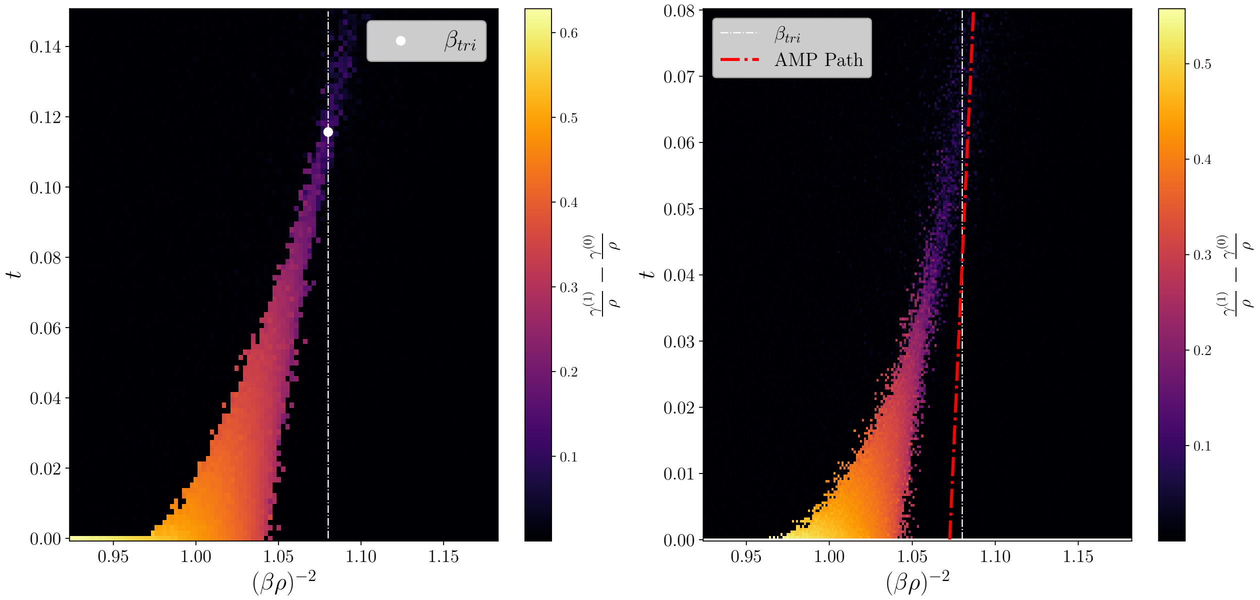


(a) (b)

Figure 5.2: Comparison of the sparse ($\rho = 0.08$) isotropic SE (left) from Eq. (5.2.14) and the proposed sparse anisotropic SE (right) from Eq. (5.2.18) with $\lambda = 2.01$. We notice that for $t = 0$, the phase transition at $\beta_d \approx 12.2$ aligns with results from Figure 3.2a, providing us with strong evidence towards the correctness of our theory.

5.4 Discussion of Results

Figure 5.1 shows that our isotropic AMP follows the SE, as expected. In the anisotropic case, there is a slight discrepancy between the expected and obtained behaviour. We believe that this could be due to implementation and precision errors, or lack of sufficient averaging over trials. Figure 5.2 demonstrates two main things. Our anisotropic SE from Eq. (5.2.18)



(a) Recreation of the phase diagram from Fig. 3.2a by finding fixed points of Eq. (5.2.14), with dimensions 100 × 100.

(b) 200 × 200 Phase diagram corresponding to fixed points discrepancies of eq. (5.2.18), computed as in Fig. 5.3a. The AMP begins at $\beta = 12.07$, showing a clear gain.

Figure 5.3: Implementation of the full phase diagrams. We highlight the tri-point of the isotropic case (left), as well as report it on the x-axis of the anisotropic phase diagram (right). The non-linear AMP trajectory clearly indicates access to lower temperatures beyond β_{tri}, demonstrating a proof-of-concept. Note that the y-axis is rescaled for convenience, implying a 'lower' tri-point in the anisotropic case.

is not only plausible, but also correctly predicts the "natural" phase transition of sparse problems. We notice that in Figure 5.2b the $t > 0$ trajectory (turquoise) is much smoother than its isotropic counterpart, indicating higher correlation early on (which is expected, due to the gain incurred by $\lambda > 2$).

Finally, Figure 5.3 shows that the tri-point for anisotropic side information could potentially be decreased. Combined with the non-linear AMP trajectory, this highly supports the superiority of DDDs for sampling within the gap $(\beta_{\text{d}}, \beta_{\text{tri}})$. We also note that numerical precision presents itself to be a significant bottleneck here, as we notice that beyond the tri-point of Fig. 5.3a, slight discrepancies still arise when they shouldn't. We conjecture that this is due to a very "flat" optimization surface, trapping Eqs. (5.2.14) and (5.2.18) in incorrect pseudo-fixed points. This carries over to the analysis of Fig. 5.3b, leading us to believe an actual decrease of β_{tri} still ensuring the validity of our claim about DDDs.

Chapter 6

Summary

Motivated by sampling beyond the tri-critical point β_{tri} of spin glass Gibbs measures, this work explored the notion of disorder-dependent diffusions (DDDs) as an extension to current diffusion-based sampling algorithms (DBSAs). While isotropic DBSAs rely on simulating the reverse SDE

$$\mathrm{d}\mathbf{y}(t) := \mathbf{m}(\mathbf{y}(t); t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t,$$

DDDs are a natural anisotropic extension thereof:

$$\mathrm{d}\mathbf{y}(t) := \mathbf{Q}\mathbf{m}(\mathbf{y}(t); \mathbf{Q}t)\mathrm{d}t + \sqrt{\mathbf{Q}}\mathrm{d}\mathbf{B}_t,$$

where $\mathbf{m}(\mathbf{y}(t); t)$ resp. $\mathbf{m}(\mathbf{y}(t); \mathbf{Q}t)$ denotes the tilted mean of a sample $\mathbf{x} \sim \mu$ we wish to produce (see Eq. (3.4.4)).

Concretely, DDDs offer an advantage over isotropic DBSAs with their non-linear trajectory through phase space (see Figs. 3.2a and 5.3) by allowing incorporation of the disorder matrix $\mathbf{A}$ into the sensing matrix $\mathbf{Q} \equiv \mathbf{Q}(\mathbf{A})$. This allows accurate computation of the drift term of the reverse SDE $\mathbf{y}(t)$ (the tilted mean) via Approximate Message Passing (AMP), effectively increasing the sampling range.

Contributions

While we have yet to produce a provably polynomial-time DBSA based on DDDs, we believe to have addressed the main bottleneck of the sampling procedure, as successful estimation of $\mathbf{m}$ directly impacts the accuracy of the produced samples. Our contribution is twofold.

- We provide an extension to current AMP theory for $\mathbb{Z}_2$-synchronization (planted SK model), enabling the analysis and formulation of state evolution (SE) in the presence of anisotropic side information $\mathbf{y}(t) = \mathbf{Q}t\bar{\mathbf{x}} + \sqrt{\mathbf{Q}t}\mathbf{z}$. We fully derive the AMP algorithm in Section 4.2, and rigorously prove its SE in Section 4.4, resulting in Theorem 4.4.1.

The key feature of this result is that this SE holds for *any* prior (under mild measure-theoretical conditions), enabling usage in open problems.

- Chapter 5 applies our new AMP analysis to the problem of the planted sparse SK model, allowing for not only recreation of results by Ghio et al. (2023), but extension thereof with anisotropic side information. We derive the new AMP algorithm and propose its associated SE, which exhibits the expected phase transitions. Furthermore, we demonstrate a clear increase of the sampling range in Fig. 5.3b, by a conjectured lowering of β_{tri}, combined with the the non-linear AMP trajectory.

Future Work

While we have laid the foundations of AMP algorithms with anisotropic side information, we still have yet to demonstrate the almost sure convergence that most AMP algorithms achieve in the literature (see Bayati and Montanari (2011); Montanari and Venkataramanan (2019); Schniter, Rangan, and Fletcher (2016)). Although some novel AMP variants are also restricted to convergence in probability (Berthier, Montanari, and Nguyen (2019)), we nonetheless expect Theorem 4.4.1 to hold almost surely.

Our AMP algorithm (Algorithm 3) recovers the Bayes / tilted mean for the SK model, but we did not address lines 6-10 of Algorithm 2 which would imply polynomial-time estimation and sampling. This would require verifying the convexity of $\mathcal{F}_{\text{TAP}}$ with anisotropic $\mathbf{y}$, which could be done using methods developed by Celentano (2022). We also expect this to work without too much extra adaptation.

Finally, a method of integration to compute the associated RS free energy from the anisotropic SE's fixed points would not only allow for more precise computing of the phase diagrams of Fig. 5.3, but would also provide yet another rigorous verification of the replica method and complete the framework.

Acknowledgements

It goes without saying that Prof. Yuansi Chen has given me not only exceptional supervision and guidance, but also a very pleasant first experience in mathematical research. I am deeply grateful and hope that we can continue to collaborate in future endeavours.

I would like to extend my thanks to Prof. Florent Krzakala and his students Davide and Yatin for hosting me at their laboratory to have interesting discussions and for providing the code to generate all the phase diagrams. I am also very grateful to the Institute for Dynamical Systems and Control for providing me with a (semi-private) workspace. The people I met while working there were all wonderful.

Bibliography

Aizenman, M. and R. Holley (1987). *Rapid Convergence to Equilibrium of Stochastic Ising Models in the Dobrushin Shlosman Regime*, pp. 1–11. New York, NY: Springer New York.

Aizenman, M., J. L. Lebowitz, and D. Ruelle (1987, March). Some rigorous results on the Sherrington-Kirkpatrick spin glass model. *Communications in Mathematical Physics 112*(1), 3–20.

Alaoui, A., A. Montanari, and M. Sellke (2021, 11). Optimization of mean-field spin glasses. *The Annals of Probability 49*.

Alaoui, A. E., A. Montanari, and M. Sellke (2023). Sampling from mean-field gibbs measures via diffusion processes.

Alaoui, A. E., A. Montanari, and M. Sellke (2024). Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization.

Anderson, B. D. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications 12*(3), 313–326.

Baik, J., G. B. Arous, and S. Péché (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability 33*(5), 1643 – 1697.

Bakry, D., I. Gentil, and M. Ledoux (2013). *Analysis and Geometry of Markov Diffusion Operators*. Grundlehren der mathematischen Wissenschaften. Springer International Publishing.

Barbier, J., M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborova (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula.

Bauerschmidt, R. and T. Bodineau (2019, April). A very simple proof of the lsi for high temperature spin systems. *Journal of Functional Analysis 276*(8), 2582–2588.

Bayati, M. and A. Montanari (2011, February). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory 57*(2), 764–785.

Benaych-Georges, F. and R. R. Nadakuditi (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics 227*(1), 494–521.

Berthier, R., A. Montanari, and P.-M. Nguyen (2019, January). State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA 9*(1), 33–79. _eprint: https://academic.oup.com/imaiai/article-pdf/9/1/33/32931175/iay021.pdf.

Bolthausen, E. (2012). An iterative construction of solutions of the tap equations for the sherrington-kirkpatrick model.

Bolthausen, E. (2019). The Thouless-Anderson-Palmer Equations in spin glass theory. https://anr-malin.sciencesconf.org/data/pages/Aussois_2.pdf. [Online; accessed 21-Jan-2025].

Bray, A. J., M. A. Moore, and A. P. Young (1984, feb). Weighted averages of tap solutions and parisi's q(x). *Journal of Physics C: Solid State Physics 17*(5), L155.

Cademartori, C. and C. Rush (2024). A non-asymptotic analysis of generalized vector approximate message passing algorithms with rotationally invariant designs. *IEEE Transactions on Information Theory 70*(8), 5811–5856.

Cavagna, A., I. Giardina, G. Parisi, and M. Mézard (2003, January). On the formal equivalence of the tap and thermodynamic methods in the sk model. *Journal of Physics A: Mathematical and General 36*(5), 1175–1194.

Celentano, M. (2022). Sudakov-fernique post-amp, and a new proof of the local convexity of the tap free energy.

Celentano, M., Z. Fan, and S. Mei (2023). Local convexity of the tap free energy and amp convergence for z2-synchronization.

Chen, S., S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang (2023). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions.

Chen, W.-K. and D. Panchenko (2018, May). On the tap free energy in the mixed p-spin models. *Communications in Mathematical Physics 362*(1), 219–252.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience.

de Almeida, J. R. L. and D. J. Thouless (1978). *Stability of the Sherrington–Kirkpatrick solution of a spin glass model*, pp. 129–136.

Deshpande, Y., E. Abbe, and A. Montanari (2015). Asymptotic mutual information for the two-groups stochastic block model.

Donoho, D. L., A. Maleki, and A. Montanari (2009, November). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences 106*(45), 18914–18919.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Eldan, R. (2013, March). Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis 23*(2), 532–569.

Eldan, R. (2019). Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation.

Fan, Z. (2022, 02). Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics 50.*

Fan, Z., S. Mei, and A. Montanari (2020). Tap free energy, spin glasses, and variational inference.

Feng, O. Y., R. Venkataramanan, C. Rush, and R. J. Samworth (2021). A unifying tutorial on approximate message passing.

Gamarnik, D., C. Moore, and L. Zdeborová (2022, nov). Disordered systems insights on computational hardness. *Journal of Statistical Mechanics: Theory and Experiment 2022*(11), 114015.

Ghio, D., Y. Dandi, F. Krzakala, and L. Zdeborová (2023). Sampling with flows, diffusion and autoregressive neural networks: A spin-glass perspective.

Guo, D., S. Shamai, and S. Verdu (2004). Mutual information and minimum mean-square error in gaussian channels.

Ho, J., A. Jain, and P. Abbeel (2020). Denoising diffusion probabilistic models.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik 31*, 253–258.

Knowles, A. and J. Yin (2013). The isotropic semicircle law and deformation of wigner matrices. *Communications on Pure and Applied Mathematics 66*(11), 1663–1749.

Lesieur, T., F. Krzakala, and L. Zdeborová (2017, July). Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment 2017*(7), 073403.

Li, Y., Z. Fan, S. Sen, and Y. Wu (2023, 01). Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature. *IEEE Transactions on Information Theory PP*, 1–36.

Liptser, R. S. and A. N. Shiryayev (1977). *Statistics of Random Processes I.* New York, NY: Springer.

Liu, L., Y. Cheng, S. Liang, J. H. Manton, and L. Ping (2023). On orthogonal approximate message passing.

Ma, J. and L. Ping (2016). Orthogonal amp. *IEEE Access 5*, 2020–2033.

Maillard, A. (2021, 08). *Fundamental limits of high-dimensional estimation : a stroll between statistical physics, probability and random matrix theory.* Ph. D. thesis, Ecole Normale Supérieure de Paris.

Marčenko, V. A. and L. A. Pastur (1967, April). DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES. *Mathematics of the USSR-Sbornik 1*(4), 457.

Mehta, M. (1991). *Random Matrices.* Academic Press.

Mezard, M. and A. Montanari (2009). *Information, Physics, and Computation.* USA: Oxford University Press, Inc.

Mezard, M., G. Parisi, and M. Virasoro (1987). *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications.* World Scientific Lecture Notes In Physics. World Scientific Publishing Company.

Mingo, J. A. and R. Speicher (2017). *Free Probability and Random Matrices*, Volume 35 of *Fields Institute Monographs.* New York, NY: Springer.

Montanari, A. (2023). Sampling, diffusions, and stochastic localization.

Montanari, A. and D. Tse (2006). Analysis of belief propagation for non-linear problems: The example of cdma (or: How to prove tanaka's formula).

Montanari, A. and R. Venkataramanan (2019). Estimation of low-rank matrices via approximate message passing.

Montanari, A. and Y. Wu (2024). Posterior sampling in high dimension via diffusion processes.

Panchenko, D. (2012, 11). *The Sherrington-Kirkpatrick Model.*

Parisi, G., F. Ricci-Tersenghi, and T. Rizzo (2014, April). Diluted mean-field spin-glass models at criticality. *Journal of Statistical Mechanics: Theory and Experiment 2014*(4), P04013.

Rangan, S., P. Schniter, and A. K. Fletcher (2016). Vector approximate message passing. *CoRR abs/1610.03082.*

Richardson, T. and R. Urbanke (2008). *Modern Coding Theory.* Cambridge University Press.

Schniter, P., S. Rangan, and A. K. Fletcher (2016). Vector approximate message passing for the generalized linear model. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 1525–1529.

Sherrington, D. and S. Kirkpatrick (1975, 12). Solvable model of a spin-glass. *Physical Review Letters 35*, 1792+.

Singer, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis 30*(1), 20–36.

Sohl-Dickstein, J., E. A. Weiss, N. Maheswaranathan, and S. Ganguli (2015). Deep unsupervised learning using nonequilibrium thermodynamics.

Sompolinsky, H. and A. Zippelius (1981, Aug). Dynamic theory of the spin-glass phase. *Phys. Rev. Lett. 47*, 359–362.

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole (2021). Score-based generative modeling through stochastic differential equations.

Takeuchi, K. (2017). Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 501–505.

Talagrand, M. (2011, 01). *Mean field models for spin glasses. Volume I: Basic examples. 2nd revised and enlarged ed.*

Tanaka, T. (2002, November). A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Transactions on Information Theory 48*(11), 2888–2910. Conference Name: IEEE Transactions on Information Theory.

Thouless, Anderson, and Palmer (1977). Solution of 'solvable model of a spin glass'. *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics 35*(3), 593–601.

Tulino, A. and S. Verdú (2004). *Random Matrix Theory and Wireless Communications.* Foundations and trends in communications and information theory. Now.

Vaart, A. W. v. d. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Zdeborová, L. and F. Krzakala (2016, August). Statistical physics of inference: thresholds and algorithms. *Advances in Physics 65*(5), 453–552.

Appendix A

Random Matrix Theory

This appendix will mainly serve as a repository of definitions and results for readers who are unfamiliar with Random Matrix Theory (RMT), or those who want a bite-sized and very rudimentary refresher. RMT studies the spectral properties of matrices with random entries, focusing on the statistical behaviour of their eigenvalues and eigenvectors as the matrix size N becomes large. A central result is the Wigner semicircle law, which describes the limiting eigenvalue density for large Hermitian or symmetric random matrices with i.i.d. entries, demonstrating that the eigenvalues are asymptotically confined to a compact interval. RMT finds applications in diverse areas of mathematics, including the study of concentration phenomena, free probability, and has strong historical connections to statistical mechanics.

A.1 A Few Random Matrices

Definition A.1.1 (Gaussian Matrix)
The simplest kind of random matrix is what we will call a ***Gaussian matrix*** $\mathbf{G} \in \mathbb{R}^{n \times n}$. *The entries of* $\mathbf{G}$ *are simply all i.i.d. Gaussian random variables, i.e.*

$$G_{ij} \sim \mathcal{N}(0, 1/n). \tag{A.1.1}$$

This definition will form the basis for the upcoming two random matrices, which are of central use in the thesis.

Definition A.1.2 (GOE Matrix)
Let $\mathbf{G} \in \mathbb{R}^{n \times n}$ *be a Gaussian matrix, and let* $\mathbf{A} = \frac{1}{2}(\mathbf{G} + \mathbf{G}^T)$. *We say that* $\mathbf{A}$ *belongs to the* ***Gaussian Orthogonal Ensemble****. The distribution of the elements of* $\mathbf{A}$ *is*

given by

$$A_{ij} = A_{ji} \sim \begin{cases} \mathcal{N}(0, 1/n), & i \neq j, \\ \mathcal{N}(0, 2/n) & i = j, \end{cases} \tag{A.1.2}$$

GOE matrices belong to a wider class of matrices, called Wigner matrices, but we will not treat these in this thesis. For more information, we refer the interested reader to the historically significant book by Mehta (1991), or the more modern approach by Mingo and Speicher (2017).

Definition A.1.3 (Wishart Matrix)
Let $\mathbf{G} \in \mathbb{R}^{n\times m}$ be a Gaussian matrix with $m < n$, and consider $\mathbf{A} = \mathbf{G}\mathbf{G}^T$. Then, $\mathbf{A}$ is called a ***Wishart matrix*** *with aspect ratio $\delta = \frac{m}{n}$.*

Finally, we present a very important type of random matrix: one that is uniformly drawn from the Haar measure of a locally compact topological group.[1]

Definition A.1.4 (Haar Matrix)
Let $\mathbb{O}(n)$ be the group of $n \times n$ orthogonal matrices, i.e. $\mathbf{O} \in \mathbb{O}(n)$ corresponds to all matrices $\mathbf{O} \in \mathbf{R}^{n\times n}$ such that

$$\mathbf{O}^T\mathbf{O} = \mathbf{O}\mathbf{O}^T = \mathbf{I}_n.$$

An $n \times n$ random matrix $\mathbf{U}$ is said to be ***Haar distribution*** *(or simpler: a* ***Haar matrix****) if it is uniformly distributed on $\mathbb{O}(n)$. An interesting property of Haar matrices is that they are closed under multiplication by any other orthogonal matrix, i.e. for $\mathbf{O}$ Haar and $\mathbf{V} \in \mathbb{O}(n)$ a deterministic matrix,*

$$\mathbf{O} \overset{d}{=} \mathbf{V}\mathbf{O} \overset{d}{=} \mathbf{O}\mathbf{V}. \tag{A.1.3}$$

Remark A.1
The above definition also applies to $\mathbb{U}(n)$, the unitary group for complex matrices. We will restrict ourselves to real matrices, but a lot of results simply transfer by considering the Haar measure on $\mathbb{U}(n)$ instead of $\mathbb{O}(n)$.

A.2 Spectral Distributions of Random Matrices

As mentioned, RMT specializes in computing the limiting eigenvalue distribution of large random matrices, as well as their fluctuations. We will introduce two central distributions, and relate them to the spectral distribution of the previously mentioned random matrices.

[1] This is way beyond the scope of the thesis. The definition is much simpler than the introduction to the topic. We kept the formal introduction for the curious readers.

Definition A.2.1 (Wigner's Semi-Circle Distribution)
A random variable S is said to follow the ***Wigner semi-circle distribution*** *with radius R if its law is given by*

$$\rho_R^{\mathrm{SC}}(s) = \frac{2}{\pi R^2}\sqrt{R^2 - x^2}. \tag{A.2.1}$$

In the case of $R = 1$, we simply omit the argument for R.

Definition A.2.2 (Marchenko-Pastur Distribution)
Let $\delta \in (0, \infty)$. Then a random variable S follows the ***Marchenko-Pastur distribution*** *of aspect ratio δ if its law is given by*

$$\rho_\delta^{\mathrm{MP}}(s) = \frac{\sqrt{(\delta_+ - s)(s - \delta_-)}}{2s\pi\delta}, \tag{A.2.2}$$

for $\delta_\pm = \left(1 \pm \sqrt{\delta}\right)^2$.

With these definitions in hand, we are ready to state two pillars of RMT.

Theorem A.2.1
Let $\mathbf{A}$ be a GOE matrix. Then its empirical spectral distribution converges almost surely to the Wigner semi-circle law with radius 1.

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\mathbf{A}) \overset{\text{a.s.}}{=} S, \tag{A.2.3}$$

where $S \sim \rho^{SC}$.

Proof Outline. This theorem is arguably the most central theorem to RMT, as the semi-circle distribution is the non-commutative equivalent to Gaussian distributions for classical probability theory.

One way to prove this result is by an application of Wick's Formula to compute the trace of integer powers of $\mathbf{A}$, giving us the Catalan numbers, which are the free moments of the semi-circle distribution, allowing us to appeal to the Free Central Limit Theorem (Theorem 5, Mingo and Speicher (2017)) by matching moments. For a full derivation, we refer the reader to the referenced material. □

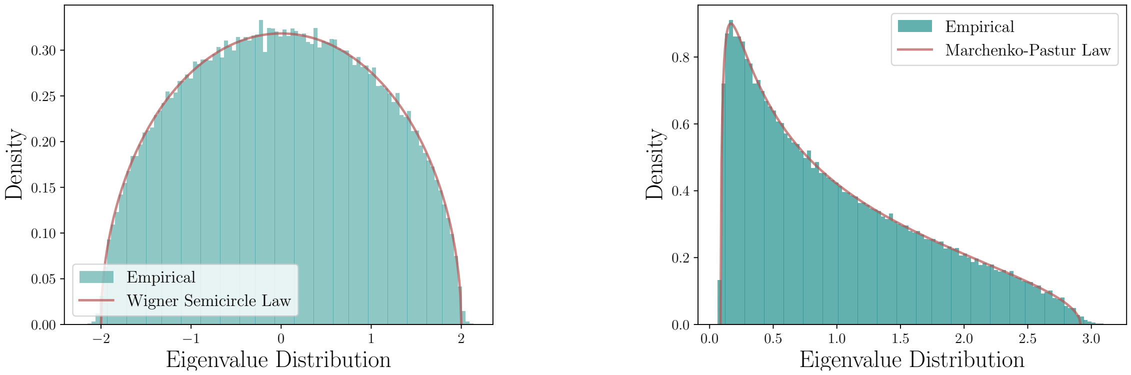
Again, the following is a well-known result. The original derivation was made by Marčenko and Pastur in 1967, but more recent texts treat this result.

Theorem A.2.2
Let $\mathbf{A}$ be a Wishart matrix with aspect ratio $\frac{m}{n} \to \delta$ as $m, n \to \infty$. Then, the empirical eigenvalue distribution of $\mathbf{A}$ converges almost surely to the Marchenko-Pastur distribution with aspect ratio δ.

$$\lim_{m,n\to\infty} \frac{1}{m} \sum_{i=1}^{m} \lambda_i(\mathbf{A}) \overset{\text{a.s.}}{=} S, \tag{A.2.4}$$

with $S \sim \rho_\delta^{MP}$.

We pictorially describe Theorems A.2.1 and A.2.2 in Figures A.1a and A.1b below, respectively.



(a)

(b)

Figure A.1: Plots of average empirical spectral distributions (blue) of GOE and Wishart(m, p) matrices of sizes 100 and 100 × 50 resp. with $\delta = \frac{m}{p} = \frac{1}{2}$ over 1000 trials against their asymptotic theoretical distribution (red).

We now use Haar matrices to characterize a certain class of random matrices.

Definition A.2.3 (Rotationally Invariant Matrices)
A random matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is said ***rotationally invariant*** *if for any $\mathbf{O} \in \mathbb{O}(n)$,*

$$\mathbf{A} \overset{d}{=} \mathbf{OAO}^T. \tag{A.2.5}$$

Example 1
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be a GOE matrix. Then,* $\mathbf{A}$ *admits the following eigendecomposition:*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{A.2.6}$$

where $\mathbf{V} \sim \mathrm{Haar}(\mathbb{O}(n))$*, and* $\mathbf{\Lambda}$ *is diagonal, such that*

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Tr}[\mathbf{\Lambda}] \overset{\text{a.s.}}{=} S, \tag{A.2.7}$$

with $S \sim \rho^{\mathrm{SC}}$*. GOE matrices are thus rotationally invariant.*

Transforms

We finally present a series of useful tools in RMT, which will culminate in the $\mathcal{R}$-transform. These transforms act on real-valued random variables, which we will consider to be limiting eigenvalue distributions of random matrices.

Definition A.2.4 (Stieltjes Transform)
Let λ *be a real random variable with distribution* $\rho_\lambda(\cdot)$*. Its associated* ***Stieltjes transform*** *evaluated at* $z \in \mathbb{C}$ *is given by*

$$\mathcal{S}_\lambda(z) := \mathbb{E}\left[\frac{1}{\lambda - z}\right] = \int_{-\infty}^{\infty} \frac{1}{s-z}\mathrm{d}\rho_\lambda(s). \tag{A.2.8}$$

Definition A.2.5 ($\mathcal{R}$-Transform)
Let λ *be a real-valued random variable, and* $\mathcal{S}_\lambda^{-1}(z)$ *denote the functional inverse of its Stieltjes transform. Then, we define* λ*'s* $\mathcal{R}$***-transform*** *evaluated at* $z \in \mathbb{C}$ *as*

$$\mathcal{R}_\lambda(s) = \mathcal{S}_\lambda^{-1}(-z) - \frac{1}{z}. \tag{A.2.9}$$

Remark A.2
We will (with some abuse of notation) define the $\mathcal{R}$*-transform of a square random matrix* $\mathbf{A}$ *as*

$$\mathcal{R}_{\mathbf{A}}(z) := \mathcal{R}_\lambda(z), \tag{A.2.10}$$

for λ *the limiting empirical eigenvalue distribution.*

Example 2
The $\mathcal{R}$-transform of

i.) an atom distribution at $x = x_0$ is

$$\mathcal{R}_{\delta(x_0)} = x_0. \tag{A.2.11}$$

ii.) a semicircle distributed random variable λ is

$$\mathcal{R}_{\lambda}(z) = z. \tag{A.2.12}$$

iii.) a Marchenko-Pastur distributed random variable ν with aspect ratio β is

$$\mathcal{R}_{\nu}(z) = \frac{1}{1 - \beta z}. \tag{A.2.13}$$

Appendix B

Proofs

B.1 Proof of Chapter 3

B.1.1 Proof of Proposition 3.4.1

We first note that our construction of $\mathbf{Q}$ is still PSD, as $\lambda_{\max}(\mathbf{A}) \to 2$ as $n \to \infty$, by properties of GOE matrices (Vershynin (2018)). The claim follows by substituting the construction for $\mathbf{Q}$:

$$\begin{aligned}
\mu_{\mathbf{A},\beta,\mathbf{y}^{\text{anis}}(t)}(\mathbf{x}) &= \frac{1}{Z(\mathbf{A},\mathbf{y}^{\text{anis}}(t),\beta)} \mu_{\mathbf{A},\beta}(\mathbf{x}) \exp\left(-\frac{1}{2t}\|\mathbf{y}^{\text{anis}}(t) - t\mathbf{Q}\mathbf{x}\|_{\mathbf{Q}^{-1}}^2\right) \\
&= \frac{1}{Z}\mu_{\mathbf{A},\beta}(\mathbf{x}) \exp\left(\langle \mathbf{y}^{\text{anis}}(t), \mathbf{x}\rangle - \frac{t}{2}\|\mathbf{Q}\mathbf{x}\|_2^2\right) \\
&= \frac{1}{Z}\mu_{\mathbf{A},\beta}(\mathbf{x}) \exp\left(\langle \mathbf{y}^{\text{anis}}(t), \mathbf{x}\rangle - \frac{\lambda t}{2}\|\mathbf{x}\|_2^2 - \frac{ct}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle\right) \\
&= \frac{1}{Z'} \exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle + \langle \mathbf{y}^{\text{anis}}(t), \mathbf{x}\rangle - \frac{\lambda t}{2}\|\mathbf{x}\|_2^2 - \frac{ct}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle\right) \\
&= \frac{1}{Z'} \exp\left(\frac{\beta - ct}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle + \langle \mathbf{y}^{\text{anis}}(t), \mathbf{x}\rangle - \frac{\lambda t}{2}\|\mathbf{x}\|_2^2\right) \qquad \text{(B.1.1)}
\end{aligned}$$

□

B.2 Proofs of Section 4.1

B.2.1 Proof of Lemma 4.1.2

The goal of the Lemma is to show that $Z_{\mathrm{pl}} = Z_{\mathrm{GOE}}$. We will first analyse the Z_{pl}-term.

$$\begin{aligned} Z_{\mathrm{pl}} &= \int_{\mathbf{A}} \exp\left(-\frac{n}{4}\|\mathbf{A} - \frac{\beta}{n}\mathbf{x}\mathbf{x}^T\|_F^2\right) \mathrm{d}\mathbf{A} \\ &= \int_{\mathbf{A}} \exp\left(-\frac{n}{4}\|\mathbf{A}\|_F^2 + \frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \frac{\beta^2 n}{4}\right) \mathrm{d}\mathbf{A} \\ &= e^{-\frac{\beta^2 n}{4}} Z_{\mathrm{GOE}} \int_{\mathbf{A}} \frac{e^{-\frac{n}{4}\|\mathbf{A}\|_F^2}}{Z_{\mathrm{GOE}}} \exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle\right) \mathrm{d}\mathbf{A} \\ &= e^{-\frac{\beta^2 n}{4}} Z_{\mathrm{GOE}} \cdot \mathbb{E}_{\mathrm{GOE}}\left[\exp\left(\frac{\beta}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle\right)\right] \end{aligned} \tag{B.2.1}$$

The expectation over the GOE ensemble can be simplified by with a useful result from Bayati and Montanari (2011). We proceed by first re-proving a result on random Gaussian matrices, and then using it combined with the law of the unconscious statistician (LOTUS).

Let $\mathbf{G}$ be a random matrix such that every entry is i.i.d. according to $\mathbf{G}_{ij} \sim \mathcal{N}(0, \frac{1}{2n})$, and $\mathbf{x} \in \mathbb{R}^n$ with bounded ℓ-2 norm. Then,

$$\mathbf{G}\mathbf{x} \overset{d}{=} \frac{\|\mathbf{x}\|_2}{\sqrt{2n}}\mathbf{z}, \tag{B.2.2}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_n)$ is a standard normal random variable, and thus $\mathbf{x}^T\mathbf{G}\mathbf{x} \overset{d}{=} \frac{\|\mathbf{x}\|_2^2}{\sqrt{2n}} Z$ with $Z \sim \mathcal{N}(0,1)$. For $\|\mathbf{x}\|_2^2 = n$, we have $\mathbf{x}^T\mathbf{G}\mathbf{x} \sim \mathcal{N}(0, \frac{n}{2})$. We know that a GOE matrix $\mathbf{A}$ can be constructed as $\mathbf{A} = \mathbf{G}^T + \mathbf{G}$, hence $\mathbf{x}^T\mathbf{A}\mathbf{x} \overset{d}{=} \sqrt{2n}Z$. Thus, by LOTUS,

$$\mathbb{E}_{\mathrm{GOE}}\left[\exp\left(\frac{\beta}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\right)\right] = \mathbb{E}_Z[\exp(\frac{\beta}{2}\sqrt{2n}Z)] = \exp\left(\frac{\beta^2}{8} \cdot 2\|\mathbf{x}\|_2^2\right) = \exp\left(\frac{\beta^2 n}{4}\right). \tag{B.2.3}$$

Hence

$$Z_{\mathrm{pl}} = Z_{\mathrm{GOE}}, \tag{B.2.4}$$

which is extremely useful.

□

B.2.2 Log-Normality of SK Normalizer

We will briefly recap the main results of Aizenman et al. (1987), as they are insightful for our analysis. We must first introduce the partition function

$$Z_n(\beta J) := 2^n \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\frac{\beta J_{ij}}{\sqrt{n}} x_i x_j\right), \tag{B.2.5}$$

where the disorder $J_{ij} \sim \mathcal{N}(0, J^2)$. For our purpose, we will let $J = 1$. We also define the disorder-averaged partition function (denoted $\langle\!\langle Z_n \rangle\!\rangle$ in Aizenman et al. (1987)).

$$\begin{aligned}
\langle\!\langle Z_n \rangle\!\rangle = \mathbb{E}_{\mathbf{J}}[Z_n] := 2^n \prod_{i<j} \mathbb{E}_{\mathbf{J}} \left[\cosh\left(\frac{\beta J_{ij}}{\sqrt{n}} \right) \right] \\
= 2^n \prod_{i<j} \mathbb{E}_{\mathbf{J}} \left[\frac{1}{2} \left(e^{\frac{\beta J_{ij}}{\sqrt{n}}} + e^{-\frac{\beta J_{ij}}{\sqrt{n}}} \right) \right] \\
= 2^n \prod_{i<j} \frac{1}{2} \left(e^{\frac{\beta^2}{2n}} + e^{\frac{\beta^2}{2n}} \right) \\
= 2^n \left(e^{\frac{\beta^2}{2n}} \right)^{\frac{n\cdot(n-1)}{2}} \\
\approx 2^n e^{\frac{\beta^2 n}{4}}, \qquad \text{(B.2.6)}
\end{aligned}$$

where the last equality holds in the large n limit. We now report the main result:

Proposition B.2.1 (Prop. 2.2, Aizenman et al. (1987))
$Z_n / \langle\!\langle Z_n \rangle\!\rangle$ asymptotically tends in distribution to a log-normal distribution, i.e.

$$\frac{Z_n}{\langle\!\langle Z_n \rangle\!\rangle} \xrightarrow{d} \exp(U) \qquad \text{(B.2.7)}$$

with $U \sim \mathcal{N}(-\frac{1}{2}\sigma_U^2, \sigma_U^2)$, and

$$\sigma_U^2 = -\frac{1}{2}\left(\log(1-\beta^2) + \beta^2 \right). \qquad \text{(B.2.8)}$$

To transfer this result on our Z_{SK} from (4.1.12), we must notice three things:

i.) That we can map our rescaled coupling A_{ij} to $\frac{J_{ij}}{\sqrt{n}}$

ii.) That the $e^{-\frac{\beta^2 n}{4}}$ appears from the normalizer

iii.) The diagonal terms (not present in Aizenman et al. (1987)) induce a shift of the variance of the limiting log-normal random variable

We begin by putting Z_n in a more convenient form.

$$
\begin{aligned}
Z_n &:= 2^n \prod_{i<j} \cosh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right) \cdot \widehat{Z}_n \\
&= 2^n \prod_{i<j} \cosh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right) \cdot 2^{-n} \sum_{\mathbf{x}\in\{+1,-1\}^n} \prod_{i<j}\left(1 + x_i x_j \tanh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right)\right) \\
&= 2^n \prod_{i<j} \cosh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right) \cdot 2^{-n} \sum_{\mathbf{x}\in\{-1,1\}^n} \prod_{i<j}\left(\frac{\cosh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right) + x_i x_j \sinh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right)}{\cosh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right)}\right) \\
&= \sum_{\mathbf{x}\in\{-1,1\}^n} \prod_{i<j}\left(\cosh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right) + x_i x_j \sinh\left(\frac{\beta J_{ij}}{\sqrt{n}}\right)\right) \\
&= \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\beta \sum_{i<j} \frac{J_{ij}}{\sqrt{n}} x_i x_j\right). \qquad \text{(B.2.9)}
\end{aligned}
$$

If we consider our model now with $A_{ij} \sim \mathcal{N}(0, \frac{1}{n})$ and $A_{ii} \sim \mathcal{N}(0, \frac{2}{n})$, as well as the symmetry constraint it is easy to see that

$$
2^{-n} \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\frac{\beta}{2}\sum_{i,j} A_{ij} x_i x_j\right) = 2^{-n} \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\beta \sum_{i<j} \frac{J_{ij}}{\sqrt{n}} x_i x_j + \frac{\beta}{2}\sum_{i=1}^{n} A_{ii} x_i^2\right). \qquad \text{(B.2.10)}
$$

With this, we can perform an analysis of Z_{SK}:

$$
\begin{aligned}
Z_{\text{SK}} &:= 2^{-n} \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\frac{\beta}{2}\sum_{i,j} A_{ij} x_i x_j - \frac{\beta^2 n}{4}\right) \\
&= 2^{-n} \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\beta \sum_{i<j} \frac{J_{ij}}{\sqrt{n}} x_i x_j + \frac{\beta}{2}\sum_{i=1}^{n} A_{ii} x_i^2\right) e^{-\frac{\beta^2 n}{4}} \\
&= \underbrace{e^{-\frac{\beta^2 n}{4}} 2^{-n} \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\beta \sum_{i<j} \frac{J_{ij}}{\sqrt{n}} x_i x_j\right)}_{:=Z_n / \langle\!\langle Z_n \rangle\!\rangle} \exp\left(\frac{\beta}{2} W\right) \qquad \text{(B.2.11)}
\end{aligned}
$$

with $\sum_{i=1}^n A_{ii} \overset{d}{=} W \sim \mathcal{N}(0,2)$. Thus by the above Proposition,

$$
Z_{\text{SK}} \xrightarrow{d} \exp(V - \frac{1}{2}\mathbb{E}[V^2]) \qquad \text{(B.2.12)}
$$

where $V = U + \frac{\beta}{2} W$ has zero mean and variance

$$
\operatorname{Var}(U) + \frac{\beta^2}{4}\operatorname{Var}(W) = -\frac{1}{2}\left(\log(1-\beta^2) + \beta^2\right) + \frac{\beta^2}{4}\cdot 2 = -\frac{1}{2}\log(1-\beta^2), \qquad \text{(B.2.13)}
$$

as posited in Alaoui et al. (2024).

□

B.3 Proof of Lemma 4.4.1

By definition, $\mathbf{m}$ is the L^2-optimal estimator of $\mathbf{x}$ given $\mathbf{y}$. A well-known property of this estimator is that we can decompose $\mathbf{x} = \mathbf{m} + \mathbf{e}$ with $\mathbf{e} := \mathbf{x} - \mathbf{m}$ the error, and it holds that

$$\mathbb{E}[\langle \mathbf{m}, \mathbf{e} \rangle] = 0. \tag{B.3.1}$$

Expanding the above gives us

$$\mathbb{E}[\langle \mathbf{m}, \mathbf{x} - \mathbf{m} \rangle] = \mathbb{E}[\langle \mathbf{m}, \mathbf{x} \rangle] - \mathbb{E}[\|\mathbf{m}\|_2^2], \tag{B.3.2}$$

thus proving the claim by expanding the square.

□

B.4 Proof of Proposition 4.4.1

Using Bayes' rule, we express:

$$\mathbb{P}(\bar{X} = 1 \mid H) = \frac{p(H \mid \bar{X} = 1)\mathbb{P}(\bar{X} = 1)}{p(H)},$$

with

$$H \mid \bar{X} \sim \mathcal{N}(a, b).$$

We can now compute the posterior

$$\mathbb{P}(\bar{X} = i \mid H) = \frac{e^{-\frac{(H-ia)^2}{2b}}}{e^{-\frac{(H-a)^2}{2b}} + e^{-\frac{(H+a)^2}{2b}}}, \quad \text{for } i \in \{1, -1\},$$

which leads to the posterior mean

$$\begin{aligned} \mathbb{E}[\bar{X}|H] &= \mathbb{P}(\bar{X} = 1|H) - \mathbb{P}(\bar{X} = -1|H) \\ &= \frac{e^{-\frac{(H-a)^2}{2b}} - e^{-\frac{(H+a)^2}{2b}}}{e^{-\frac{(H-a)^2}{2b}} + e^{-\frac{(H+a)^2}{2b}}} = \tanh\left(\frac{a}{b}H\right). \end{aligned} \tag{B.4.1}$$

□

B.5 Proof of Proposition 4.4.2

Let $\mathbf{Q} = \lambda \mathbf{I}_n + \mathbf{A}$, and consider the Taylor expansion of $\sqrt{\mathbf{Q}}$ around the diagonal matrix $\lambda \mathbf{I}_n$:

$$\sqrt{\lambda \mathbf{I}_n + \mathbf{A}} \approx \sqrt{\lambda \mathbf{I}_n} + \frac{1}{2\sqrt{\lambda}}\mathbf{A} + \mathcal{O}\left(\frac{1}{\lambda^{3/2}}\right). \tag{B.5.1}$$

Using this approximation, we can now write the Euclidean distance as

$$\frac{1}{n}\|\sqrt{\mathbf{Qz}-\sqrt{\lambda}\mathbf{z}}\|_2 \approx \frac{1}{2\sqrt{\lambda}}\frac{1}{n}\|\mathbf{Az}\|_2^2. \tag{B.5.2}$$

From standard probability theory, we can write the expected squared Euclidean norm $\mathbb{E}[\|\mathbf{Az}\|_2^2] = \mathrm{Tr}[\mathbf{A}^2]$, and approximate the Euclidean norm by taking a square root.

$$\mathbb{E}[\|\mathbf{Az}\|_2] \approx \sqrt{\mathrm{Tr}[\mathbf{A}^2]} = \sqrt{\sum_{i=1}^{n}\lambda_i^2(\mathbf{A})}. \tag{B.5.3}$$

As shown in Theorem A.2.1, the eigenvalues of a GOE matrix $\mathbf{A}$ are of order $\mathcal{O}(1)$, making the last term of Eq. (B.5.3) of order $\mathcal{O}(\sqrt{n})$. Putting this back into Eq. (B.5.2) yields

$$\frac{1}{n}\|\sqrt{\mathbf{Q}}\mathbf{z}-\sqrt{\lambda}\mathbf{z}\|_2 \approx \frac{1}{2\sqrt{\lambda n}}, \tag{B.5.4}$$

which is furthermore supported by numerical results, reported in Figure B.1.

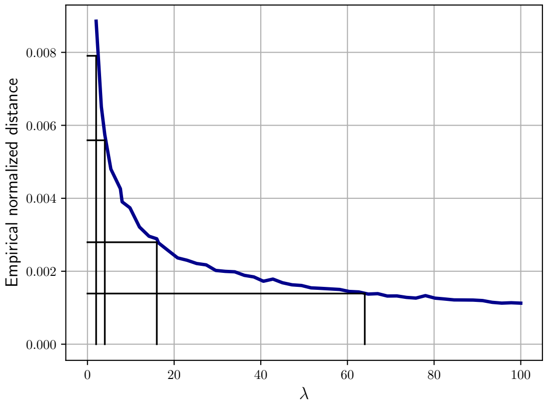


Figure B.1: Blue: plot of the normalized empirical Euclidean distance between $\sqrt{\mathbf{Q}}\mathbf{z}$ and $\sqrt{\lambda}\mathbf{z}$ for $n = 2000$. Black: samples from the function $f : \lambda \mapsto \frac{1}{2\sqrt{\lambda n}}$.

B.6 Proof of Theorem 4.4.1, i.)

Our proof will heavily depend on the proof of Theorem 1 in Montanari and Venkataramanan (2019), but with alterations. We begin by prefacing that while the algorithm runs on *fixed* instances of $\mathbf{y}, \mathbf{A}$, our analysis considers their laws, thus all statements made on convergence are to be understood in the high probability sense.

satisfies the statements of Theorem 4.4.1, and conclude using Proposition 4.4.2.

Remark B.1

We do not employ their Lemma B.3, as our spectral initialization is simply at zero. This is due to the BBP phase transition in Wigner matrices (see Baik et al. (2005);

Benaych-Georges and Nadakuditi (2011)).

The proof will consist of two main steps:

i.) Creating a surrogate AMP iteration $\mathbf{s}^{(k)}$ which will appeal to Theorem 4.3.1, allowing us to work with a formulation of SE .

ii.) Proving that the difference between $\mathbf{s}^{(k)} + \mu_k \bar{\mathbf{x}} + \mathbf{y}$ and $\mathbf{x}^{(k)}$ goes to zero in the large system limit under PL functions.

We begin by expanding Eq. (4.4.21):

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \beta \mathbf{A} f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}) \\ &= \beta \left(\mathbf{W} + \frac{\beta}{n} \bar{\mathbf{x}} \bar{\mathbf{x}}^T \right) f_k(\mathbf{x}^{(k)}) + t\bar{\mathbf{x}} + \sqrt{t}\mathbf{z} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}) \\ &= \left(\frac{\beta^2}{n} \langle f_k(\mathbf{x}^{(k)}, \bar{\mathbf{x}} \rangle + t \right) \bar{\mathbf{x}} + \mathbf{W} \beta f_k(\mathbf{x}^{(k)}) + \sqrt{t}\mathbf{z} - \mathsf{b}_k f_{k-1}(\mathbf{x}^{(k-1)}). \end{aligned} \tag{B.6.1}$$

We now define the random variable $X_k = (\alpha_k + t)\bar{X} + \sqrt{\tau_k^2 + t}G$, where $\bar{X} \sim \bar{\nu}$, $G \sim \mathcal{N}(0,1)$ independent of everything else, and $(\alpha_k, \tau_k)_{k \geq 0}$ are recursively given by

$$\alpha_{k+1} := \beta^2 \mathbb{E}[\bar{X} f_k(X_k)], \tag{B.6.2}$$

$$\tau_{k+1}^2 := \beta^2 \mathbb{E}[f_k(X_k)^2], \tag{B.6.3}$$

initialized at $\alpha_0 = \tau_0 = 0$. Our main goal will be to prove that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi(x_i^{(k)}, \bar{x}_i) \overset{a.s.}{=} \mathbb{E}[\psi(X_k, \bar{X})]. \tag{B.6.4}$$

Note 3
Our state evolution here is defined as $(\alpha_k, \tau_k)_{k \geq 0}$, but it is trivial to see that this exactly matches the state evolution present in Theorem 4.4.1 i.).

Proof of Eq. (B.6.4)

We now create our surrogate AMP iteration:

$$\mathbf{s}^{(0)} = \mathbf{0}_n, \tag{B.6.5}$$

$$\mathbf{s}^{(k+1)} = \beta \mathbf{W} f_k(\mathbf{s}^{(k)} + \alpha_k \bar{\mathbf{x}} + \mathbf{y}) - \tilde{\mathsf{b}}_k f_{k-1}(\mathbf{s}^{(k-1)} + \alpha_{k-1}\bar{\mathbf{x}} + \mathbf{y}), \tag{B.6.6}$$

$$\tilde{\mathsf{b}}_k = \frac{\beta^2}{n} \sum_{i=1}^{n} f_k'(s_i^{(k)} + \alpha_k \bar{x}_i + y_i). \tag{B.6.7}$$

We see that this iteration is that of Eq. (4.3.2), with $f = \beta f_k$. We thus appeal to Theorem 4.3.1, and may state that for any PL function of order 2 $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the following holds.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi(s_i^{(k)}, \bar{x}_i) \overset{a.s.}{=} \mathbb{E}[(\tau_k G, \bar{X})], \tag{B.6.8}$$

where it follows from the i.i.d. distribution of $\mathbf{y}$ that

$$\tau_{k+1}^2 := \beta^2 \mathbb{E}\left[f_k \left((\alpha_k + t)\bar{X} + \sqrt{\tau_k^2 + tG} \right) \right]. \tag{B.6.9}$$

Thus,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi(s_i^{(k)} + \alpha_k \bar{x}_i + y_i, \bar{x}_i) \overset{a.s.}{=} \mathbb{E}\left[\psi \left((\alpha_k + t)\, \bar{X} + \sqrt{\tau_k^2 + tG}, \bar{X} \right) \right]. \tag{B.6.10}$$

In order to prove eq. (B.6.4), we now need to show that

$$\lim_{n\to\infty} \left[\frac{1}{n} \sum_{i=1}^{n} \psi(x_i^{(k)}, \bar{x}_i) - \frac{1}{n} \sum_{i=1}^{n} \psi(s_i^{(k)} + y_i + \alpha_k \bar{x}_i, \bar{x}_i) \right] = 0 \tag{B.6.11}$$

Proof of Eq. (B.6.11)

We define the difference as

$$\mathbf{\Delta}^{(k)} := \mathbf{x}^{(k)} - \left(\mathbf{s}^{(k)} + \alpha_k \bar{\mathbf{x}} + \mathbf{y} \right), \tag{B.6.12}$$

and we prove eq (B.6.11) by induction by showing that

$$\lim_{n\to\infty} \frac{1}{n} \|\mathbf{\Delta}^{(k)}\|_2^2 = 0, \tag{B.6.13}$$

$$\limsup_{n\to\infty} \frac{1}{n} \|\mathbf{x}^{(k)}\|_2^2 < \infty, \tag{B.6.14}$$

$$\limsup_{n\to\infty} \frac{1}{n} \|\mathbf{s}^{(k)} + \alpha_k \bar{\mathbf{x}} + \mathbf{y}\|_2^2 < \infty. \tag{B.6.15}$$

The case of $k = 0$ trivially holds by the initialization assumptions. We now suppose that the above claims hold for $k \in \{0, \dots, \ell\}$, and will prove them for $k = \ell + 1$.

Proof of Eqs. (B.6.14), (B.6.15). We will first show that the ℓ_2-norms of $\mathbf{x}^{(k)}$ and $\mathbf{s}^{(k)} + \alpha_k \bar{\mathbf{x}} + \mathbf{y}$ are finite by the pseudo-Lipschitz property of $\psi(u, v, w) = u^2$. Indeed, we have by Eq. (B.6.10) that

$$\lim_{n\to\infty} \sum_{i=1}^{n} (s_i^{(k)} + \alpha_k \bar{x}_i + y_i)^2 = \mathbb{E}\left[\left(\sqrt{\tau_k^2 + tG} + (\alpha_k + t)\bar{X} \right)^2 \right] \tag{B.6.16}$$

is finite, and Eq. (B.6.11) ensure the same holds of $\|\mathbf{x}^{(k)}\|^2$. □

We may now move onto bounding $\|\mathbf{\Delta}^{(k)}\|$. Note that we first have

$$|\psi(x_i^{(k)}, \bar{x}_i) - \psi(s_i^{(k)} + \alpha_k \bar{x}_i + y_i, \bar{x}_i)| \leq C|\Delta_i^k| \left(1 + |x_i^{(k)}| + |\bar{x}_i| + |s_i^{(k)} + \alpha_k \bar{x}_i + y_i| \right),$$

which yields

$$\frac{1}{n}\left|\sum_{i=1}^{n}\psi(x_i^{(k)},\bar{x}_i)-\psi(s_i^{(k)}+\alpha_k\bar{x}_i+y_i)\right| \leq C\frac{1}{n}|\Delta_i^k|\sum_{i=1}^{n}\left[1+|\bar{x}_i|+|x_i^{(k)}|+|s_i^{(k)}+\alpha_k\bar{x}_i+y_i|\right]$$
$$\leq C\|\mathbf{\Delta}^k\|\left[\frac{1+\|\mathbf{x}^{(k)}\|_2+\|\mathbf{s}^{(k)}+\alpha_k\mathbf{v}\bar{\mathbf{x}}+\mathbf{y}\|_2}{\sqrt{n}}\right]$$

By Eqs. (B.6.1), (B.6.5), we can look at the difference between $\mathbf{x}^{(k+1)}$ and $\mathbf{s}^{(k+1)}$:

$$\begin{aligned}\mathbf{\Delta}^{(k+1)} = \beta^2\langle\bar{\mathbf{x}},f_k(\mathbf{x}^{(k)})\rangle\bar{\mathbf{x}}+\mathbf{y}+\beta\mathbf{W}f_k(\mathbf{x}^{(k)})+\mathfrak{b}_kf_{k-1}(\mathbf{x}^{(k-1)})\\ -\beta\mathbf{W}f_k(\mathbf{s}^{(k)}+\alpha_k\bar{\mathbf{x}}+\mathbf{y})-\tilde{\mathfrak{b}}_kf_{k-1}(\mathbf{s}^{(k-1)}+\alpha_{k-1}\bar{\mathbf{x}}+\mathbf{y})-\alpha_{k+1}\bar{\mathbf{x}}-\mathbf{y}.\end{aligned} \tag{B.6.17}$$

We show that $\frac{1}{n}\|\mathbf{\Delta}^{(k+1)}\|_2^2 \to 0$ by showing that

$$\lim_{n\to\infty}\frac{\beta^2\langle\bar{\mathbf{x}},f_k(\mathbf{x}^{(k)})\rangle}{n}=\alpha_{k+1}, \tag{B.6.18}$$

$$\lim_{n\to\infty}\frac{1}{n}\|\mathbf{W}\left(f_k(\mathbf{x}^{(k)})-f_k(\mathbf{s}^{(k)}-\alpha_k\bar{\mathbf{x}}+\mathbf{y})\right)\|_2^2=0, \tag{B.6.19}$$

$$\lim_{n\to\infty}\frac{1}{n}\|\tilde{\mathfrak{b}}_kf_{k-1}(\mathbf{s}^{(k-1)}+\alpha_{k-1}\bar{\mathbf{x}}+\mathbf{y})-\mathfrak{b}_kf_{k-1}(\mathbf{x}^{(k-1)})\|_2^2=0. \tag{B.6.20}$$

Proof of Eq. (B.6.18). This holds almost surely by appealing to state evolution for the pseudo-Lipschitz function $\psi(a,b)=a\cdot b$ which holds for k, giving us

$$\lim_{n\to\infty}\frac{1}{n}\beta^2\sum_{i=1}^{n}\bar{x}_if_k(x_i^{(k)}) \overset{a.s.}{=} \beta^2\mathbb{E}[\bar{X}f_k(X_k)] =: \alpha_{k+1}. \tag{B.6.21}$$

□

Proof of Eq. (B.6.19). By boundedness of the spectrum (Vershynin (2018), see also Theorem A.2.1) of $\mathbf{W}$, it holds that

$$\|\mathbf{W}\|^2 \leq \|\mathbf{W}\|_{\text{op}}^2 = 4. \tag{B.6.22}$$

Thus, it suffices to show that

$$\frac{1}{n}\|f_k(\mathbf{x}^{(k)})-f_k(\mathbf{s}^{(k)}-\alpha_k+\mathbf{y})\|_2^2 \tag{B.6.23}$$

is bounded. By the presumed Lipschitz property of f, it holds that

$$\|f_k(\mathbf{x}^{(k)})-f_k(\mathbf{s}^{(k)}-\alpha_k+\mathbf{y})\|_2^2 \leq C\|\mathbf{\Delta}^{(k)}\|_2^2 \tag{B.6.24}$$

for an absolute constant $C>0$. By the induction hypothesis, $\frac{\|\mathbf{\Delta}^{(k)}\|_2^2}{n} \to 0$ almost surely. Thus proving our claim. □

Proof of Eq. (B.6.20). For ease of notation, we will introduce $\mathbf{r}^{(k)} := \mathbf{s}^{(k)} + \alpha_k \bar{\mathbf{x}} + \mathbf{y}$, as well as drop the indices on f. We first upper bound with a triangle inequality:

$$\begin{aligned}&\frac{1}{n}\|\tilde{\mathsf{b}}_k f(\mathbf{r}^{(k-1)})) - \mathsf{b}_k f(\mathbf{x}^{(k-1)}))\|_2^2 \\ &= \frac{1}{n}\|\tilde{\mathsf{b}}_k f(\mathbf{r}^{(k-1)}) + \mathsf{b}_k f(\mathbf{r}^{(k-1)}) - \mathsf{b}_k f(\mathbf{r}^{(k-1)}) - \mathsf{b}_k f(\mathbf{x}^{(k-1)})\|_2^2 \\ &\leq \frac{1}{n}\|\tilde{\mathsf{b}}_k f(\mathbf{r}^{(k-1)}) - \mathsf{b}_k f_k(\mathbf{r}^{(k-1)})\|_2^2 + \frac{1}{n}\mathsf{b}_k^2\|f(\mathbf{r}^{(k-1)}) - f(\mathbf{x}^{(k-1)})\|_2^2. \qquad \text{(B.6.25)}\end{aligned}$$

The first term of the upper bound can be bounded by the Lipschitz property of f, such that

$$\frac{1}{n}\|\tilde{\mathsf{b}}_k f(\mathbf{r}^{(k-1)}) - \mathsf{b}_k f_k(\mathbf{r}^{(k-1)})\|^2 \leq \frac{C(\tilde{\mathsf{b}}_k - \mathsf{b}_k)}{n}\|\mathbf{r}^{(k-1)}\|_2^2. \qquad \text{(B.6.26)}$$

By the induction hypothesis, $\|\mathbf{r}^{(k-1)}\|_2^2 < \infty$. Next, we use the fact that the limiting distribution of $\mathbf{r}^{(k-1)}$ is given by $(\alpha_k + t)U + \sqrt{\tau_k^2 + t}G$, and use Lemma 5 of Bayati and Montanari (2011) for Lipschitz functions,

$$\lim_{n\to\infty} \tilde{\mathsf{b}}_k = \mathbb{E}[f'((\alpha_k + t)U + \sqrt{\tau_k^2 + t}G)]. \qquad \text{(B.6.27)}$$

This is the same as the limit of b_k (by definition), thus

$$\lim_{n\to\infty} \frac{1}{n}\|\tilde{\mathsf{b}}_k f(\mathbf{r}^{(k-1)}) - \mathsf{b}_k f_k(\mathbf{r}^{(k-1)})\|_2^2 \overset{a.s.}{=} 0, \qquad \text{(B.6.28)}$$

The second part of the upper bound is bounded by the Lipschitz property of f:

$$\frac{1}{n}\mathsf{b}_k^2\|f(\mathbf{r}^{(k-1)}) - f(\mathbf{x}^{(k-1)})\|^2 \leq \frac{\mathsf{b}_k^2}{n}C\|\mathbf{\Delta}^{(k-1)}\|_2^2 \qquad \text{(B.6.29)}$$

which goes to 0 by the induction hypothesis. Thus, we have shown that (B.6.20) goes to zero in the large system limit, and that the surrogate AMP iteration from Eq. (B.6.5) and the true AMP iteration from Eq. (4.4.21) get arbitrarily close. □

B.7 Proof of Theorem 4.4.1 ii.)

We now treat the anisotropic case. The proof follows the same spirit as Appendix B.6, but the analysis of the state evolution will require a few extra steps, and so will the convergence of the difference between the surrogate AMP and the actual AMP iteration. Before anything else, consider a slight variation of the anisotropic side-information vector,

$$\tilde{\mathbf{y}} = \mathbf{Q}t\bar{\mathbf{x}} + \sqrt{\lambda t}\mathbf{z}. \qquad \text{(B.7.1)}$$

Our proof technique will consist of proving that Theorem 4.4.1 with $\tilde{\mathbf{y}}$ holds, and then appeal to Proposition 4.4.2 to prove that AMP with $\mathbf{y}$ converges in probability to our derived result.

We again begin by decomposing the AMP iteration from Eq. (4.4.21),

$$\begin{aligned}\tilde{\mathbf{x}}^{(k+1)} &= \beta \mathbf{A} f_k(\tilde{\mathbf{x}}^{(k)}) + \tilde{\mathbf{y}} - \mathsf{b}_k f_{k-1}(\tilde{\mathbf{x}}^{(k-1)})\\ &\overset{a)}{=} \frac{\beta^2}{n}\langle f_k(\tilde{\mathbf{x}}^{(k)}), \bar{\mathbf{x}}\rangle \bar{\mathbf{x}} + \beta \mathbf{W} f_k(\mathbf{x}^{(k)}) + \tilde{\mathbf{y}} - \mathsf{b}_k f_{k-1}(\tilde{\mathbf{x}}^{(k-1)})\\ &\overset{b)}{=} \left(\frac{\beta^2}{n}\langle f_k(\tilde{\mathbf{x}}^{(k)}), \bar{\mathbf{x}}\rangle + (\lambda+\beta)t\right)\bar{\mathbf{x}} + \beta \mathbf{W} f_k(\tilde{\mathbf{x}}^{(k)}) + t\mathbf{W}\bar{\mathbf{x}} - \mathsf{b}_k f_{k-1}(\tilde{\mathbf{x}}^{(k-1)}) + \sqrt{\lambda t}\mathbf{z},\end{aligned} \tag{B.7.2}$$

where $a)$ holds by plugging the planted spiked matrix $\mathbf{A}$, $b)$ holds by plugging Eq. (B.7.1).

Again, we define an auxiliary random variable $\tilde{X}_k = (\alpha_k + c(t))\bar{X} + \sqrt{\tau_k^2 + \lambda t}G$, with $c(t) := (\lambda+\beta)t$, $\bar{X} \sim \bar{\nu}$, $G \sim \mathcal{N}(0,1)$ and independent Gaussian, and $(\alpha_k, \tau_k)_{k\geq 0}$ are recursively defined as

$$\alpha_{k+1} := \beta^2 \mathbb{E}[\bar{X} f_k(\tilde{X}_k)], \tag{B.7.3}$$

$$\tau_{k+1}^2 := \mathbb{E}[(\beta f_k(\tilde{X}_k) + t\bar{X})^2], \tag{B.7.4}$$

initialized at $\alpha_0 = \tau_0 = 0$. We now define our surrogate AMP iteration. Let

$$\mathbf{r}^{(k)} := \mathbf{s}^{(k)} + (\alpha_k + c(t))\bar{\mathbf{x}} + \sqrt{\lambda t}\mathbf{z}, \tag{B.7.5}$$

and

$$\mathbf{s}^{(0)} = \mathbf{0}_n, \tag{B.7.6}$$

$$\mathbf{s}^{(k+1)} = \mathbf{W}\left(\beta f_k(\mathbf{r}^{(k)}) + t\bar{\mathbf{x}}\right) - \tilde{\mathsf{b}}_k f_{k-1}(\mathbf{r}^{(k-1)}), \tag{B.7.7}$$

$$\tilde{\mathsf{b}}_k = \frac{\beta^2}{n}\sum_{i=1}^n f'(r_i^{(k)}). \tag{B.7.8}$$

As in Appendix B.6, we notice that $\tilde{\mathsf{b}}_k = \mathsf{b}_k$, as the addition of the constant $t\bar{\mathbf{x}}$ does not affect the derivative. Using Theorem 4.3.1, we can state that for any PL function $\psi : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ of order 2, we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \psi\left(s_i^{(k)}, \bar{x}_i\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left(\tau_k G, \bar{X}\right)\right], \tag{B.7.9}$$

and analogously to Eq. (B.6.10), we can state that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \psi\left(r_i^{(k)}, \bar{x}_i\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left((\alpha_k + c(t))\bar{X} + \sqrt{\tau_k^2 + \lambda t}G\right), \bar{X}\right]. \tag{B.7.10}$$

Remark B.2
It may seem odd (or even wrong) that the iteration from Eq. (B.7.6) appeals to the format of Theorem (4.3.1), due to the extra $t\bar{\mathbf{x}}$ that is now "included in the non-linearity". We recall the discussion in Section 4.3, and we also refer to the discussion in Section 2 of Montanari and Venkataramanan (2019). The key is that AMP's Onsager term can be expressed as $\mathbb{E}[\mathbf{W} f_{k-1}(\tilde{\mathbf{x}}^{(k-1)})|\mathscr{F}_k]$. By noticing that

$$\mathbb{E}[\mathbf{W} f_{k-1}(\tilde{\mathbf{x}}^{(k-1)})|\mathscr{F}_k] = \mathbb{E}[\mathbf{W}(f_{k-1}(\tilde{\mathbf{x}}^{(k-1)}) + t\bar{\mathbf{x}})|\mathscr{F}_k], \tag{B.7.11}$$

the issue is solved.

The remaining issue will be in showing that

$$\lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\psi\left(\tilde{x}_i^{(k)}, \bar{x}_i\right) - \frac{1}{n}\sum_{i=1}^{n}\psi\left(r_i^{(k)}, \bar{x}_i\right)\right] \overset{a.s.}{=} 0. \tag{B.7.12}$$

Proof of Eq. (B.7.12)

We again define the difference by

$$\mathbf{\Delta}^{(k)} := \tilde{\mathbf{x}}^{(k)} - \mathbf{r}^{(k)}. \tag{B.7.13}$$

To avoid repetition, we skip directly to the inductive step of proving that $\lim_{n\to\infty} n^{-1}\|\mathbf{\Delta}^{(k+1)}\|_2^2 = 0$. For a more extensive proof, see Appendix B.6 and the proofs of Eqs. (B.6.14), (B.6.15).

We expand $\mathbf{\Delta}^{(k)}$ by plugging in Eqs. (B.7.2) and (B.7.5).

$$\begin{aligned}
\mathbf{\Delta}^{(k+1)} &:= \tilde{\mathbf{x}}^{(k+1)} - \mathbf{r}^{(k+1)} \\
&= \left(\frac{\beta^2}{n}\langle\bar{\mathbf{x}}, f_k(\tilde{\mathbf{x}}^{(k)})\rangle + c(t)\right)\bar{\mathbf{x}} + \mathbf{W}\left(\beta f_k(\tilde{\mathbf{x}}^{(k)}) + t\bar{\mathbf{x}}\right) \\
&\quad + \sqrt{\lambda t}\mathbf{z} - \mathfrak{b}_k f_k(\mathbf{x}^{(k)}) - \mathbf{r}^{(k+1)} \\
&= \left(\frac{\beta^2}{n}\langle\bar{\mathbf{x}}, f_k(\tilde{\mathbf{x}}^{(k)})\rangle - \alpha_{k+1}\right)\bar{\mathbf{x}} \\
&\quad + \mathbf{W}\left(\beta f_k(\tilde{\mathbf{x}}^{(k)}) + t\bar{\mathbf{x}}\right) - \mathfrak{b}_k f_k(\tilde{\mathbf{x}}^{(k)}) - \mathbf{s}^{(k+1)}.
\end{aligned} \tag{B.7.14}$$

Again, we will show this by proving that

$$\lim_{n\to\infty}\frac{\beta^2}{n}\langle\bar{\mathbf{x}}, f_k(\tilde{\mathbf{x}}^{(k)})\rangle = \alpha_{k+1}, \tag{B.7.15}$$

$$\lim_{n\to\infty}\|\mathbf{W}(f_k(\tilde{\mathbf{x}}^{(k)}) - f_k(\mathbf{r}^{(k)}))\|_2^2 = 0, \tag{B.7.16}$$

$$\lim_{n\to\infty}\|\tilde{\mathfrak{b}}_k f_{k-1}(\mathbf{r}^{(k)}) - \mathfrak{b}_k f_{k-1}(\tilde{\mathbf{x}}^{(k)})\|_2^2 = 0, \tag{B.7.17}$$

$$\tag{B.7.18}$$

Proof of Eqs. (B.7.15), (B.7.16) *and* (B.7.17). This is completely analogous to the proofs of Eqs. (B.6.18), (B.6.19) and (B.6.20) from Appendix B.6. □

Proving Convergence in Probability

We thus have thus proven Theorem 4.4.1 for the AMP iterates defined in Eq. (B.7.2). To prove the statement, we use Proposition 4.4.2 to show that iterates defined as

$$\mathbf{x}^{(k+1)} = \beta \mathbf{A} f_k(\mathbf{x}^{(k)}) + \mathbf{y} - \hat{\mathfrak{b}}_k f_{k-1}(\mathbf{x}^{(k-1)}) \tag{B.7.19}$$

with $\hat{\mathfrak{b}}_k = \frac{\beta^2}{n} \sum_{i=1}^n f_k'(x_i^k)$ satisfy

$$\frac{1}{n} \|\mathbf{x}^{(k)} - \tilde{\mathbf{x}}^{(k)}\|_2 \in \mathcal{O}\left(\frac{1}{\sqrt{\lambda n}}\right). \tag{B.7.20}$$

To prove Eq. (B.7.20), we again define a difference

$$\mathbf{\Xi}^{(k)} := \mathbf{x}^{(k)} - \tilde{\mathbf{x}}^{(k)}, \tag{B.7.21}$$

and will inductively prove that $\lim_{n\to\infty} \frac{1}{n} \|\mathbf{\Xi}^{(k+1)}\|_2 \in \mathcal{O}\left(\frac{1}{\sqrt{\lambda n}}\right)$ for a fixed k.

The base case will be simple, as we can consider $\mathbf{x}^{(1)}$ and $\tilde{\mathbf{x}}^{(1)}$. Their difference is $\sqrt{\mathbf{Q}t}\mathbf{z} - \sqrt{\lambda t}\mathbf{z}$, which satisfies our claim by Proposition 4.4.2. We now suppose that the claim holds for $\ell \in \{1, \dots, k\}$, and prove it for $k+1$.

$$\begin{aligned} \mathbf{\Xi}^{(k+1)} = {} & \overbrace{\frac{\beta^2}{n} \langle \bar{\mathbf{x}}, f_k(\mathbf{x}^{(k)}) - f_k(\tilde{\mathbf{x}}^{(k)} \rangle \bar{\mathbf{x}}}^{:=\textcircled{1}} + \overbrace{\left(\sqrt{\mathbf{Q}t} - \sqrt{\lambda t}\right)\mathbf{z}}^{:=\textcircled{2}} \\ & + \underbrace{\mathbf{W}\beta\left(f_k(\mathbf{x}^{(k)}) - f_k(\tilde{\mathbf{x}}^{(k)})\right)}_{:=\textcircled{3}} - \underbrace{\hat{\mathfrak{b}}_k f_{k-1}(\mathbf{x}^{(k-1)}) + \mathfrak{b}_k f_{k-1}(\tilde{\mathbf{x}}^{(k-1)})}_{:=\textcircled{4}}. \end{aligned} \tag{B.7.22}$$

The claim follows by showing that the normalized ℓ_2 norms of all elements either go to zero or satisfy the square root decay.

Proof for $\textcircled{1}$. We prove this by showing that the squared ℓ_2 norm asymptotically tends to zero. By the Lipschitz property of f_k, we have

$$\frac{1}{n} \|f_k(\mathbf{x}^{(k)}) - f_k(\tilde{\mathbf{x}}^{(k)})\|_2^2 \le \frac{C}{n} \|\mathbf{\Xi}^{(k)}\|_2^2 \to 0$$

□

Proof for $\textcircled{2}$. By Proposition 4.4.2, we have that

$$\frac{1}{n} \|\sqrt{\mathbf{Q}}\mathbf{z} - \sqrt{\lambda}\mathbf{z}\|_2 \in \mathcal{O}\left(\frac{1}{\sqrt{\lambda n}}\right).$$

□

Proof for ③. By the same arguments of boundedness of $\|\mathbf{W}\|_{\text{op}}$ used in the proof of convergence of Eq. (B.6.19), we have that

$$\frac{1}{n}\left\|\mathbf{W}\beta\left(f_k(\mathbf{x}^{(k)}) - f_k(\tilde{\mathbf{x}}^{(k)})\right)\right\|_2^2 \leq \frac{4\beta^2}{n}\|f_k(\mathbf{x}^{(k)}) - f_k(\tilde{\mathbf{x}}^{(k)})\|_2^2,$$

which goes to zero by the same Lipschitz argument as for ①.

□

Proof for ④. Finally, we apply the same arguments as for Eq. (B.6.20):

$$\begin{aligned}
&\frac{1}{n}\|\mathfrak{b}_k f(\tilde{\mathbf{x}}^{(k-1)}) - \hat{\mathfrak{b}}_k f(\mathbf{x}^{(k-1)})\|_2^2 \\
&= \frac{1}{n}\|\mathfrak{b}_k f(\tilde{\mathbf{x}}^{(k-1)}) + \hat{\mathfrak{b}}_k f(\tilde{\mathbf{x}}^{(k-1)}) - \hat{\mathfrak{b}}_k f(\tilde{\mathbf{x}}^{(k-1)}) - \hat{\mathfrak{b}}_k f(\mathbf{x}^{(k-1)})\|_2^2 \\
&\leq \frac{1}{n}\|\mathfrak{b}_k f(\tilde{\mathbf{x}}^{(k-1)}) - \hat{\mathfrak{b}}_k f_k(\mathbf{x}^{(k-1)})\|_2^2 + \frac{1}{n}\hat{\mathfrak{b}}_k^2\|f(\tilde{\mathbf{x}}^{(k-1)}) - f(\mathbf{x}^{(k-1)})\|_2^2. \qquad \text{(B.7.23)}
\end{aligned}$$

We again use the same arguments to prove that the first term goes to zero by using the Lipschitz property of f_{k-1}, and the convergence of $|\mathfrak{b}_k - \hat{\mathfrak{b}}_k|$ to zero. The second term tends to zero by the Lipschitz property of f_{k-1} and the induction hypothesis that $\frac{1}{n}\|\boldsymbol{\Xi}^{(k-1)}\|_2^2 \to 0$. □

Putting all four statements together we indeed have that

$$\frac{1}{n}\|\boldsymbol{\Xi}^{(k+1)}\|_2 \in \mathcal{O}\left(\frac{1}{\sqrt{\lambda n}}\right)$$

for fixed k. Let $\xi_n^{(k)} := \frac{1}{n}\|\boldsymbol{\Xi}^{(k)}\|_2$. Convergence in probability follows by Markov's inequality:

$$\mathbb{P}\left(\xi_n^{(k)} \geq \epsilon\right) \leq \frac{\mathbb{E}[\xi_n^{(k)}]}{\epsilon}. \qquad \text{(B.7.24)}$$

We can now show that the RHS tends to zero for all fixed $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{\Xi}^{(k)}\|_2^2\right]\frac{1}{\epsilon} = \lim_{n\to\infty}\mathcal{O}\left(\frac{1}{\sqrt{\lambda n}}\right)\frac{1}{\epsilon} = 0, \qquad \text{(B.7.25)}$$

proving convergence in probability, and thus the entire claim.

□

Appendix C

Extra Results

This section covers a topic which took a significant amount of time to develop, but was found to unfortunately be unfruitful. We have simply chosen to include it in the thesis as a reflection of the thought process.

C.1 Orthogonal AMP for Pre-Processing

We consider the setup of the planted model in Eq. (4.1.13). That is, we generate a signal $\mathbf{x} \in \mathbb{R}^n$, a spiked GOE matrix $\mathbf{A} = \frac{\beta}{n}\mathbf{x}\mathbf{x}^T + \mathbf{W}$, and an anisotropic observation process $\mathbf{y}^{\text{anis}}(t) = \mathbf{Q}t\mathbf{x} + \sqrt{\mathbf{Q}t}\mathbf{z}$, where $\mathbf{Q} = \lambda\mathbf{I}_n + \mathbf{A}$. The generative processes are pictorially described in Figure C.1. In this Bayesian setting, we wish to approximate the posterior mean of $\mathbf{x}$ given $(\mathbf{A}, \mathbf{y}^{\text{anis}}(t))$.

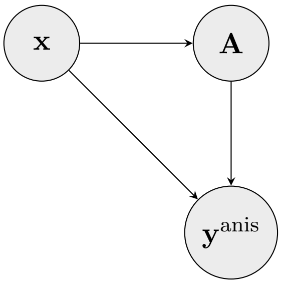


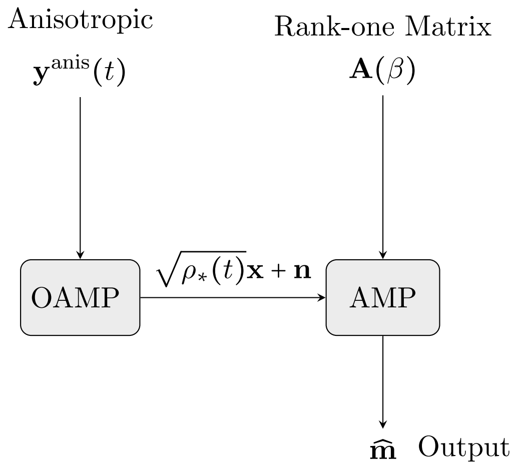
Figure C.1: Bayesian network of the planted model. Notice that $\mathbf{y}^{\text{anis}}$ is "doubly" dependent on $\mathbf{x}$, which is very different from the disorder-independent and isotropic case, in which the rank-one matrix and the observation process are conditionally independent on $\mathbf{x}$.

While similar to the disorder-independent and isotropic case of Alaoui et al. (2024), the key difference is the dependence of $\mathbf{y}^{\text{anis}}$ on $\mathbf{A}$, which greatly complicates the situation. In Alaoui et al. (2024), the posterior mean is computed by using AMP (see Algorithm 2) Montanari and Venkataramanan (2019) for matrix factorization. Unlike in Montanari and Venkataramanan (2019), the AMP algorithm is *not* spectrally initialized[1], but makes use of the isotropic observation process $\mathbf{y}^{\text{iso}}(t)$ as side information. Informally, it is like using an increasingly more informative spectral initialization at every timestep. We will propose a different estimation scheme for two reasons:

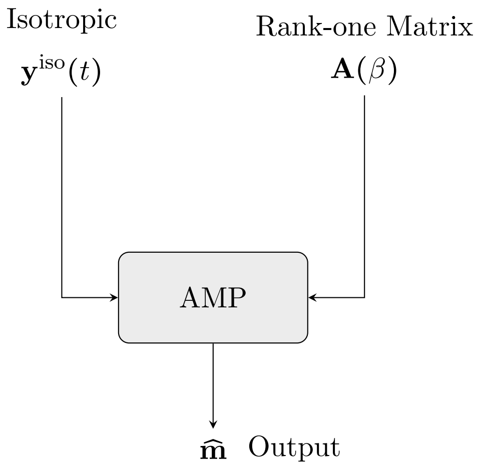
i.) The proof of state evolution for AMP has strong requirements on the form of the side information, which our anisotropic process does not satisfy

ii.) The double dependency is difficult to deal with.

Our method will be a two-step estimation of the posterior mean $\mathbb{E}[\mathbf{x}|\mathbf{y}^{\mathrm{anis}}(t), \mathbf{A}]$. In the first step, we aim to decouple the observation process into an isotropic one, which we then use as side information to run AMP. In particular, the decoupling procedure will be done by Orthogonal AMP (OAMP), introduced by Ma and Ping (2016). Informally, if the first step does not "lose information", then the resulting estimator should still be Bayes optimal. See Figure C.2 for a comparison of our method with that of Alaoui et al. (2024).


Anisotropic
$\mathbf{y}^{\mathrm{anis}}(t)$
Rank-one Matrix
$\mathbf{A}(\beta)$
OAMP
$\sqrt{\rho_*(t)}\mathbf{x} + \mathbf{n}$
AMP
$\widehat{\mathbf{m}}$ Output

(a) The anisotropic process $\mathbf{y}^{\mathrm{anis}}$ is first decoupled by OAMP, and produces an isotropic process of SINR $\rho_*(t)$. This is then fed into AMP as side-information.



(b) The isotropic process $\mathbf{y}^{\mathrm{iso}}$ is directly fed into AMP.

Figure C.2: Comparison of the OAMP+AMP procedure with isotropic AMP. The OAMP pre-processing ensures that we do not need to change anything in AMP.

OAMP is an algorithmic framework designed to address high-dimensional linear inverse problems, particularly in compressed sensing, and served as an extension to the AMP algorithm introduced by Bayati and Montanari (2011), which unfortunately struggles when considering beyond i.i.d. Gaussian measurements. OAMP was introduced as a solution to this issue, and employs orthogonalization techniques to mitigate these challenges. By incorporating a decoupling step through orthogonal projections, OAMP effectively reduces the

[1] This is also due to the presumed low SNR regime $\beta < 1$, in which any leading eigenvector estimator of $\mathbf{A}$ is asymptotically orthogonal to $\mathbf{x}$ Baik et al. (2005).

correlation between residuals and observations, enabling a more accurate approximation of the posterior distribution of $\mathbf{x}_*$. Rigorous proofs of the numerical performances were offered by Takeuchi (2017), allowing OAMP to become a viable compressed sensing algorithm.

The OAMP Algorithm

We will introduce this algorithm as presented by Ma and Ping, as well as the proofs introduced by Takeuchi. It was used as a denoiser of vector AWGN channels, where the noisy linear observation of a given signal $\mathbf{x}_* \in \mathbb{R}^n$ is given by

$$\mathbf{y} := \mathbf{A}\mathbf{x}_* + \mathbf{z}, \tag{C.1.1}$$

where $\mathbf{A}$ is rotationally invariant, and $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Before any analysis is made, we will state the assumptions under which our results hold:

Assumption 1
The signal vector $\mathbf{x}_*$ has i.i.d. non-Gaussian elements with mean zero and unit variance. Moreover, for some $\varepsilon > 0$, the elements have finite $(2 + \varepsilon)$-th moment.

Assumption 2
The sensing matrix $\mathbf{A}$ has the following properties:

- $\mathbf{A}^T\mathbf{A}$ *is unitarily invariant*
- *The empirical eigenvalue distribution of $\mathbf{A}^T\mathbf{A}$ asymptotically converges almost surely to a deterministic distribution $\rho(\lambda)$ with a compact support.*

Assumption 3
The noise vector is zero-mean i.i.d. Gaussian with variance σ^2. (This can be relaxed, see Takeuchi (2017)).

Remark C.1
To clear any confusion, we note that the unitary invariance is ***only*** *required on $\mathbf{A}^T\mathbf{A}$. This implies that the sensing matrix $\mathbf{A}$ can be relaxed to right-rotationally invariant (respectively left-rotationally invariant). This is the requirement for Vector AMP (VAMP), introduced by Rangan, Schniter, and Fletcher (2016), and actually was later shown to be equivalent to AMP.*

The Iteration

As most Approximate Message Passing (AMP) algorithms, OAMP is comprised of two steps, an affine transformation of the current iterate, and a non-linear transformation via a function $\eta_k : \mathbb{R} \to \mathbb{R}$ applied component-wise on the iterate. Following the notation of Ma and Ping (2016), the OAMP algorithm is given by

$$\mathbf{r}^t = \mathbf{s}^t + \gamma_t \mathbf{W}_t (\mathbf{y} - \mathbf{A}\mathbf{s}_t) \tag{C.1.2}$$

$$\mathbf{s}^{t+1} = \eta_t(\mathbf{r}^t) \tag{C.1.3}$$

initialized at $\mathbf{s}^0 = \mathbf{0}$. In the first line, the matrix $\mathbf{W}_k$ is the linear minimum mean-square error (LMMSE) filter, given by

$$\mathbf{W}_k := \mathbf{A}^T(\sigma^2\mathbf{I}_n + \omega_k^2\mathbf{A}\mathbf{A}^T)^{-1}, \tag{C.1.4}$$

and γ_t is a normalization coefficient, given by

$$\frac{1}{\gamma_k} := \lim_{n\to\infty} \mathrm{Tr}[\mathbf{W_k}\mathbf{A}]. \tag{C.1.5}$$

Remark C.2
Note that by Assumption 2,

$$\frac{1}{\gamma_k} \overset{a.s.}{=} \frac{1}{\gamma(\omega_k^2)}, \tag{C.1.6}$$

with

$$\frac{1}{\gamma(v)} := \int \frac{1}{\sigma^2 + v\lambda}\rho(\mathrm{d}\lambda). \tag{C.1.7}$$

Furthermore, the state evolution (SE) of the algorithm is given by

$$\chi_k^2 = \gamma(\omega_k^2) - \omega_k^2, \tag{C.1.8}$$

$$\omega_{k+1}^2 = \left(\frac{1}{\mathrm{mmse}(\chi_k^{-2})} - \frac{1}{\chi^2}\right)^{-1} \tag{C.1.9}$$

initialized at $\omega_0 = \frac{1}{n}\mathbb{E}[\|\mathbf{x}_*\|_2^2] = 1$ (see Eq.s (46)-(47) of Takeuchi (2017)).

The non-linearity $\eta_k(\cdot)$ is to be chosen depending on the application, but as we are interested in Bayesian estimation, we will consider

$$\eta_k(z){:=}\omega_{k+1}^2\left(\frac{\tilde{\eta}_k(z)}{v_{k+1}} - \frac{z}{\chi_k^2}\right), \tag{C.1.10}$$

where

$$\tilde{\eta}_k(\mathbf{r}^k) := \mathbb{E}[\mathbf{x}_*|\mathbf{x}_* + \chi_k^2\mathbf{z} = \mathbf{r}^k], \tag{C.1.11}$$

$$v_k := \frac{1}{n}\left(\mathbb{E}[\|\mathbf{x}_*\|_2^2|\mathbf{r}^k] - \|\tilde{\eta}_k(\mathbf{r}^k)\|_2^2\right). \tag{C.1.12}$$

Halting & Output

If a halting condition is met (or simply a maximal number of iterations), the algorithm does *not* output $\tilde{\eta}_k(\mathbf{r}^k)$, but rather $\chi_k^{-2}\mathbf{r}^k$. We note that this value is computable, as the algorithm must compute χ_k^{-2} at every step anyways.

Main Result

We will show that following a result from Takeuchi (2017), we can characterize the iterates of OAMP as noisy observations of the empirical limiting distribution of the input signal.

Theorem C.1.1
Under Assumptions 1, 2, 3, it holds that for any pseudo-Lipschitz functions $\psi, \phi : \mathbb{R}^{k+1} \to \mathbb{R}$*, the following holds:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(\mathbf{r}_i^1, \mathbf{r}_i^2, \ldots, \mathbf{r}_i^k, \mathbf{x}_{*,i}\right) = \mathbb{E}\left[\psi(R_1, R_2, \ldots, R_k, X)\right], \tag{C.1.13}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(\mathbf{s}_i^1, \mathbf{s}_i^2, \ldots, \mathbf{s}_i^k, \mathbf{x}_{*,i}\right) = \mathbb{E}\left[\phi(S_1, S_2, \ldots, S_k, X)\right], \tag{C.1.14}$$

where $X \sim p_{X_*}$ *is the prior,* $R_k = X + \chi_k G_k$*, where* G_k *are standard Gaussians, independent of* X*, and* $(S_1, S_2, \ldots, S_k) \sim \mathcal{N}(0, \mathbf{\Omega}_k)$*, where*

$$\mathbf{\Omega}_k = \begin{bmatrix} \omega_1^2 & \omega_1\omega_2 & \ldots & \omega_1\omega_k \\ \omega_2\omega_1 & \omega_2^2 & \ldots & \omega_2\omega_k \\ \vdots & \vdots & \ddots & \vdots \\ \omega_k\omega_1 & \omega_k\omega_2 & \ldots & \omega_k^2 \end{bmatrix}, \tag{C.1.15}$$

and $\{\chi_k, \omega_k\}_k$ *are given by the SE of the OAMP algorithm (C.1.8).*

Proof. See Appendix A of Takeuchi (2017), and Section 3 & 4 of Fan (2022). □

This characterization of the iterates allows for a statement on the estimator at each iteration.

Theorem C.1.2 (Theorem 2, Takeuchi (2017))
Let $\{\chi_k^2, \omega_k^2\}$ *as in (C.1.8). Then, the following statements hold:*

$$\lim_{n\to\infty} \frac{1}{n} \|\mathbf{r}^k - \mathbf{x}_*\|_2^2 \overset{a.s.}{=} \chi_k^2, \tag{C.1.16}$$

$$\lim_{n\to\infty} \frac{1}{n} \|\tilde{\eta}_k(\mathbf{r}^k) - \mathbf{x}_*\|_2^2 \overset{a.s.}{=} \mathrm{mmse}(\chi_k^{-2}). \tag{C.1.17}$$

We also have an important statement on the convergence of the SE.

Lemma C.1.1 (Fixed Point of SE. Theorem 3, Ma and Ping (2016))

The SE reaches a fixed point coinciding with that of the replica symmetric prediction:

$$\frac{1}{\chi^2} = \frac{1}{\sigma^2}\mathcal{R}_{\mathbf{A}^T\mathbf{A}}\left(-\frac{1}{\sigma^2}\mathrm{mmse}(\chi^{-2})\right), \tag{C.1.18}$$

where $\mathcal{R}_a(\cdot)$ denotes the R-transform (see Definition A.2.5) of a non-commutative random variable a.

With this result in hand, we provide a quick example with the traditional AMP as a sanity check.

Example 3 (Gaussian i.i.d. Matrix)
When $\mathbf{A} \in \mathbb{R}^{n\times m}$ is i.i.d. Gaussian (as in Bayati and Montanari (2011)), the empirical eigenvalue distribution of $\mathbf{A}^T\mathbf{A}$ asymptotically converges to the Marchenko-Pastur distribution (c.f. Definition A.2.2) with aspect ratio $\beta = m/n$. The $\mathcal{R}$-transform of the MP distribution is given by

$$\mathcal{R}_{\mathbf{A}^T\mathbf{A}}(z) = \frac{1}{1-\beta z}, \tag{C.1.19}$$

whence Eq. (C.1.18) yields

$$\begin{aligned}\chi^{-2} &= \frac{1}{\sigma^2}\left(\frac{1}{1+\frac{\mathrm{mmse}(\chi^2)}{\sigma^2}}\right)\\ &\iff \\ \chi^2 &= \sigma^2 + \mathrm{mmse}(\chi^2),\end{aligned} \tag{C.1.20}$$

successfully recovering the state evolution of the original AMP algorithm Eq. (1.4) of Bayati and Montanari (2011) with aspect ratio $\beta \equiv \frac{1}{\delta}$. We can further the analysis by using the I-MMSE relation (Guo, Shamai, and Verdu (2004)), integrating with respect to the SNR $\chi^{-2} \equiv \rho$:

$$\frac{1}{2}(\rho^{-1} - \sigma^2) = \frac{1}{2}\mathrm{mmse}(\rho) \tag{C.1.21}$$

implies that χ^2 is a critical point of the free energy functional

$$\Phi(\rho, \sigma^2) = \frac{\sigma^2\rho}{2} - \frac{1}{2}\ln(\rho) + \mathcal{I}(\rho), \tag{C.1.22}$$

where $\mathcal{I}(\rho)$ denotes the mutual information between a random variable X and a noise observation of itself with SNR ρ, $Y \equiv \sqrt{\rho}X + Z$, with $Z \sim \mathcal{N}(0,1)$. This is also (up to additive constant) the free energy functional described in Montanari and Wu (2024) for i.i.d. Gaussian estimation, confirming our sanity check.

Anisotropic Observations

We will now treat the case of a time-dependent observation process

$$\mathbf{y}(t) = \mathbf{Q}t\mathbf{x}_* + \sqrt{\mathbf{Q}t}\mathbf{z}, \tag{C.1.23}$$

where $\mathbf{Q}$ is a symmetric random matrix drawn from a rotationally invariant ensemble, and $t > 0$. We also add a final assumption on the sensing matrix.

Assumption 4
$\mathbf{Q}$ *is invertible.*

We first consider a time-independent anisotropic observation process:

$$\mathbf{y} = \mathbf{Q}\mathbf{x}_* + \sqrt{\mathbf{Q}}\mathbf{z}.$$

We are interested in computing the Bayes-optimal estimator of $\mathbf{x}_*$ given the observation $\mathbf{y}$ and the sensing matrix $\mathbf{Q}$. By Assumption 4, we can consider the slightly different problem:

$$\mathbb{E}[\mathbf{x}_*|\mathbf{y}] \equiv \mathbb{E}[\mathbf{x}_*|\mathbf{Q}^{-1/2}\mathbf{y} = \mathbf{Q}^{1/2}\mathbf{x}_* + \mathbf{z} =: \bar{\mathbf{y}}]. \tag{C.1.24}$$

The estimation of the signal from observation $\bar{\mathbf{y}}$ is now in the form of a noisy observation under unitary measurement. Notice that Assumptions 1, 2 and 3 are all satisfied in our simple model (C.1.23), due to the i.i.d. design of the signal, the PSD requirement on $\mathbf{Q}$ and the design of the noise.

Proposition C.1.1
By estimating $\mathbb{E}[\mathbf{x}_*|\bar{\mathbf{y}}]$ using the OAMP iteration described in Eqs. (C.1.2), we achieve Bayes-optimal error. Namely, in the replica-symmetric regime (see Proposition 3.1.1 and Section 3.3), the induced error is determined by

$$\lim_{n\to\infty} \frac{1}{n}\|\mathbb{E}[\mathbf{x}_*|\mathbf{r}^k] - \mathbf{x}_k\|_2^2 = \mathrm{mmse}(\chi_k^{-2}), \tag{C.1.25}$$

and

$$\lim_{k\to\infty}\lim_{n\to\infty} \frac{1}{n}\|\mathbb{E}[\mathbf{x}_*|\mathbf{r}^k] - \mathbf{x}_*\|_2^2 = \mathrm{mmse}(\chi_*^{-2}), \tag{C.1.26}$$

where k is the number of iterations of the algorithm. Moreover, the inverse SNR χ_*^2 is given by the fixed-point equation (C.1.18), which is given in this case by

$$\begin{aligned}\chi_*^{-2} &= \frac{1}{\sigma^2}\mathcal{R}_{(\mathbf{Q}^{1/2})^T\mathbf{Q}^{1/2}}\left(-\sigma^2\mathrm{mmse}(\chi_*^{-2})\right)\\ &= \frac{1}{\sigma^2}\mathcal{R}_{\mathbf{Q}}\left(-\sigma^2\mathrm{mmse}(\chi_*^{-2})\right). \end{aligned}\tag{C.1.27}$$

Proof. The claim follows by construction of the sensing matrix $\mathbf{Q}$, which satisfies all assumptions, allowing us to appeal to Theorem C.1.1. The results follow, and the fixed-point equation is posited first in Ma and Ping (2016), and further explained in Liu, Cheng, Liang, Manton, and Ping (2023). □

We will from here on out work with SNR rather than inverse SNR. We define $\rho_k := \chi_k^{-2}$, giving us the fixed-point equation

$$\rho_* = \frac{1}{\sigma^2}\mathcal{R}_{\mathbf{Q}}(-\sigma^2\mathrm{mmse}(\rho_*)), \tag{C.1.28}$$

which allows for easier notation.

Proposition C.1.2
For $\mathbf{Q}$ *as above, and the estimation of* $\mathbb{E}[\mathbf{x}|\sqrt{\mathbf{Q}t}\mathbf{x}_* + \mathbf{z}]$ *using OAMP, we have that*

$$\lim_{n\to\infty} \frac{1}{n} \|\mathbb{E}[\mathbf{x}_*|\mathbf{r}^k] - \mathbf{x}_*\|_2^2 = \mathrm{mmse}(\rho_k), \tag{C.1.29}$$

and

$$\lim_{k\to\infty} \lim_{n\to\infty} \frac{1}{n} \|\mathbb{E}[\mathbf{x}_*|\mathbf{r}^k] - \mathbf{x}_*\|_2^2 = \mathrm{mmse}(\rho_*), \tag{C.1.30}$$

where ρ_* *satisfies the fixed-point equation*

$$\rho_* = \lambda t - t^2 \mathrm{mmse}(\rho_*). \tag{C.1.31}$$

Proof. See Appendix C.2. □

Mutual Information Analysis

We can decompose the mutual information using the chain rule:

$$I(\mathbf{x}; \mathbf{y}^{\mathrm{anis}}(t), \mathbf{A}(\beta)) := I(\mathbf{x}; \mathbf{A}(\beta)) + I(\mathbf{x}; \mathbf{y}^{\mathrm{anis}}(t)|\mathbf{A}(\beta)) \tag{C.1.32}$$

Notice that opposed to the isotropic case, our mutual information decomposes into *two kinds of problems*: a rank-one matrix estimation and a random linear estimation problem, with a *given* sensing matrix $\mathbf{A}(\beta)$. Using results from Barbier, Dia, Macris, Krzakala, Lesieur, and Zdeborova (2016) for the first term and Theorem 1.8 of Li, Fan, Sen, and Wu (2023) for the second, we get

$$\frac{1}{n} I(\mathbf{x}; \mathbf{y}(t), \mathbf{A}(\beta)) := \frac{\beta^2}{4} + I(\rho_*) + \frac{\mathrm{mmse}(\rho_*)^2 t^2}{4}. \tag{C.1.33}$$

Conclusion

The problem is now that we have the mutual information for a system where $\mathbf{A}$ in the anisotropic observation process $\mathbf{y}(t)$ *does not contain* $\mathbf{x}$ *anymore*, by the asymptotic vanishing of the spike discussed in Appendix C.2. This in turn forbids the use of the I-MMSE relations, which take the derivative of the mutual information, as the double dependency highlighted in Figure C.1 does not allow interchange of derivative and limit, unlike in Proposition 4.7 of Alaoui et al. (2024).

C.2 Proof of Proposition C.1.2

We will break down the proof into multiple steps. We begin by showing that $\mathbb{Q}$ satisfies Assumptions 2 and 4, in order for us to appeal to Proposition C.1.1. Then, we will show the form of the fixed-point equation.

Rotational Invariance in Distribution

Clearly, for a large enough choice of λ, $\mathbf{Q}$ is positive-definite, and hence invertible. It is however with a heavy heart that we notice that $\mathbf{Q}$ is *not* unitarily invariant, due to the presence of the rank-one spike $\frac{\beta}{n}\mathbf{x}\mathbf{x}^T$ favoring one direction. There is a silver lining however. If we consider a simultaneous drawing of $(\mathbf{x}, \mathbf{W}) \sim \bar{\nu} \otimes \mu_{\text{GOE}}$ (much akin to ensemble averaging), the generated matrix $\mathbf{A}$ *does indeed preserve rotational invariance in distribution*, by

$$\begin{aligned} \mathbf{A} &= \mathbf{W} + \frac{\beta}{n}\mathbf{x}\mathbf{x}^T \\ &= \mathbf{O}\boldsymbol{\Lambda}\mathbf{O}^T + \frac{\beta}{n}\mathbf{u}\mathbf{u}^T \end{aligned} \tag{C.2.1}$$

where $\mathbf{O} \in \mathcal{U}_n$ is a Haar matrix, $\boldsymbol{\Lambda}$ is the eigenvalue matrix of $\mathbf{W}$, and $\mathbf{u} := \mathbf{O}\mathbf{x}$. By the definitions of both $\mathbf{x}$ and $\mathbf{O}$, we have that $\mathbf{u}$ is uniformly distributed on the n-dimensional unit sphere, hence is rotationally invariant in distribution. The $\lambda\mathbf{I}_n$-term is clearly rotationally invariant in distribution, whence $\mathbf{Q} \overset{d}{=} \mathbf{V}\mathbf{Q}\mathbf{V}^T$ for any orthogonal matrix $\mathbf{V} \in \mathcal{U}_n$.

Analysis on Spectrally Equivalent Model

We show that in the case of $\beta < 1$, the spike has no effect on the limiting empirical eigenvalue distribution of $\mathbf{A}$, allowing us to perform analysis on a spectrally equivalent model $\tilde{\mathbf{A}} \sim \mu_{\text{GOE}}$ with no spike.

Proposition C.2.1
Let $\mathbf{W}$ be a GOE matrix, $\mathbf{u} \in \mathbb{R}^n$ be a unit vector, and $\beta > 0$. Then, for $\beta < 1$, the rank-one perturbed GOE matrix $\mathbf{A} := \beta\mathbf{u}\mathbf{u}^T + \mathbf{W}$, the empirical eigenvalue probability distribution of $\mathbf{A}$, denoted $\mu_{\mathbf{A}}$, converges weakly towards a semicircle distribution ρ^{SC} as $n \to \infty$.

Proof of Prop. C.2.1. The proof is due to results of Benaych-Georges and Nadakuditi (2011) for additively perturbed Gaussian Wigner matrices in the subcritical BBP phase Baik et al. (2005) (corresponding to $\beta < 1$). □

It follows that $\mathbf{Q}$ satisfies both Assumptions 2 and 4, allowing us to apply the result of Proposition C.1.1 with sensing matrix $\mathbf{Q}^{1/2}$. We thus get that the empirical distribution of the iterates in OAMP will asymptotically convergence to $\bar{X}_k = X_* + \chi_k Z$, where χ_k follows the state evolution equations (C.1.8). The fixed-point equation for χ_k^2 is thus

$$\frac{1}{\chi^2} = \frac{1}{\sigma^2} \mathcal{R}_{\mathbf{Q}} \left(-\sigma^2 \mathrm{mmse}(\chi^2) \right). \tag{C.2.2}$$

We will use the formulation of the $\mathcal{R}$-transform as the formal power series of the free cumulants of the eigenvalue distribution of the given random matrix (see Theorem 39, Mingo and Speicher (2017))

$$\mathcal{R}_{\mathbf{M}}(z) = \sum_{n=0}^{\infty} \kappa_{n+1}^{(\mathbf{M})} z^n, \tag{C.2.3}$$

where $\kappa_n^{(\mathbf{M})}$ denotes the n-th free cumulant of the spectral distribution of $\mathbf{M} \in \mathbb{R}^{n \times n}$.

Remark C.3
The free cumulants of the Wigner semicircle distribution are given by

$$\kappa_n^{SC} = \begin{cases} 1, & n = 2, \\ 0, & \text{otherwise.} \end{cases} \tag{C.2.4}$$

We now extend the above well-known fact to *shifted* GOE matrices, whose eigenvalue distribution is a shifted Wigner semicircle law.

Proposition C.2.2
The free cumulants $\{\kappa_n^{\rho_\lambda}\}_{n \geq 1}$ of a λ-shifted Wigner semicircle law $\rho_\lambda(s)$,

$$\rho_\lambda(s) := \frac{1}{2\pi} \sqrt{4 - (s - \lambda)^2}, \quad s \in [\lambda - 2, \lambda + 2], \tag{C.2.5}$$

are given by

$$\kappa_n^{\rho_\lambda} = \begin{cases} \lambda, & n = 1, \\ 1, & n = 2, \\ 0, & \text{otherwise.} \end{cases} \tag{C.2.6}$$

Proof of Prop. C.2.2. The proof is via direct computation. Let Λ denote the eigenvalue distribution of a GOE matrix (i.e. Λ is Wigner semi-circle distributed). Considering the independence of Λ and an arbitrary constant $\mu > 0$, we can use the $\mathcal{R}$-transform property of (freely) independent random variables (Theorem 18, Mingo and Speicher (2017)):

$$\mathcal{R}_{\Lambda+\mu}(z) = (R)_\Lambda(z) + \mathcal{R}_\mu(z) \tag{C.2.7}$$

and the definition of the $\mathcal{R}$-transform for a unit mass $\delta_\mu(x)$ at $x = \mu$:

$$\begin{aligned}\mathcal{R}_{\Lambda+\mu}(z) &= \sum_{n=0}^{\infty} \kappa_{n+1}^{(\Lambda)} z^n + \mu \\ &= z + \mu \\ &= \mu \cdot z^0 = 1 \cdot z^1 \end{aligned} \tag{C.2.8}$$

thus proving the claim. □

Using (C.2.3), we are now equipped with everything to compute the $\mathcal{R}$-transform of $\tilde{\mathbf{Q}}$.

Corollary C.2.1
For a λ-shifted GOE matrix $\mathbf{Q}$, its $\mathcal{R}$-transform is given by

$$\mathcal{R}_{\mathbf{Q}}(z) = \lambda + z. \tag{C.2.9}$$

Thus, we may explicitly compute the fixed-point equation (C.2.2). By Proposition C.2.1, we can compute simply the R-transform of a shifted GOE matrix.

$$\begin{aligned}\frac{1}{\chi^2} &= \frac{1}{\sigma^2}\mathcal{R}_{\mathbf{Q}}\left(-\sigma^2 \mathrm{mmse}(\chi^2)\right) \\ &= \frac{1}{\sigma^2}\left(\lambda - \sigma^2 \mathrm{mmse}(\chi^2)\right) \\ &= \frac{\lambda}{\sigma^2} - \mathrm{mmse}(\chi^2). \end{aligned} \tag{C.2.10}$$

Finally, we use Theorem 2.31 of Tulino and Verdú (2004) to account for the time dependency, giving us the fixed-point equation

$$\frac{1}{\chi^2} = t \cdot \left(\frac{\lambda}{\sigma^2} - t \cdot \mathrm{mmse}(\chi^{-2})\right), \tag{C.2.11}$$

proving the entire claim.

□

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. In consultation with the supervisor, one of the following three options must be selected:

○ I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

○ I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used and cited generative artificial intelligence technologies[2].

● I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used generative artificial intelligence technologies[3]. In consultation with the supervisor, I did not cite them.

Title of paper or thesis:

Sampling for Statistical Physics with Diffusion Processes

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Kawasaki-Borruat	Victor

With my signature I confirm the following:

- I have adhered to the rules set out in the Citation Guide.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zürich - March 2nd, 2025	

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] E.g. ChatGPT, DALL E 2, Google Bard
[2] E.g. ChatGPT, DALL E 2, Google Bard
[3] E.g. ChatGPT, DALL E 2, Google Bard