
Sampling from the Random Linear Model via Stochastic Localization Up to the AMP Threshold

Han Cui
UIUC

Zhiyuan Yu
UIUC

Jingbo Liu
UIUC

Abstract

Recently, Approximate Message Passing (AMP) has been integrated with stochastic localization (diffusion model) by providing a computationally efficient estimator of the posterior mean. Existing (rigorous) analysis typically proves the success of sampling for sufficiently small noise, but determining the exact threshold involves several challenges. In this paper, we focus on sampling from the posterior in the linear inverse problem, with an i.i.d. random design matrix, and show that the threshold for sampling coincides with that of posterior mean estimation. We give a proof for the convergence in smoothed KL divergence whenever the noise variance Δ is below Δ_{AMP} , which is the computation threshold for mean estimation introduced in (Barbier et al., 2020). We also show convergence in the Wasserstein distance under the same threshold assuming a dimension-free bound on the operator norm of the posterior covariance matrix, a condition strongly suggested by recent breakthroughs on operator norm bounds in similar replica symmetric systems. A key observation in our analysis is that phase transition does not occur along the sampling and interpolation paths assuming $\Delta < \Delta_{\text{AMP}}$.

1 INTRODUCTION

Linear inverse problem is fundamental to a wide range of fields in science, engineering, and technology, including medical imaging (Qu et al., 2013; Song et al., 2022; Vlaardingerbroek and Boer, 2010), compressed sens-

ing (Bayati and Montanari, 2011, 2012; Donoho et al., 2009) and image restoration (Gonzalez and Woods, 2006). In this paper, we consider the linear inverse problem of reconstructing a signal $\theta \in \mathbb{R}^N$ from noisy measurements $\mathbf{y} \in \mathbb{R}^M$ in Gaussian random linear model

$$\mathbf{y} = \phi\theta + \sqrt{\frac{\Delta}{\alpha}}\mathbf{w}, \quad (1)$$

where $\alpha = M/N$ is the measurement rate, $\phi \in \mathbb{R}^{M \times N}$ is a known Gaussian matrix with i.i.d. entries, $(\phi_{ij})_{i \leq M, j \leq N} \sim \mathcal{N}(0, 1/M)$, \mathbf{w} is a noise vector with i.i.d. components $w_i \sim \mathcal{N}(0, 1)$ for $i \in [M]$, and Δ is a known variance. We focus on the high-dimensional setting, where $M, N \rightarrow \infty$ with α fixed.

Suppose that θ has a prior distribution $\pi = P_0^{\otimes N}$ supported on $[-L_\theta, L_\theta]^N$. Our goal is to generate a sample from the posterior distribution

$$\begin{aligned} \mu_{\mathbf{y}}(\theta) &:= \mathbb{P}(\theta|\mathbf{y}) \\ &= \frac{1}{Z(\mathbf{y})} \exp\left(-\frac{\alpha}{2\Delta}\|\phi\theta - \mathbf{y}\|_2^2\right) P_0^{\otimes N}(\theta), \end{aligned} \quad (2)$$

$$\text{where } Z(\mathbf{y}) := \int \exp\left(-\frac{\alpha}{2\Delta}\|\phi\theta - \mathbf{y}\|_2^2\right) P_0^{\otimes N}(\theta) d\theta.$$

Sampling from posterior distribution and understanding its hardness regimes is a major problem in Bayesian learning. Traditional methods such as MCMC or Langevin dynamics may be slow or require log-convexity of the prior. Specifically, MCMC tends to require substantial computational resources when applied to high-dimensional problems and suffers from slow convergence due to multimodality of the posterior distribution (Woodard and Rosenthal, 2013).

Recently, Chung et al. (2023) proposed that diffusion model can also be utilized to sample the posterior distribution. Suppose the data noising process $\theta_t, t \in [0, T]$ has the form

$$d\theta = -\frac{\beta(t)}{2}\theta dt + \sqrt{\beta(t)}d\mathbf{W}, \quad (3)$$

where \mathbf{W} is the standard N -dimensional Wiener process and $\beta(t)$ is the noise schedule of the process, typically taken to be monotonically increasing linear function of t . The data distribution satisfies $\boldsymbol{\theta}_0 \sim \mathbb{P}(\boldsymbol{\theta}|\mathbf{y})$ and $\boldsymbol{\theta}_T \sim \mathcal{N}(0, \mathbf{I}_N)$. It is possible to use the reverse SDE of (3) to sample from the posterior:

$$d\boldsymbol{\theta} = \left[-\frac{\beta(t)}{2}\boldsymbol{\theta} - \beta(t)(\nabla_{\boldsymbol{\theta}_t} \log p_t(\boldsymbol{\theta}_t) + \nabla_{\boldsymbol{\theta}_t} \log p_t(\mathbf{y}|\boldsymbol{\theta}_t)) \right] dt + \sqrt{\beta(t)} d\bar{\mathbf{W}}, \quad (4)$$

where $\bar{\mathbf{W}}$ is a standard Wiener process when time flows backwards from T to 0. In the sampling algorithm, Chung et al. (2023) use the neural network $s(\boldsymbol{\theta}_t, t)$ trained with denoising score matching to replace the score function term $\nabla_{\boldsymbol{\theta}_t} p_t(\boldsymbol{\theta}_t)$ and $p(\mathbf{y}|\mathbb{E}_{\boldsymbol{\theta}_0 \sim p(\boldsymbol{\theta}_0|\boldsymbol{\theta}_t)}[\boldsymbol{\theta}_0])$ to replace the likelihood term $p_t(\mathbf{y}|\boldsymbol{\theta}_t)$ in (4). However, $p(\mathbf{y}|\mathbb{E}_{\boldsymbol{\theta}_0 \sim p(\boldsymbol{\theta}_0|\boldsymbol{\theta}_t)}[\boldsymbol{\theta}_0])$ is only an approximation of $p_t(\mathbf{y}|\boldsymbol{\theta}_t)$. To get a more accurate estimate, we consider using the Approximate Message Passing(AMP) algorithm (Donoho et al., 2009), which is known for its low complexity and rapid convergence, often achieving near-optimal solutions in a small number of iterations. However, AMP still has some limitations. For example, ϕ should be semi-random matrix ensemble (Dudeja et al., 2023) or ϕ should satisfy orthogonal rotational invariance in law (Fan, 2022).

Barbier et al. (2020) claims that under some constraints of the prior distribution, the mean squared error (MSE) of the sample generated by AMP algorithm achieves the minimum mean squared error (MMSE) asymptotically when Δ is within a certain range and gives a rigorous proof in Barbier et al. (2019). Since Montanari (2023) pointed out that sampling based on time-reversal of diffusion processes and on stochastic localization are in fact equivalent, we extend the analysis in Barbier et al. (2020) by giving a rigorous proof of the optimality of AMP in posterior sampling with an additional stochastic localization observation $(\mathbf{z}_t)_{t \geq 0}$:

$$\mathbf{z}_t = t\boldsymbol{\theta}^* + \mathbf{G}_t \quad (5)$$

where $\boldsymbol{\theta}^*$ is drawn from the posterior distribution $\mu_{\mathbf{y}}(\boldsymbol{\theta})$, and \mathbf{G}_t is a standard N -dimensional Brownian motion and is independent of \mathbf{y} , ϕ , $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$ and \mathbf{w} . Here, we cannot directly apply the results of Barbier et al. (2020) if we add the additional stochastic localization observation $(\mathbf{z}_t)_{t \geq 0}$. We use state evolution to prove that AMP is still almost surely optimal with additional observation $(\mathbf{z}_t)_{t \geq 0}$ which is detailed in Theorem 2 in Section 3.

Recently, Montanari and Wu (2023) applied stochastic localization and AMP to reconstruct the signal $\boldsymbol{\theta}$ from spiked model $\mathbf{X} = n^{-1}\beta\boldsymbol{\theta}\boldsymbol{\theta}^\top + \mathbf{W}$, proved convergence for sufficiently large β , and conjectured that the sampling hardness threshold coincides with the infimum

β^{-1} in the hard phase of mean estimation. Since both spiked model and linear inverse problem deal with reconstructing a high-dimensional signal from the noisy observations, we adopt the same framework on linear inverse problem. However, the computation hardness threshold of the linear inverse model (1) is not obvious if we follow the idea of Montanari and Wu (2023) directly. In this paper, we verify that the sampling algorithm performs effectively when $\Delta < \Delta_{\text{AMP}}$ under the smoothed KL error metric, where Δ_{AMP} is the infimum variance at which AMP does not achieve asymptotically Bayes optimal estimation, and has been conjectured to be the computation threshold for mean estimation (Barbier et al., 2020).

Let $\mu_{\mathbf{y},t}(\boldsymbol{\theta})$ be the posterior distribution of $\boldsymbol{\theta}$ given the measurements \mathbf{y} and the additional stochastic localization observations \mathbf{z}_t at time t :

$$\begin{aligned} \mu_{\mathbf{y},t}(\boldsymbol{\theta}) &:= \mathbb{P}(\boldsymbol{\theta} | \mathbf{y}, \mathbf{z}_t) \\ &= \exp \left(-\frac{\alpha}{2\Delta} \|\phi\boldsymbol{\theta} - \mathbf{y}\|_2^2 - \frac{1}{2t} \|\mathbf{z}_t - t\boldsymbol{\theta}\|_2^2 \right) \\ &\quad \times P_0^{\otimes N}(\boldsymbol{\theta}) / Z(\mathbf{y}, t), \end{aligned} \quad (6)$$

where

$$\begin{aligned} Z(\mathbf{y}, t) &= \int \exp \left(-\frac{\alpha}{2\Delta} \|\phi\boldsymbol{\theta} - \mathbf{y}\|_2^2 - \frac{1}{2t} \|\mathbf{z}_t - t\boldsymbol{\theta}\|_2^2 \right) \\ &\quad \times P_0^{\otimes N}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \quad (7)$$

The sequence of posteriors $\mu_{\mathbf{y},t}(\boldsymbol{\theta})$ satisfies the following localization process properties (Chen and Eldan, 2022):

- The initialization is the posterior distribution $\mu_{\mathbf{y},0}(\boldsymbol{\theta}) = \mu_{\mathbf{y}}(\boldsymbol{\theta})$.
- The final is a point mass at $\boldsymbol{\theta}^*$ sampled from $\mu_{\mathbf{y}}(\boldsymbol{\theta})$, i.e., $\mu_{\mathbf{y},\infty}(\boldsymbol{\theta}) = \delta_{\boldsymbol{\theta}^*}$.
- $\mu_{\mathbf{y},t}(\boldsymbol{\theta})$ is a martingale.

The process $(\mathbf{z}_t)_{t \geq 0}$ is also the unique solution of the following stochastic differential equation (section 7.4 from Liptser and Shiriaev (1977)):

$$d\mathbf{z}_t = m'(\mathbf{z}_t, t)dt + d\mathbf{B}_t, \mathbf{z}_0 = 0, \quad (8)$$

where \mathbf{B}_t is an N -dimensional standard Brownian motion and $m'(\mathbf{z}, t)$ is the posterior mean defined as:

$$m'(\mathbf{z}, t) = \mathbb{E}[\boldsymbol{\theta}^* | t\boldsymbol{\theta}^* + \mathbf{G}_t = \mathbf{z}]. \quad (9)$$

Let

$$m(\mathbf{z}, t) = \mathbb{E}[\boldsymbol{\theta} | \phi\boldsymbol{\theta} + \sqrt{\Delta/\alpha}\mathbf{w} = \mathbf{y}, t\boldsymbol{\theta} + \mathbf{G}_t = \mathbf{z}]. \quad (10)$$

We claim that the process $(\mathbf{z}_t)_{t \geq 0}$ is the unique solution of the following stochastic differential equation, which is rewritten from (8):

$$d\mathbf{z}_t = m(\mathbf{z}_t, t)dt + d\mathbf{B}_t, \mathbf{z}_0 = 0, \quad (11)$$

since $m'(\mathbf{z}, t) = m(\mathbf{z}, t)$ (note that (9) and (10) are equivalent since $\boldsymbol{\theta}^*$ follows from the posterior distribution whereas $\boldsymbol{\theta}$ follows from the prior distribution).

Based on these facts, we can sample from the posterior distribution of $\boldsymbol{\theta}$ by tracking the process \mathbf{z}_t .

- First, we discretize the SDE (11).
- Then, we apply the AMP algorithm to approximate $m(\mathbf{z}_t, t)$ in each step.
- Finally, we use the approximation $\hat{\mathbf{z}}_T/T$ as a sample from the posterior distribution.

We prove convergence of the distribution generated by this algorithm under the smoothed KL divergence, and under the Wasserstein distance assuming a dimension-free bound on the operator norm of the covariance of the posterior distribution.

1.1 Main contributions

We adopt the same framework as in Montanari and Wu (2023) and provide a sampling algorithm for (1) by approximating the posterior mean with the Bayes-AMP solution, and generating a diffusion process. We prove L_2 convergence results for the AMP solution to the posterior mean in (11), by leveraging recent extension of the state evolution analysis (Berthier et al., 2020) to the general denoiser case. As a consequence, we prove the convergence in smoothed KL divergence between the target distribution and the distribution generated by the algorithm. Compared to similar stochastic localization sampling convergence guarantees in recent works for other statistical models (Montanari and Wu, 2023), our analysis pushed the computation threshold for sampling to Δ_{AMP} defined in Barbier et al. (2020), which is believed to be the computation threshold for mean estimation (See Section 5) and has also been proved to be the computation threshold for a certain class of first order algorithms (Celentano et al., 2020). This characterization of the threshold implies that sampling succeeds for any fixed Δ , as long as α is sufficiently large (Corollary 1), which is not obvious from previous results (Montanari and Wu, 2023) only establishing the existence of a threshold. We discuss the connection between overlap concentration and Lipschitz constant of $m(\mathbf{z}_t, t)$ with respect to \mathbf{z}_t (See Appendix B) and show that the convergence under the Wasserstein distance can be established assuming that

$m_t(\mathbf{z}_t) := m(\mathbf{z}_t, t)$ has a dimension free Lipschitz constant. A key observation allowing us to determine the threshold is Proposition 1, showing that $\Delta < \Delta_{\text{AMP}}$ implies two favorable properties: 1) in the Gaussian interpolation argument for proving MSE optimality at fixed t , no phase transition occurs; 2) while changing t in the diffusion process, no phase transition occurs, so that the posterior covariance is expected to have a dimension-free operator norm bound. In contrast, prior works tackle similar issues by either choosing very small noise Δ (possibly smaller than Δ_{AMP}) (Montanari and Wu, 2023), or choosing N -dependent time-discretization, raising further issues that are difficult to justify (see Remark 5).

1.2 Related works

Posterior mean estimation in linear models.

Over the past decade, a substantial amount of work focused on developing the approximate message passing (AMP) algorithm to solve inverse problems (Rangan, 2011)(Bayati and Montanari, 2011), which holds the promise of making Bayesian inference for high-dimensional models computationally tractable, in contrast to traditional MCMC algorithms. The AMP algorithm is an iterative algorithm, which can be viewed as an (asymptotically equivalent) approximation of the message passing algorithm on a bipartite graph (Montanari et al., 2012). For any given number of iterations and as $M, N \rightarrow \infty$, the limiting distributions of the state vectors of AMP can be tracked by the state evolution (SE) (Bayati and Montanari, 2011). Thus by appropriately choosing the denoiser in AMP, the asymptotic error may satisfy the same fixed point equation as that of the minimum mean square error (mmse), which is achieved by the posterior mean (Bayes optimal) estimator. However, it is nontrivial to establish that the error achieved by AMP actually converges to mmse, as the fixed point may not be unique (Barbier et al., 2020). It has been conjectured that the algorithmic hard phase is characterized by the parameter regime where the AMP algorithm converges to the Bayes optimal solution (e.g. the discussions in Montanari and Wu (2023) and Barbier et al. (2020)).

Variational inference is another popular class of algorithms for Bayesian inference, which leverages efficient local (first-order) optimization techniques. In naive mean-field variational Bayes (NMF), a simple product structure is assumed for the posterior distribution, enabling fast computation. While the asymptotic correctness of NMF has been established under some assumptions such as $M > N$ and well-separatedness (Mukherjee and Sen, 2022), it is expected that a higher order correction based on Bethe’s approximation (also known as the TAP free energy) holds in more general

settings (Fan et al., 2021)Krzakala et al. (2014). Interestingly, the stationary points of the TAP free energy coincide with that of the AMP stationary points, and local stationary points that are close to global optimality ensures good posterior estimation (Fan et al., 2021); however, it is nontrivial to find computationally efficient algorithms for locating such good local optimizers.

Posterior sampling. In Bayesian inference and ensemble learning, it is often useful to generate samples from the posterior distribution. Sampling is often considered harder than mean estimation, since given a good sampler (e.g. with error measured in the Wasserstein distance) one can estimate the mean by computing the average of the samples (Montanari and Wu, 2023). Stochastic localization (El Alaoui et al., 2022) recently emerged as a powerful tool for sampling using stochastic gradient methods, where the gradient is a certain conditional mean. Therefore, existing algorithms for mean estimation, such as AMP, can be utilized, with proper modifications that incorporate an additional observation of the localization process. Closely related to our work are recent deployments of the stochastic localization method in the Sherrington-Kirkpatrick (SK) Model and the spiked PCA (Montanari and Wu, 2023). These works established the convergence under the Wasserstein distance, and typically require establishing a dimension-free Lipschitz constant for the AMP estimator. These conditions can be challenging to establish, and place restrictions on the range of noise variance that convergence of the algorithm can be established. We discuss more in Section 5.

2 METHOD

Consider the data distribution model of (1). Suppose θ is drawn from $\pi = P_0^{\otimes N}$ supported on $[-L_\theta, L_\theta]^N$ where P_0 is known and independent of N . To construct a diffusion process (11) for sampling, we need to estimate the posterior mean $m(\mathbf{z}_t, t)$. To do this, we apply the AMP algorithm in equation(202-203) of Berthier et al. (2020) as follows:

$$\mathbf{r}^{(k)} = \mathbf{y} - \phi \hat{\mathbf{m}}^{(k)} + \mathbf{r}^{(k-1)} b_t^{(k)}, \quad (12)$$

$$\hat{\mathbf{m}}^{(k+1)} = \eta(\phi^\top \mathbf{r}^{(k)} + \hat{\mathbf{m}}^{(k)}; \tau_{k,t}^2, \mathbf{z}_t, t), \quad (13)$$

where the initialization is given by $\hat{\mathbf{m}}^{(0)} = \mathbf{0}$ and

$$b_t^{(k)} = \frac{1}{N\alpha} \sum_{i=1}^N [\eta'(\phi^\top \mathbf{r}^{(k-1)} + \hat{\mathbf{m}}^{(k-1)}; \tau_{k-1,t}^2, \mathbf{z}_t, t)]_i. \quad (14)$$

The denoiser $\eta(\mathbf{u}; \Sigma^2, \mathbf{z}_t, t)$ is the MMSE estimator associated with the channels $\mathbf{u} = \mathbf{x} + \mathbf{v}\Sigma$ and $\mathbf{z}_t =$

Algorithm 1 Sampling Algorithm

Input: Data \mathbf{y} , Matrix $\phi \in \mathbb{R}^{M \times N}$, parameters (K, T, δ) ;

- 1: Set $\tilde{\mathbf{z}}_0 = \mathbf{0}_N$;
- 2: Compute v , the prior mean of $\|\theta\|^2$ divided by N ;
- 3: **for** $l = 0, 1, \dots, T/\delta - 1$ **do**
- 4: Draw $\mathbf{B}_{l+1} \sim N(0, \mathbf{I}_N)$ independent of everything so far;
- 5: Let $\hat{m}(\tilde{\mathbf{z}}_{t_l}, t_l)$ be the output of Bayes AMP algorithm (12), (13) with K iterations;
- 6: Update $\tilde{\mathbf{z}}_{t_{l+1}} = \tilde{\mathbf{z}}_{t_l} + \hat{m}(\tilde{\mathbf{z}}_{t_l}, t_l)\delta + \sqrt{\delta}\mathbf{B}_{l+1}$;
- 7: **end for**
- 8: **return** $\theta^{\text{alg}} = \tilde{\mathbf{z}}_T/T$

Output: θ^{alg} .

$t\mathbf{x} + \mathbf{g}_t$ where $\mathbf{u}, \mathbf{x}, \mathbf{v}, \mathbf{z}_t, \mathbf{g}_t \in \mathbb{R}^N$, $\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_N)$, and \mathbf{g}_t is a standard N -dimensional Brownian motion. The l -th component of the the N -dimensional vector $\eta(\mathbf{u}; \Sigma^2, \mathbf{z}_t, t)$ is given by:

$$\frac{\int \mathbf{x}_l P_0(\mathbf{x}_l) \exp[-\frac{(\mathbf{z}_{t,l} - t\mathbf{x}_l)^2}{2t} - \frac{(\mathbf{u}_l - \mathbf{x}_l)^2}{2\Sigma^2}] d\mathbf{x}_l}{\int P_0(\mathbf{x}_l) \exp[-\frac{(\mathbf{z}_{t,l} - t\mathbf{x}_l)^2}{2t} - \frac{(\mathbf{u}_l - \mathbf{x}_l)^2}{2\Sigma^2}] d\mathbf{x}_l}, \quad (15)$$

and $[\eta'(\mathbf{u}; \Sigma^2, \mathbf{z}_t, t)]_i$ denotes the the i -th scalar component of the gradient of η w.r.t. its first argument.

$\tau_{k,t}$ is a sequence of variances defined by the following recursion:

$$\tau_{0,t}^2 = \frac{\Delta + v}{\alpha}, \text{ where } v = \mathbb{E}_\pi(\|\theta\|_2^2/N), \quad (16)$$

$$\tau_{k+1,t}^2 = \frac{\Delta + \text{mmse}(\tau_{k,t}^{-2}, t)}{\alpha}, \quad (17)$$

where $\text{mmse}(\Sigma^{-2}, t) := \mathbb{E}_{\theta, \tilde{G}, \tilde{W}}[\|\theta - E[\theta|\theta + \Sigma\tilde{G}, t\theta + \sqrt{t}\tilde{W}]\|^2]$ in which $\tilde{G} \sim N(0, 1)$ and $\tilde{W} \sim N(0, 1)$ are independent. Moreover, we can write $\text{mmse}(\Sigma^{-2}, t)$ in another way as follows:

$$\text{mmse}^*(\Sigma^{-2}) = \mathbb{E}_{\theta, \tilde{G}}[\|\theta - E[\theta|\theta + \Sigma\tilde{G}]\|^2], \quad (18)$$

$$\text{mmse}(\Sigma^{-2}, t) = \text{mmse}^*(\Sigma^{-2} + t). \quad (19)$$

As mmse^* has an easier form, we will use it instead of mmse for calculation. Actually, $\tau_{k,t}$ satisfies the following state evolution recursion by Lemma 3:

$$\tau_{0,t}^2 = \frac{\Delta + v}{\alpha}, \text{ where } v = \mathbb{E}_\pi(\|\theta\|_2^2/N), \quad (20)$$

$$\tau_{k+1,t}^2 = \lim_{N \rightarrow \infty} \frac{1}{\alpha N} \mathbb{E}_{\theta, \tilde{\mathbf{Z}}}[\|\eta(\theta + \tau_{k,t}\tilde{\mathbf{Z}}; \tau_{k,t}^2, \mathbf{z}_t, t) - \theta\|_2^2] + \frac{\Delta}{\alpha}, \quad (21)$$

where $\tilde{\mathbf{Z}} \sim N(0, I_N)$.

Under suitable conditions for Δ , AMP algorithm outputs an asymptotically equivalent estimate of $m(\mathbf{z}_t, t)$ estimate as $N \rightarrow \infty$. Denote $\hat{m}(\mathbf{z}_t, t)$ as the approximation of $m(\mathbf{z}_t, t)$. Let $T = L\delta$ and $t_l = l\delta, l = 1, \dots, L$. Now, we construct the following three processes with the true posterior mean and its approximation given by AMP:

$$d\mathbf{z}_t = m(\mathbf{z}_t, t)dt + d\mathbf{W}_t, \quad (22)$$

$$d\hat{\mathbf{z}}_t = \hat{m}(\hat{\mathbf{z}}_t, t)dt + d\mathbf{W}_t, \quad (23)$$

$$\begin{aligned} \tilde{\mathbf{z}}_{t_l} - \tilde{\mathbf{z}}_{t_{l-1}} &= \hat{m}(\tilde{\mathbf{z}}_{t_{l-1}}, t_{l-1})\delta + \mathbf{W}_{t_l} - \mathbf{W}_{t_{l-1}} \\ &\quad \text{for } l = 1, 2, \dots, L, \end{aligned} \quad (24)$$

where \mathbf{W}_t is a standard brownian motion, \hat{m} is the output of AMP algorithm (12), (13) with K iterations.

The first process is the same process as that defined in (11). The second process is a continuous process with drift term as estimated posterior mean $\hat{m}(\hat{\mathbf{z}}_t, t)$ replacing the true posterior mean $m(\mathbf{z}_t, t)$ in the first process (22). And the last process is the discrete version of the second process (23), which is the update step in our algorithm.

3 CONVERGENCE GUARANTEES

Now we provide a bound of the KL divergence between the distribution generated by our algorithm and by the first process(22).

Theorem 1. Consider the linear observation model in (1), with fixed measurement rate α and noise level Δ , and suppose that $\boldsymbol{\theta}$ has a product prior compactly supported on $[-L_\theta, L_\theta]^N$. Suppose that $\Delta < \Delta_{\text{AMP}}$, where Δ_{AMP} is the threshold for mean estimation by the AMP algorithm explicitly defined after (28). Then for any $\epsilon > 0$ and $T > 0$, there exist $K \in \mathbb{N}$ and $\delta \in \mathbb{R}^+$ such that Algorithm 1 satisfies the following with probability $1 - o_N(1)$:

$$\frac{1}{N} D_{\text{KL}}(P_{\mathbf{z}_{T/T}} \| P_{\boldsymbol{\theta}^{\text{alg}}}) \leq \epsilon. \quad (25)$$

Remark 1. $P_{\mathbf{z}_{T/T}}$ is the distribution of $\boldsymbol{\theta}^* + T^{-1/2}\mathbf{G}$, $\mathbf{G} \sim \mathcal{N}(0, I_N)$ rather than the true posterior. When T is large enough, we can consider it as a smoothed version of the target distribution. In fact, if the true posterior density is smooth, it is possible to show that $P_{\mathbf{z}_{T/T}}$ is close to the true posterior using the Fourier analysis (see e.g. (Zhang et al., 2024, Theorem E.1)).

Remark 2. Here we used Theorems 3.1 and 3.2 of Barbier et al. (2020), which is stated for discrete priors. However, as pointed out in Barbier et al. (2020, Remark 3.6), the prior could be extended to more generalized distributions. The constraint of the prior is

relaxed in Barbier et al. (2019). Therefore, we can consider this more general setting, a prior with bounded support.

Remark 3. We also mention here that under a regularity assumption that the derivative of $\text{mmse}^*(\Sigma^{-2})$ is continuous at $\Sigma^2 = 0$, Theorem 1 remains valid for any fixed Δ , when α is sufficiently large. We state this as a corollary and will prove it in the appendix.

Corollary 1. If the derivative of $\text{mmse}^*(\Sigma^{-2})$ is continuous at $\Sigma^2 = 0$, the the sampling algorithm succeeds for any fixed Δ when α is sufficiently large.

Sketch of proof for Theorem 1

Step 1 For any Δ and t , we can show that AMP convergence implies that MSE equals the solution of a certain fixed point equation. Denote the asymptotic MSE obtained by AMP at time t and iteration k as

$$E_{t,\Delta}^{(k)} := \lim_{N \rightarrow \infty} \frac{1}{N} \|\boldsymbol{\theta} - \hat{m}^{(k)}(\mathbf{z}_t, t)\|_2^2 \quad (26)$$

where $\hat{m}^{(k)}(\mathbf{z}_t, t)$ is the output of AMP algorithm at time t and iteration k . This asymptotic MSE can actually be tracked by the state evolution recursion by Lemma 4 given by:

$$E_{t,\Delta}^{(k+1)} = \text{mmse}^* \left(\frac{\alpha}{\Delta + E_{t,\Delta}^{(k)}} + t \right). \quad (27)$$

The function mmse^* defined in (18) is a monotone function which indicates that $E_{t,\Delta}^{(k)}$ is a decreasing sequence w.r.t. k . Therefore, $\lim_{k \rightarrow \infty} E_{t,\Delta}^{(k)} = E_{t,\Delta}^{(\infty)}$ exists and is the fixed point of the equation given as follows:

$$E = \text{mmse}^* \left(\frac{\alpha}{\Delta + E} + t \right). \quad (28)$$

Moreover, following Barbier et al. (2020), we define Δ_{AMP} as the supremum of all Δ such that the state evolution fixed point equation (28) has a unique solution for all noise value in $[0, \Delta]$ when $t = 0$. With the definition of Δ_{AMP} , it can be shown that $E_{t,\Delta}^{(\infty)}$ is unique for all t if $\Delta \in [0, \Delta_{\text{AMP}}]$, which indicates that Δ_{AMP} is also the threshold for AMP applied in Algorithm 1:

Proposition 1. For given (Δ, t) , suppose that (Δ_0, E_0) satisfies

$$E_0 = \text{mmse}^* \left(\frac{\alpha}{\Delta + E_0} + t \right) = \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + E_0} \right). \quad (29)$$

Then $\Delta_0 < \Delta_{\text{AMP}}$ is a sufficient condition for $E = E_0$ to be the unique solution to

$$E = \text{mmse}^* \left(\frac{\alpha}{\Delta + E} + t \right). \quad (30)$$

In particular, $\Delta < \Delta_{\text{AMP}}$ is a sufficient condition for $E = E_0$ to be the unique solution.

Proof. Since $\Delta_0 < \Delta < \Delta_{\text{AMP}}$, E_0 is the unique solution to

$$E = \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + E} \right). \quad (31)$$

Since $0 < \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + 0} \right)$ and $\infty > \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + \infty} \right)$, using the fact that $E \mapsto \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + E} \right)$ is increasing, we see that

$$E < \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + E} \right), \quad \forall E < E_0 \quad (32)$$

and

$$E > \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + E} \right), \quad \forall E > E_0. \quad (33)$$

Now if $E_1 > E_0$ is another fixed point for (30), we have $\frac{\alpha}{\Delta + E_1} + t > \frac{\alpha}{\Delta_0 + E_1}$ because $\frac{\alpha}{\Delta + E_0} + t = \frac{\alpha}{\Delta_0 + E_0}$ and $\Delta > \Delta_0$. Therefore

$$\text{mmse}^* \left(\frac{\alpha}{\Delta + E_1} + t \right) \leq \text{mmse}^* \left(\frac{\alpha}{\Delta_0 + E_1} \right), \quad (34)$$

which, combined with (33), implies that $E_1 > \text{mmse}^* \left(\frac{\alpha}{\Delta + E_1} + t \right)$, a contradiction. Similarly, we can also show that it is impossible to have another fixed point $E_1 < E_0$. \square

Step 2 Next, we construct the family of observation models in Barbier et al. (2020, (37)), which is just our model if we let $\tilde{\mathbf{y}} = \mathbf{z}_t/t$:

$$\mathbf{y} = \phi \boldsymbol{\theta} + \sqrt{\frac{\Delta}{\alpha}} \mathbf{w}; \quad (35)$$

$$\tilde{\mathbf{y}} = \boldsymbol{\theta} + \frac{1}{\sqrt{t}} \mathbf{G}_t. \quad (36)$$

where $G_t \sim \mathcal{N}(0, I)$ is independent of $(\mathbf{w}, \phi, \boldsymbol{\theta})$. We can then derive the mutual information of the above model, denoted as $i_{t,\Delta}$ as:

$$i_{t,\Delta} = -\left(\frac{\alpha}{2} + 1\right) - \frac{1}{N} \mathbb{E}[\ln Z_{t,\Delta}(\mathbf{y}, \tilde{\mathbf{y}})]. \quad (37)$$

where $Z_{t,\Delta}$ is the partition function defined as follows:

$$Z_{t,\Delta}(\mathbf{y}, \tilde{\mathbf{y}}) = \int d\mathbf{x} \exp(-H_{t,\Delta}(\mathbf{x} | \mathbf{y}, \tilde{\mathbf{y}})) P_0^{\otimes N}(\mathbf{x}), \quad (38)$$

$$H_{t,\Delta}(\mathbf{x} | \mathbf{y}, \tilde{\mathbf{y}}) = \frac{\alpha}{2\Delta} \|\phi \bar{\mathbf{x}} - \sqrt{\frac{\Delta}{\alpha}} \mathbf{w}\|_2^2 + \frac{t}{2} \|\bar{\mathbf{x}} - \frac{1}{\sqrt{t}} \mathbf{G}_t\|_2^2. \quad (39)$$

Here we use the notation $\bar{\mathbf{x}} = \mathbf{x} - \boldsymbol{\theta}$. We can then calculate the derivative of $i_{t,\Delta}$ w.r.t. t by the quantities above. Also, by Lemma 4.5 and Lemma 4.6 of Barbier et al. (2020), the relationship between $\text{mmse}_{t,\Delta}$ and MI is given, where $\text{mmse}_{t,\Delta}$ and $\text{ymmse}_{t,\Delta}$ are given as follows:

$$E_{t,\Delta} := \text{mmse}_{t,\Delta} = \frac{1}{N} \mathbb{E}[\|\boldsymbol{\theta} - \langle \mathbf{X} \rangle_{t,\Delta}\|_2^2] \quad (40)$$

$$\text{ymmse}_{t,\Delta} = \frac{1}{M} \mathbb{E}[\|\phi(\boldsymbol{\theta} - \langle \mathbf{X} \rangle_{t,\Delta})\|_2^2]. \quad (41)$$

Lemma 1. Model (35) verifies the following I-MMSE relation

$$\frac{\partial i_{t,\Delta}}{\partial \Delta^{-1}} = \frac{\alpha}{2} \text{ymmse}_{t,\Delta}. \quad (42)$$

Lemma 2. For a.e. t ,

$$\text{ymmse}_{t,\Delta} = \frac{E_{t,\Delta}}{1 + E_{t,\Delta}/\Delta} + o_N(1). \quad (43)$$

The construction of $i_{t,\Delta}$ as a function of $E_{t,\Delta}$ suggests that $E_{t,\Delta}$ satisfies the following PDE:

$$\alpha \frac{\partial}{\partial t} \left(\frac{E}{1 + E/\Delta} \right) = \frac{\partial E}{\partial \Delta^{-1}}, \quad (44)$$

which is the property shared by the fixed point of (28). Given $E_{0,\Delta} = E_{0,\Delta}^{(\infty)}$, which is shown in Barbier et al. (2020), as the initialization condition for the PDE, we can prove $E_{t,\Delta} = E_{t,\Delta}^{(\infty)}$ combining the above equation.

Step 3 Note that we have

$$\begin{aligned} & \frac{1}{N} \mathbb{E}[\|\hat{m}^{(k)}(\mathbf{z}_t, t) - m(\mathbf{z}_t, t)\|_2^2] \\ &= \frac{1}{N} \mathbb{E}[\|\hat{m}^{(k)}(\mathbf{z}_t, t) - \boldsymbol{\theta}\|_2^2] - \frac{1}{N} \mathbb{E}[\|\boldsymbol{\theta} - m(\mathbf{z}_t, t)\|_2^2]. \end{aligned} \quad (45)$$

Note that $\frac{1}{N} \mathbb{E}[\|\boldsymbol{\theta} - m(\mathbf{z}_t, t)\|_2^2]$ converges to $E_{t,\Delta}$ defined in Step 2. And for most \mathbf{z} , $\frac{1}{N} \mathbb{E}[\|\hat{m}^{(k)}(\mathbf{z}_t, t) - \boldsymbol{\theta}\|_2^2]$ converges to the fixed point equation solution, $E_{t,\Delta}^{(\infty)}$, shown in Step 1, so it converges to the same quantity. Thus (45) converges to 0 as $k \rightarrow \infty$. We summarize this result in the following theorem:

Theorem 2. For $\forall t$, if $\Delta < \Delta_{\text{AMP}}$, then AMP is almost surely optimal, namely,

$$\lim_{k \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{N} \|\hat{m}^{(k)}(\mathbf{z}, t) - m(\mathbf{z}, t)\|_2^2 = 0 \quad (46)$$

where $\hat{m}^{(k)}$ is the output of AMP algorithm at time t after k iterations.

Step 4 Now we show how to bound the KL divergence between the distribution generated by our algorithm and by the first process(22). First, we bound it by using the Girsanov's theorem (Oko et al., 2023) as follows:

$$\begin{aligned} & \frac{1}{N} D_{\text{KL}}(P_{\mathbf{z}_T/T} \| P_{\tilde{\mathbf{z}}_T/T}) \\ & \lesssim \sum_{l=1}^L \mathbb{E} \int_{t_{l-1}}^{t_l} \frac{1}{N} \|m(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_{t_l}, t_l)\|_2^2 dt. \end{aligned} \quad (47)$$

It is easy to note that the right hand side of (47) can be bounded by the summation of three terms:

$$\begin{aligned} & \sum_{l=1}^L \mathbb{E} \int_{t_{l-1}}^{t_l} \frac{1}{N} (\|m(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_t, t)\|_2^2 \\ & \quad + \|\hat{m}(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_{t_l}, t)\|_2^2 \\ & \quad + \|\hat{m}(\mathbf{z}_{t_l}, t) - \hat{m}(\mathbf{z}_{t_l}, t_l)\|_2^2) dt. \end{aligned} \quad (48)$$

For the first term, it is just the result of Theorem 2 given above. Then, we prove that the maps $\mathbf{z}_t \mapsto \hat{m}(\mathbf{z}_t, t)$ and $t \mapsto \hat{m}(\mathbf{z}_t, t)$ are Lipschitz continuous with high probability shown in Lemma 8 and Lemma 10. The second and third term can then be bounded by using the definition of diffusion process. We have:

$$\frac{1}{N} \mathbb{E} \|\hat{m}(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_{t_l}, t)\|_2^2 \quad (49)$$

$$\lesssim \frac{1}{N} \mathbb{E} \|\mathbf{z}_t - \mathbf{z}_{t_l}\|_2^2 \quad (50)$$

$$= \frac{1}{N} \mathbb{E} \left\| \int_{t_{l-1}}^t m(\mathbf{z}_s, s) ds + W_t - W_{t_l} \right\|_2^2. \quad (51)$$

Therefore, we can use the bound of support L_θ and the property of Brownian Motion to bound the second term. For the last term, using the Lipschitz property of \hat{m} with respect to t , we can bound it in the same way as follows:

$$\sum_{l=1}^L \int_{t_{l-1}}^{t_l} \frac{1}{N} \|\hat{m}(\mathbf{z}_{t_l}, t) - \hat{m}(\mathbf{z}_{t_l}, t_l)\|_2^2 dt \quad (52)$$

$$\lesssim \sum_{l=1}^L \int_{t_{l-1}}^{t_l} (t - t_{l-1})^2 dt \leq \frac{T\delta^2}{3}. \quad (53)$$

Therefore, for fixed T , we can choose δ small enough such that the summation of these two terms is less than $\epsilon/2$.

However, if we consider other metrics such as the Wasserstein distance, the results may not hold under the same assumptions as Theorem 1. Under a stronger assumption of the Lipschitz continuity of $m(\cdot, t)$, we can prove that the convergence of W_2 distance still holds.

Theorem 3. Assume that θ is compactly supported on $[-L_\theta, L_\theta]^N$, $\Delta < \Delta_{\text{AMP}}$ and $m(z, t)$ is L_m -Lipschitz w.r.t. z where L_m is bounded uniformly in N . Then for any fixed T , there exist $K \in \mathbb{N}$ and $\delta \in \mathbb{R}^+$ such that the following holds:

$$\lim_{\delta \rightarrow 0} \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}} W_2(P_{\mathbf{z}_T/T}, P_{\tilde{\mathbf{z}}_T/T}) = 0. \quad (54)$$

Remark 4. Here, we established the proof assuming that $m(z, t)$ has a dimension-free Lipschitz constant. The validity of this assumption is supported by recent breakthrough results of Brennecke et al. (2023) and El Alaoui and Gaitonde (2024) about the operator norm of covariance matrices for replica symmetric systems; see Appendix B.

Remark 5. In Theorem 3, we choose the time precision δ to be a constant independent of N . In contrast, Ghio et al. (2024) mentioned that the Lipschitz constant L_m could be $\Theta(N)$ when phase transition occurs, and suggested choosing $\delta = O(1/N)$. However, a closer inspection reveals that the accumulated error scales as $L_m \delta k$ where k is the number of steps (see for example Ghio et al. (2024, Lemma 4)), so choosing small δ won't compensate for diverging L_m , as k also increases. After consulting the authors of Ghio et al. (2024) about this issue, they suggested that small δ may only be needed during a phase transition, but since a proof that the phase transition only occurs in an $O(1/N)$ time interval is lacking, it remains an interesting challenge to provide a rigorous proof. By contrast, our Proposition 1 shows that phase transition does not occur while changing t as along as $\Delta < \Delta_{\text{AMP}}$, thus avoiding choosing an N -dependent δ .

4 NUMERICAL EXPERIMENTS

In our experiments, we first consider the following discrete prior distribution:

$$P_0(\theta) = \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_1. \quad (55)$$

For the first experiment, we set $M = 1000, N = 1250, K = 20, L = 2000, \delta = 0.1, v = 1$. In Figure 2 in Appendix C, we plot the trajectories of the first and the second coordinates of the samples generated by Algorithm 1: $\tilde{\mathbf{z}}_{t_l}/t_l, l \in \{1, \dots, L\}$ with different Δ . We choose $\Delta = 0.01, 1, 10$. For each Δ , we run four independent simulations. In each graph, the circle marker represents the initial value of the sample, and the star marker represents the final value of the sample. Notice that $(\tilde{\mathbf{z}}_{t_l}/t_l)_1$ and $(\tilde{\mathbf{z}}_{t_l}/t_l)_2$ converge to $+1$ or -1 as $l \rightarrow \infty$.

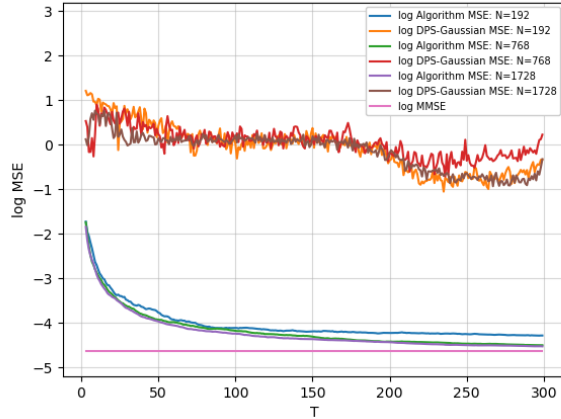


Figure 1: log Algorithm MSE and log DPS-Gaussian MSE for different values of N . For this experiment, we let the prior be standard normal distribution and set $\Delta = 0.01, \alpha = 2, K = 50, T = 300, \delta = 0.1$.

Then we consider the following continuous prior distribution which is a standard normal distribution:

$$P_0(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right). \quad (56)$$

For the second experiment, we fix $\alpha = 2$ and set $K = 50, T = 300, \delta = 0.1, v = 1$. For $\Delta = 0.01$, we plot the log MSE of the samples generated by our algorithm with different values of N and M . We choose $N = 192, 768, 1728$, which implies that $M = 384, 1536, 3456$ since $\alpha = 2$. We perform comparison with DPS-Gaussian method (Chung et al., 2023).

In Figure 1, we observe a gap between the MSE of the samples generated by DPS-Gaussian and MMSE since Chung et al. (2023) used $p(\mathbf{y}|\mathbb{E}_{\theta_0 \sim p(\theta_0|\theta_t)}[\theta_0])$ to approximate the likelihood $p_t(\mathbf{y}|\theta_t)$. In contrast, for our Algorithm, we find that within each fixed N , the algorithm MSE will gradually approach the MMSE as T increases.

5 DISCUSSION

Closely related to our work are recent deployments of the stochastic localization method in the Sherrington-Kirkpatrick (SK) Model and the spiked PCA (Montanari and Wu, 2023). These works established the convergence under the Wasserstein distance, and typically require establishing a dimension-free Lipschitz constant for the AMP estimator. These conditions can be challenging to establish, and place restrictions on the range of noise variance that convergence of the algorithm can be established. The challenges include:

- 1) Ensure that $m(\cdot, t)$ is Lipschitz continuous; see the choice of β_0 in Section G of Montanari and Wu (2023).
- 2) The noise variance also needs to be small enough such that a fixed point equation involving the localization time t has a unique solution for all $t > 0$ (Montanari and Wu, 2023, Proposition D.1). This condition may appear stronger than assuming the uniqueness of the fixed point equation in the mean estimation problem, which is just the special case of $t = 0$. It is conjectured that the threshold for efficient sampling is the infimum variance at which AMP does not achieve asymptotically Bayes optimal estimation (Montanari and Wu, 2023, Remark 3.2). This conjecture is natural since the additional stochastic localization process effectively reduces the noise variance. In this paper we partially resolve this conjecture (for the setting of linear models), by adopting slightly different proof strategies and error metrics that bypass or resolve the above mentioned challenges in establishing Lipschitz continuity and uniqueness of the fixed point. For challenge 1), we leverage a monotone convergence property of the mmse in state evolution, which bypasses the more sophisticated AMP contraction property, and we use the Girsanov theorem to prove convergence in the smoothed KL divergence, as opposed to the Wasserstein distance. For 2), we make an observation in Proposition 1 that the uniqueness of the fixed point equation involving the stochastic localization process can be guaranteed under the uniqueness of the fixed point equation without the stochastic localization process.

For comparison, in Appendix B we provide a proof of convergence under the Wasserstein distance, following a similar style of arguments in Montanari and Wu (2023), under an assumption of the Lipschitz continuity of $m(\cdot, t)$. In Montanari and Wu (2023), a similar condition for the Lipschitz continuity of $\hat{m}^{(k)}(\cdot, t)$ for large k was adopted, which relies on sophisticated results on contractive properties of the AMP iteration from prior works, which is one of the reasons why Montanari and Wu (2023) requires sufficiently small noise variance. In Appendix B.1 we note a connection between Lipschitz continuity of $m(\cdot, t)$ and the well-studied problem of concentration of the *overlap*, which suggests that such Lipschitz conditions are likely to be true in the “Bayes optimal” setting, but may not hold in more general settings. We remark that the Overlap Gap Property (Gamarnik, 2021) has recently been proposed as a theory for computation hardness.

References

- Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear

- models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.
- Barbier, J., Macris, N., Dia, M., and Krzakala, F. (2020). Mutual information and optimality of approximate message-passing in random linear estimation. *IEEE Transactions on Information Theory*, 66(7):4270–4303.
- Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.
- Bayati, M. and Montanari, A. (2012). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.
- Berthier, R., Montanari, A., and Nguyen, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79.
- Brennecke, C., Xu, C., and Yau, H.-T. (2023). Operator norm bounds on the correlation matrix of the sk model at high temperature. *arXiv preprint arXiv:2307.12535*.
- Celentano, M., Montanari, A., and Wu, Y. (2020). The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR.
- Chen, Y. and Eldan, R. (2022). Localization schemes: A framework for proving mixing bounds for markov chains (extended abstract). In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122. IEEE.
- Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. (2023). Diffusion posterior sampling for general noisy inverse problems. *International Conference on Learning Representations*.
- Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.
- Dudeja, R., M. Lu, Y., and Sen, S. (2023). Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683.
- El Alaoui, A. and Gaitonde, J. (2024). Bounds on the covariance matrix of the sherrington–kirkpatrick model. *Electronic Communications in Probability*, 29:1–13.
- El Alaoui, A., Montanari, A., and Sellke, M. (2022). Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE.
- Fan, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224.
- Fan, Z., Mei, S., and Montanari, A. (2021). Tap free energy, spin glasses and variational inference.
- Gamarnik, D. (2021). The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41):e2108492118.
- Ghio, D., Dandi, Y., Krzakala, F., and Zdeborová, L. (2024). Sampling with flows, diffusion, and autoregressive neural networks from a spin-glass perspective. *Proceedings of the National Academy of Sciences*, 121(27):e2311810121.
- Gonzalez, R. C. and Woods, R. E. (2006). *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc.
- Krzakala, F., Manoel, A., Tramel, E. W., and Zdeborová, L. (2014). Variational free energies for compressed sensing. In *2014 IEEE International Symposium on Information Theory*, pages 1499–1503. IEEE.
- Liptser, R. S. and Shiriaev, A. N. (1977). *Statistics of random processes: General theory*, volume 394. Springer.
- Montanari, A. (2023). Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*.
- Montanari, A., Eldar, Y., and Kutyniok, G. (2012). Graphical models concepts in compressed sensing. *Compressed Sensing*, pages 394–438.
- Montanari, A. and Wu, Y. (2023). Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*.
- Mukherjee, S. and Sen, S. (2022). Variational inference in high-dimensional linear regression. *The Journal of Machine Learning Research*, 23(1):13703–13758.
- Oko, K., Akiyama, S., and Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582.
- Qiu, J. and Sen, S. (2023). The tap free energy for high-dimensional linear regression. *The Annals of Applied Probability*, 33(4):2643–2680.
- Qu, X., Hou, Y., Lam, F., Zhong, J., and Chen, Z. (2013). Magnetic resonance image reconstruction from undersampled measurements using a patch-based nonlocal operator. *Medical image analysis*, 18.
- Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE.

- Song, Y., Shen, L., Xing, L., and Ermon, S. (2022). Solving inverse problems in medical imaging with score-based generative models. *International Conference on Learning Representations*.
- Talagrand, M. (2010). *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media.
- Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Vlaardingerbroek, M. T. and Boer, J. A. (2010). Magnetic resonance imaging: theory and practice. *Springer Berlin, Heidelberg*.
- Woodard, D. B. and Rosenthal, J. S. (2013). Convergence rate of Markov chain methods for genomic motif discovery. *The Annals of Statistics*, 41(1):91 – 124.
- Zhang, K., Yin, H., Liang, F., and Liu, J. (2024). Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *International Conference on Machine Learning*.

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
 - (b) Complete proofs of all theoretical results. [Yes]
 - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Not Applicable]
 - (b) The license information of the assets, if applicable. [Not Applicable]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOFS

A.1 Proof of Theorem 2

To prove this theorem, we need to calculate the state evolution recursion to show how AMP converges. The state evolution here has a simple intuitive explanation. Suppose we replace ϕ by i.i.d. sampled random matrix ϕ^k for each iteration. \mathbf{y}^k is replaced by $\phi^k \mathbf{m}_0 + \sqrt{\Delta/\alpha} \mathbf{w}$. Then, the iterations become:

$$\mathbf{r}^{(k)} = \mathbf{y}^k - \phi^k \hat{\mathbf{m}}^{(k)} \quad (57)$$

$$\hat{\mathbf{m}}^{(k+1)} = \eta_k(\phi^{k\top} \mathbf{r}^{(k)} + \hat{\mathbf{m}}^{(k)}). \quad (58)$$

By eliminating $\mathbf{r}^{(k)}$, the recursion can be given by:

$$\hat{\mathbf{m}}^{(k+1)} = \eta_k(\phi^{k\top} \mathbf{y}^k + (I - \phi^{k\top} \phi^k) \hat{\mathbf{m}}^{(k)}) \quad (59)$$

$$= \eta_k(m_0 + \sqrt{\Delta/\alpha} \phi^{k\top} \mathbf{w} + (I - \phi^{k\top} \phi^k)(\hat{\mathbf{m}}^{(k)} - m_0)). \quad (60)$$

Therefore, if we denote $\hat{\tau}_k = \lim_{N \rightarrow \infty} \frac{1}{N} \|\hat{\mathbf{m}}^{(k)} - m_0\|_2^2$, we can use $\hat{\tau}_k$ to describe the convergence of each term in (60). We have $(I - \phi^{k\top} \phi^k)(\hat{\mathbf{m}}^{(k)} - m_0)$ converges to a vector with i.i.d. entry with mean 0 and variance $\hat{\tau}_k/\alpha$. Besides, by law of large numbers $\phi^{k\top} \mathbf{w}$ converges to $1/\alpha$. Thus, we have

$$\tau_k = \frac{\hat{\tau}_k}{\alpha} + \frac{\Delta}{\alpha} \quad (61)$$

$$\lim_{N \rightarrow \infty} \hat{\mathbf{m}}^{(k+1)} = \eta_k(m_0 + \tau_k \mathbf{Z}) \quad (62)$$

$$\hat{\tau}_{k+1} = \mathbb{E}(\eta_k(m_0 + \tau_k \mathbf{Z}) - m_0)^2 \quad (63)$$

which is just the state evolution given in (20) and (21). However, the correlations between ϕ and $\hat{\mathbf{m}}^{(k)}$ can not be neglected. The Onsager terms are then used to asymptotically eliminate these correlations shown as $b_t^{(k)}$. Remark 7.2 in Berthier et al. (2020) shows that it can be given by:

$$b_t^{(k)} = \frac{1}{N\alpha} \sum_{i=1}^N [\eta'(\phi^T \mathbf{r}^{(k-1)} + \hat{\mathbf{m}}^{(k-1)}; \tau_{k-1,t}^2, \mathbf{z}_t, t)]_i. \quad (64)$$

Now, we will show that the state evolution here coincides with $\tau_{k,t}$ defined in (16).

Lemma 3. $\tau_{k,t}$ satisfies the state evolution:

$$\tau_{0,t}^2 = \frac{\Delta}{\alpha} + \frac{v}{\alpha}, \text{ where } v = \mathbb{E}_\pi(\|\boldsymbol{\theta}\|_2^2/N), \quad (65)$$

$$\tau_{k+1,t}^2 = \frac{\Delta}{\alpha} + \lim_{N \rightarrow \infty} \frac{1}{\alpha N} \mathbb{E}_{\boldsymbol{\theta}, \tilde{\mathbf{Z}}} [\|\eta(\boldsymbol{\theta} + \tau_{k,t} \tilde{\mathbf{Z}}; \tau_{k,t}^2, \mathbf{z}_t, t) - \boldsymbol{\theta}\|_2^2]. \quad (66)$$

Proof. We apply induction method to prove the result. For $k = 0$, we can find that (65) is just the definition of $\tau_{k,0}$ in (16).

Supposing (66) holds for k , let's consider the case for $k + 1$:

$$\frac{\Delta}{\alpha} + \lim_{N \rightarrow \infty} \frac{1}{\alpha N} \sum_{i=1}^N \mathbb{E}_{\theta_i, \tilde{Z}_i} [\|\eta(\theta_i + \tau_{k,t} \tilde{Z}_i; \tau_{k,t}^2, z_{t,i}, t) - \theta_i\|^2] \quad (67)$$

$$\stackrel{a.s.}{=} \frac{\Delta}{\alpha} + \frac{1}{\alpha} \mathbb{E}_{z_{t,i}} \mathbb{E}_{\theta_i, \tilde{Z}_i} [\|\eta(\theta_i + \tau_{k,t} \tilde{Z}_i; \tau_{k,t}^2, z_{t,i}, t) - \theta_i\|^2] \quad (68)$$

$$= \frac{\Delta}{\alpha} + \frac{1}{\alpha} \mathbb{E}_{\theta_i, \tilde{Z}_i, \tilde{W}} [\|\theta_i - E[\theta_i | \theta_i + \tau_{k,t} \tilde{Z}_i, t \theta_i + \sqrt{t} \tilde{W}]\|^2] \quad (69)$$

$$= \frac{\Delta}{\alpha} + \frac{\text{mmse}(\tau_{k,t}^{-2}, t)}{\alpha} \quad (70)$$

$$= \tau_{k+1,t}^2, \quad (71)$$

where the first equality holds by the Strong Law of Large Numbers. Thus, the state evolution holds for all k . \square

With the state evolution, we can now describe the asymptotic MSE of AMP by $\tau_{k,t}$.

Lemma 4. Recall $E_{t,\Delta}^{(k)}$ in (26), we have:

$$E_{t,\Delta}^{(k+1)} = \text{mmse}^* \left(\frac{\alpha}{\Delta + E_{t,\Delta}^{(k)}} + t \right). \quad (72)$$

Proof. By Lemma 3 and remark 7.1 in Berthier et al. (2020), we have

$$E_{t,\Delta}^{(k+1)} = \lim_{N \rightarrow \infty} \frac{1}{N} \|\hat{m}^{(k+1)} - \boldsymbol{\theta}\|_2^2 \quad (73)$$

$$\xrightarrow{p} \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\boldsymbol{\theta}, \tilde{\mathbf{Z}}} [\|\eta(\boldsymbol{\theta} + \tau_{k,t} \tilde{\mathbf{Z}}; \tau_{k,t}^2, \mathbf{z}_t, t) - \boldsymbol{\theta}\|_2^2] \quad (74)$$

$$= \alpha \left(\tau_{k+1,t}^2 - \frac{\Delta}{\alpha} \right) \quad (75)$$

$$= \text{mmse}^*(\tau_{k,t}^{-2} + t) \quad (76)$$

$$= \text{mmse}^* \left(\frac{\alpha}{\Delta + E_{t,\Delta}^{(k)}} + t \right). \quad (77)$$

\square

With Lemma 4 and the fact that mmse^* is a monotonically decreasing function, we have $\lim_{k \rightarrow \infty} E_{t,\Delta}^{(k)} = E_{t,\Delta}^{(\infty)}$ exists and is the fixed point of the equation given as follows:

$$E = \text{mmse}^*(\alpha/(\Delta + E) + t). \quad (78)$$

This $E_{t,\Delta}^{(\infty)}$ is actually the true MMSE under some assumptions. To show this, we would apply Lemma 1 and Lemma 2. With the property of relationship between MI and MMSE, we can derive the following theorem with Proposition 1, which indicates the convergence of AMP algorithm.

Theorem 4. For $\Delta < \Delta_{AMP}$ and a.e. t , we have

$$E_{t,\Delta} = E_{t,\Delta}^{(\infty)}, \quad (79)$$

where $E_{t,\Delta}$ denotes the limiting normalized mmse for the observation model in (35)-(36), and $E_{t,\Delta}^{(\infty)}$ denotes the limiting normalized mse of the AMP algorithm. Therefore, the \hat{m} generated by the AMP algorithm will converge to m when the number of iterations grow to infinity, i.e.:

$$\lim_{k \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{N} \|\hat{m}^{(k)}(z, t) - m(z, t)\|_2^2 = 0. \quad (80)$$

Proof. First, let's consider a special case when $t = 0$. That is we only have the first observation (35). This is the case proved in Barbier et al. (2020), and we have

$$\frac{d}{d\Delta^{-1}} \lim_{N \rightarrow \infty} i_{0,\Delta} = \frac{\alpha}{2} \frac{E_{0,\Delta}}{1 + E_{0,\Delta}/\Delta} = \frac{\alpha}{2} \frac{E_{0,\Delta}^{(\infty)}}{1 + E_{0,\Delta}^{(\infty)}/\Delta}. \quad (81)$$

For general $t > 0$, we can write $i_{t,\Delta}$ as:

$$i_{t,\Delta} = i_{0,\Delta} + \int_0^t \frac{\partial i_{s,\Delta}}{\partial s} ds. \quad (82)$$

Denote the expectation w.r.t. the Gibbs measure as:

$$\langle g(\mathbf{X}) \rangle_{t,\Delta} := \int d\mathbf{x} P_t(\boldsymbol{\theta} = \mathbf{x} \mid \mathbf{y}, \tilde{\mathbf{y}}) g(\mathbf{x}). \quad (83)$$

where P_t is the posterior of $\boldsymbol{\theta}$ given \mathbf{y} and $\tilde{\mathbf{y}}$. By computing the derivative of $i_{t,\Delta}$ defined as (37), we can find that:

$$\frac{\partial i_{t,\Delta}}{\partial t} = \frac{1}{2N} \sum_{i=1}^N \mathbb{E}[\langle \bar{X}_i^2 - \frac{1}{\sqrt{t}} \bar{X}_i G_i \rangle_{t,\Delta}]. \quad (84)$$

Applying integration by parts w.r.t. G_i , we can simplify (84) as:

$$\lim_{N \rightarrow \infty} \frac{\partial i_{t,\Delta}}{\partial t} = \lim_{N \rightarrow \infty} \frac{1}{2N} \sum_{i=1}^N \mathbb{E}[\langle \bar{X}_i \rangle_{t,\Delta}^2] = \frac{1}{2} E_{t,\Delta}. \quad (85)$$

Taking the derivative of (82) with respect to Δ^{-1} on both sides when $N \rightarrow \infty$, by Lemma 1 and Lemma 2, we have

$$\frac{\alpha}{2} \frac{E_{t,\Delta}}{1 + E_{t,\Delta}/\Delta} = \frac{\alpha}{2} \frac{E_{0,\Delta}}{1 + E_{0,\Delta}/\Delta} + \frac{1}{2} \int_0^t \frac{\partial E_{s,\Delta}}{\partial \Delta^{-1}} ds. \quad (86)$$

The proof proceeds by showing that $E_{t,\Delta}^{(\infty)}$ and $E_{t,\Delta}$ satisfy the same PDE in Lemma 5. First, we observe that if $E := E(t, \Delta)$ is any fixed point of (28), then

$$\frac{\alpha}{2} \frac{\partial}{\partial t} \left(\frac{E}{1 + E/\Delta} \right) = \frac{1}{2} \frac{\partial E}{\partial \Delta^{-1}}. \quad (87)$$

Indeed, denote

$$A := (\text{mmse}^*)' \left(\frac{\alpha}{\Delta + E} + t \right). \quad (88)$$

Since $E = E(t, \Delta)$ satisfies

$$E = \text{mmse}^* \left(\frac{\alpha}{\Delta + E} + t \right), \quad (89)$$

by the chain rule, we have

$$\frac{\partial E}{\partial \Delta^{-1}} = A \alpha \frac{\Delta^2 - \frac{\partial E}{\partial \Delta^{-1}}}{(\Delta + E)^2}; \quad (90)$$

$$\frac{\partial E}{\partial t} = A \left(\frac{-\alpha \frac{\partial E}{\partial t}}{(\Delta + E)^2} + 1 \right), \quad (91)$$

which implies

$$\frac{\partial E}{\partial \Delta^{-1}} = \frac{A \alpha \Delta^2}{(\Delta + E)^2 + \alpha A}; \quad (92)$$

$$\frac{\partial E}{\partial t} = \frac{A(\Delta + E)^2}{(\Delta + E)^2 + \alpha A}. \quad (93)$$

Plugging them into both sides of (87), we obtain

$$\text{L.H.S.} = \frac{\alpha}{2} \frac{\partial E / \partial t}{(1 + E/\Delta)^2} = \frac{\alpha}{2} \frac{1}{(1 + E/\Delta)^2} \frac{A(\Delta + E)^2}{(\Delta + E)^2 + \alpha A} = \frac{\alpha}{2} \frac{A \Delta^2}{(\Delta + E)^2 + \alpha A}; \quad (94)$$

$$\text{R.H.S.} = \frac{1}{2} \frac{A \alpha \Delta^2}{(\Delta + E)^2 + \alpha A}, \quad (95)$$

which verifies (87). This suggests that $\frac{E_{t,\Delta}^{(\infty)}}{1 + E_{t,\Delta}^{(\infty)}/\Delta}$ follows the same track as $\frac{E_{t,\Delta}}{1 + E_{t,\Delta}/\Delta}$, i.e.,

$$\frac{\alpha}{2} \left(\frac{E_{t,\Delta}^{(\infty)}}{1 + E_{t,\Delta}^{(\infty)}/\Delta} - \frac{E_{0,\Delta}^{(\infty)}}{1 + E_{0,\Delta}^{(\infty)}/\Delta} \right) = \frac{\alpha}{2} \int_0^t \frac{\partial}{\partial s} \left(\frac{E_{s,\Delta}^{(\infty)}}{1 + E_{s,\Delta}^{(\infty)}/\Delta} \right) ds = \frac{1}{2} \int_0^t \frac{\partial E_{s,\Delta}^{(\infty)}}{\partial \Delta^{-1}} ds. \quad (96)$$

As $E_{t,\Delta}^{(\infty)}$ is the unique fixed point of (28) and initial condition $E_{0,\Delta} = E_{0,\Delta}^{(\infty)}$ holds for any $\Delta < \Delta_{\text{AMP}}$, we have $E_{t,\Delta} = E_{t,\Delta}^{(\infty)}$ as a result of the common property in PDE. \square

The property is given as follows:

Lemma 5. The solution of the following PDE is unique:

$$\begin{cases} \alpha \frac{\partial}{\partial t} \left(\frac{E}{1 + E/\Delta} \right) = \frac{\partial E}{\partial \Delta^{-1}}, \\ E(\delta, 0) = E_{0,\Delta}. \end{cases} \quad (97)$$

Proof. For any solution E of the PDE (97), we have

$$\alpha \frac{\partial E}{\partial t} \frac{1}{(\Delta + E)^2} = - \frac{\partial E}{\partial \Delta}. \quad (98)$$

$\partial E / \partial t = 0$ indicates that $\partial E / \partial \Delta = 0$, then E is a constant function which is unique. Now we suppose $\partial E / \partial t \neq 0$.

Suppose Δ is a function of t such that $E(\Delta(t), t) = E(\Delta(0), 0)$ for all t . Then

$$0 = \frac{\partial E(\Delta(t), t)}{\partial t} \quad (99)$$

$$= \frac{\partial E}{\partial \Delta} \frac{\partial \Delta}{\partial t} + \frac{\partial E}{\partial t} \quad (100)$$

$$= \frac{\partial E}{\partial t} \left(1 - \frac{\alpha}{(\Delta + E)^2} \frac{\partial \Delta}{\partial t} \right) \quad (101)$$

implies

$$\frac{\alpha}{(\Delta + E)^2} \frac{\partial \Delta}{\partial t} - 1 = 0. \quad (102)$$

This indicates that $\alpha / (\Delta + E) + t$ is a constant as a function of t since

$$\frac{\partial}{\partial t} \left(\frac{\alpha}{\Delta(t) + E(\Delta(t), t)} + t \right) \quad (103)$$

$$= - \frac{\alpha}{(\Delta + E)^2} \frac{\partial \Delta}{\partial t} + 1 \quad (104)$$

$$= 0. \quad (105)$$

Therefore

$$\frac{\alpha}{\Delta(0) + E(\Delta(0), 0)} = \frac{\alpha}{\Delta(t) + E(\Delta(t), t)} + t. \quad (106)$$

If we let

$$\Delta(t) = \frac{\alpha}{\frac{\alpha}{\Delta(0) + E(\Delta(0), 0)} - t} - E(\Delta(0), 0), \quad (107)$$

then for any solution E of PDE (97), there always exists $\Delta(t)$ such that

$$E(\Delta(t), t) = E(\Delta(0), 0) \quad \text{for all } t, \quad (108)$$

where $\Delta(0)$ and $E(\Delta(0), 0)$ are given previously.

Suppose the PDE (97) has two solutions E_1 and E_2 . Then there exist Δ_1 and Δ_2 such that

$$E_1(\Delta_1(t), t) = E_1(\Delta_1(0), 0) \quad (109)$$

$$E_2(\Delta_2(t), t) = E_2(\Delta_2(0), 0). \quad (110)$$

If we let $\Delta_1(0) = \Delta_2(0) = \delta$, then $E_1(\Delta_1(0), 0) = E_1(\delta, 0) = E_{0,\Delta} = E_2(\delta, 0) = E_2(\Delta_2(0), 0)$ which implies $\Delta_1(t) = \Delta_2(t)$. Since any (Δ, t) can be tracked by (107), we have $E_1 \equiv E_2$. \square

A.2 Proof of Theorem 1

In order to prove Theorem 1, we introduce the following lemmas.

Lemma 6. With probability at least $1 - o_N(1)$, the following inequality holds for some $C_1(\epsilon, T), C_2(\epsilon, T)$:

$$\frac{1}{\sqrt{N}} \sup_{t \in [0, T]} \|W_t\|_2 \leq C_1, \quad (111)$$

$$\frac{1}{N} \int_{t_{l-1}}^{t_l} \|W_t - W_{t_{l-1}}\|_2^2 dt \leq C_2 \delta^2, \text{ for } l = 1, 2, \dots, L. \quad (112)$$

Proof. Actually, as W_t is a N dimensional vector and all the dimensions are independent from each other, we can write $\|W_t\|_2^2$ as the sum of $W_{t,i}^2, i = 1, 2, \dots, M$, which are M i.i.d. squared normal distributions. We have:

$$\frac{1}{N} \sup_{t \in [0, T]} \|W_t\|_2^2 = \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^M W_{t,i}^2 \quad (113)$$

$$\leq \frac{M}{N} \frac{1}{M} \sum_{i=1}^M \sup_{t \in [0, T]} W_{t,i}^2 \quad (114)$$

By S.L.L.N, $\frac{1}{M} \sum_{i=1}^M \sup_{t \in [0, T]} W_{t,i}^2 \rightarrow T^2 \mathbb{E}\{\sup_{t \in [0, T]} W_{t,1}^2\}$ almost surely. And as $M/N \rightarrow \alpha$, we can always find C_α to bound M/N . Thus, we only need to let

$$C_1 = C_\alpha T^2 \mathbb{E}\{\sup_{t \in [0, T]} W_{t,1}^2\} \quad (115)$$

For the second constant C_2 , if we write $\|W_t - W_{t_{l-1}}\|_2^2 = \sum_{i=1}^M (W_{t,i} - W_{t_{l-1},i})^2$ and use the fact that $\int_{t_{l-1}}^{t_l} \|W_t - W_{t_{l-1}}\|_2^2 dt \leq \delta \sup_{t \in [t_{l-1}, t_l]} \|W_t - W_{t_{l-1}}\|_2^2$. We can get C_2 in the same way as C_1 by S.L.L.N. \square

Lemma 7. With probability $1 - o_N(1)$, we have

$$\|\phi\|_{\text{op}} \leq C(1 + \frac{2}{\sqrt{\alpha}}) := C_\phi \quad (116)$$

where C is a constant.

Proof. Applying Theorem 4.4.5 in Vershynin (2018). \square

Lemma 8. When condition (111) satisfies, there exists a constant C_3 only depends on C_1 such that:

$$\frac{1}{\sqrt{N}} \|\hat{m}(\hat{z}_t, t) - \hat{m}(\hat{z}_t, t_{l-1})\|_2 \leq C_3 |t - t_{l-1}|, \text{ for } t \in [t_{l-1}, t_l], l = 1, 2, \dots, L. \quad (117)$$

Proof. For $t \in [t_{l-1}, t_l]$, $\|\hat{z}_t\|_2 = \|\int_0^t \hat{m}(\hat{z}_t, t) dt + W_t\|_2 \leq \|\int_0^t \hat{m}(\hat{z}_t, t) dt\|_2 + \|W_t\|_2 \leq \sqrt{N} L_\theta T + \sqrt{N} C_1 = \sqrt{N}(L_\theta T + C_1)$.

Consider the i -th component of \hat{m} , Let

$$\langle A(\theta) \rangle_z = \frac{\int A(\theta) \exp(-\frac{(\hat{z}_{t,i} - \theta)^2(\tau_{k,t}^{-2} + t)}{2}) \pi(\theta) d\theta}{\int \exp(-\frac{(\hat{z}_{t,i} - \theta)^2(\tau_{k,t}^{-2} + t)}{2}) \pi(\theta) d\theta} \quad (118)$$

Since $\hat{m}(\hat{z}_{t,i}, t) = E[\theta | \theta + (\tau_{k,t}^{-2} + t)^{-1/2} G = \hat{z}_{t,i}]$, then $|\frac{d\hat{m}(\hat{z}_{t,i}, t)}{dt}| = |-\frac{1}{2}(\langle \theta^3 \rangle_z - \langle \theta \rangle_z \langle \theta^2 \rangle_z - 2\hat{z}_{t,i}(\langle \theta^2 \rangle_z - \langle \theta \rangle_z^2))| \leq \frac{1}{2}(2L_\theta^3 + 2|\hat{z}_{t,i}|L_\theta^2) = L_\theta^3 + |\hat{z}_{t,i}|L_\theta^2$ which means $|\hat{m}(\hat{z}_{t,i}, t) - \hat{m}(\hat{z}_{t,i}, t_{l-1})| \leq (L_\theta^3 + |\hat{z}_{t,i}|L_\theta^2)|t - t_{l-1}| \leq \sqrt{(L_\theta^3 + |\hat{z}_{t,i}|^2)(L_\theta^3 + L_\theta^4)}|t - t_{l-1}|$. Therefore,

$$\frac{1}{\sqrt{N}} \|\hat{m}(\hat{z}_t, t) - \hat{m}(\hat{z}_t, t_{l-1})\|_2 \quad (119)$$

$$= \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^N |\hat{m}(\hat{z}_{t,i}, t) - \hat{m}(\hat{z}_{t,i}, t_{l-1})|^2} \quad (120)$$

$$\leq \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^N (L_\theta^3 + |\hat{z}_{t,i}|^2)(L_\theta^3 + L_\theta^4)(t - t_{l-1})^2} \quad (121)$$

$$= \frac{1}{\sqrt{N}} \sqrt{(L_\theta^3 + L_\theta^4) \sum_{i=1}^N (L_\theta^3 + |\hat{z}_{t,i}|^2) |t - t_{l-1}|} \quad (122)$$

$$\leq \frac{1}{\sqrt{N}} \sqrt{(L_\theta^3 + L_\theta^4)(NL_\theta^3 + \|\hat{z}_t\|_2^2) |t - t_{l-1}|} \quad (123)$$

$$\leq \sqrt{(L_\theta^3 + L_\theta^4)(L_\theta^3 + (L_\theta T + C_1)^2) |t - t_{l-1}|}. \quad (124)$$

Finally, to complete the proof, we only need to let $C_3 = \sqrt{(L_\theta^3 + L_\theta^4)(L_\theta^3 + (L_\theta T + C_1)^2)}$. \square

Lemma 9. Suppose $f(x)$ is L_f -Lipschitz, and $g(x)$ is L_g -Lipschitz. Moreover, if $\|f\|_2 \leq M_f$ and $\|g\|_2 \leq M_g$, then

$$f(x)g(x) \text{ is } L_{fg}\text{-Lipschitz} \quad (125)$$

where $L_{fg} = L_f M_g + L_g M_f$.

Lemma 10. When condition (116) satisfies, there exists a constant $C_4(K, T)$ which only depends on K, T such that:

$$\|\hat{m}(\mathbf{z}_1, t) - \hat{m}(\mathbf{z}_2, t)\|_2 \leq C_4 \|\mathbf{z}_1 - \mathbf{z}_2\|_2. \quad (126)$$

Proof. First, for $1 \leq i \leq N$, notice that $\eta(\mathbf{y}_i^*; \tau_{k,t}^2) := E[\theta | \theta + (\tau_{k,t}^{-2} + t)^{-1/2} G = \mathbf{y}_i^*]$, where $G \sim N(0, 1)$, we have $\eta'(\mathbf{y}_i^*; \tau_{k,t}^2) = d\eta(\mathbf{y}_i^*; \tau_{k,t}^2)/d\mathbf{y}_i^* = (\tau_{k,t}^{-2} + t) \text{Var}(\theta | \theta + (\tau_{k,t}^{-2} + t)^{-1/2} G = \mathbf{y}_i^*) \leq (\alpha/\Delta + T)L_\theta^2$. Therefore, $\eta(\mathbf{y}_i^*; \tau_{k,t}^2)$ is $(\alpha/\Delta + T)L_\theta^2$ -Lipschitz, and consequently, $\eta(\mathbf{y}^*; \tau_{k,t}^2)$ is $(\alpha/\Delta + T)L_\theta^2$ -Lipschitz.

Moreover, let

$$\langle A(\theta) \rangle_{\mathbf{y}_i^*} = \frac{\int A(\theta) \exp(-\frac{(\mathbf{y}_i^* - \theta)^2(\tau_{k,t}^{-2} + t)}{2}) \pi(\theta) d\theta}{\int \exp(-\frac{(\mathbf{y}_i^* - \theta)^2(\tau_{k,t}^{-2} + t)}{2}) \pi(\theta) d\theta} \quad (127)$$

then $|d\eta'(\mathbf{y}_i^*; \tau_{k,t}^2)/d\mathbf{y}_i^*| = |(\tau_{k,t}^{-2} + t)^2(\langle \theta^3 \rangle_{\mathbf{y}_i^*} - \langle \theta^2 \rangle_{\mathbf{y}_i^*} \langle \theta \rangle_{\mathbf{y}_i^*} - 2\langle \theta \rangle_{\mathbf{y}_i^*} \langle \theta^2 \rangle_{\mathbf{y}_i^*} + 2\langle \theta \rangle_{\mathbf{y}_i^*}^3)| \leq 6(\alpha/\Delta + T)^2 L_\theta^3$ which means that $\eta'(\mathbf{y}_i^*; \tau_{k,t}^2)$ is $6(\alpha/\Delta + T)^2 L_\theta^3$ -Lipschitz. Therefore,

$$|\frac{1}{N\alpha} \sum_{i=1}^N [\eta'(\mathbf{y}_i^*; \tau_{k,t}^2)]_i - \frac{1}{N\alpha} \sum_{i=1}^N [\eta'(\tilde{\mathbf{y}}_i; \tau_{k,t}^2)]_i| \quad (128)$$

$$= |\frac{1}{N\alpha} \sum_{i=1}^N \eta'(\mathbf{y}_i^*; \tau_{k,t}^2) - \frac{1}{N\alpha} \sum_{i=1}^N \eta'(\tilde{\mathbf{y}}_i; \tau_{k,t}^2)| \quad (129)$$

$$\leq \frac{1}{N\alpha} \sum_{i=1}^N |\eta'(\mathbf{y}_i^*; \tau_{k,t}^2) - \eta'(\tilde{\mathbf{y}}_i; \tau_{k,t}^2)| \quad (130)$$

$$\leq \frac{1}{N\alpha} \sum_{i=1}^N 6(\alpha/\Delta + T)^2 L_\theta^3 |\mathbf{y}_i^* - \tilde{\mathbf{y}}_i| \quad (131)$$

$$\leq \frac{6(\alpha/\Delta + T)^2 L_\theta^3}{\sqrt{N}\alpha} \|\mathbf{y}^* - \tilde{\mathbf{y}}\|_2 \quad (132)$$

which means that $\frac{1}{N\alpha} \sum_{i=1}^N [\eta'(\mathbf{y}_i^*; \tau_{k,t}^2)]_i$ is $\frac{6(\alpha/\Delta + T)^2 L_\theta^3}{\sqrt{N}\alpha}$ -Lipschitz.

Denote $C_b := \frac{6(\alpha/\Delta + T)^2 L_\theta^3}{\sqrt{N}\alpha}$ and let $b_t^{(k+1)}(\mathbf{z}, t) := \frac{1}{N\alpha} \sum_{i=1}^N [\eta'(\phi^T \mathbf{r}^{(k)}(\mathbf{z}, t) + \hat{m}^{(k)}(\mathbf{z}, t); \tau_{k,t}^2)]_i$.

Now we use induction to prove the following inequalities:

$$\|\hat{m}^{(k)}(\mathbf{z}_1, t) - \hat{m}^{(k)}(\mathbf{z}_2, t)\|_2 \leq C(k, T) \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (133)$$

$$\|\mathbf{r}^{(k)}(\mathbf{z}_1, t) - \mathbf{r}^{(k)}(\mathbf{z}_2, t)\|_2 \leq C^*(k, T) \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (134)$$

$$\|\mathbf{r}^{(k)}(\mathbf{z}_1, t) \prod_{i=0}^v b_t^{(k+i)}(\mathbf{z}_1, t) - \mathbf{r}^{(k)}(\mathbf{z}_2, t) \prod_{i=0}^v b_t^{(k+i)}(\mathbf{z}_2, t)\|_2 \leq C'(k, v, T) \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \text{ for all } v \geq 0 \quad (135)$$

where

$$C(k+1, T) = (\alpha/\Delta + T) L_\theta^2 (C_\phi C^*(k, T) + C(k, T)) \quad (136)$$

$$C^*(k+1, T) = (C_\phi C(k+1, T, N) + C'(k, 0, T)) \quad (137)$$

$$\begin{aligned} C'(k+1, v, T) = & \{ \|y\|_2 (\frac{2L_\theta^2}{\alpha})^v \frac{6(\alpha/\Delta + T)^2 L_\theta^3}{\alpha} + C_\phi [(v-1) (\frac{L_\theta^2}{\alpha})^{v-1} \frac{6(\alpha/\Delta + T)^2 L_\theta^4}{\alpha} \\ & + (\frac{L_\theta^2}{\alpha})^v C(k+1, T)] + C'(k, v+1, T) \}. \end{aligned} \quad (138)$$

Suppose (133), (134) and (135) holds for $l = 0, \dots, k$, where $1 \leq k \leq K-1$.

For $l = k+1$, we have

$$\|\hat{m}^{(k+1)}(\mathbf{z}_1, t) - \hat{m}^{(k+1)}(\mathbf{z}_2, t)\|_2 \quad (139)$$

$$= \|\eta(\phi^T \mathbf{r}^{(k)}(\mathbf{z}_1, t) + \hat{m}^{(k)}(\mathbf{z}_1, t); \tau_{k,t}^2) - \eta(\phi^T \mathbf{r}^{(k)}(\mathbf{z}_2, t) + \hat{m}^{(k)}(\mathbf{z}_2, t); \tau_{k,t}^2)\|_2 \quad (140)$$

$$\leq (\alpha/\Delta + T) L_\theta^2 \|\phi^T(\mathbf{r}^{(k)}(\mathbf{z}_1, t) - \mathbf{r}^{(k)}(\mathbf{z}_2, t)) + \hat{m}^{(k)}(\mathbf{z}_1, t) - \hat{m}^{(k)}(\mathbf{z}_2, t)\|_2 \quad (141)$$

$$\leq (\alpha/\Delta + T) L_\theta^2 (\|\phi\|_{\text{op}} C^*(k, T) + C(k, T)) \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (142)$$

$$\leq (\alpha/\Delta + T) L_\theta^2 (C_\phi C^*(k, T) + C(k, T)) \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (143)$$

$$= C(k+1, T) \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (144)$$

and

$$\|\mathbf{r}^{(k+1)}(\mathbf{z}_1, t) - \mathbf{r}^{(k+1)}(\mathbf{z}_2, t)\| \quad (145)$$

$$= \|\phi(\hat{m}^{(k+1)}(\mathbf{z}_1, t) - \hat{m}^{(k+1)}(\mathbf{z}_2, t)) + \mathbf{r}^{(k)}(\mathbf{z}_1, t)b_t^{(k)}(\mathbf{z}_1, t) - \mathbf{r}^{(k)}(\mathbf{z}_2, t)b_t^{(k)}(\mathbf{z}_2, t)\|_2 \quad (146)$$

$$\leq (\|\phi\|_{\text{op}}C(k+1, T) + C'(k, 0, T))\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (147)$$

$$\leq (C_\phi C(k+1, T) + C'(k, 0, T))\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (148)$$

$$= C^*(k+1, T)\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (149)$$

and

$$\|\mathbf{r}^{(k+1)}(\mathbf{z}_1, t) \prod_{i=0}^v b_t^{(k+1+i)}(\mathbf{z}_1, t) - \mathbf{r}^{(k+1)}(\mathbf{z}_2, t) \prod_{i=0}^v b_t^{(k+1+i)}(\mathbf{z}_2, t)\|_2 \quad (150)$$

$$\begin{aligned} &= \|\mathbf{y}(\prod_{i=0}^v b_t^{(k+1+i)}(\mathbf{z}_1, t) - \prod_{i=0}^v b_t^{(k+1+i)}(\mathbf{z}_2, t)) \\ &\quad - \phi(\hat{m}^{(k+1)}(\mathbf{z}_1, t) \prod_{i=0}^v b_t^{(k+1+i)}(\mathbf{z}_1, t) - \hat{m}^{(k+1)}(\mathbf{z}_2, t) \prod_{i=0}^v b_t^{(k+1+i)}(\mathbf{z}_2, t)) \\ &\quad + \mathbf{r}^{(k)}(\mathbf{z}_1, t) \prod_{i=0}^{v+1} b_t^{(k+i)}(\mathbf{z}_1, t) - \mathbf{r}^{(k)}(\mathbf{z}_2, t) \prod_{i=0}^{v+1} b_t^{(k+i)}(\mathbf{z}_2, t)\|_2 \end{aligned} \quad (151)$$

$$\leq \{\|\mathbf{y}\|_2(v+1)(\frac{L_\theta^2}{\alpha})^v C_b + \|\phi\|_{\text{op}}\}\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad (152)$$

$$\begin{aligned} &\leq \{\|\mathbf{y}\|_2(\frac{2L_\theta^2}{\alpha})^v C_b + C_\phi[(v-1)(\frac{L_\theta^2}{\alpha})^{v-1} C_b \sqrt{N} L_\theta + (\frac{L_\theta^2}{\alpha})^v C(k+1, T)] \\ &\quad + C'(k, v+1, T)\}\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \end{aligned} \quad (153)$$

$$\begin{aligned} &= \{\|\mathbf{y}\|_2(\frac{2L_\theta^2}{\alpha})^v \frac{6(\alpha/\Delta + T)^2 L_\theta^3}{\alpha} + C_\phi[(v-1)(\frac{L_\theta^2}{\alpha})^{v-1} \frac{6(\alpha/\Delta + T)^2 L_\theta^4}{\alpha} + (\frac{L_\theta^2}{\alpha})^v C(k+1, T)] \\ &\quad + C'(k, v+1, T)\}\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \end{aligned} \quad (154)$$

$$= C'(k+1, v, T)\|\mathbf{z}_1 - \mathbf{z}_2\|_2. \quad (155)$$

Finally, we prove (133), (134) and (135) holds for $k = 0$.

Obviously, (133), (134) hold since $\hat{m}^{(0)}(\mathbf{z}_1, t) = \hat{m}^{(0)}(\mathbf{z}_2, t) = 0$ and $\mathbf{r}^{(0)}(\mathbf{z}_1, t) - \mathbf{r}^{(0)}(\mathbf{z}_2, t) = \mathbf{y}$.

Moreover,

$$\|\mathbf{r}^{(0)}(\mathbf{z}_1, t) \prod_{i=0}^v b_t^{(i)}(\mathbf{z}_1, t) - \mathbf{r}^{(0)}(\mathbf{z}_2, t) \prod_{i=0}^v b_t^{(i)}(\mathbf{z}_2, t)\|_2 \quad (156)$$

$$= \|\mathbf{y}\|_2 \|\prod_{i=0}^v b_t^{(i)}(\mathbf{z}_1, t) - \prod_{i=0}^v b_t^{(i)}(\mathbf{z}_2, t)\|_2 \quad (157)$$

$$\leq \|\mathbf{y}\|_2 (2\frac{L_\theta^2}{\alpha})^v C_b \quad (158)$$

$$\leq \|\mathbf{y}\|_2 (2\frac{L_\theta^2}{\alpha})^v \frac{6(\alpha/\Delta + T)^2 L_\theta^3}{\alpha}. \quad (159)$$

We only need to let $C'(0, v, T) = \|\mathbf{y}\|_2 (2\frac{L_\theta^2}{\alpha})^v \frac{6(\alpha/\Delta + T)^2 L_\theta^3}{\alpha}$ to complete the proof.

□

Proof of Theorem 1.

$$\frac{1}{N} D_{\text{KL}}(P_{\mathbf{z}_T/T} \| P_{\hat{\mathbf{z}}_T/T}) \quad (160)$$

$$\lesssim \sum_{l=1}^L \mathbb{E} \int_{t_{l-1}}^{t_l} \frac{1}{N} \|m(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_{t_l}, t_l)\|_2^2 dt \quad (161)$$

$$\lesssim \sum_{l=1}^L \mathbb{E} \int_{t_{l-1}}^{t_l} \frac{1}{N} (\|m(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_t, t)\|_2^2 + \|\hat{m}(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_{t_l}, t)\|_2^2 + \|\hat{m}(\mathbf{z}_{t_l}, t) - \hat{m}(\mathbf{z}_{t_l}, t_l)\|_2^2) dt \quad (162)$$

where (161) follows by Girsanov's theorem (see for example (Oko et al., 2023, Proposition D.1)).

For the first term of (162),

$$\lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \sum_{l=1}^L \mathbb{E} \int_{t_{l-1}}^{t_l} \frac{1}{N} \|m(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_t, t)\|_2^2 dt \quad (163)$$

$$= \lim_{K \rightarrow \infty} \sum_{l=1}^L \int_{t_{l-1}}^{t_l} \lim_{N \rightarrow \infty} \mathbb{E} \frac{1}{N} \|m(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_t, t)\|_2^2 dt \quad (164)$$

$$= 0 \quad (165)$$

where (164) follows from Dominated Convergence Theorem, and (165) follows from Theorem 2. Therefore, we could choose K such that $1/N \sum_{l=1}^L \mathbb{E} \int_{t_{l-1}}^{t_l} \|m(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_t, t)\|_2^2 dt \leq \epsilon/2$ for large N . For such K , δ could be given as follows:

For the second term of (162), by Lemma 10, with probability $1 - o_N(1)$, we have

$$\frac{1}{N} \mathbb{E} \|\hat{m}(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_{t_l}, t)\|_2^2 \quad (166)$$

$$\leq \frac{1}{N} C_4^2 \mathbb{E} \|\mathbf{z}_t - \mathbf{z}_{t_l}\|_2^2 \quad (167)$$

$$= \frac{1}{N} C_4^2 \mathbb{E} \left\| \int_{t_{l-1}}^t m(\mathbf{z}_s, s) ds + W_t - W_{t_l} \right\|_2^2 \quad (168)$$

$$\leq \frac{2}{N} C_4^2 \left(\mathbb{E} \left\| \int_{t_{l-1}}^t m(\mathbf{z}_s, s) ds \right\|_2^2 + \mathbb{E} \|W_t - W_{t_l}\|_2^2 \right) \quad (169)$$

$$\leq 2C_4^2 (L_\theta^2 (t - t_{l-1})^2 + t - t_{l-1}). \quad (170)$$

$$(171)$$

Therefore,

$$\frac{1}{N} \sum_{l=1}^L \mathbb{E} \int_{t_{l-1}}^{t_l} \|\hat{m}(\mathbf{z}_t, t) - \hat{m}(\mathbf{z}_{t_l}, t)\|_2^2 dt \quad (172)$$

$$\leq 2C_4^2 \sum_{l=1}^L \int_{t_{l-1}}^{t_l} (L_\theta^2 (t - t_{l-1})^2 + t - t_{l-1}) dt \quad (173)$$

$$= 2C_4^2 \sum_{l=1}^L \left(\frac{L_\theta^2 \delta^3}{3} + \frac{\delta^2}{2} \right) \quad (174)$$

$$= 2C_4^2 \left(\frac{L_\theta^2 T \delta}{3} + \frac{T}{2} \right) \delta. \quad (175)$$

For the third term of (162), by Lemma 8, with probability $1 - o_N(1)$, we have

$$\sum_{l=1}^L \int_{t_{l-1}}^{t_l} \frac{1}{N} \|\hat{m}(\mathbf{z}_{t_l}, t) - \hat{m}(\mathbf{z}_{t_l}, t_l)\|_2^2 dt \quad (176)$$

$$\leq \sum_{l=1}^L \int_{t_{l-1}}^{t_l} C_3^2 (t - t_{l-1})^2 dt \quad (177)$$

$$\leq \sum_{l=1}^L C_3^2 \delta^3 / 3 \quad (178)$$

$$= \frac{C_3^2 T \delta^2}{3}. \quad (179)$$

By (175) and (179), (1) holds if we let

$$\delta \leq \min \left\{ 1, \frac{\epsilon/2}{2C_4^2(\frac{L_\theta^2 T}{3} + \frac{T}{2}) + \frac{C_3^2 T}{3}} \right\}. \quad (180)$$

□

Meanwhile, we also prove another condition for the theorem to hold, which we stated before as Corollary 1. This can be derived directly from the following lemma.

Lemma 11. If the derivative of $\Sigma^2 \mapsto \text{mmse}^*(\Sigma^{-2})$ is continuous at $\Sigma^2 = 0$, the fixed point equation (28) will always have a unique solution for any fixed Δ when α is sufficiently large.

Proof. As $\alpha \rightarrow \infty$, any fixed point of (28), E , will converge to 0. Suppose there are two different fixed points, $E_1 < E_2$. Then, there exists one $E_3 \in [E_1, E_2]$ such that

$$1 = \frac{d}{dE} \text{mmse}^*\left(\frac{\alpha}{\Delta + E}\right) \Big|_{E=E_3}, \quad (181)$$

which is the derivative of the fixed point equation. Thus, by the regularity assumption, the derivative of mmse^* can be bounded near 0. That is, we can find $\delta > 0$, $\frac{d}{d\Sigma^2} \text{mmse}^*(\Sigma^{-2}) < \delta$ holds for all Σ near 0.

Thus, as $E_3 \rightarrow 0$ when $\alpha \rightarrow \infty$, the right hand side of (181) can be bounded as follows:

$$\frac{d}{dE} \text{mmse}^*\left(\frac{\alpha}{\Delta + E}\right) \Big|_{E=E_3} \quad (182)$$

$$= \frac{d}{d(\frac{\Delta+E}{\alpha})} \text{mmse}^*\left(\frac{\alpha}{\Delta + E}\right) \frac{d}{dE} \left(\frac{\Delta + E}{\alpha} \right) \Big|_{E=E_3} \quad (183)$$

$$\leq \frac{\delta}{\alpha} \rightarrow 0, \quad (184)$$

which leads to a contradiction, as the L.H.S. of (181) equals to 1. □

B BOUNDING THE WASSERSTEIN DISTANCE BETWEEN z_T/T AND \tilde{z}_T/T

Here we improve our main result Theorem 1 by giving the bound of $1/\sqrt{N}W_2(P_{z_T/T}, P_{\tilde{z}_T/T})$. We use two terms to bound it by applying triangle inequality. The first term is $1/\sqrt{N}W_2(P_{\hat{z}_T/T}, P_{\tilde{z}_T/T})$ where \hat{z}_T and \tilde{z}_T are generated by continuous AMP process(23) and discretized AMP process(24). We show that the gap between these two AMP processes is small by using the properties that the maps $\mathbf{z}_t \mapsto \hat{m}(\mathbf{z}_t, t)$ and $t \mapsto \hat{m}(\mathbf{z}_t, t)$ are Lipschitz continuous shown in Lemma 8 and Lemma 10.

Lemma 12. If $\delta < 1$ and (126), (117) and (112) hold, the gap between two AMP processes is small, i.e.

$$\sum_{l=1}^L \frac{1}{N} \int_{t_{l-1}}^{t_l} \|\hat{m}(\hat{z}_t, t) - \hat{m}(\tilde{z}_{t_{l-1}}, t_{l-1})\|_2^2 dt \rightarrow 0 \quad \text{as } \delta \rightarrow 0. \quad (185)$$

Proof. Let $B_l = \frac{1}{N} \int_{t_{l-1}}^{t_l} \|\hat{m}(\hat{z}_t, t) - \hat{m}(\tilde{z}_{t_{l-1}}, t_{l-1})\|_2^2 dt$ and $D_l = \|\hat{z}_{t_l} - \tilde{z}_{t_l}\|_2$ for $l = 1, 2, \dots, L$. We claim that there exist constants $C_5(T, N, K, \epsilon)$ and $C_6(T, N, K, \epsilon)$ such that

$$B_l \leq C_5 \delta^2 (C_6 \delta + 1)^{l-1} \quad \text{for } l = 1, 2, \dots, L. \quad (186)$$

Then $\sum_{l=1}^L B_l \leq \frac{C_5}{C_6} \delta ((C_6 \delta + 1)^L - 1) \leq \delta \frac{C_5}{C_6} e^{C_6 T}$ which converges to 0 as $\delta \rightarrow 0$.

Now we use induction to prove (186). Suppose (186) holds for $i = 1, 2, \dots, l-1$. For B_l , it can be calculated as:

$$B_l \leq \frac{1}{N} \int_{t_{l-1}}^{t_l} (2\|\hat{m}(\hat{z}_t, t) - \hat{m}(\hat{z}_t, t_{l-1})\|_2^2 + 2\|\hat{m}(\hat{z}_t, t_{l-1}) - \hat{m}(\tilde{z}_{t_{l-1}}, t_{l-1})\|_2^2) dt \quad (187)$$

$$\leq \frac{2C_3^2}{3} \delta^3 + \frac{2C_4^2}{N} \int_{t_{l-1}}^{t_l} \|\hat{z}_t - \tilde{z}_{t_{l-1}}\|_2^2 dt \quad (188)$$

$$\leq \frac{2C_3^2}{3} \delta^3 + \frac{4C_4^2}{N} \int_{t_{l-1}}^{t_l} \|\hat{z}_t - \hat{z}_{t_{l-1}}\|_2^2 dt + \frac{4C_4^2}{N} \int_{t_{l-1}}^{t_l} D_{l-1}^2 dt, \quad (189)$$

where the first inequality is from (126) and (117). Then, consider the second term in the last equation, by (112), we have:

$$\int_{t_{l-1}}^{t_l} \|\hat{z}_t - \hat{z}_{t_{l-1}}\|_2^2 dt = \int_{t_{l-1}}^{t_l} \left\| \int_{t_{l-1}}^t \hat{m}(\hat{z}_s, s) ds + dW_s \right\|_2^2 dt \quad (190)$$

$$\leq \int_{t_{l-1}}^{t_l} 2 \left(\left\| \int_{t_{l-1}}^t \hat{m}(\hat{z}_s, s) ds \right\|_2^2 + \|W_t - W_{t_{l-1}}\|_2^2 \right) dt \quad (191)$$

$$\leq \int_{t_{l-1}}^{t_l} (2L_\theta^2 \delta^2 + 2\|W_t - W_{t_{l-1}}\|_2^2) dt \quad (192)$$

$$\leq 2L_\theta^2 \delta^3 + 2NC_2 \delta^2. \quad (193)$$

Now, we consider the third term.

$$D_l - D_{l-1} \leq \|\hat{z}_{t_l} - \hat{z}_{t_{l-1}} - (\tilde{z}_{t_l} - \tilde{z}_{t_{l-1}})\|_2 \quad (194)$$

$$\leq \int_{t_{l-1}}^{t_l} \|\hat{m}(\hat{z}_t, t) - \hat{m}(\tilde{z}_{t_{l-1}}, t_{l-1})\|_2 dt \quad (195)$$

$$\leq (\delta N B_l)^{1/2}. \quad (196)$$

Since $D_0 = 0$, we have

$$D_l \leq \sqrt{\delta N} \sum_{s=1}^l B_s^{1/2} \leq \sqrt{\delta N l} \left(\sum_{s=1}^l B_s \right)^{1/2} \leq \sqrt{NT} \left(\sum_{s=1}^l B_s \right)^{1/2} \quad (197)$$

Combining (189), (193) and (197), we have

$$B_l \leq \frac{2C_3^2}{3} \delta^3 + \frac{4C_4^2}{N} (2L_\theta^2 \delta^3 + 2NC_2 \delta^2) + \frac{4C_4^2}{N} \int_{t_{l-1}}^{t_l} NT \left(\sum_{s=1}^{l-1} B_s \right) \quad (198)$$

$$\leq \left[\frac{2C_3^2}{3} + \frac{4C_4^2}{N} (2L_\theta^2 + 2NC_2) \right] \delta^2 + 4C_4^2 T \delta \left(\sum_{s=1}^{l-1} B_s \right). \quad (199)$$

Let $C_5 = \frac{2C_3^2}{3} + \frac{4C_4^2}{N}(2L_\theta^2 + 2NC_2)$ and $C_6 = 4TC_4^2$, by (186), we have

$$B_l \leq C_5\delta^2 + C_6\delta\left(\sum_{s=1}^{l-1} B_s\right) \quad (200)$$

$$\leq C_5\delta^2 + C_6\delta\left[\sum_{s=1}^{l-1} C_5\delta^2(C_6\delta + 1)^{l-1}\right] \quad (201)$$

$$= C_5\delta^2(C_6\delta + 1)^l. \quad (202)$$

Moreover, we need to prove the case when $l = 1$. Actually, in (189), the third term equals to 0 when $l = 1$. Thus, by (199), we have,

$$B_1 \leq (2L_\theta^2 + 2NC_2)\delta^2 = C_5\delta^2. \quad (203)$$

□

The second term is $1/\sqrt{N}W_2(P_{z_T/T}, P_{\hat{z}_T/T})$ where z_T and \hat{z}_T are generated by diffusion process(22) with drift term $m(z_t, t)$ and continuous AMP process(23). We show that the second term will converge to 0 under the assumption that $z_t \mapsto m(z_t, t)$ is Lipschitz continuous. Finally, we give the following theorem.

Theorem 5. If $m(z_t, t)$ is L_m -Lipschitz w.r.t z_t ,

$$\lim_{\delta \rightarrow 0} \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}} W_2(P_{z_T/T}, P_{\hat{z}_T/T}) = 0. \quad (204)$$

Proof.

$$\lim_{\delta \rightarrow 0} \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}} W_2(P_{z_T/T}, P_{\hat{z}_T/T}) \quad (205)$$

$$\leq \lim_{\delta \rightarrow 0} \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}} W_2(P_{\hat{z}_T/T}, P_{\tilde{z}_T/T}) + \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}} W_2(P_{z_T/T}, P_{\hat{z}_T/T}). \quad (206)$$

For the first term of (206), we have

$$\frac{1}{\sqrt{N}} W_2(P_{\hat{z}_T/T}, P_{\tilde{z}_T/T}) \quad (207)$$

$$\leq \frac{1}{\sqrt{NT}} (\mathbb{E} \|\tilde{z}_T - \hat{z}_T\|_2^2)^{1/2} \quad (208)$$

$$\leq \frac{1}{\sqrt{NT}} (\mathbb{E} \left\| \sum_{l=1}^L \int_{t_{l-1}}^{t_l} (\hat{m}(\hat{z}_t, t) - \hat{m}(\tilde{z}_{t_{l-1}}, t_{l-1})) dt \right\|_2^2)^{1/2} \quad (209)$$

$$\leq \frac{1}{\sqrt{N}} (\mathbb{E} \sum_{l=1}^L \int_{t_{l-1}}^{t_l} \|\hat{m}(\hat{z}_t, t) - \hat{m}(\tilde{z}_{t_{l-1}}, t_{l-1})\|_2^2 dt)^{1/2} \quad (210)$$

which converges to 0 as $\delta \rightarrow 0$ by applying Lemma 12.

For the second term of (206), we first consider $1/N\mathbb{E}\|z_T - \hat{z}_T\|_2^2$

$$\frac{1}{N}\mathbb{E}\|z_T - \hat{z}_T\|_2^2 \quad (211)$$

$$= \frac{1}{N}\mathbb{E}\left\|\int_0^T (m(z_t, t) - \hat{m}(\hat{z}_t, t))dt\right\|_2^2 \quad (212)$$

$$\leq \frac{1}{N}\mathbb{E}\int_0^T \|m(z_t, t) - \hat{m}(\hat{z}_t, t)\|_2^2 dt \quad (213)$$

$$\lesssim \frac{1}{N}\mathbb{E}\int_0^T \|m(z_t, t) - m(\hat{z}_t, t)\|_2^2 dt + \frac{1}{N}\mathbb{E}\int_0^T \|m(\hat{z}_t, t) - \hat{m}(\hat{z}_t, t)\|_2^2 dt \quad (214)$$

$$\leq \frac{1}{N}\mathbb{E}\int_0^T L_m^2 \|z_t - \hat{z}_t\|_2^2 dt + o_{K,N} \quad (215)$$

where $\lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} o_{K,N} \rightarrow 0$ by using Theorem 2.

Let $f(t) = 1/N\mathbb{E}\|z_t - \hat{z}_t\|_2^2$, then we get $f(t) \lesssim L_m^2 \int_0^T f(t)dt + o_{K,N}$ which implies $f(t) \lesssim o_{K,N} \exp(L_m^2 t)$.

Therefore,

$$\lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{N} W_2^2(P_{z_T/T}, P_{\hat{z}_T/T}) \quad (216)$$

$$\leq \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{NT^2} \mathbb{E}\|z_T - \hat{z}_T\|_2^2 \quad (217)$$

$$\lesssim \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{T^2} (o_{K,N} e^{L_m^2 T}) \quad (218)$$

$$= 0. \quad (219)$$

□

B.1 Connection between overlap concentration and Lipschitz constant of m_t

In Appendix B, we showed that convergence under the Wasserstein distance can be established assuming that $m_t(z_t) := m(z_t, t)$ has a dimension free Lipschitz constant. In this section, we point out a connection between such a condition and a concentration property of the overlap.

For the Gibbs measure, a fundamental quantity is the overlap:

$$R_{1,2} := \frac{1}{p} \theta^{1\top} \theta^2 \quad (220)$$

where θ^1 and θ^2 are two independent random vectors (spins) drawn from the Gibbs measure. we write $q_p := \langle R_{1,2} \rangle$, where $\langle \cdot \rangle$ denotes the average with respect to the Gibbs measure (note that (Talagrand, 2010, (1.74)) defined q , which is the a.s. limit of q_p). For the SK model, it is well known that $R_{1,2}$ concentrates around q only in a certain parameter regime, whose boundary is conjectured to be the Almeida-Thouless (AT) line (Talagrand, 2010, Section 1.8).

In this section, we argue that if $R_{1,2}$ does not concentrate, then m_t cannot have a dimension-free Lipschitz constant:

Proposition 2. Suppose that the prior of each coordinate of θ is supported on $[-1, 1]$. Let $\lambda_m := \lambda_{\max}(\nabla m_t) = \lambda_{\max}(\langle \theta \theta^\top \rangle - \langle \theta \rangle \langle \theta \rangle^\top)$, then

$$\frac{p}{2} \langle |R_{1,2} - q_p|^2 \rangle \leq \lambda_m \leq p \sqrt{\langle |R_{1,2} - q_p|^2 \rangle}. \quad (221)$$

As a consequence, $\lambda_m = O(1)$ only if $\langle |R_{1,2} - q_p|^2 \rangle = O(1/p)$, which is the typical order of the variance of the overlap in the high-temperature regime (Talagrand, 2010). Conversely, if $\langle |R_{1,2} - q_p|^2 \rangle = O(1/p)$, then by (221), $\lambda_m \leq \sqrt{p}$, which is sharper than the *a priori estimate* $\lambda_m = O(p)$ using the information $\|\theta\|_2^2 \leq p$.

Proof. Note that

$$\langle |R_{1,2} - q_p|^2 \rangle = \frac{1}{p^2} \langle |\theta^{1\top} \theta^2|^2 \rangle - \frac{1}{p^2} |\langle \theta \rangle^\top \langle \theta \rangle|^2; \quad (222)$$

and

$$\langle \theta^\top (\langle \theta \theta^\top \rangle - \langle \theta \rangle \langle \theta \rangle^\top) \theta \rangle = \langle |\theta^{1\top} \theta^2|^2 \rangle - \text{Tr} \left(\langle \theta \rangle \langle \theta \rangle^\top \langle \theta \theta^\top \rangle \right); \quad (223)$$

and moreover

$$\langle \theta \rangle^\top (\langle \theta \theta^\top \rangle - \langle \theta \rangle \langle \theta \rangle^\top) \langle \theta \rangle = \text{Tr} \left(\langle \theta \rangle \langle \theta \rangle^\top \langle \theta \theta^\top \rangle \right) - |\langle \theta \rangle^\top \langle \theta \rangle|^2. \quad (224)$$

Note that (223) and (224) are both nonnegative and their sum is $p^2 \langle |R_{1,2} - q_p|^2 \rangle$, so one of them is at least $\frac{p^2}{2} \langle |R_{1,2} - q_p|^2 \rangle$. Since the norm of θ is at most \sqrt{p} , the lower bound on the eigenvalue follows.

Similarly, by Jensen's inequality we have $\langle \theta \rangle^\top \langle \theta \theta^\top \rangle \langle \theta \rangle \geq (\langle \theta \rangle^\top \langle \theta \rangle)^2$, hence

$$\lambda_{\max}^2(\langle \theta \theta^\top \rangle - \langle \theta \rangle \langle \theta \rangle^\top) \leq \text{Tr} \left[\left(\langle \theta \theta^\top \rangle - \langle \theta \rangle \langle \theta \rangle^\top \right)^2 \right] \leq p^2 \langle |R_{1,2} - q_p|^2 \rangle. \quad (225)$$

□

While the non-concentration of the overlap is indeed possible for the SK model in some parameter regimes, it still remains to address if (or when) does non-concentration of the overlap happens in the linear-inverse problem. For “good” prior distributions such as Gaussian, we clearly have $\langle |R_{1,2} - q_p|^2 \rangle = O(1/p)$ and indeed $\lambda_{\max}(\nabla m_t) = O(1)$ as the posterior is simply a Gaussian distribution with covariance matrix smaller than that of the prior. On the other hand, consider the example where the prior is supported on $\{\pm 1\}$, $\phi_{ij} \sim \mathcal{N}(0, \frac{1}{n})$, and $\delta = n/p \rightarrow \infty$. Then since $|\theta_j|$ is constant, we have

$$\nu(\theta) \propto \exp \left(\frac{1}{\Delta} y^\top \phi \theta - \frac{1}{2\Delta} \theta^\top (\phi^\top \phi - I) \theta + z^\top \theta \right). \quad (226)$$

It is well-known that the law of $\sqrt{\delta}(\phi^\top \phi - I)$ converges (in eigen-spectrum) to that of the Wigner random matrix when n/p is large, so that ν converges to the Gibbs measure of the SK model

$$\nu_{SK}(\theta) \propto \exp \left(\frac{\beta}{\sqrt{p}} \sum_{i < j \leq p} g_{ij} \theta_i \theta_j + \sum_{i \leq p} h_i \theta_i \right) \quad (227)$$

where $\beta = \frac{1}{\Delta \sqrt{\delta}}$ and $h = z + \frac{1}{\Delta} y^\top \phi$. This suggests that m_t may not have a dimension independent Lipschitz constant, at least in some parameter regime and for some (y, z) (where y does not necessarily follow the linear model (1)). This is in sharp contrast with the fact that $m_{K,t}$ does have a dimension free (depending on K) Lipschitz constant.

In the special case of (226) where y does follow the linear model (1) (the so-called “planted” or “Bayes optimal” setting), Barbier et al. (2020) used the Nishimori identity to establish certain overlap concentration results. Furthermore, the recent breakthroughs (Brennecke et al., 2023)(El Alaoui and Gaitonde, 2024) proved that for the SK model, the interior of the overlap concentration regime also satisfies $O(1)$ operator norm bound for the covariance matrix. These results strongly suggest the correctness our hypothesis about an $O(1)$ Lipschitz constant of our m_t , although a complete proof does not immediately follow, since (Brennecke et al., 2023)(El Alaoui and Gaitonde, 2024) relied on a TAP characterization of the free energy, which has been conjectured but not rigorously proved for the linear inverse problem in the general setting (Qiu and Sen, 2023).

C ADDITIONAL EXPERIMENTS

In this section we present some results of applying Algorithm 1 to sample from the posterior distribution in model (1). The prior distribution and the parameter settings are described in Section 4.

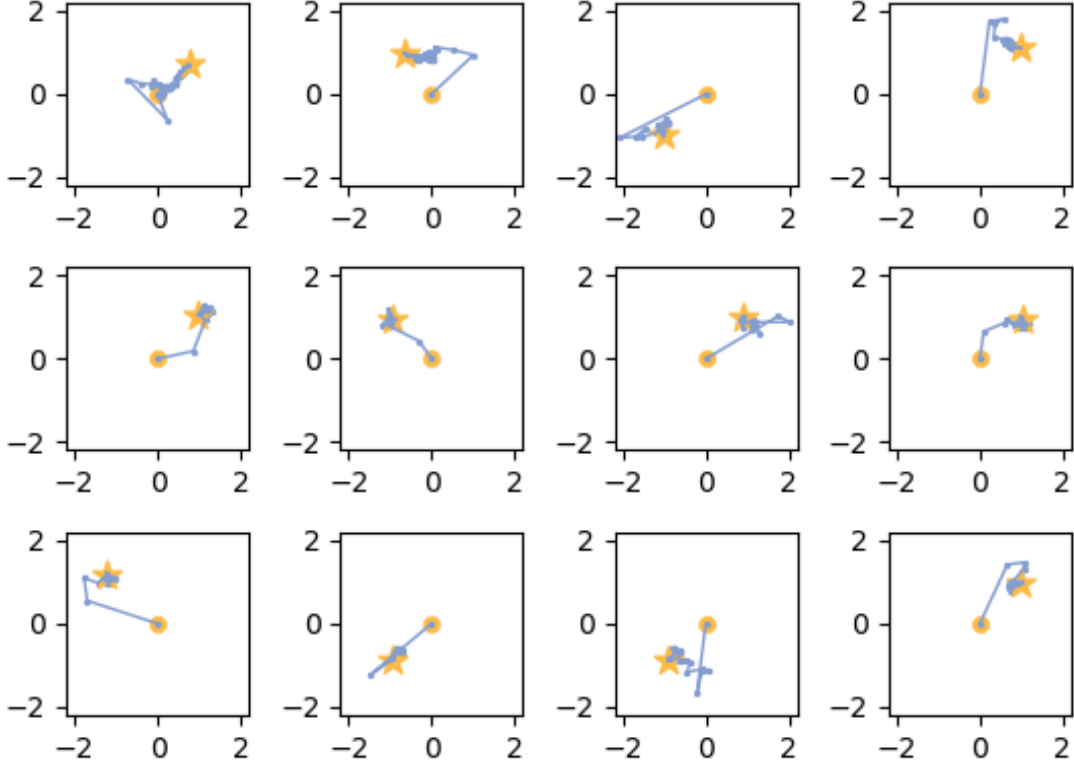


Figure 2: Trajectories of the first and second coordinates of the samples generated by Algorithm 1. In this experiment, the prior distribution is $P_0(\theta) = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, and we set $M = 1000, N = 1250, K = 20, T = 200, \delta = 0.1$. The three rows correspond to $\Delta=0.01, 1, 10$ respectively.

Additionally, in Figure 1, we calculate the Algorithm MSE and MMSE as follows:

$$\text{Algorithm MSE} = \frac{1}{2N} \|\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{alg}}\|_2^2, \quad (228)$$

$$\text{MMSE} = \frac{1}{N} \mathbb{E}[\|\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} | \phi\boldsymbol{\theta} + \sqrt{\Delta/\alpha}\mathbf{w}]\|_2^2]. \quad (229)$$

We use $1/(2N)$ rather than $1/N$ in (228), because if $\boldsymbol{\theta}^{\text{alg}}$ is a perfect posterior sample, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{\text{alg}}$ are i.i.d. conditioned on (\mathbf{X}, \mathbf{Y}) , and hence $\mathbb{E}[\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{alg}}\|_2^2] = 2\mathbb{E}[\|\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}]\|_2^2]$.

For the second experiment, we also compare the sampling time between our algorithm and DPS-Gaussian method. From Table 1, we observe that our algorithm is more efficient.

Table 1: Sampling Time

Method	N=192	N=768	N=1728
Our Algorithm	8s	53s	390s
DPS-Gaussian	184s	427s	882s