Deep Networks as Denoising Algorithms: Sample-Efficient Learning of Diffusion Models in High-Dimensional Graphical Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate the efficiency of deep neural networks for approximating scoring functions in diffusion-based generative modeling. While existing approximation theories leverage the smoothness of score functions, they suffer from the curse of dimensionality for intrinsically high-dimensional data. This limitation is pronounced in graphical models such as Markov random fields, where the approximation efficiency of score functions remains unestablished.

To address this, we note score functions can often be well-approximated in graphical models through variational inference denoising algorithms. Furthermore, these algorithms can be efficiently represented by neural networks. We demonstrate this through examples, including Ising models, conditional Ising models, restricted Boltzmann machines, and sparse encoding models. Combined with off-the-shelf discretization error bounds for diffusion-based sampling, we provide an efficient sample complexity bound for diffusion-based generative modeling when the score function is learned by deep neural networks.

1 Introduction

In recent years, diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020, Song and Ermon, 2019, Song et al., 2020] have emerged as a leading approach for generative modeling, achieving state-of-the-art results across diverse domains. Given a dataset of n independent and identically distributed samples $\{\boldsymbol{x}_i\}_{i=1}^n$ drawn from an unknown distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$, diffusion models aim to learn a generative model that produces new samples $\hat{\boldsymbol{x}} \sim \hat{\mu}$ that match this distribution. Popular diffusion models such as DDPM [Ho et al., 2020] achieve this through a two-step procedure:

- **Step 1.** Fit approximate score functions $\hat{\boldsymbol{s}}_t : \mathbb{R}^d \to \mathbb{R}^d$ for $t \in [0, T]$ by minimizing the following empirical risk over a neural network class $\mathcal{F}$:

$$\hat{\boldsymbol{s}}_t = \arg\min_{\mathrm{NN}\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left\| \sigma_t^{-1} \boldsymbol{g}_i + \mathrm{NN}(\lambda_t \boldsymbol{x}_i + \sigma_t \boldsymbol{g}_i) \right\|_2^2. \tag{ERM}$$

  In the above display, $\boldsymbol{g}_i \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $(\lambda_t, \sigma_t^2) = (e^{-t}, 1 - e^{-2t})$.

- **Step 2.** Discretize the following stochastic differential equation (SDE) from Gaussian initialization, whose drift term is given by the fitted approximate score functions:

$$\mathrm{d}\boldsymbol{Y}_t = \big(\boldsymbol{Y}_t + 2\hat{\boldsymbol{s}}_{T-t}(\boldsymbol{Y}_t)\big)\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t, \quad t \in [0, T], \quad \boldsymbol{Y}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \tag{SDE}$$

  and take the approximate sample $\hat{\boldsymbol{x}} = \boldsymbol{Y}_T \in \mathbb{R}^d$.

Submitted to NeurIPS 2023 Workshop on Diffusion Models. Do not distribute.

Score functions $\boldsymbol{s}_t(\boldsymbol{z})$ are central to the diffusion model framework. Given infinite data and model capacity, the minimizer of the empirical risk in Eq. (ERM) yields the score function,

$$\boldsymbol{s}_t(\boldsymbol{z}) = \nabla_{\boldsymbol{z}} \log \mu_t(\boldsymbol{z}), \quad \mu_t(\boldsymbol{z}) : \text{density of } \boldsymbol{z}, \boldsymbol{x} \sim \mu \text{ and } [\boldsymbol{z}|\boldsymbol{x}] \sim \mathcal{N}(\lambda_t \boldsymbol{x}, \sigma_t^2 \mathbf{I}_d). \tag{Score}$$

The sample quality from diffusion models relies on two key factors: (1) how well $\hat{\boldsymbol{s}}_t$ approximates $\boldsymbol{s}_t$; and (2) how accurately the SDE discretization scheme approximates process (SDE). Recent work has made substantial progress on controlling the SDE discretization error in diffusion models, assuming access to a good score function estimator [Chen et al., 2022a, 2023a, Lee et al., 2023, Li et al., 2023a, Benton et al., 2023]. However, understanding when neural networks can accurately estimate the score function itself remains less explored. Some analyses rely on strong distributional assumptions for score function realizability [Shah et al., 2023, Yuan et al., 2023], while others exploit the smoothness of score functions, incurring the curse of dimensionality [Oko et al., 2023, Chen et al., 2023b]. These results do not cover many common high-dimensional graphical models for images and text, such as Markov random fields or restricted Boltzmann machines [Geman and Graffigne, 1986, Ranzato et al., 2010, Conroy and O'leary, 2001].

A new perspective on score function approximation. We provide a new perspective on approximating diffusion model score functions with neural networks. First, we observe that by Tweedie's formula, score functions $\boldsymbol{s}_t$ are related to denoising functions $\boldsymbol{m}_t$:

$$\boldsymbol{s}_t(\boldsymbol{z}) = (\lambda_t \cdot \boldsymbol{m}_t(\boldsymbol{z}) - \boldsymbol{z})/\sigma_t^2, \qquad \boldsymbol{m}_t(\boldsymbol{z}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{g})\sim\mu\otimes\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\boldsymbol{x}|\lambda_t \boldsymbol{x} + \sigma_t \boldsymbol{g} = \boldsymbol{z}]. \tag{Denoiser}$$

Our key insight is that if the data distribution μ arises from a graphical model, these denoisers $\boldsymbol{m}_t(\boldsymbol{z})$ can often be approximated by variational inference (VI) algorithms, which takes the form

$$\boldsymbol{m}_t(\boldsymbol{z}) \approx \boldsymbol{f}_{\rm out}(\boldsymbol{u}^{(L)}), \quad \boldsymbol{u}^{(\ell)} = \boldsymbol{f}_\ell(\boldsymbol{u}^{(\ell-1)}), \quad \ell \in \{1, \ldots, L\}, \quad \boldsymbol{u}^{(0)} = \boldsymbol{f}_{\rm in}(\boldsymbol{z}). \tag{VI}$$

For instance, when μ is an Ising model, $\boldsymbol{m}_t$ can be approximated by an iterative algorithm that minimizes a VI objective [Jordan et al., 1999, Wainwright et al., 2008]. Each update step $\boldsymbol{f}_\ell$ is composed of simple operations, including matrix-vector multiplication and pointwise nonlinearity, which can be captured by a two-layer neural network $\boldsymbol{f}_\ell(\boldsymbol{u}) \approx \boldsymbol{u} + \boldsymbol{W}_1 \cdot {\rm ReLU}(\boldsymbol{W}_2 \boldsymbol{u})$. By comparing updates (VI) and residual network forms (ResNet), we can see how the iterative variational inference steps directly translate to residual block approximations. This establishes a clear connection between variational inference in graphical models and score approximation in diffusion models.

2 Preliminaries: the DDPM sampling scheme

Algorithm 1 The DDPM sampling scheme

Require: $\{\boldsymbol{x}_i\}_{i\in[n]}$, (d, D, L, M, B), $(N, T, \delta, \{t_k\}_{0\le k\le N})$ with $0 = t_0 < \cdots < t_N = T - \delta$.

1: // Computing the approximate score function
2: Sample $\{\boldsymbol{g}_i\}_{i\in[n]} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.
3: **for** $t \in \{T - t_k\}_{0\le k\le N-1}$ **do**
4: Solve the ERM problem below for $t = T - t_k$:

$$\widehat{\boldsymbol{W}}_t = \arg\min_{\boldsymbol{W}\in\mathcal{W}_{d,D,L,M,B}} \frac{1}{n}\sum_{i=1}^n \Big\| \sigma_t^{-1}\boldsymbol{g}_i + \mathsf{P}_t[{\rm ResN}_{\boldsymbol{W}}](\lambda_t \boldsymbol{x}_i + \sigma_t \boldsymbol{g}_i) \Big\|_2^2. \tag{1}$$

5: Take the approximate score function to be $\hat{\boldsymbol{s}}_t(\boldsymbol{z}) = \mathsf{P}_t[{\rm ResN}_{\widehat{\boldsymbol{W}}_t}](\boldsymbol{z})$.
6: // Sampling by discretizing the stochastic differential equation
7: Sample $\widehat{\boldsymbol{Y}}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.
8: **for** $k = 0, \cdots, N-1$ **do**
9: Sample $\boldsymbol{G}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Calculate $\widehat{\boldsymbol{Y}}_{k+1}$ using the exponential integrator scheme:

$$\widehat{\boldsymbol{Y}}_{k+1} = e^{\gamma_k} \cdot \widehat{\boldsymbol{Y}}_k + 2(e^{\gamma_k} - 1) \cdot \hat{\boldsymbol{s}}_{T-t_k}(\widehat{\boldsymbol{Y}}_k) + \sqrt{e^{2\gamma_k} - 1} \cdot \boldsymbol{G}_k, \quad \gamma_k = t_{k+1} - t_k. \tag{2}$$

Return: $\hat{\boldsymbol{x}} = \widehat{\boldsymbol{Y}}_N$.

This section provides details on the two-step DDPM sampling scheme in Algorithm 1. The inputs of the algorithm are n IID samples $\{\boldsymbol{x}_i\}_{i\in[n]}$ from μ. The algorithm also receives parameters

(d, D, L, M, B) for specifying the ResNet class, and $(N, T, \delta, \{t_k\}_{0\le k\le N})$ for specifying the time discretization scheme. The first step of the algorithm performs empirical risk minimization to compute the approximate score functions $\hat{\boldsymbol{s}}_t$ (lines 2-5). The second step generates a sample by discretizing the reverse-time SDE using the fitted score functions (lines 7-9). We discuss the score learning and SDE discretization steps in more detail below.

ERM and the ResNet class. The first step of Algorithm 1 solves an ERM problem (1) to fit the score functions. This regresses standard Gaussian noises $\{\boldsymbol{g}_i\}_{i\in[n]}$ on the noisy samples $\{\lambda_t \boldsymbol{x}_i + \sigma_t \boldsymbol{g}_i\}_{i\in[n]}$, using a standard ResNet architecture $\mathrm{ResN}_{\boldsymbol{W}} : \mathbb{R}^d \to \mathbb{R}^d$. The ResNet is parameterized by a set of weight matrices $\boldsymbol{W} = \{\boldsymbol{W}_1^{(\ell)} \in \mathbb{R}^{D\times M}, \boldsymbol{W}_2^{(\ell)} \in \mathbb{R}^{M\times D}\}_{\ell\in[L]} \cup \{\boldsymbol{W}_{\rm in} \in \mathbb{R}^{(d+1)\times D}, \boldsymbol{W}_{\rm out} \in \mathbb{R}^{D\times d}\}$ with embedding dimension D, number of layers L, and hidden-layer width M. It applies iterative residual blocks with ReLU nonlinearities to map an input $\boldsymbol{z}$ to an output in $\mathbb{R}^d$:

$$\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) = \boldsymbol{W}_{\rm out}\boldsymbol{u}^{(L)}, \quad \boldsymbol{u}^{(\ell)} = \boldsymbol{u}^{(\ell-1)} + \boldsymbol{W}_1^{(\ell)}\mathrm{ReLU}(\boldsymbol{W}_2^{(\ell)}\boldsymbol{u}^{(\ell-1)}), \quad \boldsymbol{u}^{(0)} = \boldsymbol{W}_{\rm in}[\boldsymbol{z}; 1]. \tag{ResNet}$$

The minimization in (1) is over the ResNets whose weights are contained in a B-bounded set, specified by parameters (d, D, L, M, B)

$$\mathcal{W}_{d,D,L,M,B} := \Big\{\boldsymbol{W} = \{\boldsymbol{W}_1^{(\ell)}, \boldsymbol{W}_2^{(\ell)}\}_{\ell\in[L]} \cup \{\boldsymbol{W}_{\rm in}, \boldsymbol{W}_{\rm out}\} : |||\boldsymbol{W}||| \le B\Big\}. \tag{3}$$

Here the norm of ResNet weights is defined as

$$|||\boldsymbol{W}||| := \max_{\ell\in[L]}\big\{\|\boldsymbol{W}_1^{(\ell)}\|_{\rm op} + \|\boldsymbol{W}_2^{(\ell)}\|_{\rm op}\big\} \vee \max\big\{\|\boldsymbol{W}_{\rm in}\|_{\rm op}, \|\boldsymbol{W}_{\rm out}\|_{\rm op}\big\}. \tag{4}$$

For technical reasons, we truncate the ResNet output using P_t. Given a function $f : \mathbb{R}^d \to \mathbb{R}^d$, we define $\mathsf{P}_t[f](\boldsymbol{z}) = \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(f(\boldsymbol{z}) + \sigma_t^{-2}\boldsymbol{z}) - \sigma_t^{-2}\boldsymbol{z}$, where $\mathrm{proj}_R(\boldsymbol{z})$ is the projector of $\boldsymbol{z} \in \mathbb{R}^d$ into the R-Euclidean ball. Note that when $f(\boldsymbol{z})$ is a score function, $f(\boldsymbol{z}) + \sigma_t^{-2}\boldsymbol{z}$ is a rescaled denoising function and should be bounded for data distribution with compact support. This operator is a technical detail that could be eliminated in practice — it is only used to control the generalization error of the empirical risk minimization problem.

Choice of the discretization scheme. We choose a particular scheme that uses a uniform grid in the first phase and an exponential decaying grid in the second phase. As shown in Benton et al. [2023], such a scheme provides a sharp sampling error control. We delay the detailed description of our discretization scheme to Appendix A.1.

The conditional diffusion model. In conditional generative modeling tasks, we observe IID samples $\{(\boldsymbol{x}_i, \boldsymbol{\theta}_i)\}_{i\in[n]} \sim_{iid} \mu$, and our goal is to learn a model to generate new samples $\hat{\boldsymbol{x}}$ from the conditional distribution $\mu(\boldsymbol{x}|\boldsymbol{\theta})$ for a given $\boldsymbol{\theta}$.

The DDPM sampling scheme can be simply adapted to solve conditional generative modeling tasks, as per Algorithm 2. Specifically, we modify the ResNet in empirical risk minimization to take the form (ResNet-Conditional), admitting inputs $(\lambda_t \boldsymbol{x}_i + \sigma_t \boldsymbol{g}_i, \boldsymbol{\theta}_i) \in \mathbb{R}^d \times \mathbb{R}^m$. The approximated score functions $\hat{\boldsymbol{s}}_t(\boldsymbol{z})$ become conditional $\hat{\boldsymbol{s}}_t(\boldsymbol{z}; \boldsymbol{\theta}) = \mathsf{P}_t[\mathrm{ResN}_{\widehat{\boldsymbol{W}}_t}](\boldsymbol{z}, \boldsymbol{\theta})$, estimating the conditional score functions $\boldsymbol{s}_t(\boldsymbol{z}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{z}} \log \mu_t(\boldsymbol{z}, \boldsymbol{\theta})$, where μ_t is the joint density of $(\boldsymbol{z}, \boldsymbol{\theta})$ when $(\boldsymbol{x}, \boldsymbol{\theta}) \sim \mu$ and $[\boldsymbol{z}|\boldsymbol{x}] \sim \mathcal{N}(\lambda_t \boldsymbol{x}, \sigma_t^2 \mathbf{I}_d)$. Details of the conditional algorithm are provided in Appendix D.1.

3 Diffusion models for Ising models

The Ising model $\mu \in \mathcal{P}(\{\pm 1\}^d)$ is a distribution over the discrete hypercube, with probability mass function characterized by an energy function of spin configurations. Specifically,

$$\mu(\boldsymbol{x}) = Z^{-1}\exp\{\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2\}, \quad \boldsymbol{x} \in \{\pm 1\}^d, \quad Z = \textstyle\sum_{\boldsymbol{x}\in\{\pm 1\}^d}\exp\{\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2\}. \tag{Ising}$$

The Ising model stands as one of the most fundamental graphical models; it belongs to the exponential family, yet its normalizing constant, Z, does not possess an analytic expression.

Consider the task of generative modeling where the input consists of IID samples $\{\boldsymbol{x}_i\}_{i\in[n]} \sim \mu$ derived from the Ising model. To demonstrate that Algorithm 1 outputs valid samples, we need to control the estimation error $\mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]$. Recall that $\boldsymbol{s}_t$ relates to $\boldsymbol{m}_t$. To calculate $\boldsymbol{m}_t(\boldsymbol{z})$, one often seeks to minimize certain type of free energy, for instance, the naive variational Bayes free energy [Wainwright et al., 2008]. To establish our main result, we will assume the consistency of a free energy minimizer with the denoiser.

Assumption 1. *Let $\boldsymbol{x} \sim \mu(\boldsymbol{\sigma}) \propto \exp\{\langle \boldsymbol{\sigma}, \boldsymbol{A}\boldsymbol{\sigma}\rangle/2\}$ and $\boldsymbol{z} \sim \mathcal{N}(\lambda_t \boldsymbol{x}, \sigma_t^2 \mathbf{I}_d)$. Denote the marginal distribution of $\boldsymbol{z}$ by μ_t. For any fixed t, assume that there exists $\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}) < \infty$ and $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{A}, t)$ with $\|\boldsymbol{K} - \boldsymbol{A}\|_{\mathrm{op}} \le A < 1$, such that*

$$\mathbb{E}_{\boldsymbol{z}\sim\mu_t}[\|\hat{\boldsymbol{m}}_t(\boldsymbol{z}) - \boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d \le \varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}),$$
$$\hat{\boldsymbol{m}}_t(\boldsymbol{z}) = \operatorname{argmin}_{\boldsymbol{m}\in[-1,1]^d}\Big\{\sum_{i=1}^d -\mathsf{h}_{\mathrm{bin}}(m_i) - \frac{1}{2}\langle \boldsymbol{m}, \boldsymbol{A}\boldsymbol{m}\rangle - \frac{\lambda_t}{\sigma_t^2}\langle \boldsymbol{z}, \boldsymbol{m}\rangle + \frac{1}{2}\langle \boldsymbol{m}, \boldsymbol{K}\boldsymbol{m}\rangle\Big\}.$$

In Appendix A.2, we will discuss cases in which the VI approximation error $\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A})$ can be well-controlled. Given Assumption 1 holds, we are ready to provide a bound on the estimation error of the approximate score function. We give a proof outline in Appendix A.4 and the full proof in Appendix E.

Theorem 1. *Let Assumption 1 hold. Let $\{\hat{\boldsymbol{s}}_{T-t_k}\}_{0\le k\le N-1}$ be the approximate score function given by Algorithm 1 in which we take*

$$D = 3d, \quad M \ge 4d, \quad B \ge 7\cdot(M/d)\cdot\log(M) + 1/\min_k\{T - t_k\} + \sqrt{d}.$$

Then with probability at least $1-\eta$, for any $t \in \{T - t_k\}_{0\le k\le N-1}$, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mu_t}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \lesssim \lambda_t^2\sigma_t^{-4}\cdot\Big(\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}) + \varepsilon_{\mathrm{ResN}}^2 + \varepsilon_{\mathrm{gen}}^2\Big), \tag{5}$$

where

$$\varepsilon_{\mathrm{ResN}}^2 = \frac{d^2}{M^2(1-A)^2} + A^{2L}, \quad \varepsilon_{\mathrm{gen}}^2 = \sqrt{\frac{(MdL + d^2)[T + L\log(BL)] + \log(N/\eta)}{n}}. \tag{6}$$

Combining Theorem 1 with off-the-shelf results on the DDPM discretization error [Benton et al., 2023], we obtain the following bound on the sampling error in terms of KL divergence:

Corollary 1. *Let Assumption 1 hold. Consider the two-phase discretization scheme as in Definition 1. Denote the distribution of the output of Algorithm 1 as $\hat{\mu}$. Then, with probability at least $1-\eta$, we have*

$$\mathrm{KL}(\mu_\delta, \hat{\mu})/d \lesssim \varepsilon_{\mathrm{score}}^2 + \varepsilon_{\mathrm{disc}}^2, \tag{7}$$

where

$$\varepsilon_{\mathrm{score}}^2 \le \delta^{-1}\cdot\Big(\sup_{0\le k\le N-1}\varepsilon_{\mathrm{VI},T-t_k}^2 + \varepsilon_{\mathrm{ResN}}^2 + \varepsilon_{\mathrm{gen}}^2\Big), \quad \varepsilon_{\mathrm{disc}}^2 \le \kappa^2 N + \kappa T + e^{-2T}. \tag{8}$$

Equation (7) provides control on the KL divergence between μ_δ and $\hat{\mu}$ normalized by dimension d. If the right-hand side is small, this guarantees the two distributions are close in an average per-coordinate sense: for two d-dimensional product distributions $\mu = \mathcal{N}(0,1)^{\otimes d}$ and $\nu = \mathcal{N}(\varepsilon,1)^{\otimes d}$ that are close per coordinate, their KL divergence scales as $\mathrm{KL}(\mu,\nu) \asymp d\cdot\varepsilon^2$, growing linearly with d. Furthermore, it is possible to derive bounds on the distance between the original distribution μ (instead of μ_δ) and the learned distribution $\hat{\mu}$ using other DDPM discretization analyses such as Chen et al. [2022a, 2023a], Li et al. [2023a]. We provide additional discussions of our results in Appendix A.3. We discuss other related works and future directions in Appendix C.

Generalization to other high-dimensional graphical models To demonstrate the flexibility of our proposed framework, we also generalize the results in this section to other high-dimensional graphical models in Appendix B. Specifically, we consider latent variable Ising models (Appendix B.1), the conditional Ising models for the conditional generative modeling task (Appendix B.2), and the sparse coding models (Appendix B.3).

References

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Jean Barbier, Mohamad Dia, Nicolas Macris, and Florent Krzakala. The mutual information in random linear estimation. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 625–632. IEEE, 2016.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Mark Borgerding, Philip Schniter, and Sundeep Rangan. Amp-inspired deep networks for sparse linear inverse problems. *IEEE Transactions on Signal Processing*, 65(16):4293–4308, 2017.

Philippe Carmona and Yueyun Hu. Universality in sherrington–kirkpatrick's spin glass model. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 42, pages 215–222. Elsevier, 2006.

Michael Celentano. Sudakov-fernique post-amp, and a new proof of the local convexity of the tap free energy. *arXiv preprint arXiv:2208.09550*, 2022.

Michael Celentano, Zhou Fan, and Song Mei. Local convexity of the tap free energy and amp convergence for z2-synchronization. *arXiv preprint arXiv:2106.11428*, 2021.

Michael Celentano, Zhou Fan, Licong Lin, and Song Mei. Mean-field variational inference with the tap free energy: Geometric and statistical properties in linear models. In *in preparison*, 2023+.

Sourav Chatterjee. Spin glasses and stein's method. *Probability theory and related fields*, 148(3-4): 567–600, 2010.

Sourav Chatterjee and Amir Dembo. Nonlinear large deviations. *Advances in Mathematics*, 299: 396–450, 2016.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*, 2023b.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022a.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023c.

Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023d.

Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Theoretical linear convergence of unfolded ista and its practical weights and thresholds. *Advances in Neural Information Processing Systems*, 31, 2018.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022b.

John M Conroy and Dianne P O'leary. Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407, 2001.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constructive Approximation*, 33(1):125–143, 2011.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

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

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. 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, 2022.

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

Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. *Geometric and Functional Analysis*, 28(6):1548–1596, 2018.

Ronen Eldan. Analysis of high-dimensional distributions using pathwise methods. *Proceedings of ICM, to appear*, 2022.

Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature ising models. *Probability theory and related fields*, 182(3-4):1035–1051, 2022.

Zhou Fan, Song Mei, and Andrea Montanari. Tap free energy, spin glasses and variational inference. 2021.

Zhou Fan, Yufan Li, and Subhabrata Sen. Tap equations for orthogonally invariant spin glasses at high temperature. *arXiv preprint arXiv:2202.09325*, 2022.

Oliver Y Feng, Ramji Venkataramanan, Cynthia Rush, Richard J Samworth, et al. A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15(4):335–536, 2022.

Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the international congress of mathematicians*, volume 1, page 2. Berkeley, CA, 1986.

Davide Ghio, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Sampling with flows, diffusion and autoregressive neural networks: A spin-glass perspective. *arXiv preprint arXiv:2308.14085*, 2023.

Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. In *International conference on machine learning*, pages 2221–2231. PMLR, 2019.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pages 399–406, 2010.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kurt Hornik. Some new results on neural network approximation. *Neural networks*, 6(8):1069–1072, 1993.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Vishesh Jain, Frederic Koehler, and Elchanan Mossel. The mean-field approximation: Information inequalities, algorithms, and complexity. In *Conference On Learning Theory*, pages 1326–1347. PMLR, 2018.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

Marc Lelarge and Leo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173:859–929, 2019.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023a.

Yufan Li, Zhou Fan, Subhabrata Sen, and Yihong Wu. Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 156–161. IEEE, 2023b.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022a.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022b.

Yao Ma, Xiaorui Liu, Tong Zhao, Yozen Liu, Jiliang Tang, and Neil Shah. A unified view on graph neural networks as graph signal denoising. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1202–1211, 2021.

Antoine Maillard, Laura Foini, Alejandro Lage Castellanos, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. High-temperature expansions and message passing algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(11):113301, 2019.

Tanya Marwah, Zachary Lipton, and Andrej Risteski. Parametric complexity bounds for approximating pdes with neural networks. *Advances in Neural Information Processing Systems*, 34: 15044–15055, 2021.

Tanya Marwah, Zachary Chase Lipton, Jianfeng Lu, and Andrej Risteski. Neural network approximations of pdes beyond linearity: A representational perspective. In *International Conference on Machine Learning*, pages 24139–24172. PMLR, 2023.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648, 2007.

Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

Sumit Mukherjee and Subhabrata Sen. Variational inference in high-dimensional linear regression. *The Journal of Machine Learning Research*, 23(1):13703–13758, 2022.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*, 2023.

Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017.

Judea Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 129–138. 1982.

Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

Timm Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, 15(6):1971, 1982.

Marc'Aurelio Ranzato, Alex Krizhevsky, and Geoffrey Hinton. Factored 3-way restricted boltzmann machines for modeling natural images. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 621–628. JMLR Workshop and Conference Proceedings, 2010.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *arXiv preprint arXiv:2307.01178*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Michel Talagrand. *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer Science & Business Media, 2003.

David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022.

E Weinan, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

Qiang Wu. Thouless-anderson-palmer equations for the multi-species sherrington-kirkpatrick model. *arXiv preprint arXiv:2308.09099*, 2023.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. *arXiv preprint arXiv:2306.01129*, 2023a.

Yaodong Yu, Tianzhe Chu, Shengbang Tong, Ziyang Wu, Druv Pai, Sam Buchanan, and Yi Ma. Emergence of segmentation with minimalistic white-box transformers. *arXiv preprint arXiv:2308.16271*, 2023b.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *arXiv preprint arXiv:2307.07055*, 2023.

Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1828–1837, 2018.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

Contents

A More details for Section 3

A.1 Discretization scheme

Definition 1 (Two-phase discretization scheme [Benton et al., 2023])**.** *The two-phase discretization scheme has parameters $(\kappa, N_0, N, T, \delta) \in (0,1) \times \mathbb{N} \times \mathbb{N} \times \mathbb{R} \times (0,1)$, where (κ, N_0, N) are free parameters and (T, δ) are fully determined by (κ, N_0, N). In the first uniform phase, the N_0 time steps have equal length κ. In the second exponential phase, the $N - N_0$ steps decay with rate $1/(1+\kappa) \in (0,1)$. The last time step t_N has a gap $\delta = (1+\kappa)^{N_0 - N} \in (0,1)$ to T.*

Specifically, we take $t_0 = 0$, $t_k = k\kappa$ for $k \le N_0$, $t_{N_0} = N_0\kappa = T - 1$, $t_{N_0+k} = T - (1+\kappa)^{-k}$ for $0 \le k \le N - N_0$, and $t_N = T - (1+\kappa)^{N_0 - N} = T - \delta$. Defining $\gamma_k = t_{k+1} - t_k$, we have $\gamma_k = \kappa$ for $k \le N_0 - 1$, and $\gamma_{N_0+k} = \kappa/(1+\kappa)^{k+1}$ for $0 \le k \le N - N_0 - 1$. See [Benton et al., 2023, Figure 1] for a pictorial illustration of this scheme.

A.2 Verifying the assumption in examples

This section provides examples that admit controlled VI approximation error $\varepsilon^2_{\mathrm{VI},t}$. The results in this section are proved in Appendix E.3.

Ising model in the VB consistency regime. There is a line of work studying the consistency of the naive mean-field variational Bayes (VB) free energy in Ising models under high-temperature conditions [Chatterjee and Dembo, 2016, Eldan, 2018, Jain et al., 2018, Mukherjee and Sen, 2022]. We build on this by providing a quantitative bound on the variational inference approximation error for a general coupling matrix $\boldsymbol{A}$ in this regime.

Lemma 1. *Assume $\|\boldsymbol{A}\|_{\mathrm{op}} < 1/2$. Then for any t, Assumption 1 is satisfied for $\boldsymbol{K} = \boldsymbol{0}$, and*

$$\varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}) \le \frac{4}{1 - 2\|\boldsymbol{A}\|_{\mathrm{op}}} \frac{\|\boldsymbol{A}\|_F^2}{d}. \tag{9}$$

As an example, for the ferromagnetic Ising model we have $\boldsymbol{A} = \beta \mathbf{1}\mathbf{1}^\mathsf{T}/d$, giving $\varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}) \le [4\beta^2/(1-2\beta)]/d$. This shows the VI approximation error vanishes as $\beta < 1/2$ and $d \to \infty$. However, this is not a particularly interesting regime for Ising models, since they can be well-approximated by a product distribution when β is small [Chatterjee and Dembo, 2016, Eldan, 2018].

The Sherrington-Kirkpatrick model. The Sherrington-Kirkpatrick model assumes $\boldsymbol{A} = \beta \boldsymbol{J}$, where $\boldsymbol{J} \sim \mathrm{GOE}(d)$ is a symmetric Gaussian random matrix with off-diagonal entries that are IID Gaussian with variance $1/d$. Prior work has shown that the VB free energy does not provide consistent estimation in this model [Ghorbani et al., 2019, Fan et al., 2021]. Instead, the variational objective that yields a consistent estimator of the Gibbs mean is the Thouless-Anderson-Palmer (TAP) free energy [Thouless et al., 1977, Fan et al., 2021, El Alaoui et al., 2022]. Using results on the TAP free energy, the variational inference (VI) approximation error can be controlled for this model when $\beta < 1/4$.

Lemma 2. *[Corollary of Lemma 4.10 of El Alaoui et al. [2022]] Assume $\boldsymbol{A} = \beta \boldsymbol{J}$ where $\boldsymbol{J} \sim \mathrm{GOE}(d)$ and $\beta < 1/4$. Then for any t, there exists matrices $\boldsymbol{K} = c_t \mathbf{I}_d$ for some c_t, such that with high probability, $\|\boldsymbol{A} - c_t \mathbf{I}_d\|_{\mathrm{op}} \le A < 1$ and*

$$\varepsilon_{\mathrm{VI},t}(\beta \boldsymbol{J}) \xrightarrow{p} 0, \quad \text{as } d \to \infty.$$

Lemma 2 provides a qualitative result on the consistency of variational inference (VI) for the Sherrington-Kirkpatrick model, but does not give a non-asymptotic error bound. To establish a non-asymptotic guarantee, one could potentially leverage tools like the smart path method [Talagrand, 2003, Theorem 2.4.20] or Stein's method [Chatterjee, 2010]. We conjecture it is possible to prove a quantitative error bound of order $C(\beta)/d$ using these techniques, as illustrated in [Talagrand, 2003, Theorem 2.4.20].

Other Ising models. We conjecture that Lemma 2 could extend to a variety of other models including non-Gaussian Wigner matrices [Carmona and Hu, 2006], heterogeneous variances [Wu, 2023], orthogonally invariant spin glasses [Fan et al., 2022], and spiked matrix models with non-Rademacher priors [Fan et al., 2021, Lelarge and Miolane, 2019]:

- *Non-Gaussian Wigner matrices.* We have $\boldsymbol{A} = \beta \boldsymbol{J}$ where $\boldsymbol{J}$ is a symmetric random matrix whose off-diagonal elements are independent with variance $1/d$ and satisfy some moment condition. This generalizes GOE matrices to non-Gaussian distributions. Since these matrices have similar properties to GOE matrices [Carmona and Hu, 2006], we conjecture Lemma 2 should hold.
- *Heterogeneous variance: multi-species Sherrington-Kirkpatrick models.* We have $\boldsymbol{A} = \beta \boldsymbol{J}$ where $\boldsymbol{J}$ is a random matrix with independent entries but heterogeneous variance. An example is the bipartite Sherrington-Kirkpatrick model specified by a set $S \subseteq [d]$, with $J_{ij} = J_{ji} \sim \mathcal{N}(0, 1/d)$ for $i \in S$ and $j \in S^c$, and $J_{ij} = 0$ for $i, j \in S$ or $i, j \in S^c$. The TAP equations verifying Assumption 1 has been shown to hold in similar models [Wu, 2023] in the high-temperature regime $\beta \le \beta_0$.
- *Orthogonally invariant spin glass models.* We have $\boldsymbol{A} = \beta \boldsymbol{J}$, where $\boldsymbol{J} = \boldsymbol{O}\boldsymbol{E}\boldsymbol{O}^\mathsf{T} \in \mathbb{R}^{d \times d}$. Here, $\boldsymbol{O} \sim \mathrm{Haar}(\mathrm{SO}(d))$ is a uniform random orthogonal matrix and $\boldsymbol{E} = \mathrm{diag}(e_1, \ldots, e_d) \in \mathbb{R}^{d \times d}$ is a diagonal matrix. The TAP equations have been shown for related models [Fan et al., 2022] in the high-temperature regime.
- *Spiked matrix models.* Suppose we observe $\boldsymbol{Y} = \boldsymbol{u}\boldsymbol{u}^\mathsf{T} + \boldsymbol{J}$ where $\boldsymbol{J} \sim \mathrm{GOE}(d)$ and $\boldsymbol{u} \in \mathbb{R}^d$ with $u_i \sim_{iid} \pi_0$ for some distribution $\pi_0 \in \mathcal{P}(\mathbb{R})$. The posterior distribution of $\boldsymbol{u}$ given observation $\boldsymbol{Y}$ is given by $\mu(\boldsymbol{x}) \propto \exp\{\langle \boldsymbol{x}, \boldsymbol{Y}\boldsymbol{x}\rangle/2 - \|\boldsymbol{x}\|_2^4/(4n)\}\pi_0^d(\boldsymbol{x})$. Taking this μ as the sample distribution, we conjecture that Assumption 1 can be verified for this model Fan et al. [2021], Lelarge and Miolane [2019].

A.3 Discussions

More explicit sample complexity bounds. Corollary 1 provides a sampling error bound in terms of the KL divergence of μ_δ and $\hat{\mu}$. To interpret this bound, assume $\hat{\mu}$ satisfies a dimension-free transportation-information inequality, i.e., $W_1^2(\mu_\delta, \hat{\mu}) \lesssim \mathrm{KL}(\mu_\delta, \hat{\mu})$. Further assume

$\sup_t \varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}) \lesssim 1/d$ (conjectured to hold for the SK model when $\beta < 1$). Since $W_1^2(\mu_\delta, \mu)/d \lesssim \delta$, this implies

$$W_1^2(\mu, \hat{\mu})/d \lesssim W_1^2(\mu, \mu_\delta)/d + \mathrm{KL}(\mu_\delta, \hat{\mu})/d \lesssim \delta + \varepsilon^2_{\mathrm{score}} + \varepsilon^2_{\mathrm{disc}}.$$

By the formulation of $\varepsilon^2_{\mathrm{score}}$ and $\varepsilon^2_{\mathrm{disc}}$ in Eq. (6) and (8) and by $\sup_t \varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}) \lesssim 1/d$, to ensure $W_1^2(\mu, \hat{\mu})/d \lesssim \varepsilon^2$, it suffices to take

$$\begin{aligned} &\delta \asymp \varepsilon^2, && T \asymp \log(1/\varepsilon), && \kappa \asymp \varepsilon^2/\log(1/\varepsilon), && N \asymp \log^2(1/\varepsilon)/\varepsilon^2, \\ &d \asymp 1/\varepsilon^4, && M \asymp 1/\varepsilon^6, && L \asymp \log(1/\varepsilon), && n \asymp \log^3(1/\varepsilon)/\varepsilon^{18}. \end{aligned}$$

The role of dimensionality. In contrast to existing results [Oko et al., 2023, Chen et al., 2023b] in which the score estimation error bounds exhibit a curse of dimensionality, our result seems to demonstrate a "blessing of dimensionality". Specifically, the term $\varepsilon^2_{\mathrm{VI,t}}$ in Theorem 1 is independent of ResNet size, sample size, and will typically vanish as dimension d goes to infinity. However, we cannot conclude that score estimation actually becomes easier for higher-dimensional Ising models, since our result only provides an upper bound on the estimation error. Whether score approximation truly simplifies with increasing dimensions is an open question deserving further investigation.

Generalizing Assumption 1. While Assumption 1 provides a sufficient condition for efficient score approximation, it is stronger than necessary. For example, in the Sherrington-Kirkpatrick model when $\boldsymbol{A} = \beta \boldsymbol{J}$ where $\boldsymbol{J} \sim \mathrm{GOE}(d)$, an efficient sampling algorithm is known when $\beta < 1$ [Celentano, 2022]. However, we can only verify Assumption 1 for $\beta \le \beta_0$ for some $1/4 < \beta_0 < 1/2$. Nevertheless, we believe one can weaken our assumption to show score estimation is efficient for any $\beta < 1$ by leveraging local convexity of the TAP free energy of the SK model, proved in El Alaoui et al. [2022], Celentano [2022].

The choice of sampling scheme and discretization scheme. Importantly, our score estimation error bound in Theorem 1 can combine with sampling schemes beyond DDPM, as it does not rely on a specific diffusion model. For instance, stochastic localization schemes [Eldan, 2013, El Alaoui et al., 2022, Montanari and Wu, 2023, Montanari, 2023] estimate the denoiser rather than the score, and our analysis can be adapted to bound the denoiser estimation error, enabling sampling guarantees for stochastic localization. Additionally, the discretization scheme and sampling error bound in Corollary 1 may not be optimal. The analysis could likely be sharpened, or the discretization improved, to provide tighter error guarantees.

A.4 Proof outline of Theorem 1

Here, we outline the proof of Theorem 1, with full details in Appendix E.

Recall that we have $\hat{\boldsymbol{s}}_t(\boldsymbol{z}) = \mathsf{P}_t[\mathrm{ResN}_{\widehat{\boldsymbol{W}}}](\boldsymbol{z})$, where $\widehat{\boldsymbol{W}} = \mathrm{argmin}_{\boldsymbol{W}\in\mathcal{W}} \hat{\mathbb{E}}[\|\mathsf{P}_t \mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]$ for $\mathcal{W} = \mathcal{W}_{d,D,L,M,B}$. Here, $\hat{\mathbb{E}}$ denotes averaging over the empirical data distribution. By standard error decomposition analysis in empirical risk minimization theory, we have:

$$\begin{aligned} \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\widehat{\boldsymbol{W}}}](\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]/d \le{}& \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]/d \\ &+ 2\sup_{\boldsymbol{W}\in\mathcal{W}} \Big|\hat{\mathbb{E}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d - \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d\Big|. \end{aligned}$$

Furthermore, a standard identity in diffusion model theory shows:

$$\mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d = \mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]/d + C, \quad C = \mathbb{E}[\|\boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d - \mathbb{E}[\|\sigma_t^{-1}\boldsymbol{g}\|_2^2]/d.$$

Combining the above yields:

$$\mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \le \bar{\varepsilon}^2_{\mathrm{app}} + \bar{\varepsilon}^2_{\mathrm{gen}},$$

where $\bar{\varepsilon}^2_{\mathrm{app}}$ is the approximation error and $\bar{\varepsilon}^2_{\mathrm{gen}}$ is the generalization error,

$$\begin{aligned} \bar{\varepsilon}^2_{\mathrm{app}} &= \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d, \\ \bar{\varepsilon}^2_{\mathrm{gen}} &= 2\sup_{\boldsymbol{W}\in\mathcal{W}} \Big|\hat{\mathbb{E}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d - \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d\Big|. \end{aligned}$$

The generalization error $\bar{\varepsilon}_{\rm gen}^2$ can be controlled by a standard empirical process analysis. We simply use a parameter counting argument to control this term, which can be found in Proposition 6. This gives rise to the term $\varepsilon_{\rm gen}^2$ in (6).

To control the approximation error $\bar{\varepsilon}_{\rm app}^2$, we note that $\boldsymbol{s}_t(\boldsymbol{z}) = (\lambda_t \cdot \boldsymbol{m}_t(\boldsymbol{z}) - \boldsymbol{z})/\sigma_t^2$, where $\boldsymbol{m}_t(\boldsymbol{z}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{g})\sim\mu\otimes\mathcal{N}(\boldsymbol{0},\mathbf{I}_d)}[\boldsymbol{x}|\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g} = \boldsymbol{z}]$ is the denoiser. Thus, approximating the score function reduces to approximating $\boldsymbol{m}_t(\boldsymbol{z})$ using a ResNet. By Assumption 1, the denoiser $\boldsymbol{m}_t$ can be approximated by the minimizer of a variational free energy $\mathcal{F}_t^{\rm VI}$. This minimizer can be found by a fixed point iteration, which can further be approximated by a ResNet.

More specifically, simple calculus shows that the minimizer $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}_t$ of the variational free energy $\mathcal{F}_t^{\rm VI}$ satisfies the fixed point equation

$$\hat{\boldsymbol{m}} = \tanh(\boldsymbol{U}\hat{\boldsymbol{m}} + \boldsymbol{h}), \quad \boldsymbol{U} = \boldsymbol{A} - \boldsymbol{K}, \quad \boldsymbol{h} = \lambda_t\sigma_t^{-2}\boldsymbol{z}.$$

When $\|\boldsymbol{U}\|_{\rm op} < 1$, this can be efficiently solved by fixed point iteration

$$\hat{\boldsymbol{m}} \approx \boldsymbol{m}^L, \qquad \boldsymbol{m}^{\ell+1} = \tanh(\boldsymbol{U}\boldsymbol{m}^\ell + \boldsymbol{h}), \qquad \boldsymbol{m}^0 = \boldsymbol{0}.$$

This fixed point iteration can further be approximated by the ResNet structure (ResNet), where $\tanh$ is approximated by a linear combination of ReLU activations. Lemma 5 and 6 analyze this approximation error $\varepsilon_{\rm ResN}^2$. Our analysis shows that the total approximation error $\bar{\varepsilon}_{\rm app}^2$ is controlled by $\varepsilon_{\rm VI}^2 + \varepsilon_{\rm ResN}^2$. Adding the generalization error yields the overall score estimation error bound in Eq. (5).

B Generalization to other high-dimensional graphical models

B.1 Diffusion models for latent variable Ising models

In the latent variable Ising model μ, we have a coupling matrix $\boldsymbol{A} = [\boldsymbol{A}_{11}, \boldsymbol{A}_{12}; \boldsymbol{A}_{12}^\mathsf{T}, \boldsymbol{A}_{22}] \in \mathbb{R}^{(d+m)\times(d+m)}$ (where $\boldsymbol{A}_{11} \in \mathbb{R}^{d\times d}$, $\boldsymbol{A}_{12} \in \mathbb{R}^{d\times m}$, and $\boldsymbol{A}_{22} \in \mathbb{R}^{m\times m}$), specifying a joint distribution over $(\boldsymbol{x}, \boldsymbol{\theta}) \in \{\pm 1\}^{d+m}$,

$$\mu(\boldsymbol{x}, \boldsymbol{\theta}) \propto \exp\{\langle \boldsymbol{x}, \boldsymbol{A}_{11}\boldsymbol{x}\rangle/2 + \langle \boldsymbol{x}, \boldsymbol{A}_{12}\boldsymbol{\theta}\rangle + \langle \boldsymbol{\theta}, \boldsymbol{A}_{22}\boldsymbol{\theta}\rangle/2\}, \quad \boldsymbol{x} \in \{\pm 1\}^d, \boldsymbol{\theta} \in \{\pm 1\}^m. \tag{10}$$

Note that the joint distribution over $(\boldsymbol{x}, \boldsymbol{\theta})$ is still an Ising model. However, here we will treat $\boldsymbol{\theta}$ as a latent variable and consider generative modeling for the marginal distribution $\mu(\boldsymbol{x}) = \sum_{\boldsymbol{\theta}} \mu(\boldsymbol{x}, \boldsymbol{\theta})$ when $\boldsymbol{\theta}$ is unobserved. When $\boldsymbol{A}_{11} = \boldsymbol{0}$ and $\boldsymbol{A}_{22} = \boldsymbol{0}$, this model reduces to a restricted Boltzmann machine, which is often used to model natural image distributions [Ranzato et al., 2010].

We still consider the generative modeling task where we observe $\{\boldsymbol{x}_i\}_{i\in[n]} \sim_{iid} \mu$, and our goal is to sample a new $\hat{\boldsymbol{x}} \sim \hat{\mu}$ with $\hat{\mu} \approx \mu$. To show the DDPM scheme (Algorithm 1) provides a controlled error bound, we need to bound the score estimation error [Benton et al., 2023]. This estimation error can be controlled if we assume the denoiser minimizes a VI objective.

Assumption 2 (Consistency of the free energy minimizer in marginal Ising models)**.** *Let $\boldsymbol{\sigma} = (\boldsymbol{x}, \boldsymbol{\theta}) \sim \mu(\boldsymbol{x}, \boldsymbol{\theta}) \propto \exp\{\langle \boldsymbol{\sigma}, \boldsymbol{A}\boldsymbol{\sigma}\rangle/2\}$ and $\boldsymbol{z} \sim \mathcal{N}(\lambda_t\boldsymbol{x}, \sigma_t^2\mathbf{I}_d)$. For any fixed t, assume that there exists $\varepsilon_{{\rm VI},t}^2(\boldsymbol{A}) < \infty$ and $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{A}, t) \in \mathbb{R}^{(d+m)\times(d+m)}$ with $\|\boldsymbol{K} - \boldsymbol{A}\|_{\rm op} \le A < 1$, such that*

$$\mathbb{E}_{\boldsymbol{z}\sim\mu_t}[\|\hat{\boldsymbol{m}}_t(\boldsymbol{z}) - \boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d \le \varepsilon_{{\rm VI},t}^2(\boldsymbol{A}), \qquad \hat{\boldsymbol{m}}_t(\boldsymbol{z}) = [\hat{\boldsymbol{\omega}}_t(\boldsymbol{z})]_{1:d},$$
$$\hat{\boldsymbol{\omega}}_t(\boldsymbol{z}) = \mathrm{argmin}_{\boldsymbol{\omega}\in[-1,1]^{d+m}}\Big\{\sum_{i=1}^{d+m} -\mathsf{h}_{\rm bin}(\omega_i) - \frac{1}{2}\langle \boldsymbol{\omega}, \boldsymbol{A}\boldsymbol{\omega}\rangle - \frac{\lambda_t}{\sigma_t^2}\langle \boldsymbol{z}, \boldsymbol{\omega}_{1:d}\rangle + \frac{1}{2}\langle \boldsymbol{\omega}, \boldsymbol{K}\boldsymbol{\omega}\rangle\Big\}.$$

Assumption 2 can be verified in concrete examples. Lemma 1 still applies in this model: when $\|\boldsymbol{A}\|_{\rm op} < 1/2$, taking $\boldsymbol{K} = \boldsymbol{0}$ gives $\varepsilon_{{\rm VI},t}^2(\boldsymbol{A}) \le 4d^{-1}(1 - 2\|\boldsymbol{A}\|_{\rm op})^{-1}\|\boldsymbol{A}\|_F^2$. We conjecture that for $\boldsymbol{A}$ being spin glass models like the Sherrington-Kirkpatrick model at high temperature, there exists $\boldsymbol{K}$ such that $\mathbb{E}[\varepsilon_{{\rm VI},t}^2(\boldsymbol{A})] \to 0$ as $d, m \to \infty$. Given Assumption 2, the following theorem provides a score estimation error bound and a sampling error bound in latent variable Ising models, proved in Appendix F.1.

Theorem 2. *Let Assumption 2 hold. Let $\{\hat{\boldsymbol{s}}_{T-t_k}\}_{0\le k\le N-1}$ be the approximate score function given by Algorithm 1 in which we take*

$$D = 3(d+m), \quad M \ge 4(d+m), \quad B \ge 7\cdot(M/(d+m))\cdot\log(M) + \sqrt{d+m} + 1/\min_k\{T-t_k\}.$$

Then with probability at least $1-\eta$, for any $t \in \{T-t_k\}_{0\le k\le N-1}$, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mu_t}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \lesssim \lambda_t^2\sigma_t^{-4}\cdot\Big(\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}) + \varepsilon_{\mathrm{ResN}}^2 + \varepsilon_{\mathrm{gen}}^2\Big), \tag{11}$$

where $\varepsilon_{\mathrm{VI},t}^2$ is given in Assumption 2, and

$$\begin{aligned} \varepsilon_{\mathrm{ResN}}^2 &= \frac{d+m}{d}\Big(\frac{(d+m)^2}{M^2(1-A)^2} + A^{2L}\Big), \\ \varepsilon_{\mathrm{gen}}^2 &= \sqrt{\frac{(ML+d)(d+m)[T+L\log(BL)] + \log(N/\eta)}{n}}. \end{aligned} \tag{12}$$

Furthermore, consider the two-phase discretization scheme as in Definition 1, we have with probability $1-\eta$ that

$$\mathrm{KL}(\mu_\delta, \hat{\mu})/d \lesssim \delta^{-1}\cdot\Big(\sup_{0\le k\le N-1}\varepsilon_{\mathrm{VI},T-t_k}^2 + \varepsilon_{\mathrm{ResN}}^2 + \varepsilon_{\mathrm{gen}}^2\Big) + \kappa^2 N + \kappa T + e^{-2T}. \tag{13}$$

B.2 Conditional diffusion models for Ising models

In the conditional Ising model, we also have a coupling matrix $\boldsymbol{A} = [\boldsymbol{A}_{11}, \boldsymbol{A}_{12}; \boldsymbol{A}_{12}^\mathsf{T}, \boldsymbol{A}_{22}] \in \mathbb{R}^{(d+m)\times(d+m)}$, specifying a joint distribution over $(\boldsymbol{x}, \boldsymbol{\theta}) \in \{\pm1\}^{d+m}$ as in Eq. (10). However, we now consider the conditional generative modeling task where we observe $\{(\boldsymbol{x}_i, \boldsymbol{\theta}_i)\}_{i\in[n]} \sim_{iid} \mu$. The goal is to sample $\hat{\boldsymbol{x}} \sim \hat{\mu}(\cdot|\boldsymbol{\theta}) \approx \mu(\cdot|\boldsymbol{\theta})$ for a given $\boldsymbol{\theta}$. Such problems naturally arise in image imputation tasks, where $(\boldsymbol{x}, \boldsymbol{\theta})$ represents a full image, $\boldsymbol{\theta}$ is the observed part, and $\boldsymbol{x}$ is the missing part to impute.

The conditional generative modeling task can be solved using the conditional DDPM scheme (Algorithm 2 as described in Appendix D.1). To bound the error, we need to control the estimation error of the conditional score $\boldsymbol{s}_t(\boldsymbol{z}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{z}}\log\mu_t(\boldsymbol{z}, \boldsymbol{\theta})$. By Tweedie's formula, we have $\boldsymbol{s}_t(\boldsymbol{z}; \boldsymbol{\theta}) = (\lambda_t\boldsymbol{m}_t(\boldsymbol{z}; \boldsymbol{\theta}) - \boldsymbol{z})/\sigma_t^2$, where $\boldsymbol{m}_t(\boldsymbol{z}; \boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{g})\sim\mu\otimes\mathcal{N}(0,1)}[\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{z} = \lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}]$ is the conditional denoiser. We assume the following about $\boldsymbol{m}_t(\boldsymbol{z}; \boldsymbol{\theta})$.

Assumption 3 (Consistency of the free energy minimizer in conditional Ising models)**.** *Let $(\boldsymbol{x}, \boldsymbol{\theta}) \sim \mu(\boldsymbol{x}, \boldsymbol{\theta}) \propto \exp\{\langle\boldsymbol{\sigma}, \boldsymbol{A}\boldsymbol{\sigma}\rangle/2\}$ and $\boldsymbol{z} \sim \mathcal{N}(\lambda_t\boldsymbol{x}, \sigma_t^2\mathbf{I}_d)$. For any fixed t, assume that there exists $\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}) < \infty$ and $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{A}, t) \in \mathbb{R}^{d\times d}$ with $\|\boldsymbol{K} - \boldsymbol{A}_{11}\|_{\mathrm{op}} \le A < 1$, such that*

$$\mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{z})}[\|\hat{\boldsymbol{m}}_t(\boldsymbol{z}; \boldsymbol{\theta}) - \boldsymbol{m}_t(\boldsymbol{z}; \boldsymbol{\theta})\|_2^2]/d \le \varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}),$$

$$\hat{\boldsymbol{m}}_t(\boldsymbol{z}; \boldsymbol{\theta}) = \mathrm{argmin}_{\boldsymbol{m}\in[-1,1]^d}\Big\{\sum_{i=1}^d -\mathsf{h}_{\mathrm{bin}}(m_i) - \frac{1}{2}\langle\boldsymbol{m}, \boldsymbol{A}_{11}\boldsymbol{m}\rangle - \langle\boldsymbol{m}, \boldsymbol{A}_{12}\boldsymbol{\theta}\rangle - \frac{\lambda_t}{\sigma_t^2}\langle\boldsymbol{z}, \boldsymbol{m}\rangle + \frac{1}{2}\langle\boldsymbol{m}, \boldsymbol{K}\boldsymbol{m}\rangle\Big\}.$$

Assumption 3 can be verified in concrete examples. Lemma 1 still applies in this model: when $\|\boldsymbol{A}_{11}\|_{\mathrm{op}} < 1/2$, taking $\boldsymbol{K} = \boldsymbol{0}$ gives $\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}) \le 4d^{-1}(1-2\|\boldsymbol{A}_{11}\|_{\mathrm{op}})^{-1}\|\boldsymbol{A}_{11}\|_F^2$. We conjecture that $\mathbb{E}[\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A})] \to 0$ as $d, m \to \infty$ for $\boldsymbol{A}$ being spin glass models at high temperature. Given Assumption 3, the following theorem provides a conditional score estimation error bound and a conditional sampling error bound in conditional Ising models, proved in Appendix F.2.

Theorem 3. *Let Assumption 3 hold. Let $\{\hat{\boldsymbol{s}}_{T-t_k}\}_{0\le k\le N-1}$ be the approximate score function given by Algorithm 2 in which we take*

$$D = 4d, \quad M \ge 4d, \quad B \ge 7\cdot(M/d)\cdot\log(M) + \sqrt{d} + 1/\min_k\{T-t_k\} + \|\boldsymbol{A}_{12}\|_{\mathrm{op}}\cdot(M/d+1).$$

Then with probability at least $1-\eta$, for any $t \in \{T-t_k\}_{0\le k\le N-1}$, we have

$$\mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{z})}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}; \boldsymbol{\theta}) - \boldsymbol{s}_t(\boldsymbol{z}; \boldsymbol{\theta})\|_2^2]/d \lesssim \lambda_t^2\sigma_t^{-4}\cdot\Big(\varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}) + \varepsilon_{\mathrm{ResN}}^2 + \varepsilon_{\mathrm{gen}}^2\Big),$$

where $\varepsilon^2_{\mathrm{VI},t}$ is given in Assumption 3, and

$$\begin{aligned}\varepsilon^2_{\mathrm{ResN}} &= \frac{d^2}{M^2(1-A)^2} + A^{2L},\\ \varepsilon^2_{\mathrm{gen}} &= \sqrt{\frac{(MdL + d(d+m))[T + L\log(BLd^{-1}(m+d))] + \log(N/\eta)}{n}}.\end{aligned} \tag{14}$$

Furthermore, consider the two-phase discretization scheme as in Definition 1, we have with probability $1-\eta$ that

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mu}[\mathrm{KL}(\mu_\delta(\cdot|\boldsymbol{\theta}), \hat{\mu}(\cdot|\boldsymbol{\theta}))/d] \lesssim \delta^{-1}\cdot\Big(\sup_{0\le k\le N-1}\varepsilon^2_{\mathrm{VI},T-t_k} + \varepsilon^2_{\mathrm{ResN}} + \varepsilon^2_{\mathrm{gen}}\Big) + \kappa^2 N + \kappa T + e^{-2T}.$$

We note the score estimation and sampling error bounds in Theorem 3 are averaged over $\boldsymbol{\theta}\sim\mu(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in\{\pm1\}^d}\mu(\boldsymbol{x},\boldsymbol{\theta})$, the marginal of $\boldsymbol{\theta}$. These do not ensure error bounds for any fixed $\boldsymbol{\theta}$.

B.3 Diffusion models for sparse coding

In sparse coding, there is a fixed dictionary $\boldsymbol{A}\in\mathbb{R}^{d\times m}$. Our observations are noisy, sparse linear combinations of the columns of the dictionary: $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i$ for $i\in[n]$. Here $\boldsymbol{\varepsilon}_i \sim_{iid} \mathcal{N}(\boldsymbol{0}, \tau^2\mathbf{I}_d)$ are noise vectors, and $\boldsymbol{\theta}_i \sim_{iid} \pi_0^{\otimes m}$ are sparse coefficient vectors, with $\pi_0\in\mathcal{P}(\mathbb{R})$ having a Dirac delta mass at 0. Given observations $\{\boldsymbol{x}_i\}_{i\in[n]}$, sparse coding typically aims to recover $\boldsymbol{A}$ and estimate $\{\boldsymbol{\theta}_i\}_{i\in[n]}$. Instead, we consider the generative modeling problem — learning a model to generate new samples $\hat{\boldsymbol{x}}$ resembling the observations $\{\boldsymbol{x}_i\}_{i\in[n]}$.

The generative modeling task for sparse coding can be solved by the DDPM sampling scheme (Algorithm 1). To control the score estimation error, we make the following assumption on the following denoising function $\boldsymbol{e}_t$, which requires a little modification in the sparse coding setting:

$$\boldsymbol{e}_t(\boldsymbol{z}_*) := \mathbb{E}_{(\boldsymbol{z}_*,\boldsymbol{\theta})}\left[\boldsymbol{\theta}\mid\boldsymbol{z}_*\right], \qquad \boldsymbol{z}_* = \boldsymbol{A}\boldsymbol{\theta} + \bar{\boldsymbol{\varepsilon}}, \qquad \bar{\varepsilon}_j \sim_{iid} \mathcal{N}(0, \tau^2 + \sigma_t^2/\lambda_t^2). \tag{15}$$

Assumption 4 (Consistency of the free energy minimizer in sparse coding)**.** *Fix $\boldsymbol{A}\in\mathbb{R}^{d\times m}$. Consider the Bayesian linear model $\boldsymbol{z}_* = \boldsymbol{A}\boldsymbol{\theta} + \bar{\boldsymbol{\varepsilon}}\in\mathbb{R}^d$, $\bar{\varepsilon}_j\sim_{iid}\mathcal{N}(0,\bar{\tau}_t^2)$ where $\bar{\tau}_t^2 = \tau^2 + \sigma_t^2/\lambda_t^2$ and $\theta_i \sim_{iid} \pi_0$ where $\pi_0\in\mathcal{P}([-\Pi,\Pi])$. Assume that for any $t>0$, there exist $(\nu_t, \boldsymbol{K}_t, \varepsilon^2_{\mathrm{VI},t})$ that depend on $(\pi_0, \boldsymbol{A}, \tau, t)$ with $\|\boldsymbol{A}^\mathsf{T}\boldsymbol{A}/\bar{\tau}_t^2 - \boldsymbol{K}_t\|_{\mathrm{op}} \le A < 1/\Pi^2$, such that*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}\sim\mu_t}[\|\hat{\boldsymbol{e}}_t(\boldsymbol{z}_*) - \boldsymbol{e}_t(\boldsymbol{z}_*)\|_2^2]/m \le \varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}),\\ &\hat{\boldsymbol{e}}_t(\boldsymbol{z}_*) = \mathrm{argmin}_{\boldsymbol{e}\in[-\Pi,\Pi]^m}\Big\{\sum_{i=1}^m \max_\lambda\Big[\lambda m_i - \log\mathbb{E}_{\beta\sim\pi_0}[e^{\lambda\beta-\beta^2\nu_t/2}]\Big] + \frac{1}{2\bar{\tau}_t^2}\|\boldsymbol{z}_* - \boldsymbol{A}\boldsymbol{e}\|_2^2 - \frac{1}{2}\langle\boldsymbol{e}, \boldsymbol{K}_t\boldsymbol{e}\rangle\Big\}.\end{aligned}$$

We also use a different truncation operator in Algorithm 1, replacing P_t by $\bar{\mathsf{P}}_t$:

$$\bar{\mathsf{P}}_t[f](\boldsymbol{z}) = \mathrm{proj}_{\sqrt{m}\,\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(f(\boldsymbol{z}) + (\sigma_t^2+\tau^2\lambda_t^2)^{-1}\boldsymbol{z}) - (\sigma_t^2+\tau^2\lambda_t^2)^{-1}\boldsymbol{z}.$$

Given Assumption 4, the following theorem provides a score estimation error bound in sparse coding models, proved in Appendix F.3.

Theorem 4. *Let Assumption 4 hold. Let $\{\hat{\boldsymbol{s}}_{T-t_k}\}_{0\le k\le N-1}$ be the approximate score function given by Algorithm 1 in which we take*

$$\begin{aligned}&D = 3m + d, \quad M \ge 4m,\\ &B \ge (M/m)\cdot\big(A + 1 + 2\Pi^2 + w_\star\big) + 2\Pi + 6 + (\|\boldsymbol{A}\|_{\mathrm{op}} + 1)/\min_k\{T - t_k\} + \tau^{-2}\|\boldsymbol{A}\|_{\mathrm{op}} + \sqrt{m},\end{aligned}$$

where $w_\star$ is defined in Eq. (65). Then with probability at least $1-\eta$, when $n\ge\log(2/\eta)$, for any $t\in\{T-t_k\}_{0\le k\le N-1}$, we have the following score estimation error bound

$$\begin{aligned}&\mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{z})}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z};\boldsymbol{\theta}) - \boldsymbol{s}_t(\boldsymbol{z};\boldsymbol{\theta})\|_2^2]/d\\ &\lesssim \lambda_t^2\|\boldsymbol{A}\|_{\mathrm{op}}^2(1+\tau^{-4})\cdot\frac{m}{d}\cdot\Big(\varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}) + \varepsilon^2_{\mathrm{ResN}}\Big) + \Big(\lambda_t^2\|\boldsymbol{A}\|_{\mathrm{op}}^2(1+\tau^{-4})\Pi^2\cdot\frac{m}{d} + \frac{\lambda_t^2}{\sigma_t^2}(1+\tau^2)\Big)\varepsilon^2_{\mathrm{gen}},\end{aligned}$$

for $\varepsilon^2_{\mathrm{VI},t}$ as given in Assumption 4, and

$$\varepsilon^2_{\mathrm{ResN}} = \Pi^2 \cdot (\Pi^2 A)^{2L} + \frac{m^2\Pi^2}{(1-\Pi^2 A)^2 M^2}, \quad \varepsilon^2_{\mathrm{gen}} = \sqrt{\frac{(dD + LDM)\cdot(T+L)\cdot\iota}{n}}. \tag{16}$$

where $\iota = \log(LBnmT(1+\tau)(1+\|\boldsymbol{A}\|_{\mathrm{op}}\Pi)\tau^{-1}N\eta^{-1})$.

Theorem 4 can be further combined with an off-the-shelf discretization bound as in Theorem 5 to derive a sampling error bound.

Verifying the assumption. The VI approximation error $\varepsilon^2_{\mathrm{VI}}$ in Assumption 4 converges to 0 as $d, m \to \infty$ when $\boldsymbol{A}$ is a rotationally invariant design matrix, by choosing the variational objective to be the TAP free energy [Thouless et al., 1977]. Specifically, assume the SVD decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{O}^\mathsf{T}$ where $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{O} \in \mathbb{R}^{m\times m}$ are orthonormal, and $\boldsymbol{D} \in \mathbb{R}^{d\times m}$ is diagonal. Assume that $\boldsymbol{O} \sim \mathrm{Haar}(\mathrm{SO}(m))$ is independent of everything else, and the diagonal elements of $\boldsymbol{D}$ have certain empirical distribution converging to a bounded distribution D. As an example, $\boldsymbol{A}$ with IID Gaussian entries of variance $1/m$ is rotationally invariant. Under the assumption that $\boldsymbol{A}$ is rotationally invariant, a corollary of [Li et al., 2023b, Theorem 1.11] gives the following lemma, with proof contained in Appendix F.4.

Lemma 3 (Corollary of Li et al. [2023b] Theorem 1.11)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{d\times m}$ be a rotationally invariant design matrix and let Assumption 7 hold. Then for any π_0, $\alpha = d/m$, and limiting distribution D, there exists $\tau^2 > 0$, such that for any t, there exists matrices $\boldsymbol{K} = c_t \mathbf{I}_d$ for some c_t, such that*

$$\varepsilon_{\mathrm{VI},t}(\boldsymbol{A}) \overset{a.s.}{\to} 0, \quad d, m \to \infty, \quad d/m \to \alpha.$$

Although Lemma 3 does not provide non-asymptotic control of the VI approximation error, we believe this could be obtained through more refined analysis.

C Other related work and future directions

Score function approximation in diffusion models. Neural network-based score function approximation has been recently studied in Oko et al. [2023], Chen et al. [2023b], Yuan et al. [2023], Shah et al. [2023]. Oko et al. [2023] assumes that the data distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ has a density with s-order bounded derivatives and shows that estimating the score to precision ε requires network size and sample complexity at least $\varepsilon^{-d/s}$. This suffers from the curse of dimensionality unless the data distribution is very smooth ($s \asymp d$). Oko et al. [2023], Chen et al. [2023b] avoid the curse of dimensionality by assuming that the data distribution has a low-dimensional structure, but this assumption does not apply to high-dimensional graphical models. Shah et al. [2023] considers Gaussian mixture models where the score function has a closed form, enabling parameterized by a small shallow network.

In contrast, we assume the data distribution is a graphical model, common for images and text [Blei et al., 2003, Mnih and Hinton, 2007, Geman and Graffigne, 1986]. Assuming the efficiency of variational inference approximation, we show that the score can be well-approximated by a network polynomial in dimension, enabling efficient learning from polynomial samples. Our graphical model assumption and algorithm unrolling of variational inference perspective circumvent dimensionality issues faced by prior work.

Discretizing the diffusion process. Recent work has studied the convergence rates of the discretized reverse SDEs/ODEs for diffusion models [Liu et al., 2022b, Li et al., 2023a, Lee et al., 2023, Chen et al., 2022b, 2023d, 2022a, 2023c,a, Benton et al., 2023]. In particular, Chen et al. [2023a], Benton et al. [2023] provide minimal assumptions to quantitatively control the KL divergence between the perturbed and data distributions. These assumptions include the second moment bound and the controlled score estimation error. Our work focuses on controlling the score estimation error, a goal that is orthogonal to analyzing discretization schemes. Specifically, we directly leverage the result of Benton et al. [2023] to provide an end-to-end error bound.

Stochastic localization. Stochastic localization, proposed by Eldan [2013, 2022], is another sampling scheme similar to diffusion models. Recent works have developed algorithmic sampling techniques based on stochastic localization [El Alaoui et al., 2022, Montanari and Wu, 2023, Celentano, 2022]. Montanari [2023] shows the equivalence of stochastic localization to the DDPM sampling scheme in the Gaussian setting and proposes various ways of generalizing stochastic localization schemes. While we present our results in the diffusion model framework, our methods can also provide sampling error bound for stochastic localization schemes.

Neural network approximation theory. Classical neural network approximation theory typically relies on assumptions that the target function is smooth or hierarchically smooth [Cybenko, 1989, Hornik et al., 1989, Hornik, 1993, Pinkus, 1999, DeVore et al., 2011, Weinan et al., 2019, Yarotsky, 2017, Barron, 1993, Bach, 2017, DeVore et al., 2021]. These enable overcoming the curse of dimensionality for higher-order smooth or low-dimensional target functions [Barron, 1993, Weinan et al., 2019, Bach, 2017]. However, when applying them to score function approximation in diffusion models, it is unclear whether such assumptions hold for the score function of high-dimensional graphical models.

A recent line of work investigated the expressiveness of neural networks through an algorithm approximation viewpoint [Wei et al., 2022, Bai et al., 2023, Giannou et al., 2023, Liu et al., 2022a, Marwah et al., 2021, 2023]. Wei et al. [2022], Bai et al. [2023], Giannou et al. [2023], Liu et al. [2022a] show that transformers can efficiently approximate several algorithm classes, such as gradient descent and Turing machines. Marwah et al. [2021, 2023] demonstrate that deep networks can efficiently approximate PDE solutions by approximating the gradient dynamics. We also adopt this algorithmic perspective for neural network approximation but apply it to score function approximation for diffusion models.

Variational inference in graphical models. Variational inference is commonly used to approximate the marginal statistics of graphical models [Pearl, 1982, Jordan et al., 1999, Minka, 2013, Mezard and Montanari, 2009, Wainwright et al., 2008, Blei et al., 2017]. In certain regimes, such as graphical models in the high temperature, naive variational Bayes has been shown to yield consistent posterior estimates [Chatterjee and Dembo, 2016, Eldan, 2018, Jain et al., 2018, Mukherjee and Sen, 2022]. For high dimensional statistical models in the low signal-to-noise ratio regime, approximate message passing [Donoho et al., 2009, Feng et al., 2022] and equivalently TAP variational inference [Thouless et al., 1977, Ghorbani et al., 2019, Fan et al., 2021, Celentano et al., 2021, Celentano, 2022, Celentano et al., 2023+], can achieve consistent estimation of the Bayes posterior. Our paper directly adopts results developed for variational inference methods in spin glass models and Bayesian linear models [Talagrand, 2003, Chatterjee, 2010, Barbier et al., 2019, 2016, Fan et al., 2021, 2022, Li et al., 2023b, Celentano et al., 2021, Celentano, 2022, Celentano et al., 2023+].

Algorithm unrolling. A line of work has focused on neural network denoising by unrolling iterative denoising algorithms into deep networks [Gregor and LeCun, 2010, Zheng et al., 2015, Zhang and Ghanem, 2018, Papyan et al., 2017, Ma et al., 2021, Chen et al., 2018, Borgerding et al., 2017, Monga et al., 2021, Yu et al., 2023a,b]. These approaches include unrolling ISTA for LASSO into recurrent nets [Gregor and LeCun, 2010, Zhang and Ghanem, 2018, Papyan et al., 2017, Borgerding et al., 2017], unrolling belief propagation for Markov random fields into recurrent nets [Zheng et al., 2015], and unrolling graph denoising algorithms into graph neural nets [Ma et al., 2021]. Our work also adopts this algorithm unrolling viewpoint, but with a different goal: while the prior literature has mainly focused on devising better denoising algorithms, our work uses this perspective to provide neural network approximation theories for diffusion-based generative models.

Algorithmic hard phase. The algorithm unrolling perspective can also shed light on the failure mode of score approximation, namely when score functions cannot be efficiently represented by neural networks. For example, we can conclude that the score function of the Sherrinton-Kirkpatrick model with $\beta > 1$ cannot be efficiently represented by a neural network, as it was proven in El Alaoui et al. [2022] that there is no stable algorithm to sample the SK Gibbs measure for $\beta > 1$. More generally, the relationship between hardness of sampling, hardness of diffusion-based sampling, and hardness of score approximation deserves further investigation. Recent work such as Ghio et al. [2023] provides a valuable discussion on this important topic.

Future directions. Our work leaves open several interesting questions. One issue is that for fixed dimension d, our score approximation error does not decay as the network size and sample size increase, and is lower bounded by the variational inference approximation error $\varepsilon_{\mathrm{VI}}^2$. To resolve this, one approach could consider a hierarchy of variational inference algorithms, such as Plefka's expansion [Plefka, 1982, Maillard et al., 2019], which provide increasingly accurate approximations. Using these hierarchical approximations within our framework could potentially reduce the score approximation error.

Another open question is understanding the algorithms that diffusion neural networks like U-nets and transformers implement in diffusion models for image tasks. One hypothesis is that U-nets with convolution layers are implementing some form of variational inference denoising on graphical models with certain locality and invariance structures. It would be interesting to test this hypothesis on real image datasets.

Finally, an exciting direction is leveraging the algorithmic unrolling perspective to design improved neural network architectures for diffusion models. The resulting architectures could potentially be more interpretable and achieve better emergent capabilities, as illustrated by recent works like Yu et al. [2023a,b].

D Technical preliminaries

D.1 DDPM conditional sampling scheme

We provide the details of the DDPM conditional sampling scheme (Algorithm 2) as mentioned in Section 2. The algorithm still has two steps, with minor modifications from unconditional DDPM (Algorithm 1). In the first step, empirical risk minimization (Eq. (18)) fits manually-generated noises $\{\boldsymbol{g}_i\}_{i\in[n]}$ using the noisy samples and conditioning variables $\{(\lambda_t \boldsymbol{x}_i + \sigma_t \boldsymbol{g}_i; \boldsymbol{\theta}_i)\}_{i\in[n]}$. The ResNet $\mathrm{ResN}_{\boldsymbol{W}} : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is parameterized by $\boldsymbol{W} = \{\boldsymbol{W}_1^{(\ell)} \in \mathbb{R}^{D\times M}, \boldsymbol{W}_2^{(\ell)} \in \mathbb{R}^{M\times D}\}_{\ell\in[L]} \cup \{\boldsymbol{W}_{\mathrm{in}} \in \mathbb{R}^{(d+m+1)\times D}, \boldsymbol{W}_{\mathrm{out}} \in \mathbb{R}^{D\times d}\}$ and is defined iteratively as

$$\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}, \boldsymbol{\theta}) = \boldsymbol{W}_{\mathrm{out}} \boldsymbol{u}^{(L)}, \quad \boldsymbol{u}^{(\ell)} = \boldsymbol{u}^{(\ell-1)} + \boldsymbol{W}_1^{(\ell)} \mathrm{ReLU}(\boldsymbol{W}_2^{(\ell)} \boldsymbol{u}^{(\ell-1)}), \quad \boldsymbol{u}^{(0)} = \boldsymbol{W}_{\mathrm{in}}[\boldsymbol{z}; \boldsymbol{\theta}; 1]. \tag{ResNet-Conditional}$$

The only difference between (ResNet) and (ResNet-Conditional) is the input dimension. Minimization is over the ResNets with weights in the set (for parameters d, m, D, L, M, B):

$$\begin{aligned} \mathcal{W}_{d,m,D,L,M,B} &:= \Big\{ \boldsymbol{W} = \{\boldsymbol{W}_1^{(\ell)}, \boldsymbol{W}_2^{(\ell)}\}_{\ell\in[L]} \cup \{\boldsymbol{W}_{\mathrm{in}}, \boldsymbol{W}_{\mathrm{out}}\} : \|\!|\boldsymbol{W}|\!\| \le B \Big\}, \\ \|\!|\boldsymbol{W}|\!\| &:= \max_{\ell\in[L]} \big\{ \|\boldsymbol{W}_1^{(\ell)}\|_{\mathrm{op}} + \|\boldsymbol{W}_2^{(\ell)}\|_{\mathrm{op}} \big\} \vee \max\big\{ \|\boldsymbol{W}_{\mathrm{in}}\|_{\mathrm{op}}, \|\boldsymbol{W}_{\mathrm{out}}\|_{\mathrm{op}} \big\}. \end{aligned} \tag{17}$$

We still truncate the ResNet output using P_t: for $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$, we define $\mathsf{P}_t[f](\boldsymbol{z}, \boldsymbol{\theta}) = \mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(f(\boldsymbol{z}, \boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z}) - \sigma_t^{-2}\boldsymbol{z}$, where proj_R projects $\boldsymbol{z} \in \mathbb{R}^d$ into the R-Euclidean ball.

The second step of Algorithm 2 still discretizes the backward SDE through the exponential integrator scheme (19) and the two-phase discretization scheme (Definition 1). However, we replace the score function $\hat{\boldsymbol{s}}_t(\widehat{\boldsymbol{Y}}_k)$ with the conditional score function $\hat{\boldsymbol{s}}_t(\widehat{\boldsymbol{Y}}_k; \boldsymbol{\theta}) = \mathsf{P}_t[\mathrm{ResN}_{\widehat{\boldsymbol{W}}_t}](\widehat{\boldsymbol{Y}}_k, \boldsymbol{\theta})$.

D.2 Sampling error bound of the DDPM scheme

In this section, we state a result from Benton et al. [2023], which establishes the convergence of the DDPM discretization scheme, when evaluated using Kullback-Leibler (KL) divergence, with only minimal assumptions required. A slight generalization of the result in Benton et al. [2023] is necessary, generalizing the identity covariance assumption to a general covariance matrix. The proof requires little modification, but we present a proof sketch here for completeness.

Suppose we are interested in drawing samples from μ in $\mathbb{R}^d$. The forward process that evolves according to the Ornstein-Uhlenbeck (OU) process is defined as the following SDE:

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \qquad X_0 \sim \mu, \qquad 0 \le t \le T. \tag{20}$$

In the above display, $(B_t)_{0\le t\le T}$ is a standard Brownian motion in $\mathbb{R}^d$. We denote by μ_t the distribution of X_t. One can check that $X_t \overset{d}{=} e^{-t}X_0 + \sqrt{1-e^{-2t}}\boldsymbol{g}$ for $\boldsymbol{g} \sim \mathcal{N}(, \mathbf{I}_d)$ that is

Algorithm 2 The DDPM conditional sampling scheme

Require: Samples $\{(\boldsymbol{x}_i, \boldsymbol{\theta}_i)\}_{i\in[n]} \subseteq \mathbb{R}^d \times \mathbb{R}^m$. Conditional latent variable $\boldsymbol{\theta}$. ResNet parameters (d, m, D, L, M, B). Discretization scheme parameters $(N, T, \delta, \{t_k\}_{0\le k\le N})$ with $0 = t_0 < \cdots < t_N = T - \delta$. Denote $\gamma_k = t_{k+1} - t_k$.

// Computing the approximate conditional score function

Sample $\{\boldsymbol{g}_i\}_{i\in[n]} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

for $t \in \{T - t_k\}_{0\le k\le N}$ **do**

  Solve the ERM problem below for $t = T - t_k$:

$$\widehat{\boldsymbol{W}}_t = \arg\min_{\boldsymbol{W}\in\mathcal{W}_{d,m,D,L,M,B}} \frac{1}{n}\sum_{i=1}^{n} \left\| \sigma_t^{-1}\boldsymbol{g}_i + \mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\lambda_t \boldsymbol{x}_i + \sigma_t \boldsymbol{g}_i, \boldsymbol{\theta}_i)\right\|_2^2. \tag{18}$$

  Take the approximate score function to be $\hat{\boldsymbol{s}}_t(\boldsymbol{z}; \boldsymbol{\theta}) = \mathsf{P}_t[\mathrm{ResN}_{\widehat{\boldsymbol{W}}_t}](\boldsymbol{z}, \boldsymbol{\theta})$.

// Sampling by discretizing the stochastic differential equation

Sample $\widehat{\boldsymbol{Y}}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

for $k = 0, \cdots, N-1$ **do**

  Sample $\boldsymbol{G}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Calculate $\widehat{\boldsymbol{Y}}_{k+1}$ using the exponential integrator scheme: (here $\boldsymbol{\theta}$ is provided as an input)

$$\widehat{\boldsymbol{Y}}_{k+1} = e^{\gamma_k} \cdot \widehat{\boldsymbol{Y}}_k + 2(e^{\gamma_k} - 1) \cdot \hat{\boldsymbol{s}}_{T-t_k}(\widehat{\boldsymbol{Y}}_k; \boldsymbol{\theta}) + \sqrt{e^{2\gamma_k} - 1} \cdot \boldsymbol{G}_k. \tag{19}$$

Return: $\hat{\boldsymbol{x}} = \widehat{\boldsymbol{Y}}_N$.

independent of X_0. The reverse process that corresponds to process (20) is defined via the SDE

$$\mathrm{d}Y_t = \{Y_t + 2\nabla \mu_{T-t}(Y_t)\}\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t', \qquad Y_0 \sim \mu_T. \tag{21}$$

An approximation to continuous-time process (21) is obtained via performing time discretization, which directly leads to a sampling algorithm. More precisely, for $0 = t_0 < t_1 < \cdots < t_N = T - \delta$, we let

$$\mathrm{d}\hat{Y}_t = \{\hat{Y}_t + 2\hat{s}_{T-t_k}(\hat{Y}_{t_k})\}\mathrm{d}t + \mathrm{d}\hat{B}_t \quad \text{for } t_k \le t \le t_{k+1}, \qquad \hat{Y}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \tag{22}$$

where $\hat{s}_{T-t}(\cdot)$ is an estimate of the true score function $s_{T-t}(\cdot) = \nabla \log \mu_{T-t}(\cdot)$. We denote by p_t the marginal distribution of $\hat{Y}_t$, and set $\gamma_k = t_{k+1} - t_k$. In addition, we assume there exists $\kappa > 0$, such that $\gamma_k \le \kappa \cdot \min\{1, T - t_{k+1}\}$.

Next, we state the assumptions required to establish the discretization error bound of the DDPM sampling scheme.

Assumption 5 (Rescaled version of Benton et al. [2023] Assumption 1)**.** *The score function estimator $\hat{s}_t$ satisfies*

$$\sum_{k=0}^{N-1} \gamma_k \mathbb{E}_{\boldsymbol{x}\sim\mu_{T-t_k}} \left[\|\nabla \log \mu_{T-t_k}(\boldsymbol{x}) - \hat{s}_{T-t_k}(\boldsymbol{x})\|_2^2\right] \le d \cdot \varepsilon_{\text{score}}^2.$$

Assumption 6. *The data distribution μ has finite second moment: $\mathbb{E}_{\boldsymbol{x}_0\sim\mu}[\|\boldsymbol{x}_0\|_2^2] \le d \cdot B$, where $B \ge 1$ is a fixed constant.*

With Assumptions 5 and 6, we are ready to state the main theorem for this part.

Theorem 5. *[Theorem 1 of Benton et al. [2023]] Let Assumptions 5 and 6 hold. Then there exists a numerical constant $C_0 > 0$, such that*

$$\mathsf{KL}(\mu_\delta, p_{t_N}) \le C_0 \cdot d \cdot \left(\varepsilon_{\text{score}}^2 + \kappa^2 NB + \kappa TB + e^{-2T}B\right).$$

Proof sketch of Theorem 5.

Part 1. We first control the quantity

$$E_{s,t} = \mathbb{E}\left[\|\nabla \log \mu_{T-t}(Y_t) - \nabla \log \mu_{T-s}(Y_s)\|_2^2\right],$$

where $0 \le s \le t \le T$. According to Lemma 2 of Benton et al. [2023], we have

$$\begin{aligned}
&\mathrm{d}\left(\|\nabla \log \mu_{T-t}(Y_t) - \nabla \log \mu_{T-s}(Y_s)\|_2^2\right) = -2\|\nabla \log \mu_{T-t}(Y_t) - \nabla \log \mu_{T-s}(Y_s)\|_2^2 \mathrm{d}t \\
&- 2\left\{\nabla \log \mu_{T-t}(Y_t) - \nabla \log \mu_{T-s}(Y_s)\right\} \cdot \nabla \log \mu_{T-s}(Y_s)\mathrm{d}t + 2\|\nabla^2 \log \mu_{T-t}(Y_t)\|_F^2 \mathrm{d}t \\
&+ 2\sqrt{2}\left\{\nabla \log \mu_{T-t}(Y_t) - \nabla \log \mu_{T-s}(Y_s)\right\} \cdot \nabla^2 \log \mu_{T-t}(Y_t) \cdot \mathrm{d}B'_t.
\end{aligned} \tag{23}$$

In the above display, s is fixed and t varies. Taking expectation and integrate over $[s, t]$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla \log \mu_{T-t}(Y_t) - \nabla \log \mu_{T-s}(Y_s)\|^2\right] = \mathbb{E}\int_s^t -2\|\nabla \log \mu_{T-r}(Y_r) - \nabla \log \mu_{T-s}(Y_s)\|_2^2 \mathrm{d}r \\
&- \mathbb{E}\int_s^t 2\left\{\nabla \log \mu_{T-r}(Y_r) - \nabla \log \mu_{T-s}(Y_s)\right\} \cdot \nabla \log \mu_{T-s}(Y_s)\mathrm{d}r \\
&+ \mathbb{E}\int_s^t 2\|\nabla^2 \log \mu_{T-r}(Y_r)\|_F^2 \mathrm{d}r.
\end{aligned}$$

Observe that all terms above are integrable. Hence, we may apply Fubini's theorem and interchange integration and expectation, which gives

$$\begin{aligned}
&\frac{\mathrm{d}E_{s,t}}{\mathrm{d}t} = -2\mathbb{E}\left[\|\nabla \log \mu_{T-t}(Y_t) - \nabla \log \mu_{T-s}(Y_s)\|_2^2\right] \\
&+ 2\mathbb{E}\left[\left\{\nabla \log \mu_{T-s}(Y_s) - \nabla \log \mu_{T-t}(Y_t)\right\} \cdot \nabla \log \mu_{T-s}(Y_s)\right] + 2\mathbb{E}\left[\|\nabla^2 \log \mu_{T-t}(Y_t)\|_F^2\right].
\end{aligned}$$

Invoking Cauchy-Schwartz inequality, we have

$$\frac{\mathrm{d}E_{s,t}}{\mathrm{d}t} \le \mathbb{E}\left[\|\nabla \log \mu_{T-s}(Y_s)\|_2^2\right] + 2\mathbb{E}\left[\|\nabla^2 \log \mu_{T-t}(Y_t)\|_F^2\right]. \tag{24}$$

Next, we upper bound $\mathbb{E}\left[\|\nabla \log \mu_{T-s}(Y_s)\|_2^2\right]$ and $\mathbb{E}\left[\|\nabla^2 \log \mu_{T-t}(Y_t)\|_F^2\right]$, respectively.

Lemma 3 of Benton et al. [2023] gives

$$\begin{aligned}
\nabla \log \mu_t(\boldsymbol{x}_t) &= -\sigma_t^{-2}\boldsymbol{x}_t + e^{-t}\sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{x}_t), \\
\nabla^2 \log \mu_t(\boldsymbol{x}_t) &= -\sigma_t^{-2}\mathbf{I} + e^{-2t}\sigma_t^{-4}\boldsymbol{\Sigma}_t(\boldsymbol{x}_t),
\end{aligned} \tag{25}$$

where $\sigma_t^2 = 1 - e^{-2t}$, $\boldsymbol{m}_t(\boldsymbol{x}_t) = \mathbb{E}_{\mu_0|\mu_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}[\boldsymbol{x}_0]$, and $\boldsymbol{\Sigma}_t(\boldsymbol{x}_t) = \mathrm{Cov}_{\mu_0|\mu_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}[\boldsymbol{x}_0]$. By Eq. (25), we see that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}\left[\|\nabla \log \mu_t(\boldsymbol{x}_t)\|_2^2\right] \\
=&\sigma_t^{-4}\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}\left[\|\boldsymbol{x}_t\|_2^2\right] - 2e^{-t}\sigma_t^{-4}\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}\left[\boldsymbol{x}_t \cdot \boldsymbol{m}_t(\boldsymbol{x}_t)\right] + e^{-2t}\sigma_t^{-4}\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}\left[\|\boldsymbol{m}_t(\boldsymbol{x}_t)\|_2^2\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}[\boldsymbol{x}_t \cdot \boldsymbol{m}_t(\boldsymbol{x}_t)] = \mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}[\boldsymbol{x}_t \cdot \boldsymbol{x}_0] = e^{-t}\mathbb{E}_{\boldsymbol{x}_0\sim\mu_0}\left[\|\boldsymbol{x}_0\|^2\right] \le dBe^{-t}, \\
&\mathrm{Tr}(\boldsymbol{\Sigma}_t(\boldsymbol{x}_t)) = \mathbb{E}[\|\boldsymbol{x}_0\|^2 \mid \boldsymbol{x}_t] - \|\boldsymbol{m}_t(\boldsymbol{x}_t)\|_2^2,
\end{aligned}$$

hence

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}\left[\|\nabla \log \mu_t(\boldsymbol{x}_t)\|^2\right] \\
&= \sigma_t^{-4} \cdot \left(e^{-2t}\mathbb{E}[\|\boldsymbol{x}_0\|^2] + \sigma_t^2 d\right) - 2e^{-2t}\sigma_t^{-4}\mathbb{E}[\|\boldsymbol{x}_0\|^2] + e^{-2t}\sigma_t^{-4} \cdot \left(\mathbb{E}[\|\boldsymbol{x}_0\|^2] - \mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}_t(\boldsymbol{x}_t))]\right) \\
&= \sigma_t^{-2}d - e^{-2t}\sigma_t^{-4}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}_t(\boldsymbol{x}_t))] \le d\sigma_t^{-2}.
\end{aligned} \tag{26}$$

That is to say, we have $\mathbb{E}[\|\nabla \log \mu_{T-s}(Y_s)\|_2^2] \le d\sigma_{T-s}^{-2}$. We write $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_t(\boldsymbol{x}_t)$ for short. The second part of Eq. (25) implies that

$$\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}\left[\|\nabla^2 \log \mu_t(\boldsymbol{x}_t)\|_F^2\right] = \sigma_t^{-4}d - 2\sigma_t^{-6}e^{-2t}\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Sigma}_t)\right] + e^{-4t}\sigma_t^{-8}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}_t^2)]. \tag{27}$$

Lemma 1 of Benton et al. [2023] gives

$$\frac{e^{2t}\sigma_t^4}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\boldsymbol{\Sigma}_t\right] = \mathbb{E}[\boldsymbol{\Sigma}_t^2]. \tag{28}$$

Putting together Eq. (27) and (28), we obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_t\sim\mu_t}\left[\|\nabla^2\log\mu_t(\boldsymbol{x}_t)\|_F^2\right] &= d\sigma_t^{-4}-2\sigma_t^{-6}e^{-2t}\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Sigma}_t)\right]+\frac{e^{-2t}\sigma_t^{-4}}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\mathrm{Tr}[\boldsymbol{\Sigma}_t]] \\ &\le d\sigma_t^{-4}+\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\sigma_t^{-4}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}_t)]\right).\end{aligned} \tag{29}$$

Putting together Eq. (26) and (29), we get

$$\begin{aligned}&\mathbb{E}\left[\|\nabla\log\mu_{T-s}(Y_s)\|_2^2\right]+2\mathbb{E}\left[\|\nabla^2\log\mu_{T-t}(Y_t)\|_F^2\right]\\ &\le\sigma_{T-s}^{-2}d+2d\sigma_{T-t}^{-4}-\frac{\mathrm{d}}{\mathrm{d}r}\left(\sigma_{T-r}^{-4}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}_{T-r})]\right)\big|_{r=t}.\end{aligned}$$

We define

$$E_{s,t}^{(1)}:=d\sigma_{T-s}^{-2}+2d\sigma_{T-t}^{-4},\qquad E_{s,t}^{(2)}:=-\frac{\mathrm{d}}{\mathrm{d}r}\left(\sigma_{T-r}^{-4}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}_{T-r})]\right)\big|_{r=t}.$$

According to Eq. (24) and notice that $E_{t_k,t_k}=0$, we have

$$E_{t_k,t}\le\int_{t_k}^{t}\left\{\mathbb{E}\left[\|\nabla\log\mu_{T-t_k}(Y_{t_k})\|_2^2\right]+2\mathbb{E}\left[\|\nabla^2\log\mu_{T-s}(Y_s)\|_F^2\right]\right\}\mathrm{d}s\le\int_{t_k}^{t}\left(E_{t_k,s}^{(1)}+E_{t_k,s}^{(2)}\right)\mathrm{d}s.$$

Following exactly the same procedure as in Benton et al. [2023], we conclude that there exists a positive numerical constant C_0, such that

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\|\nabla\log\mu_{T-t}(Y_t)-\nabla\log\mu_{T-t_k}(Y_{t_k})\|^2\right]\le C_0(\kappa^2dNB+\kappa dTB).$$

Part 2. We denote by Q the distribution of Y_{t_N} derived from process (21), and P^{μ_T} the distribution of process (22) at time t_N initialized at μ_T. By Proposition 3 of Benton et al. [2023], we obtain

$$\mathsf{KL}(Q\,||\,P^{\mu_T})\le\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\|\nabla\log\mu_{T-t}(Y_t)-\hat{s}_{T-t_k}(Y_{t_k})\|_2^2\right]\mathrm{d}t,$$

which by triangle inequality is no smaller than

$$\begin{aligned}&2\sum_{k=0}^{N-1}\gamma_k\mathbb{E}\left[\|\nabla\log\mu_{T-t_k}(Y_{t_k})-\hat{s}_{T-t_k}(Y_{t_k})\|_2^2\right]\mathrm{d}t\\ &+2\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\|\nabla\log\mu_{T-t_k}(Y_{t_k})-\nabla\log\mu_{T-t}(Y_t)\|_2^2\right]\\ &\le\,2\,d\cdot\varepsilon_{\text{score}}^2+2C_0(\kappa^2dNB+\kappa dTB).\end{aligned}$$

We denote by P the distribution of process (22) at time t_N initialized at $\mathcal{N}(\mathbf{0},\mathbf{I}_d)$. By Eq. (19) of Benton et al. [2023], we have

$$\mathsf{KL}(Q\,||\,P)=\mathsf{KL}(Q\,||\,P^{\mu_T})+\mathsf{KL}(\mu_T\,||\,\mathcal{N}(\mathbf{0},\mathbf{I}_d)).$$

Proposition 4 of Benton et al. [2023] gives $\mathsf{KL}(\mu_T\,||\,\mathcal{N}(\mathbf{0},\mathbf{I}_d))\lesssim dBe^{-2T}$. Putting together the above upper bounds, we arrive at the following conclusion:

$$\mathsf{KL}(\mu_\delta||\,p_{t_N})\le C_0\cdot d\cdot\left(Be^{-2T}+\kappa^2NB+\kappa TB+\varepsilon_{\text{score}}^2\right),$$

thus concluding the proof of Theorem 5. □

D.3 Generalization error of empirical risk minimization over ResNets

D.3.1 Result for Ising models

Note that the conditional (and unconditional) DDPM methods estimate the score function $\hat{\boldsymbol{s}}_t=\mathsf{P}_t\mathrm{ResN}_{\widehat{\boldsymbol{W}}_t}$ by solving the following ERM problem:

$$\begin{aligned}\widehat{\boldsymbol{W}}_t&=\mathrm{argmin}_{\boldsymbol{W}\in\mathcal{W}_{d,m,D,L,M,B}}\widehat{R}_n(\boldsymbol{W}),\\ \widehat{R}_n(\boldsymbol{W})&=\frac{1}{nd}\sum_{i=1}^{n}\left\|\sigma_t^{-1}\boldsymbol{g}_i+\mathsf{P}_t(\mathrm{ResN}_{\boldsymbol{W}}(\lambda_t\boldsymbol{x}_i+\sigma_t\boldsymbol{g}_i,\boldsymbol{\theta}_i))\right\|_2^2.\end{aligned} \tag{30}$$

Here, $\boldsymbol{x}_i, \boldsymbol{g}_i \in \mathbb{R}^d$, and $\boldsymbol{\theta}_i \in \mathbb{R}^m$ follow $\{(\boldsymbol{x}_i, \boldsymbol{\theta}_i, \boldsymbol{z}_i)\}_{i\in[n]} \sim_{iid} \mu \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Recall that the truncation operator gives $\mathsf{P}_t[f](\boldsymbol{z}, \boldsymbol{\theta}) = \mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(f(\boldsymbol{z}, \boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z}) - \sigma_t^{-2}\boldsymbol{z}$. In cases where $\boldsymbol{\theta}_i$ does not exist (unconditional DDPM), we simply set $m = 0$. The population risk gives

$$R(\boldsymbol{W}) := \frac{1}{d}\mathbb{E}_{(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{g})\sim\mu\otimes\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\Big[\big\|\sigma_t^{-1}\boldsymbol{g}_0 + \mathsf{P}_t(\mathrm{ResN}_{\boldsymbol{W}}([\lambda_t \boldsymbol{x} + \sigma_t \boldsymbol{g}, \boldsymbol{\theta}]))\big\|_2^2\Big].$$

In the proposition below, we provide a uniform upper bound for $|\widehat{R}(\boldsymbol{W}) - R(\boldsymbol{W})|$ over $\mathcal{W}_{d,m,D,L,M,B}$, where the ResNet class is given by Eq. (17).

Proposition 6. *Assume that $\mu \in \mathcal{P}([-1,1]^{d+m})$. There exists a numerical constant $C > 0$, such that with probability at least $1-\eta$,*

$$\begin{aligned}&\sup_{\boldsymbol{W}\in\mathcal{W}_{d,m,D,L,M,B}} \Big|\widehat{R}(\boldsymbol{W}) - R(\boldsymbol{W})\Big| \\ &\le C \cdot \frac{\lambda_t^2}{\sigma_t^4} \cdot \sqrt{\frac{[(d+m)D + LDM] \cdot [L \cdot \log(LB(m+d)/d) + \log(\lambda_t^{-1})] + \log(1/\eta)}{n}}.\end{aligned}$$

Proof of Proposition 6. The proof of this proposition uses the following lemma.

Lemma 4 (Proposition A.4 of Bai et al. [2023])**.** *Suppose that $\{X_w\}_{w\in\Theta}$ is a zero-mean random process given by*

$$X_w \equiv \frac{1}{n}\sum_{i=1}^n f(z_i; w) - \mathbb{E}_z[f(z; w)],$$

where $z_1, \cdots, z_n$ are i.i.d samples from a distribution $\mathbb{P}_z$ such that the following assumption holds:

(a) The index set Θ is equipped with a distance ρ and diameter B. Further, assume that for some constant A, for any ball Θ' of radius r in Θ, the covering number admits upper bound $\log N(\Delta; \Theta', \rho) \le d\log(2Ar/\Delta)$ for all $0 < \Delta \le 2r$.

(b) For any fixed $w \in \Theta$ and z sampled from $\mathbb{P}_z$, the random variable $f(z; w) - \mathbb{E}_z[f(z; w)]$ is a σ-sub-Gaussian random variable ($\mathbb{E}[e^{\lambda[f(z;w) - \mathbb{E}_{z'}[f(z';w)]]}] \le e^{\lambda^2\sigma^2/2}$ for any $\lambda \in \mathbb{R}$).

(c) For any $w, w' \in \Theta$ and z sampled from $\mathbb{P}_z$, the random variable $f(z; w) - f(z; w')$ is a $\sigma'\rho(w, w')$-sub-Gaussian random variable ($\mathbb{E}[e^{\lambda[f(z;w) - f(z;w')]}] \le e^{\lambda^2(\sigma')^2\rho^2(w,w')/2}$ for any $\lambda \in \mathbb{R}$).

Then with probability at least $1 - \eta$, it holds that

$$\sup_{w\in\Theta}|X_w| \le C\sigma\sqrt{\frac{d \cdot \log(2A(1 + B\sigma'/\sigma)) + \log(1/\eta)}{n}},$$

where C is a universal constant.

In Lemma 4, we can take $z = (\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\theta})$, $w = \boldsymbol{W}$, $\Theta = \mathcal{W}_{d,m,D,L,M,B}$, $\rho(w, w') = \|\!|\boldsymbol{W} - \boldsymbol{W}'|\!\|$, and $f(z_i; w) = d^{-1}\|\sigma_t^{-1}\boldsymbol{g}_i + \mathsf{P}_t(\mathrm{ResN}_{\boldsymbol{W}}(\lambda_t \boldsymbol{x}_i + \sigma_t \boldsymbol{g}_i, \boldsymbol{\theta}_i))\|_2^2$. Therefore, to show Proposition 6, we just need to apply Lemma 4 by checking (a), (b), (c).

Check (a). We note that the index set $\Theta = \mathcal{W}_{d,m,D,L,M,B}$ equipped with $\rho(w, w') = \|\!|\boldsymbol{W} - \boldsymbol{W}'|\!\|$ has diameter $2B$. Further note that $\mathcal{W}_{d,m,D,L,M,B}$ has dimension bounded by $4(d+m)D + 2LDM$. According to Example 5.8 of Wainwright [2019], it holds that $\log N(\Delta; \mathcal{W}_{d,m,D,L,M,r}, \|\!|\cdot|\!\|) \le [4(d+m)D + 2LDM] \cdot \log(1 + 2r/\Delta)$ for any $0 < \Delta \le 2r$. This verifies (a).

Check (b). By the definition of the projection operator that $\mathsf{P}_t[f](\boldsymbol{z}) = \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(f(\boldsymbol{z}) + \sigma_t^{-2}\boldsymbol{z}) - \sigma_t^{-2}\boldsymbol{z}$ and that $\boldsymbol{z} = \lambda_t \boldsymbol{x} + \sigma_t \boldsymbol{g}$, we have

$$\begin{aligned}0 \le f(z; w) &= d^{-1}\|\sigma_t^{-1}\boldsymbol{g} + \mathsf{P}_t(\mathrm{ResN}_{\boldsymbol{W}}(\lambda_t \boldsymbol{x} + \sigma_t \boldsymbol{g}, \boldsymbol{\theta}))\|_2^2 \\ &= d^{-1}\| - \lambda_t\sigma_t^{-2}\boldsymbol{x} + \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}_1}(\lambda_t \boldsymbol{x} + \sigma_t \boldsymbol{g}, \boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z})\|_2^2 \\ &\le 4\lambda_t^2\sigma_t^{-4}.\end{aligned}$$

As a consequence, $f(z,w) - \mathbb{E}_z[f(z,w)]$ is a $\sigma = 4\lambda_t^2\sigma_t^{-4}$ sub-Gaussian random variable.

Check (c). Direct calculation yields

$$
\begin{aligned}
&|f(z;w_1) - f(z;w_2)| \\
=& \frac{1}{d}\Big| \|\sigma_t^{-1}\boldsymbol{g} + \mathsf{P}_t(\mathrm{ResN}_{\boldsymbol{W}_1}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}, \boldsymbol{\theta}))\|_2^2 - \|\sigma_t^{-1}\boldsymbol{g} + \mathsf{P}_t(\mathrm{ResN}_{\boldsymbol{W}_2}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}, \boldsymbol{\theta}))\|_2^2 \Big| \\
=& \frac{1}{d}\Big| \| -\lambda_t\sigma_t^{-2}\boldsymbol{x} + \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}_1}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}, \boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z})\|_2^2 \\
&- \| -\lambda_t\sigma_t^{-2}\boldsymbol{x} + \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}_2}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}, \boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z})\|_2^2 \Big| \\
\le& \frac{8\lambda_t}{\sigma_t^2\sqrt{d}} \cdot \big\| \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}_1}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}, \boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z}) \\
&- \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}_2}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}, \boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z}) \big\|_2 \\
\lesssim& \frac{2\lambda_t L(B^2+1)^L}{\sigma_t^2} \cdot \frac{1}{\sqrt{d}} \Big(\lambda_t\|\boldsymbol{x}\|_2 + \sigma_t\|\boldsymbol{g}\|_2 + \|\boldsymbol{\theta}\|_2 \Big) \cdot |||\boldsymbol{W}_1 - \boldsymbol{W}_2|||.
\end{aligned}
$$

Notice that $(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{g}) \sim \mu \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and note that $\mu \in \mathcal{P}([-1,1]^{d+m})$, we have that $\|\boldsymbol{\theta}\|_2/\sqrt{d}$ is $\sqrt{m/d}$-bounded and is thus $\mathcal{O}(\sqrt{m/d})$-sub-Gaussian, $\|\boldsymbol{x}\|_2/\sqrt{d}$ is 1-bounded and is thus $\mathcal{O}(1)$-sub-Gaussian, and $\|\boldsymbol{g}\|_2/\sqrt{d}$ is $\mathcal{O}(1)$-sub-Gaussian. As a consequence, $f(z;w_1) - f(z;w_2)$ is $\sigma'\rho(w_1,w_2) = C \cdot \lambda_t\sigma_t^{-2}L(B^2+1)^L\sqrt{(m+d)/d} \cdot |||\boldsymbol{W}_1 - \boldsymbol{W}_2|||$ sub-Gaussian.

Therefore, we apply Lemma 4, and use the fact that

$$
\begin{aligned}
&\log(2(1 + B\sigma'/\sigma)) \\
=&\log(2(1 + (C/2)B\lambda_t^{-1}\sigma_t^2 L(B^2+1)^L\sqrt{(m+d)/d})) \lesssim L\log(LB(m+d)/d) + \log(\lambda_t^{-1}).
\end{aligned}
$$

This concludes the proof of Proposition 6. □

D.3.2 Result for Sparse coding

In the setting of sparse coding, we assume a fixed dictionary $\boldsymbol{A} \in \mathbb{R}^{d\times m}$. The model $\boldsymbol{x} \sim \mu$ is given by $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}_d)$ is independent of anything else and $\theta_i \sim_{iid} \pi_\theta \in \mathcal{P}([-\Pi, \Pi])$ for $i \in [m]$. Assume that we have $\{(\boldsymbol{x}_i, \boldsymbol{g}_i)\}_{i\in[n]} \sim_{iid} \mu \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. We are interested in estimating the score function $\hat{\boldsymbol{s}}_t = \bar{\mathsf{P}}_t\mathrm{ResN}_{\widehat{\boldsymbol{W}}_t}$ by solving the following ERM problem:

$$
\begin{aligned}
\widehat{\boldsymbol{W}}_t &= \mathrm{argmin}_{\boldsymbol{W} \in \mathcal{W}_{d,D,L,M,B}} \widehat{R}_n(\boldsymbol{W}), \\
\widehat{R}_n(\boldsymbol{W}) &= \frac{1}{nd}\sum_{i=1}^n \big\| \sigma_t^{-1}\boldsymbol{g}_i + \bar{\mathsf{P}}_t(\mathrm{ResN}_{\boldsymbol{W}}(\lambda_t\boldsymbol{x}_i + \sigma_t\boldsymbol{g}_i)) \big\|_2^2.
\end{aligned} \tag{31}
$$

Here, the truncation operator gives $\bar{\mathsf{P}}_t[f](\boldsymbol{z}) = \mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(f(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}) - (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}$. The corresponding population risk gives

$$
R(\boldsymbol{W}) := \frac{1}{d}\mathbb{E}_{(\boldsymbol{x},\boldsymbol{g})\sim\mu\otimes\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\Big[\big\| \sigma_t^{-1}\boldsymbol{g} + \bar{\mathsf{P}}_t(\mathrm{ResN}_{\boldsymbol{W}}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g})) \big\|_2^2 \Big].
$$

In the proposition below, we provide a uniform upper bound for $|\widehat{R}(\boldsymbol{W}) - R(\boldsymbol{W})|$ over $\mathcal{W}_{d,D,L,M,B}$ in the sparse coding setup, where the ResNet class is given by Eq. (3).

Proposition 7. *Under the setting of sparse coding stated above, there exists a numerical constant $C > 0$, such that with probability at least $1-\eta$, for $n \ge \log(2/\eta)$, we have*

$$
\begin{aligned}
\sup_{\boldsymbol{W}\in\mathcal{W}_{d,D,L,M,B}} \Big| \widehat{R}(\boldsymbol{W}) - R(\boldsymbol{W}) \Big| &\lesssim \Big(\lambda_t^2\|\boldsymbol{A}\|_{\mathrm{op}}^2\Pi^2(\tau^{-4}+1)\frac{m}{d} + \frac{\lambda_t^2}{\sigma_t^2}(1+\tau^2) \Big) \\
&\times \sqrt{\frac{(dD + LDM)\cdot[T + L\log(LB) + \log(nmT(\tau+1)(\|\boldsymbol{A}\|_{\mathrm{op}}\Pi+1)\tau^{-1})] + \log(2/\eta)}{n}}.
\end{aligned}
$$

Proof of Proposition 7. Note that $\{(\boldsymbol{x}_i, \boldsymbol{g}_i)\}_{i\in[n]} \sim_{iid} \mu \times \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ where μ is the sparse coding model. Then we must have $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i$ for some $(\boldsymbol{\theta}_i, \boldsymbol{\varepsilon}_i) \sim_{iid} \pi_0^m \times \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}_d)$. Denote $z = (\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\varepsilon})$, $w = \boldsymbol{W}$, and

$$f(z; w) = d^{-1}(\|\sigma_t^{-1}\boldsymbol{g} + \bar{\mathsf{P}}_t(\mathrm{ResN}_{\boldsymbol{W}}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}))\|_2^2 - \|(\sigma_t^{-1} - \sigma_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1})\boldsymbol{g} - \lambda_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{\varepsilon}\|_2^2).$$

We further denote $\boldsymbol{z} = \lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}$. Note that we have

$$\begin{aligned}
&|f(z; w_1) - f(z; w_2)| \\
=& \frac{1}{d}\Big|\|\sigma_t^{-1}\boldsymbol{g} + \bar{\mathsf{P}}_t(\mathrm{ResN}_{\boldsymbol{W}_1}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}))\|_2^2 - \|\sigma_t^{-1}\boldsymbol{g} + \bar{\mathsf{P}}_t(\mathrm{ResN}_{\boldsymbol{W}_2}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}))\|_2^2\Big| \\
\leq& \frac{1}{d}\Big|\|\mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(\mathrm{ResN}_{\boldsymbol{W}_1}(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}) - (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z} + \sigma_t^{-1}\boldsymbol{g}\|_2^2 \\
&- \|\mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(\mathrm{ResN}_{\boldsymbol{W}_2}(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}) - (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z} + \sigma_t^{-1}\boldsymbol{g}\|_2^2\Big| \\
\lesssim& \Big(\frac{\sqrt{m}\lambda_t\Pi\|\boldsymbol{A}\|_{\mathrm{op}}}{d(\sigma_t^2 + \tau^2\lambda_t^2)} + \frac{\lambda_t}{d(\sigma_t^2 + \tau^2\lambda_t^2)}\|\boldsymbol{\varepsilon}\|_2 + \frac{\tau^2\lambda_t^2}{d\sigma_t(\sigma_t^2 + \tau^2\lambda_t^2)}\|\boldsymbol{g}\|_2\Big) \\
&\times \big\|\mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(\mathrm{ResN}_{\boldsymbol{W}_1}(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}) \\
&- \mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(\mathrm{ResN}_{\boldsymbol{W}_2}(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z})\big\|_2 \\
\lesssim& \frac{L(B^2+1)^L\lambda_t}{(\sigma_t^2 + \tau^2\lambda_t^2)} \cdot \Big(\frac{\sqrt{m}\Pi\|\boldsymbol{A}\|_{\mathrm{op}}}{\sqrt{d}} + \frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{d}} + \frac{\tau^2\lambda_t\|\boldsymbol{g}\|_2}{\sqrt{d}\sigma_t}\Big) \times \\
&\frac{1}{\sqrt{d}}\Big(\lambda_t\|\boldsymbol{A}\boldsymbol{\theta}\|_2 + \lambda_t\|\boldsymbol{\varepsilon}\|_2 + \sigma_t\|\boldsymbol{g}\|_2\Big) \cdot |||\boldsymbol{W}_1 - \boldsymbol{W}_2|||.
\end{aligned}$$

Therefore, we denote by $\mathcal{N}(\Delta; \mathcal{W}_{d,D,L,M,B}, \rho)$ a Δ-covering of $\mathcal{W}_{d,D,L,M,B}$ under metric $\rho(\boldsymbol{W}_1, \boldsymbol{W}_2) = |||\boldsymbol{W}_1 - \boldsymbol{W}_2|||$ for some $\Delta > 0$. Then

$$\begin{aligned}
&\sup_{w\in\mathcal{W}_{d,D,L,M,B}}\Big|\frac{1}{n}\sum_{i=1}^n f(z_i; w) - \mathbb{E}[f(z; w)]\Big| \\
\leq& \sup_{w\in\mathcal{N}(\Delta;\mathcal{W}_{d,D,L,M,B},\rho)}\Big|\frac{1}{n}\sum_{i=1}^n f(z_i; w) - \mathbb{E}[f(z; w)]\Big| + \frac{L(B^2+1)^L\lambda_t}{(\sigma_t^2 + \tau^2\lambda_t^2)} \cdot \Delta \cdot (L_n + \mathbb{E}[L_n]),
\end{aligned}$$

where

$$L_n = \frac{1}{nd}\sum_{i=1}^n \big(\sqrt{m}\Pi\|\boldsymbol{A}\|_{\mathrm{op}} + \|\boldsymbol{\varepsilon}_i\|_2 + \sigma_t^{-1}\tau^2\lambda_t\|\boldsymbol{g}_i\|_2\big) \cdot \Big(\lambda_t\|\boldsymbol{A}\boldsymbol{\theta}_i\|_2 + \lambda_t\|\boldsymbol{\varepsilon}_i\|_2 + \sigma_t\|\boldsymbol{g}_i\|_2\Big).$$

Since $(\boldsymbol{\theta}_i, \boldsymbol{\varepsilon}_i, \boldsymbol{g}_i) \sim \pi_0^m \otimes \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}_d) \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $\sigma_t^2 \leq 1$ and $\lambda_t^2 \leq 1$, we have $\mathbb{E}[L_n] \leq \overline{L}$ and $L_n - \mathbb{E}[L_n]$ is $\mathrm{SE}(\overline{L}/\sqrt{n}, \overline{L})$, for $\overline{L} = (m/d)\Pi^2\|\boldsymbol{A}\|_{\mathrm{op}}^2 + \sigma_t^{-2}(\tau^4 + 1)$. By Bernstein's inequality, we conclude that with probability at least $1 - \eta/2$, we have

$$\begin{aligned}
L_n + \mathbb{E}[L_n] &\leq C \cdot \overline{L}(1 + \sqrt{\log(2/\eta)/n} + \log(2/\eta)/n) \leq C \cdot \overline{L}(1 + \log(2/\eta)) \\
&= C \cdot \big((m/d)\Pi^2\|\boldsymbol{A}\|_{\mathrm{op}}^2 + \sigma_t^{-2}(\tau^4 + 1)\big) \cdot (1 + \log(2/\eta)).
\end{aligned}$$

for some numerical constant C.

Furthermore, note that we have

$$\begin{aligned}
&f(z; w) \\
=& d^{-1}\|\sigma_t^{-1}\boldsymbol{g} + \bar{\mathsf{P}}_t(\mathrm{ResN}_{\boldsymbol{W}}(\lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}))\|_2^2 - d^{-1}\|(\sigma_t^{-1} - \sigma_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1})\boldsymbol{g} - \lambda_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{\varepsilon}\|_2^2 \\
=& d^{-1}\|\sigma_t^{-1}\boldsymbol{g} + \mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}) - (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}\|_2^2 \\
&- d^{-1}\|(\sigma_t^{-1} - \sigma_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1})\boldsymbol{g} - \lambda_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{\varepsilon}\|_2^2 \\
=& d^{-1}\|\mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}) - \lambda_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{A}\boldsymbol{\theta}\|_2^2 \\
&+ 2d^{-1}\big\langle(\sigma_t^{-1} - \sigma_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1})\boldsymbol{g} - \lambda_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{\varepsilon}, \\
&\quad \mathrm{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\mathrm{op}}\Pi\cdot\lambda_t(\sigma_t^2+\tau^2\lambda_t^2)^{-1}}(\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + (\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{z}) - \lambda_t(\sigma_t^2 + \tau^2\lambda_t^2)^{-1}\boldsymbol{A}\boldsymbol{\theta}\big\rangle.
\end{aligned}$$

As a consequence, $f(z;w) - \mathbb{E}_z[f(z,w)]$ is sub-Gaussian with variance proxy

$$C^2 \cdot \left(\frac{m\|\boldsymbol{A}\|_{\mathrm{op}}^2 \Pi^2 \lambda_t^2}{d(\sigma_t^2 + \tau^2\lambda_t^2)^2} + \frac{\tau^2 \lambda_t^2}{\sigma_t^2(\sigma_t^2 + \lambda_t^2\tau^2)} \right)^2$$

for some other numerical constant C. Therefore, with probability at least $1-\eta/2$, by sub-Gaussian tail bound and by the bound $\log|\mathcal{N}(\Delta; \mathcal{W}_{d,D,L,M,B}, \rho)| \le [4dD + 2LDM] \cdot \log(1 + 2B/\Delta)$, we have

$$\begin{aligned} &\sup_{\boldsymbol{W} \in \mathcal{N}(\Delta; \mathcal{W}_{d,D,L,M,B}, \rho)} \left| \frac{1}{n} \sum_{i=1}^n f(z_i; w_i) - \mathbb{E}[f(z;w)] \right| \\ &\lesssim \Big(\frac{m\|\boldsymbol{A}\|_{\mathrm{op}}^2 \Pi^2 \lambda_t^2}{d(\sigma_t^2 + \tau^2\lambda_t^2)^2} + \frac{\tau^2 \lambda_t^2}{\sigma_t^2(\sigma_t^2 + \lambda_t^2\tau^2)} \Big) \cdot \sqrt{\frac{[4dD + 2LDM]\cdot \log(1+2B/\Delta) + \log(2/\eta)}{n}}. \end{aligned}$$

Setting

$$\Delta = \Big(\frac{m\|\boldsymbol{A}\|_{\mathrm{op}}^2 \Pi^2 \lambda_t^2}{d(\sigma_t^2 + \tau^2\lambda_t^2)^2} + \frac{\tau^2 \lambda_t^2}{\sigma_t^2(\sigma_t^2 + \lambda_t^2\tau^2)} \Big) \cdot \frac{(\sigma_t^2 + \tau^2\lambda_t^2)}{nL(B^2+1)^L \lambda_t \cdot \big(md^{-1}\Pi^2\|\boldsymbol{A}\|_{\mathrm{op}}^2 + \sigma_t^{-2}(\tau^4+1)\big)},$$

we conclude that with probability at least $1-\eta$, when $n \ge \log(2/\eta)$, we have

$$\begin{aligned} &\sup_{\boldsymbol{W} \in \mathcal{N}(\mathcal{W}_{d,D,L,M,B}, \rho, \Delta)} \left| \frac{1}{n} \sum_{i=1}^n f(z_i; w_i) - \mathbb{E}[f(z;w)] \right| \\ &\lesssim n^{-1} \cdot \Big(\frac{m\|\boldsymbol{A}\|_{\mathrm{op}}^2 \Pi^2 \lambda_t^2}{d(\sigma_t^2 + \tau^2\lambda_t^2)^2} + \frac{\tau^2 \lambda_t^2}{\sigma_t^2(\sigma_t^2 + \lambda_t^2\tau^2)} \Big) \cdot (\log(2/\eta) + 1) + \left\{ \frac{m\|\boldsymbol{A}\|_{\mathrm{op}}^2 \Pi^2 \lambda_t^2}{d(\sigma_t^2 + \tau^2\lambda_t^2)^2} + \frac{\tau^2 \lambda_t^2}{\sigma_t^2(\sigma_t^2 + \lambda_t^2\tau^2)} \right\} \\ &\quad \times \sqrt{\frac{(dD + LDM)\cdot [T + L\log(LB) + \log(nmT(\tau+1)(\|\boldsymbol{A}\|_{\mathrm{op}}\Pi + 1)\tau^{-1})] + \log(2/\eta)}{n}} \\ &\lesssim \Big(\lambda_t^2 \|\boldsymbol{A}\|_{\mathrm{op}}^2 \Pi^2 (\tau^{-4}+1) \frac{m}{d} + \frac{\lambda_t^2}{\sigma_t^2}(1+\tau^2) \Big) \\ &\quad \times \sqrt{\frac{(dD + LDM)\cdot [T + L\log(LB) + \log(nmT(\tau+1)(\|\boldsymbol{A}\|_{\mathrm{op}}\Pi + 1)\tau^{-1})] + \log(2/\eta)}{n}}, \end{aligned}$$

where the inequalities above uses the definition that $\lambda_t = e^{-t}$, $\sigma_t^2 = 1 - e^{-2t}$ and $t \le T$. This concludes the proof of Proposition 7. $\square$

D.4 Uniform approximation of the denoiser

The lemma below tells us that denoiser functions can be uniformly approximated with a linear combination of $\mathrm{ReLU}(\cdot)$ with changing intercepts. Furthermore, such approximation can achieve arbitrary precision.

Lemma 5. *Assume π_0 is a probability distribution over $\mathbb{R}$ that has bounded support, and $\gamma > 0$ is a fixed constant. Define $F(\lambda) := \mathbb{E}_{(\beta,z)\sim\pi_0\otimes\mathcal{N}(0,1)}[\,\beta \mid \beta + \gamma^{-1/2}z = \lambda\gamma^{-1}]$. Let $\Pi_{\min} := \inf_\lambda F(\lambda)$, $\Pi_{\max} := \sup_\lambda F(\lambda)$, $\Pi := \max\{\,|\Pi_{\max}|, |\Pi_{\min}|\,\}$, and $\Delta := \Pi_{\max} - \Pi_{\min}$. One can verify that $F(\cdot)$ is Π^2-Lipschitz continuous and non-decreasing. For any $\zeta > 0$, we define*

$$w_\zeta := \inf\Big\{ w: \textit{ for all } \lambda_1 > \lambda_2 \ge w \textit{ or } \lambda_1 < \lambda_2 \le -w \textit{ we have } |F(\lambda_1) - F(\lambda_2)| < \Delta/\lceil \Delta\zeta^{-1}\rceil \Big\}. \tag{32}$$

Then there exists $\{a_j\}_{j\in\{0\}\cup[\lceil\Delta\zeta^{-1}\rceil - 1]}$ and $\{w_j\}_{j\in[\lceil\Delta\zeta^{-1}\rceil-1]}$, such that

$$\sup_{\lambda\in\mathbb{R}} |F(\lambda) - f(\lambda)| \le \zeta, \quad \textit{where} \quad f(\lambda) = \sum_{j=1}^{\lceil\Delta\zeta^{-1}\rceil - 1} a_j \mathrm{ReLU}(\lambda - w_j) + a_0. \tag{33}$$

Furthermore, we have $\sup_{j\in[\lceil\Delta\zeta^{-1}\rceil-1]}|w_j| \le w_\zeta$, $|a_0| \le \Pi$, and $|a_j| \le 2\Pi^2$ for all $j \in [\lceil\Delta\zeta^{-1}\rceil - 1]$.

Proof of Lemma 5. When π_0 is a Dirac measure, we simply take $a_0 = \mathbb{E}[\beta]$. In other cases, one can verify that $F(\cdot)$ is strictly increasing, hence $\Pi_{\max} > \Pi_{\min}$. Then for any $\alpha \in (\Pi_{\min}, \Pi_{\max})$, there exists a unique $\mu_\alpha \in \mathbb{R}$, such that $F(\mu_\alpha) = \alpha$.

Let $a_0 = \Pi_{\min} + \Delta\lceil\Delta\zeta^{-1}\rceil^{-1}$. For $j \in [[\Delta\zeta^{-1}] - 1]$, we let

$$w_j = \mu_{-\Pi_{\min}+j\Delta/\lceil\Delta\zeta^{-1}\rceil}, \qquad a_j = \frac{\Delta}{\lceil\Delta\zeta^{-1}\rceil(w_{j+1} - w_j)} - \frac{\Delta}{\lceil\Delta\zeta^{-1}\rceil(w_j - w_{j-1})}.$$

In the above equations, we make the convention that $w_0 = w_{\lceil\Delta\zeta^{-1}\rceil} = \infty$. With $\{a_j\}_{j\in\{0\}\cup[\lceil\Delta\zeta^{-1}\rceil-1]}$ and $\{w_j\}_{j\in[\lceil\Delta\zeta^{-1}\rceil-1]}$ defined as above, one can verify that Eq. (33) is true. Furthermore, since $\|F'\|_\infty \le \Pi^2$, we have $|\Delta/\lceil\Delta\zeta^{-1}\rceil(w_{j+1} - w_j)| \le \Pi^2$ for all possible j. This gives $|a_j| \le 2\Pi^2$ for every j.

□

Remark 1. *When $\pi_0 = \mathrm{Unif}(\{\pm 1\})$, one can check that for any $\gamma > 0$, we have $F(x) = \tanh(x)$. In this case, one can verify that $|w_\zeta| \le \log\lceil\zeta^{-1}\rceil$. In addition, we can further guarantee that $\sum_{j\in[\lceil\Delta\zeta^{-1}\rceil-1]} |a_j| \le 2$.*

D.5 Approximation error of fixed point iteration

Lemma 6. *Assume that $\boldsymbol{h} \in \mathbb{R}^d$, $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ with $\|\boldsymbol{U}\|_{\mathrm{op}} \le A < \Pi^{-2}$ for some $\Pi > 0$. Further assume that $f_* : \mathbb{R} \mapsto \mathbb{R}$ is Π^2-Lipschitz continuous and $f : \mathbb{R} \mapsto \mathbb{R}$ is a function satisfying*

$$\sup_{u\in\mathbb{R}} |f(u) - f_*(u)| \le \zeta. \tag{34}$$

Let $\hat{\boldsymbol{m}} \in \mathbb{R}^d$ satisfying $\|\hat{\boldsymbol{m}}\|_2 \le \Pi\sqrt{d}$ be the unique fixed point of

$$\hat{\boldsymbol{m}} = f_*(\boldsymbol{U}\hat{\boldsymbol{m}} + \boldsymbol{h}). \tag{35}$$

Let $\tilde{\boldsymbol{m}}^0 = \boldsymbol{0}$ and

$$\tilde{\boldsymbol{m}}^k = f(\boldsymbol{U}\tilde{\boldsymbol{m}}^{k-1} + \boldsymbol{h}). \tag{36}$$

Then we have

$$\frac{1}{\sqrt{d}}\|\tilde{\boldsymbol{m}}^k - \hat{\boldsymbol{m}}\|_2 \le \Pi \cdot (\Pi^2 A)^k + \frac{\zeta}{1 - \Pi^2 A}. \tag{37}$$

Proof of Lemma 6. By Eq. (34) and (36), we have

$$\tilde{\boldsymbol{m}}^k = f_*(\boldsymbol{U}\tilde{\boldsymbol{m}}^{k-1} + \boldsymbol{h}) + \boldsymbol{\zeta}^k,$$

where $\|\boldsymbol{\zeta}^k\|_2 \le \sqrt{d}\zeta$. Comparing with Eq. (35), we get

$$\|\tilde{\boldsymbol{m}}^k - \hat{\boldsymbol{m}}\|_2 \le \Pi^2\|\boldsymbol{U}\|_{\mathrm{op}}\|\tilde{\boldsymbol{m}}^{k-1} - \hat{\boldsymbol{m}}\|_2 + \|\boldsymbol{\zeta}^k\|_2 \le \Pi^2 A \cdot \|\tilde{\boldsymbol{m}}^{k-1} - \hat{\boldsymbol{m}}\|_2 + \sqrt{d}\zeta.$$

By the fact that $\|\tilde{\boldsymbol{m}}^0 - \hat{\boldsymbol{m}}\|_2 = \|\hat{\boldsymbol{m}}\|_2 \le \Pi\sqrt{d}$, this gives Eq. (37), which concludes the proof of the lemma. □

D.6 Properties of two-phase time discretization scheme

The lemma below provides a bound related to the two-phase time discretization scheme that appears to be useful when deriving the sampling error bound.

Lemma 7. *Consider the two-phase discretization scheme $(\kappa, N_0, N, T, \delta, \{t_k\}_{0\le k\le N})$ and recall that $\gamma_k = t_{k+1} - t_k$ (Definition 1). Recall the definition $\lambda_t = e^{-t}$ and $\sigma_t^2 = 1 - e^{-2t}$. Then we have*

$$\sum_{0\le k\le N-1} \gamma_k \cdot \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4} \lesssim 1 + \delta^{-1}. \tag{38}$$

Proof of Lemma 7. Simple algebra yields

$$\sigma_t^{-2} = 1/[1 - e^{-2t}] \le 10 \cdot [1 \vee (1/t)].$$

Note that $T - t_k \le 1$ for all $k \ge N_0$ and $T - t_k \ge 1$ for all $k \le N_0 - 1$ (c.f. Definition 1 for N_0). Then the summation in the first phase has bound (we use the fact that $\kappa < 1$)

$$\sum_{0 \le k \le N_0 - 1} \gamma_k \cdot \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4} \le 100\kappa \sum_{0 \le k \le M-1} e^{-2(T-t_k)} \le 100\kappa e^{-2} \sum_{k \ge 0} e^{-2k\kappa}$$
$$\le 100\kappa e^{-2} \frac{1}{1 - e^{-2\kappa}} \le 100.$$

Furthermore, the summation in the second phase yields (recall from Definition 1 that for $k \ge N_0$, we have $T - t_{N_0+k} = (1+\kappa)^{-k}$, $\gamma_{N_0+k} = \kappa/(1+\kappa)^{k+1}$, and $\delta = (1+\kappa)^{N_0 - N}$)

$$\sum_{N_0 \le k \le N-1} \gamma_k \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4} \le 100 \sum_{N_0 \le k \le N-1} \gamma_k/(T - t_k)^2$$
$$= 100 \sum_{0 \le k \le N - N_0 - 1} [\kappa/(1+\kappa)^{k+1}] \cdot (1+\kappa)^{2k} = 100 \frac{\kappa}{\delta} \sum_{0 \le k \le N - N_0 - 1} (1+\kappa)^{-2-k}$$
$$\le 100 \frac{\kappa}{\delta} \sum_{k=1}^{\infty} (1+\kappa)^{-k} = 100/\delta.$$

Combining the two inequalities above proves Eq. (38) and concludes the proof. $\square$

E Proofs for Section 3: Ising models

E.1 Proof of Theorem 1

Approximate the minimizer of the free energy via an iterative algorithm

We first show that we can approximate the minimizer of $\mathcal{F}_t^{\rm VI}$ using a simple iterative algorithm. Calculating the Hessian of $\mathcal{F}_t^{\rm VI}$, we obtain

$$\nabla_{\boldsymbol{m}}^2 \mathcal{F}_t^{\rm VI}(\boldsymbol{m}; \boldsymbol{z}) = \mathrm{diag}\{(1 - m_i^2)_{i \in [d]}\} - \boldsymbol{A} + \boldsymbol{K} \succeq (1 - A) \cdot \mathbf{I}_d \succ 0, \quad \forall \boldsymbol{m} \in [-1, 1]^d,$$

where the inequalities are due to the fact that $\mathrm{diag}\{(1 - m_i^2)_{i \in [d]}\} \succeq \mathbf{I}_d$ and the assumption that $\|\boldsymbol{K} - \boldsymbol{A}\|_{\rm op} \le A < 1$. Therefore, $\mathcal{F}_t^{\rm VI}(\cdot, \boldsymbol{z})$ is strongly convex in its first coordinate for all $\boldsymbol{z} \in \mathbb{R}^d$, hence the critical equation

$$\nabla_{\boldsymbol{m}} \mathcal{F}_t^{\rm VI}(\boldsymbol{m}; \boldsymbol{z}) = \tanh^{-1}(\boldsymbol{m}) - \boldsymbol{A}\boldsymbol{m} - \lambda_t \sigma_t^{-2} \boldsymbol{z} + \boldsymbol{K}\boldsymbol{m} = \boldsymbol{0},$$

can have at most one solution on $[-1, 1]^d$. Furthermore, $\nabla_{\boldsymbol{m}} \mathcal{F}_t^{\rm VI}(\boldsymbol{m}; \boldsymbol{z}) = \boldsymbol{0}$ is equivalent to the fixed point equation

$$\boldsymbol{m} = \tanh((\boldsymbol{A} - \boldsymbol{K})\boldsymbol{m} + \lambda_t \sigma_t^{-2} \boldsymbol{z}),$$

and $T(\boldsymbol{m}) = \tanh((\boldsymbol{A} - \boldsymbol{K})\boldsymbol{m} + \lambda_t \sigma_t^{-2} \boldsymbol{z})$ is a continuous mapping from $[-1, 1]^d$ to itself. Therefore, there exists a solution of $\boldsymbol{m} = T(\boldsymbol{m})$ by Brouwer's fixed-point theorem. This implies that the above fixed point equation has a unique solution $\hat{\boldsymbol{m}}_t(\boldsymbol{z}) \in [-1, 1]^d$.

Take $f : \mathbb{R} \to \mathbb{R}$ to be the function as derived by Lemma 5 achieving ζ-uniform approximation to $\tanh(\cdot)$. We write $f(x) = \sum_{j=1}^{\lceil 2\zeta^{-1} \rceil - 1} a_j \mathrm{ReLU}(x - w_j) + a_0$. Define iterative algorithm $\{\tilde{\boldsymbol{m}}^\ell\}_{\ell \ge 0}$ by

$$\tilde{\boldsymbol{m}}^0 = \boldsymbol{0}, \qquad \tilde{\boldsymbol{m}}^\ell(\boldsymbol{z}) = \tilde{\boldsymbol{m}}^\ell = f((\boldsymbol{A} - \boldsymbol{K})\tilde{\boldsymbol{m}}^{\ell-1} + \lambda_t \sigma_t^{-2} \boldsymbol{z}). \tag{39}$$

Then by Lemma 6 with $\Pi = 1$, we obtain that

$$\|\tilde{\boldsymbol{m}}^\ell(\boldsymbol{z}) - \hat{\boldsymbol{m}}_t(\boldsymbol{z})\|_2/\sqrt{d} \le A^\ell + \zeta \cdot (1 - A)^{-1}. \tag{40}$$

Represent the iterative algorithm as a ResNet

Next, we show that $\tilde{\boldsymbol{m}}^\ell(\boldsymbol{z})$ defined as above takes the form of a ResNet.

Lemma 8. *For all $\ell \in \mathbb{N}_+$ and $\delta \le t \le T$, there exists $\boldsymbol{W} \in \mathcal{W}_{d,D,\ell,M,B}$ with*

$$D = 3d, \quad M = (\lceil 2\zeta^{-1}\rceil + 3)d,$$
$$B = (\lceil 2\zeta^{-1}\rceil - 1)(4 + \log\lceil\zeta^{-1}\rceil) + 8 + (1 - e^{-2\delta})^{-1} + \sqrt{d},$$

such that $(\lambda_t \tilde{\boldsymbol{m}}^\ell(\boldsymbol{z}) - \boldsymbol{z})/\sigma_t^2 = \mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z})$, where $\tilde{\boldsymbol{m}}^\ell$ is as defined in Eq. (39).

Proof of Lemma 8. Recall the definition of f as an approximation of $\tanh$ as in Lemma 5. Recall that a ResNet takes the form (ResNet). We shall choose the weight matrices appropriately such that $\boldsymbol{u}^{(\ell)} = [\tilde{\boldsymbol{m}}^\ell; \sigma_t^{-2}\boldsymbol{z}; \mathbf{1}_d]^\mathsf{T} \in \mathbb{R}^{3d}$. In particular, for $\ell = 0$, we set

$$\boldsymbol{W}_{\mathrm{in}} = \begin{bmatrix} \mathbf{0}_{d\times d} & \sigma_t^{-2}\mathbf{I}_d & \mathbf{0}_{d\times d} \\ \mathbf{0}_{1\times d} & \mathbf{0}_{1\times d} & \mathbf{1}_{1\times d} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{3d\times(d+1)}.$$

For $\ell \ge 1$, we set

$$\boldsymbol{W}_1^{(\ell)} = \begin{bmatrix} a_i\mathbf{I}_d & \cdots & a_{\lceil 2\zeta^{-1}\rceil - 1}\mathbf{I}_d & -\mathbf{I}_d & \mathbf{I}_d & a_0\mathbf{I}_d & -a_0\mathbf{I}_d \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \end{bmatrix} \in \mathbb{R}^{3d\times(\lceil 2\zeta^{-1}\rceil + 3)d},$$

$$\boldsymbol{W}_2^{(\ell)} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{K} & \cdots & \boldsymbol{A} - \boldsymbol{K} & \mathbf{I}_d & -\mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \lambda_t\mathbf{I}_d & \cdots & \lambda_t\mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ -w_1\mathbf{I}_d & \cdots & -w_{\lceil 2\zeta^{-1}\rceil - 1}\mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{I}_d & -\mathbf{I}_d \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{(\lceil 2\zeta^{-1}\rceil + 3)d\times 3d}.$$

Finally, we take $\boldsymbol{W}_{\mathrm{out}} = [\lambda_t\sigma_t^{-2}\mathbf{I}_d, -\mathbf{I}_d, \mathbf{0}_{d\times d}] \in \mathbb{R}^{d\times 3d}$.

By Lemma 5 and Remark 1, we have $\sum_{j=1}^{\lceil 2\zeta^{-1}\rceil - 1} |a_j| \le 2$, $|a_0| \le 1$, and $|w_j| \le \log\lceil\zeta^{-1}\rceil$. Therefore, $\|\boldsymbol{W}_{\mathrm{in}}\|_{\mathrm{op}} \le \sqrt{d} + \sigma_t^{-2}$, $\|\boldsymbol{W}_{\mathrm{out}}\|_{\mathrm{op}} \le 1 + \lambda_t\sigma_t^{-2}$, $\|\boldsymbol{W}_1^{(\ell)}\|_{\mathrm{op}} \le 2\lceil 2\zeta^{-1}\rceil + 2$ and $\|\boldsymbol{W}_2^{(\ell)}\|_{\mathrm{op}} \le (\lceil 2\zeta^{-1}\rceil - 1)(2 + \log\lceil\zeta^{-1}\rceil) + 4$. Hence, $\|\!|\boldsymbol{W}|\!\| \le (\lceil 2\zeta^{-1}\rceil - 1)(4 + \log\lceil\zeta^{-1}\rceil) + 8 + \sigma_t^{-2} + \sqrt{d}$. Note that for $\delta \le t \le T$, it holds that $\sigma_t^{-2} \le (1 - e^{-2\delta})^{-1}$. Therefore, we have

$$\|\!|\boldsymbol{W}|\!\| \le B = (\lceil 2\zeta^{-1}\rceil - 1)(3 + \log\lceil\zeta^{-1}\rceil) + 8 + (1 - e^{-2\delta})^{-1} + \sqrt{d}.$$

This completes the proof of Lemma 8. $\square$

Proof of Theorem 1

Recall that we have $\hat{\boldsymbol{s}}_t(\boldsymbol{z}) = \mathsf{P}_t[\mathrm{ResN}_{\widehat{\boldsymbol{W}}}](\boldsymbol{z})$, where $\widehat{\boldsymbol{W}} = \mathrm{argmin}_{\boldsymbol{W}\in\mathcal{W}}\hat{\mathbb{E}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]$ for $\mathcal{W} = \mathcal{W}_{d,D,L,M,B}$. Here, $\hat{\mathbb{E}}$ denotes averaging over the empirical data distribution. By standard error decomposition analysis in empirical risk minimization theory, we have:

$$\begin{aligned} \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\widehat{\boldsymbol{W}}}](\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]/d &\le \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]/d \\ &\quad + 2\sup_{\boldsymbol{W}\in\mathcal{W}} \Big|\hat{\mathbb{E}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d - \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d\Big|. \end{aligned}$$

Furthermore, a standard identity in diffusion model theory shows:

$$\mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d = \mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) + \sigma_t^{-1}\boldsymbol{g}\|_2^2]/d + C, \quad C = \mathbb{E}[\|\boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d - \mathbb{E}[\|\sigma_t^{-1}\boldsymbol{g}\|_2^2]/d.$$

Combining the above yields:

$$\mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \le \bar{\varepsilon}_{\mathrm{app}}^2 + \bar{\varepsilon}_{\mathrm{gen}}^2, \tag{41}$$

where $\bar{\varepsilon}_{\mathrm{app}}^2$ is the approximation error and $\bar{\varepsilon}_{\mathrm{gen}}^2$ is the generalization error,

$$\begin{aligned} \bar{\varepsilon}_{\mathrm{app}}^2 &= \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d, \\ \bar{\varepsilon}_{\mathrm{gen}}^2 &= 2\sup_{\boldsymbol{W}\in\mathcal{W}} \Big|\hat{\mathbb{E}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d - \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d\Big|. \end{aligned}$$

By Proposition 6 and take $D = 3d$ and $m = 0$, with probability at least $1-\eta$, simultaneously for any $t \in \{T - t_k\}_{0 \le k \le N-1}$, we have

$$\bar{\varepsilon}_{\rm gen}^2 \lesssim \frac{\lambda_t^2}{\sigma_t^4} \cdot \sqrt{\frac{[d^2 + LdM] \cdot [L \cdot \log(LB) + \log(\lambda_t^{-1})] + \log(N/\eta)}{n}}. \tag{42}$$

To bound $\bar{\varepsilon}_{\rm app}^2$, by the identity that $\boldsymbol{s}_t(\boldsymbol{z}) = (\lambda_t \boldsymbol{m}_t(\boldsymbol{z}) - \boldsymbol{z})/\sigma_t^2$ and $\mathsf{P}_t \mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) = \mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + \sigma_t^{-2}\boldsymbol{z}) - \sigma_t^{-2}\boldsymbol{z}$, recalling $\tilde{\boldsymbol{m}}^L(\boldsymbol{z})$ as defined in Eq. (39), and by Lemma 8, we have

$$\begin{aligned}
\bar{\varepsilon}_{\rm app}^2 &= \inf_{\boldsymbol{W} \in \mathcal{W}} \mathbb{E}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \\
&= \inf_{\boldsymbol{W} \in \mathcal{W}} \mathbb{E}[\|\mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + \sigma_t^{-2}\boldsymbol{z}) - \lambda_t \sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d \\
&\le \mathbb{E}[\|\mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\lambda_t \sigma_t^{-2}\tilde{\boldsymbol{m}}^L(\boldsymbol{z})) - \lambda_t \sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d \\
&\lesssim \mathbb{E}[\|\mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\lambda_t \sigma_t^{-2}\tilde{\boldsymbol{m}}^L(\boldsymbol{z})) - \mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\lambda_t \sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z}))\|_2^2]/d \\
&\quad + \mathbb{E}[\|\mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\lambda_t \sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z})) - \lambda_t \sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d
\end{aligned} \tag{43}$$

where the last inequality uses the triangle inequality. By Eq. (40) and the 1-Lipschitzness of $\mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}$, the first quantity in the right-hand side is controlled by

$$\begin{aligned}
&\mathbb{E}[\|\mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\lambda_t \sigma_t^{-2}\tilde{\boldsymbol{m}}^L(\boldsymbol{z})) - \mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\lambda_t \sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z}))\|_2^2]/d \\
\lesssim\ & \frac{\lambda_t^2}{\sigma_t^4} \cdot (A^{2L} + \zeta^2 (1-A)^{-2}) \lesssim \frac{\lambda_t^2}{\sigma_t^4} \cdot \Big(A^{2L} + \frac{d^2}{(1-A)^2 M^2}\Big),
\end{aligned} \tag{44}$$

where the last inequality is by the fact that we can choose ζ such that $M = d \cdot (\lceil 2\zeta^{-1} \rceil + 3)$, which gives $\zeta \le 6d/M$. Furthermore, by Assumption 1 and by $\|\hat{\boldsymbol{m}}(\boldsymbol{z})\|_2 \le \sqrt{d}$, the second quantity in the right-hand side is controlled by

$$\mathbb{E}[\|\mathrm{proj}_{\lambda_t \sigma_t^{-2}\sqrt{d}}(\lambda_t \sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z})) - \lambda_t \sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d \lesssim \frac{\lambda_t^2}{\sigma_t^4} \cdot \varepsilon_{\mathrm{VI},t}^2(\boldsymbol{A}). \tag{45}$$

Combining Eq. (41), (42), (43), (44), (45) completes the proof of Theorem 1.

E.2 Proof of Corollary 1

Corollary 1 is a direct consequence of Theorem 1, Theorem 5, and Lemma 7.

E.3 Proofs for Section A.2

E.3.1 Proof of Lemma 1

Lemma 1 is a direct consequence of Lemma 9 below. Given Lemma 9, Lemma 1 holds by observing that when $\|\boldsymbol{A}\|_{\rm op} < 1/2$, we have $(1 - \|\boldsymbol{A}\|_{\rm op})^{-2} \le 4$.

Lemma 9. *Let $\boldsymbol{h} \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be symmetric with $\|\boldsymbol{A}\|_{\rm op} < 1/2$. Consider the Ising model $\mu(\sigma) \propto \exp\{\langle \boldsymbol{\sigma}, \boldsymbol{A}\boldsymbol{\sigma}\rangle/2 + \langle \boldsymbol{\sigma}, \boldsymbol{h}\rangle\}$ and denote $\boldsymbol{m} = \mathbb{E}_{\boldsymbol{\sigma}\sim\mu}[\boldsymbol{\sigma}]$. Let $\hat{\boldsymbol{m}}$ be the unique minimizer of the naive VB free energy*

$$\hat{\boldsymbol{m}} = \mathrm{argmin}_{\boldsymbol{m} \in [-1,1]^d} \Big\{ \sum_{i=1}^d -\mathsf{h}_{\rm bin}(m_i) - \langle \boldsymbol{m}, \boldsymbol{A}\boldsymbol{m}\rangle/2 - \langle \boldsymbol{m}, \boldsymbol{h}\rangle \Big\}.$$

Then we have

$$\frac{1}{d}\|\boldsymbol{m} - \hat{\boldsymbol{m}}\|_2^2 \le \frac{1}{(1 - 2\|\boldsymbol{A}\|_{\rm op})(1 - \|\boldsymbol{A}\|_{\rm op})^2} \frac{\|\boldsymbol{A}\|_F^2}{d}.$$

Proof of Lemma 9. Denote $\ell_i(\boldsymbol{\sigma}) = \sum_{j \ne i} A_{ij}\sigma_j + h_i$. Simple calculations yields $\mathbb{E}_\mu[\sigma_i | \{\sigma_j\}_{j\ne i}] = \tanh(\ell_i(\boldsymbol{\sigma}))$, which implies that

$$\mathbb{E}_\mu[\sigma_i] = \mathbb{E}_\mu[\tanh(\ell_i(\boldsymbol{\sigma}))].$$

By the fact that $\sup_{x\in\mathbb{R}} |(\mathrm{d}^2/\mathrm{d}x^2)\tanh(x)| \le 1$ and by Taylor's expansion, we have

$$|\mathbb{E}_\mu[\tanh(\ell_i(\boldsymbol{\sigma}))] - \tanh(\mathbb{E}_\mu[\ell_i(\boldsymbol{\sigma})])|^2 \le \mathrm{Var}_\mu(\ell_i(\boldsymbol{\sigma})).$$

By Theorem 1 of Eldan et al. [2022], the Ising model satisfies a Poincare's Inequality with Poincare's coefficient to be $1/(1-2\|\boldsymbol{A}\|_{\mathrm{op}})$ (we need to translate the Ising model to their setting, which leads to an additional 2 coefficient in front of $\|\boldsymbol{A}\|_{\mathrm{op}}$). Therefore, the Poincare's inequality implies that

$$\mathrm{Var}_\mu(\ell_i(\boldsymbol{\sigma})) \le \frac{1}{1-2\|\boldsymbol{A}\|_{\mathrm{op}}}\sum_{j\neq i} A_{ij}^2.$$

Combining the equations above, we get

$$\frac{1}{d}\Big\|\boldsymbol{m} - \tanh(\boldsymbol{A}\boldsymbol{m}+\boldsymbol{h})\Big\|_2^2 = \frac{1}{d}\sum_{i=1}^d \big(\mathbb{E}_\mu[\sigma_i] - \tanh(\mathbb{E}_\mu[\ell_i(\boldsymbol{\sigma})])\big)^2 \le \frac{1}{1-2\|\boldsymbol{A}\|_{\mathrm{op}}}\frac{\|\boldsymbol{A}\|_F^2}{d} \equiv \varepsilon^2.$$

Furthermore, notice that $\hat{\boldsymbol{m}}$ is the unique minimizer of the naive VB free energy implies that $\hat{\boldsymbol{m}} = \tanh(\boldsymbol{A}\hat{\boldsymbol{m}}+\boldsymbol{h})$. Therefore, by the equation above, we get

$$\begin{aligned}
\varepsilon &\ge \frac{1}{\sqrt{d}}\big\|(\boldsymbol{m}-\hat{\boldsymbol{m}}) - (\tanh(\boldsymbol{A}\boldsymbol{m}+\boldsymbol{h}) - \tanh(\boldsymbol{A}\hat{\boldsymbol{m}}+\boldsymbol{h}))\big\|_2 \\
&\ge \frac{1}{\sqrt{d}}\Big(\|\boldsymbol{m}-\hat{\boldsymbol{m}}\|_2 - \|\tanh(\boldsymbol{A}\boldsymbol{m}+\boldsymbol{h}) - \tanh(\boldsymbol{A}\hat{\boldsymbol{m}}+\boldsymbol{h})\|_2\Big) \\
&\ge (1-\|\boldsymbol{A}\|_{\mathrm{op}})\cdot\frac{1}{\sqrt{d}}\|\boldsymbol{m}-\hat{\boldsymbol{m}}\|_2.
\end{aligned}$$

Combining the equations above concludes the proof of Lemma 9. □

E.3.2 Proof of Lemma 2

Lemma 2 is a direct consequence of the lemma below.

Lemma 10 (Lemma 4.10 and Proposition 4.2 of El Alaoui et al. [2022])**.** *Let $\boldsymbol{J} \sim \mathrm{GOE}(d)$ and $\beta < 1/2$. Let $\boldsymbol{x} \sim \mu(\boldsymbol{x}) \propto \exp\{\beta\langle\boldsymbol{x},\boldsymbol{J}\boldsymbol{x}\rangle/2\}$ on $\{\pm1\}^d$ and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0},\mathbf{I}_d)$ independently. Let $\boldsymbol{z} = \lambda\boldsymbol{x} + \sigma\boldsymbol{g}$. Consider the posterior measure*

$$\mu(\boldsymbol{x}|\boldsymbol{z}) \propto \exp\{\beta\langle\boldsymbol{x},\boldsymbol{J}\boldsymbol{x}\rangle/2 + (\lambda/\sigma^2)\langle\boldsymbol{x},\boldsymbol{z}\rangle\},$$

and define $\boldsymbol{m}(\boldsymbol{z}) = \sum_{\boldsymbol{x}\in\{\pm1\}^d}\boldsymbol{x}\mu(\boldsymbol{x}|\boldsymbol{z})$. Furthermore, consider the TAP free energy

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m};\boldsymbol{z},q) = \sum_{i=1}^d -\mathsf{h}_{\mathrm{bin}}(m_i) - \frac{\beta}{2}\langle\boldsymbol{m},\boldsymbol{J}\boldsymbol{m}\rangle - \frac{\lambda}{\sigma^2}\langle\boldsymbol{z},\boldsymbol{m}\rangle + \frac{\beta^2(1-q)}{2}\|\boldsymbol{m}\|_2^2,$$

take $q_\star = q_\star(\beta,\lambda,\sigma)$ to be the unique solution of

$$q_\star = \mathbb{E}_{G\sim\mathcal{N}(0,1)}\big[\tanh^2(\beta^2 q_\star + (\lambda^2/\sigma^2) + \sqrt{\beta^2 q_\star + (\lambda^2/\sigma^2)}G)\big],$$

and define $\hat{\boldsymbol{m}}(\boldsymbol{z}) = \mathrm{argmin}_{\boldsymbol{m}\in[-1,1]^d}\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m};\boldsymbol{z},q)$ to be the unique minimizer. Then we have

$$\|\boldsymbol{m}(\boldsymbol{z}) - \hat{\boldsymbol{m}}(\boldsymbol{z})\|_2^2/d \xrightarrow{p} 0.$$

Remark 2. *We discuss the several seeming differences between Lemma 10 and [El Alaoui et al., 2022, Lemma 4.10].*

- *The parameter λ^2/σ^2 in Lemma 10 maps to the parameter t in [El Alaoui et al., 2022, Lemma 4.10]. The variable $(\lambda/\sigma^2)\boldsymbol{z} = (\lambda^2/\sigma^2)\boldsymbol{x} + (\lambda/\sigma)\boldsymbol{g}$ in Lemma 10 maps to the variable $\boldsymbol{y} \stackrel{d}{=} t\boldsymbol{x} + \sqrt{t}\cdot\boldsymbol{g}$ in [El Alaoui et al., 2022, Lemma 4.10].*
- *Lemma 10 takes $\hat{\boldsymbol{m}}(\boldsymbol{z})$ to be the unique global minimizer of $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m};\boldsymbol{z},q_\star)$, whereas [El Alaoui et al., 2022, Lemma 4.10] takes $\hat{\boldsymbol{m}}(\boldsymbol{z})$ to be a particular local minimizer of $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m};\boldsymbol{z},q_\star)$. However, when $\beta < 1/2$, it can be shown that $\mathcal{F}_{\mathrm{TAP}}$ is strongly convex with high probability, and hence the local minimizer is the global minimizer with high probability.*

- *[El Alaoui et al., 2022, Lemma 4.10] is proven under a different joint distribution of $(\boldsymbol{J}, \boldsymbol{z})$. However, [El Alaoui et al., 2022, Proposition 4.2] shows that the distribution for [El Alaoui et al., 2022, Lemma 4.10] is contiguous to the distribution for Lemma 10, and hence the high probability event under the sampling distribution of [El Alaoui et al., 2022, Lemma 4.10] can be translated to the corresponding high probability event under the sampling distribution of Lemma 10.*

To prove Lemma 2, we take $c_t = \beta^2(1-q_t)$ where q_t is the unique solution of

$$q_t = \mathbb{E}_{G\sim\mathcal{N}(0,1)}\big[\tanh^2(\beta^2 q_t + (\lambda_t^2/\sigma_t^2) + \sqrt{\beta^2 q_t + (\lambda_t^2/\sigma_t^2)}G)\big].$$

Hence by Lemma 10, for any $\beta < 1/2$, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mu_t}[\|\hat{\boldsymbol{m}}_t(\boldsymbol{z}) - \boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d \xrightarrow{p} 0, \quad d\to\infty.$$

Furthermore, note that $c_t \le \beta^2$ and $\|\beta\boldsymbol{J}\|_{\rm op} \le 2\beta + \varepsilon$ with high probability for arbitrarily small ε. This ensures that $\|\beta\boldsymbol{J} - c_t\mathbf{I}_d\|_{\rm op} \le \|\beta\boldsymbol{J}\|_{\rm op} + \beta^2 < 1$ when $\beta \le 1/4$. This proves Lemma 2.

F Proofs for Section B: Generalization to other models

F.1 Proof of Theorem 2

Approximate the minimizer of the free energy via iterative algorithms

Once again, we first prove that we can approximately minimize the free energy by implementing a simple iterative algorithm. Recall that

$$\begin{aligned}
&\hat{\boldsymbol{\omega}}_t(\boldsymbol{z}) = \mathrm{argmin}_{\boldsymbol{\omega}\in[-1,1]^{d+m}}\mathcal{F}_t^{\rm marginal}(\boldsymbol{\omega};\boldsymbol{z}),\\
&\mathcal{F}_t^{\rm marginal}(\boldsymbol{\omega};\boldsymbol{z}) := \Big\{\sum_{i=1}^{d+m} -\mathsf{h}_{\rm bin}(\omega_i) - \frac{1}{2}\langle\boldsymbol{\omega},\boldsymbol{A}\boldsymbol{\omega}\rangle - \frac{\lambda_t}{\sigma_t^2}\langle\boldsymbol{z},\boldsymbol{\omega}_{1:d}\rangle + \frac{1}{2}\langle\boldsymbol{\omega},\boldsymbol{K}\boldsymbol{\omega}\rangle\Big\}.
\end{aligned}$$

Taking the gradient and the Hessian of $\mathcal{F}_t^{\rm marginal}(\boldsymbol{\omega};\boldsymbol{z})$, we obtain

$$\begin{aligned}
&\nabla_{\boldsymbol{\omega}}\mathcal{F}_t^{\rm marginal}(\boldsymbol{\omega};\boldsymbol{z}) = \tanh^{-1}(\boldsymbol{\omega}) + (\boldsymbol{K}-\boldsymbol{A})\boldsymbol{\omega} - \frac{\lambda_t}{\sigma_t^2}[\boldsymbol{z};\boldsymbol{0}_m]^{\mathsf{T}},\\
&\nabla_{\boldsymbol{\omega}}^2\mathcal{F}_t^{\rm marginal}(\boldsymbol{\omega};\boldsymbol{z}) = \mathrm{diag}\{(1-\omega_i^2)^{-1}\}_{i\in[d+m]} + \boldsymbol{K} - \boldsymbol{A}.
\end{aligned}$$

Since $\|\boldsymbol{K}-\boldsymbol{A}\|_{\rm op} \le A < 1$, we can then conclude that $\nabla_{\boldsymbol{w}}^2\mathcal{F}_t^{\rm marginal}(\boldsymbol{w};\boldsymbol{z}) \succeq (1-A)\mathbf{I}_{d+m}$ for all $\boldsymbol{z}\in\mathbb{R}^d$, hence $\mathcal{F}_t^{\rm marginal}(\cdot;\boldsymbol{z})$ is strongly-convex for all $\boldsymbol{z}\in\mathbb{R}^d$. This further implies that the fixed-point equation below

$$\boldsymbol{\omega} = \tanh\left((\boldsymbol{A}-\boldsymbol{K})\boldsymbol{\omega} + \frac{\lambda_t}{\sigma_t^2}[\boldsymbol{z};\boldsymbol{0}_m]^{\mathsf{T}}\right)$$

has a unique solution. By Lemma 6, we obtain that if we run the iteration

$$\tilde{\boldsymbol{\omega}}^0(\boldsymbol{z}) = \boldsymbol{0}, \qquad \tilde{\boldsymbol{\omega}}^k(\boldsymbol{z}) = f((\boldsymbol{A}-\boldsymbol{K})\tilde{\boldsymbol{\omega}}^{k-1}(\boldsymbol{z}) + \lambda_t\sigma_t^{-2}[\boldsymbol{z};\boldsymbol{0}_m]^{\mathsf{T}}), \tag{46}$$

where $\|f(\cdot)-\tanh(\cdot)\|_\infty \le \zeta$, then

$$\frac{1}{\sqrt{d+m}}\|\tilde{\boldsymbol{\omega}}^k(\boldsymbol{z}) - \hat{\boldsymbol{\omega}}_t(\boldsymbol{z})\|_2 \le A^k + \zeta(1-A)^{-1}. \tag{47}$$

In particular, we require that $f(\cdot)$ is the function that we construct in Lemma 5.

Represent the iterative algorithm as a ResNet

Recall that $\hat{\boldsymbol{m}}_t(\boldsymbol{z}) = [\hat{\boldsymbol{\omega}}_t(\boldsymbol{z})]_{1:d}$. We define $\tilde{\boldsymbol{m}}^\ell(\boldsymbol{z}) := [\tilde{\boldsymbol{\omega}}^{(\ell)}(\boldsymbol{z})]_{1:d}$. In what follows, we show that $(\lambda_t\tilde{\boldsymbol{m}}^\ell(\boldsymbol{z}) - \boldsymbol{z})/\sigma_t^2$ can be expressed as a ResNet that takes input $\boldsymbol{z}$.

Lemma 11. *For all $\ell \in \mathbb{N}_+$ and $\delta \le t \le T$, there exists $\boldsymbol{W} \in \mathcal{W}_{d,D,\ell,M,B}$ with*

$$D = 3(d+m), \quad M = (\lceil 2\zeta^{-1} \rceil + 1)(d+m),$$
$$B = (\lceil 2\zeta^{-1} \rceil - 1) \cdot (\log\lceil \zeta^{-1} \rceil + 4) + 8 + \sqrt{d+m} + (1 - e^{-2\delta})^{-1},$$

such that $(\lambda_t \tilde{\boldsymbol{m}}^\ell(\boldsymbol{z}) - \boldsymbol{z})/\sigma_t^2 = \mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z})$, where $\tilde{\boldsymbol{m}}^\ell$ is as defined in Eq. (46).

Proof of Lemma 11. Recall the definition of f as an approximation of $\tanh$ as in Lemma 5. The proof of this lemma is similar to that of Lemma 8. To be specific, we will select the weight matrices $\{\boldsymbol{W}_1^{(\ell)}, \boldsymbol{W}_2^{(\ell)}, \boldsymbol{W}_{\mathrm{in}}, \boldsymbol{W}_{\mathrm{out}}\}$ appropriately such that $\boldsymbol{u}^{(\ell)} = [\tilde{\boldsymbol{\omega}}^\ell(\boldsymbol{z}); \sigma_t^{-2}[\boldsymbol{z}; \mathbf{0}_m]^\mathsf{T}; \mathbf{1}_{d+m}]^\mathsf{T} \in \mathbb{R}^{3(d+m)}$. When $\ell = 0$, this can be achieved by setting

$$\boldsymbol{W}_{\mathrm{in}} = \begin{bmatrix} \mathbf{0}_{d\times(d+m)} & \sigma_t^{-2}[\mathbf{I}_d, \mathbf{0}_{d\times m}] & \mathbf{0}_{d\times(d+m)} \\ \mathbf{0}_{1\times(d+m)} & \mathbf{0}_{1\times(d+m)} & \mathbf{1}_{1\times(d+m)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times 3(d+m)}.$$

Also, recall that $f(x) = \sum_{j=1}^{\lceil 2\zeta^{-1}\rceil - 1} \mathrm{ReLU}(x - w_j) + a_0$. Therefore, for $\ell \in \mathbb{N}_+$, we simply set

$$\boldsymbol{W}_1^{(\ell)} = \begin{bmatrix} a_i\mathbf{I}_d & \cdots & a_{\lceil 2\zeta^{-1}\rceil - 1}\mathbf{I}_{d+m} & -\mathbf{I}_{d+m} & \mathbf{I}_{d+m} & a_0\mathbf{I}_{d+m} & -a_0\mathbf{I}_{d+m} \\ \mathbf{0}_{(d+m)\times(d+m)} & \cdots & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} \\ \mathbf{0}_{(d+m)\times(d+m)} & \cdots & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} \end{bmatrix}$$
$$\in \mathbb{R}^{3(d+m)\times(\lceil 2\zeta^{-1}\rceil + 3)(d+m)},$$

$$\boldsymbol{W}_2^{(\ell)} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{K} & \cdots & \boldsymbol{A} - \boldsymbol{K} & \mathbf{I}_{d+m} & -\mathbf{I}_{d+m} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} \\ \lambda_t\mathbf{I}_{d+m} & \cdots & \lambda_t\mathbf{I}_{d+m} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} \\ -w_1\mathbf{I}_{d+m} & \cdots & -w_{\lceil 2\zeta^{-1}\rceil - 1}\mathbf{I}_{d+m} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{0}_{(d+m)\times(d+m)} & \mathbf{I}_{d+m} & -\mathbf{I}_{d+m} \end{bmatrix}^\mathsf{T}$$
$$\in \mathbb{R}^{(\lceil 2\zeta^{-1}\rceil + 3)(d+m)\times 3(d+m)}.$$

Finally, we take $\boldsymbol{W}_{\mathrm{out}} = [\lambda_t\sigma_t^{-2}\mathbf{I}_d, \mathbf{0}_{d\times m}, -\mathbf{I}_d, \mathbf{0}_{d\times(d+2m)}] \in \mathbb{R}^{d\times 3(d+m)}$.

Next, we upper bound the norm of the residual network. By Lemma 5 and Remark 1, we have $\sum_{j=1}^{\lceil 2\zeta^{-1}\rceil - 1} |a_j| \le 2$, $|a_0| \le 1$, $|w_j| \le \log\lceil \zeta^{-1}\rceil$. Therefore,

$$\|\boldsymbol{W}_{\mathrm{in}}\|_{\mathrm{op}} \le \sqrt{d+m} + \sigma_t^{-2}, \qquad \|\boldsymbol{W}_{\mathrm{out}}\|_{\mathrm{op}} \le 1 + \lambda_t\sigma_t^{-2},$$
$$\|\boldsymbol{W}_1^{(\ell)}\|_{\mathrm{op}} \le 2\lceil 2\zeta^{-1}\rceil + 2, \qquad \|\boldsymbol{W}_2^{(\ell)}\|_{\mathrm{op}} \le (\lceil 2\zeta^{-1}\rceil - 1) \cdot (\log\lceil \zeta^{-1}\rceil + 2) + 4.$$

This implies that

$$\|\!|\boldsymbol{W}|\!\| \le B = (\lceil 2\zeta^{-1}\rceil - 1) \cdot (\log\lceil \zeta^{-1}\rceil + 4) + 8 + \sqrt{d+m} + (1 - e^{-2\delta})^{-1}.$$

This completes the proof of Lemma 11. □

Proof of Theorem 2

Similar to the proof of Theorem 1, we obtain

$$\mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \le \bar{\varepsilon}_{\mathrm{app}}^2 + \bar{\varepsilon}_{\mathrm{gen}}^2, \tag{48}$$

where $\bar{\varepsilon}_{\mathrm{app}}^2$ is the approximation error and $\bar{\varepsilon}_{\mathrm{gen}}^2$ is the generalization error,

$$\bar{\varepsilon}_{\mathrm{app}}^2 = \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\mathsf{P}_t\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d,$$
$$\bar{\varepsilon}_{\mathrm{gen}}^2 = 2\sup_{\boldsymbol{W}\in\mathcal{W}} \Big|\hat{\mathbb{E}}[\|\mathsf{P}_t\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d - \mathbb{E}[\|\mathsf{P}_t\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d\Big|.$$

By Proposition 6 and take $D = 3(d+m)$, with probability at least $1 - \eta$, simultaneously for any $t \in \{T - t_k\}_{0\le k\le N-1}$, we have

$$\bar{\varepsilon}_{\mathrm{gen}}^2 \lesssim \frac{\lambda_t^2}{\sigma_t^4} \cdot \sqrt{\frac{[ML + d](d+m) \cdot [L \cdot \log(LB) + \log(\lambda_t^{-1})] + \log(N/\eta)}{n}}. \tag{49}$$

To bound $\bar{\varepsilon}^2_{\rm app}$, by the identity that $\boldsymbol{s}_t(\boldsymbol{z}) = (\lambda_t \boldsymbol{m}_t(\boldsymbol{z}) - \boldsymbol{z})/\sigma_t^2$ and $\mathsf{P}_t\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) = \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + \sigma_t^{-2}\boldsymbol{z}) - \sigma_t^{-2}\boldsymbol{z}$, recalling $\tilde{\boldsymbol{m}}^L(\boldsymbol{z}) = \tilde{\boldsymbol{\omega}}^L_{1:d}(\boldsymbol{z})$ as defined in Eq. (46), and by Lemma 11, we have

$$\begin{aligned}\bar{\varepsilon}^2_{\rm app} = &\inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\mathsf{P}_t\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \\ &\lesssim \mathbb{E}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\tilde{\boldsymbol{m}}^L(\boldsymbol{z})) - \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z}))\|_2^2]/d \\ &\quad + \mathbb{E}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z})) - \lambda_t\sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d.\end{aligned} \tag{50}$$

By Eq. (47), the 1-Lipschitzness of $\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}$, and the definition that $\hat{\boldsymbol{m}}_t(\boldsymbol{z}) = [\hat{\boldsymbol{\omega}}_t(\boldsymbol{z})]_{1:d}$ and $\tilde{\boldsymbol{m}}^\ell(\boldsymbol{z}) = [\tilde{\boldsymbol{\omega}}^{(\ell)}(\boldsymbol{z})]_{1:d}$, the first quantity on the right-hand side is controlled by

$$\begin{aligned}&\mathbb{E}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\tilde{\boldsymbol{m}}^L(\boldsymbol{z})) - \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z}))\|_2^2]/d \\ \lesssim &\frac{d+m}{d}\cdot\frac{\lambda_t^2}{\sigma_t^4}\cdot(A^{2L} + \zeta^2(1-A)^{-2}) \lesssim \frac{d+m}{d}\cdot\frac{\lambda_t^2}{\sigma_t^4}\cdot\Big(A^{2L} + \frac{(d+m)^2}{(1-A)^2M^2}\Big),\end{aligned} \tag{51}$$

where the last inequality is by the fact that we can choose ζ such that $\zeta \le 6(d+m)/M$. Furthermore, by Assumption 2 and by $\|\hat{\boldsymbol{m}}(\boldsymbol{z})\|_2 \le \sqrt{d}$, the second quantity in the right-hand side is controlled by

$$\mathbb{E}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z})) - \lambda_t\sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z})\|_2^2]/d \lesssim \frac{\lambda_t^2}{\sigma_t^4}\cdot\varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}). \tag{52}$$

Combining Eq. (48), (49), (50), (51), (52) completes the proof of the score estimation result in Theorem 2. The KL divergence bound is a direct consequence of score estimation error, Theorem 5, and Lemma 7. This concludes the proof.

F.2 Proof of Theorem 3

Approximate the minimizer of the free energy via iterative algorithm

We define

$$\mathcal{F}_t^{\rm cond}(\boldsymbol{m};\boldsymbol{z},\boldsymbol{\theta}) := \sum_{i=1}^d -\mathrm{h}_{\rm bin}(m_i) - \frac{1}{2}\langle\boldsymbol{m},\boldsymbol{A}_{11}\boldsymbol{m}\rangle - \langle\boldsymbol{m},\boldsymbol{A}_{12}\boldsymbol{\theta}\rangle - \frac{\lambda_t}{\sigma_t^2}\langle\boldsymbol{z},\boldsymbol{m}\rangle + \frac{1}{2}\langle\boldsymbol{m},\boldsymbol{K}\boldsymbol{m}\rangle.$$

Taking the gradient and the Hessian of $\mathcal{F}_t^{\rm cond}$, we obtain

$$\begin{aligned}\nabla_{\boldsymbol{m}}\mathcal{F}_t^{\rm cond}(\boldsymbol{m};\boldsymbol{z},\boldsymbol{\theta}) &= \tanh^{-1}(\boldsymbol{m}) + (\boldsymbol{K}-\boldsymbol{A}_{11})\boldsymbol{m} - \boldsymbol{A}_{12}\boldsymbol{\theta} - \frac{\lambda_t}{\sigma_t^2}\boldsymbol{z}, \\ \nabla^2_{\boldsymbol{m}}\mathcal{F}_t^{\rm cond}(\boldsymbol{m};\boldsymbol{z},\boldsymbol{\theta}) &= \mathrm{diag}\{((1-m_i^2)^{-1})_{i\in[d]}\} + \boldsymbol{K} - \boldsymbol{A}_{11}.\end{aligned}$$

When $\|\boldsymbol{K}-\boldsymbol{A}_{11}\|_{\rm op} \le A < 1$, we always have $\nabla^2_{\boldsymbol{m}}\mathcal{F}_t^{\rm cond}(\boldsymbol{m};\boldsymbol{z},\boldsymbol{\theta}) \succeq (1-A)\mathbf{I} \succ 0$. That is to say, $\mathcal{F}_t^{\rm cond}(\cdot;\boldsymbol{z},\boldsymbol{\theta})$ is strongly convex, hence

$$\boldsymbol{m} = \tanh\left((\boldsymbol{A}_{11}-\boldsymbol{K})\boldsymbol{m} + \boldsymbol{A}_{12}\boldsymbol{\theta} + \frac{\lambda_t}{\sigma_t^2}\boldsymbol{z}\right)$$

has a unique solution. We then can apply Lemma 6, and conclude that if we run iteration

$$\tilde{\boldsymbol{m}}^0(\boldsymbol{z};\boldsymbol{\theta}) = \mathbf{0}, \qquad \tilde{\boldsymbol{m}}^\ell(\boldsymbol{z};\boldsymbol{\theta}) = f((\boldsymbol{A}_{11}-\boldsymbol{K})\tilde{\boldsymbol{m}}^{\ell-1}(\boldsymbol{z};\boldsymbol{\theta}) + \boldsymbol{A}_{12}\boldsymbol{\theta} + \lambda_t\sigma_t^{-2}\boldsymbol{z}) \tag{53}$$

for some $\|f-\tanh\|_\infty \le \zeta$, it then holds that

$$\frac{1}{\sqrt{d}}\|\tilde{\boldsymbol{m}}^\ell(\boldsymbol{z};\boldsymbol{\theta}) - \hat{\boldsymbol{m}}_t(\boldsymbol{z};\boldsymbol{\theta})\|_2 \le A^\ell + \zeta(1-A)^{-1}. \tag{54}$$

As usual, we require $f(\cdot)$ satisfies all other conditions from Lemma 5.

Represent the iterative algorithm as a ResNet

Next, we show that $(\lambda_t \tilde{\boldsymbol{m}}^\ell(\boldsymbol{z};\boldsymbol{\theta}) - \boldsymbol{z})/\sigma_t^2$ can be expressed as a ResNet as in (ResNet-Conditional) that has input $(\boldsymbol{z}, \boldsymbol{\theta})$.

Lemma 12. *For all $\ell \in \mathbb{N}_+$ and $\delta \le t \le T$, there exists $\boldsymbol{W} \in \mathcal{W}_{d,m,D,\ell,M,B}$ with*

$$D = 4d, \quad M = (\lceil 2\zeta^{-1} \rceil + 3)d,$$
$$B = (\lceil 2\zeta^{-1} \rceil - 1) \cdot (\log\lceil \zeta^{-1} \rceil + 4 + \|\boldsymbol{A}_{12}\|_{\rm op}) + 8 + (1 - e^{-2\delta})^{-1} + \|\boldsymbol{A}_{12}\|_{\rm op} + \sqrt{d},$$

such that $(\lambda_t \tilde{\boldsymbol{m}}^\ell(\boldsymbol{z};\boldsymbol{\theta}) - \boldsymbol{z})/\sigma_t^2 = \mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z}, \boldsymbol{\theta})$, where $\tilde{\boldsymbol{m}}^\ell$ is as defined in Eq. (53).

Proof of Lemma 12. Recall the definition of f as an approximation of $\tanh$ as in Lemma 5. We shall choose the weight matrices such that $\boldsymbol{u}^{(\ell)} = [\tilde{\boldsymbol{m}}^\ell(\boldsymbol{z};\boldsymbol{\theta}); \sigma_t^{-2}\boldsymbol{z}; \boldsymbol{A}_{12}\boldsymbol{\theta}; \mathbf{1}_d] \in \mathbb{R}^{4d}$. For $\ell = 0$, we simply set

$$\boldsymbol{W}_{\rm in} = \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times 1} \\ \sigma_t^{-2}\mathbf{I}_d & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times 1} \\ \mathbf{0}_{d\times d} & \boldsymbol{A}_{12} & \mathbf{0}_{d\times 1} \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times m} & \mathbf{1}_{d\times 1} \end{bmatrix} \in \mathbb{R}^{4d\times(d+m+1)}.$$

For $\ell \ge 1$, we let

$$\boldsymbol{W}_1^{(\ell)} = \begin{bmatrix} a_i\mathbf{I}_d & \cdots & a_{\lceil 2\zeta^{-1}\rceil - 1}\mathbf{I}_d & -\mathbf{I}_d & \mathbf{I}_d & a_0\mathbf{I}_d & -a_0\mathbf{I}_d \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \end{bmatrix} \in \mathbb{R}^{4d\times(\lceil 2\zeta^{-1}\rceil+3)d},$$

$$\boldsymbol{W}_2^{(\ell)} = \begin{bmatrix} \boldsymbol{A}_{11} - \boldsymbol{K} & \cdots & \boldsymbol{A}_{11} - \boldsymbol{K} & \mathbf{I}_d & -\mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \lambda_t\mathbf{I}_d & \cdots & \lambda_t\mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{12} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ -w_1\mathbf{I}_d & \cdots & -w_{\lceil 2\zeta^{-1}\rceil - 1}\mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{I}_d & -\mathbf{I}_d \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{(\lceil 2\zeta^{-1}\rceil+3)d\times 4d}.$$

Finally, we let $\boldsymbol{W}_{\rm out} = [\lambda_t\sigma_t^{-2}\mathbf{I}_d, -\mathbf{I}_d, \mathbf{0}_{d\times d}, \mathbf{0}_{d\times d}] \in \mathbb{R}^{d\times 4d}$. By Lemma 5 and Remark 1, we have $\sum_{j=1}^{\lceil 2\zeta^{-1}\rceil - 1} |a_j| \le 2$, $|a_0| \le 1$, $|w_j| \le \log\lceil \zeta^{-1}\rceil$. Therefore,

$$\|\boldsymbol{W}_{\rm out}\|_{\rm op} \le \lambda_t\sigma_t^{-2} + 1, \qquad \|\boldsymbol{W}_{\rm in}\|_{\rm op} \le \sqrt{d} + \sigma_t^{-2} + \|\boldsymbol{A}_{12}\|_{\rm op},$$
$$\|\boldsymbol{W}_1^{(\ell)}\|_{\rm op} \le 2\lceil 2\zeta^{-1}\rceil + 2, \qquad \|\boldsymbol{W}_2^{(\ell)}\|_{\rm op} \le (\lceil 2\zeta^{-1}\rceil - 1)\cdot(\log\lceil\zeta^{-1}\rceil + 2 + \|\boldsymbol{A}_{12}\|_{\rm op}) + 4.$$

As a result, we conclude that

$$|||\boldsymbol{W}||| \le B = (\lceil 2\zeta^{-1}\rceil - 1)\cdot(\log\lceil\zeta^{-1}\rceil + 4 + \|\boldsymbol{A}_{12}\|_{\rm op}) + 8 + (1 - e^{-2\delta})^{-1} + \|\boldsymbol{A}_{12}\|_{\rm op} + \sqrt{d}.$$

We have completed the proof of Lemma 12. □

Proof of Theorem 3

Similar to the proof of Theorem 1, we obtain

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z};\boldsymbol{\theta}) - \boldsymbol{s}_t(\boldsymbol{z};\boldsymbol{\theta})\|_2^2]/d \le \bar{\varepsilon}_{\rm app}^2 + \bar{\varepsilon}_{\rm gen}^2, \tag{55}$$

where $\bar{\varepsilon}_{\rm app}^2$ is the approximation error and $\bar{\varepsilon}_{\rm gen}^2$ is the generalization error,

$$\bar{\varepsilon}_{\rm app}^2 = \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z},\boldsymbol{\theta}) - \boldsymbol{s}_t(\boldsymbol{z};\boldsymbol{\theta})\|_2^2]/d,$$
$$\bar{\varepsilon}_{\rm gen}^2 = 2\sup_{\boldsymbol{W}\in\mathcal{W}} \Big|\hat{\mathbb{E}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z},\boldsymbol{\theta}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d - \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z},\boldsymbol{\theta}) + \boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d\Big|.$$

By Proposition 6 and take $D = 4d$, with probability at least $1 - \eta$, simultaneously for any $t \in \{T - t_k\}_{0\le k\le N-1}$, we have

$$\bar{\varepsilon}_{\rm gen}^2 \lesssim \frac{\lambda_t^2}{\sigma_t^4}\cdot\sqrt{\frac{(MdL + d(d+m))\cdot[L\cdot\log(LBd^{-1}(m+d)) + \log(\lambda_t^{-1})] + \log(N/\eta)}{n}}. \tag{56}$$

To bound $\bar{\varepsilon}^2_{\rm app}$, by the identity that $\boldsymbol{s}_t(\boldsymbol{z};\boldsymbol{\theta}) = (\lambda_t \boldsymbol{m}_t(\boldsymbol{z};\boldsymbol{\theta}) - \boldsymbol{z})/\sigma_t^2$ and $\mathsf{P}_t\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z},\boldsymbol{\theta}) = \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z},\boldsymbol{\theta}) + \sigma_t^{-2}\boldsymbol{z}) - \sigma_t^{-2}\boldsymbol{z}$, recalling $\tilde{\boldsymbol{m}}^L(\boldsymbol{z})$ as defined in Eq. (53), and by Lemma 12, we have

$$\begin{aligned}\bar{\varepsilon}^2_{\rm app} &= \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\mathsf{P}_t[\mathrm{ResN}_{\boldsymbol{W}}](\boldsymbol{z},\boldsymbol{\theta}) - \boldsymbol{s}_t(\boldsymbol{z};\boldsymbol{\theta})\|_2^2]/d \\ &\lesssim \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\tilde{\boldsymbol{m}}^L(\boldsymbol{z};\boldsymbol{\theta})) - \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z};\boldsymbol{\theta}))\|_2^2]/d \\ &\quad + \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z};\boldsymbol{\theta})) - \lambda_t\sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z};\boldsymbol{\theta})\|_2^2]/d.\end{aligned} \tag{57}$$

By Eq. (54) and the 1-Lipschitzness of $\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}$, the first quantity on the right-hand side is controlled by

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\tilde{\boldsymbol{m}}^L(\boldsymbol{z};\boldsymbol{\theta})) - \mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z};\boldsymbol{\theta}))\|_2^2]/d \\ &\lesssim \frac{\lambda_t^2}{\sigma_t^4}\cdot(A^{2L} + \zeta^2(1-A)^{-2}) \lesssim \frac{\lambda_t^2}{\sigma_t^4}\cdot\Big(A^{2L} + \frac{d^2}{(1-A)^2M^2}\Big),\end{aligned} \tag{58}$$

where the last inequality is by the fact that we can choose ζ such that $\zeta \le 6d/M$. Furthermore, by Assumption 3 and by $\|\hat{\boldsymbol{m}}(\boldsymbol{z};\boldsymbol{\theta})\|_2 \le \sqrt{d}$, the second quantity in the right-hand side is controlled by

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[\|\mathrm{proj}_{\lambda_t\sigma_t^{-2}\sqrt{d}}(\lambda_t\sigma_t^{-2}\hat{\boldsymbol{m}}(\boldsymbol{z};\boldsymbol{\theta})) - \lambda_t\sigma_t^{-2}\boldsymbol{m}_t(\boldsymbol{z};\boldsymbol{\theta})\|_2^2]/d \lesssim \frac{\lambda_t^2}{\sigma_t^4}\cdot\varepsilon^2_{\mathrm{VI},t}(\boldsymbol{A}). \tag{59}$$

Combining Eq. (55), (56), (57), (58), (59) completes the proof of the score estimation result in Theorem 3. The KL divergence bound is a direct consequence of score estimation error, Theorem 5, and Lemma 7. This concludes the proof.

To prove the second result of the bound of the expected KL divergence, we simply notice that by Theorem 5, conditioning on every $\boldsymbol{\theta}$ we have

$$\frac{1}{d}\mathrm{KL}(\mu_\delta(\cdot|\boldsymbol{\theta}), \hat{\mu}(\cdot|\boldsymbol{\theta})) \lesssim \varepsilon^2 + \kappa^2 N + \kappa T + e^{-2T},$$

where

$$\varepsilon^2 = \frac{1}{d}\sum_{k=0}^{N-1}\gamma_k\mathbb{E}\left[\|\hat{\boldsymbol{s}}_{T-t_k}(\boldsymbol{z};\boldsymbol{\theta}) - \boldsymbol{s}_{T-t_k}(\boldsymbol{z};\boldsymbol{\theta})\|_2^2 \mid \boldsymbol{\theta}\right].$$

The proof is complete of Theorem 3 by simply integrating over $\boldsymbol{\theta}$.

F.3 Proof of Theorem 4

Relationship of the score function $\boldsymbol{s}_t$ to the denoiser $\boldsymbol{e}_t$

We first compute the score function $\boldsymbol{s}_t(\boldsymbol{z}) = \nabla_{\boldsymbol{z}}\log\mu_t(\boldsymbol{z})$, for $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{z} = \lambda_t\boldsymbol{x} + \sigma_t\boldsymbol{g}$, where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}_d, \mathbf{I}_d)$ is independent of $(\boldsymbol{\theta},\boldsymbol{\varepsilon}) \sim \pi_0^m \otimes \mathcal{N}(\boldsymbol{0}, \tau^2\mathbf{I}_d)$. Note that

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}\mid\boldsymbol{z}] &= \mathbb{E}[\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \mid \lambda_t\boldsymbol{A}\boldsymbol{\theta} + \lambda_t\boldsymbol{\varepsilon} + \sigma_t\boldsymbol{g}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{\theta}\mid\boldsymbol{z}] + \mathbb{E}[\boldsymbol{\varepsilon} \mid \lambda_t\boldsymbol{A}\boldsymbol{\theta} + \lambda_t\boldsymbol{\varepsilon} + \sigma_t\boldsymbol{g}] \\ &= \boldsymbol{A}\mathbb{E}[\boldsymbol{\theta}\mid\boldsymbol{z}] + \frac{\lambda_t\tau^2}{\lambda_t^2\tau^2 + \sigma_t^2}\mathbb{E}[\lambda_t\boldsymbol{\varepsilon} + \sigma_t\boldsymbol{g} \mid \lambda_t\boldsymbol{A}\boldsymbol{\theta} + \lambda_t\boldsymbol{\varepsilon} + \sigma_t\boldsymbol{g}] \\ &= \boldsymbol{A}\mathbb{E}[\boldsymbol{\theta}\mid\boldsymbol{z}] + \frac{\lambda_t\tau^2}{\lambda_t^2\tau^2 + \sigma_t^2}\cdot(\boldsymbol{z} - \lambda_t\boldsymbol{A}\mathbb{E}[\boldsymbol{\theta}\mid\boldsymbol{z}]) = \frac{\sigma_t^2}{\lambda_t^2\tau^2 + \sigma_t^2}\boldsymbol{A}\mathbb{E}[\boldsymbol{\theta}\mid\boldsymbol{z}] + \frac{\lambda_t\tau^2}{\lambda_t^2\tau^2 + \sigma_t^2}\boldsymbol{z}.\end{aligned}$$

By (Denoiser), we obtain

$$\boldsymbol{s}_t(\boldsymbol{z}) = \frac{\lambda_t}{\sigma_t^2}\mathbb{E}[\boldsymbol{x}\mid\boldsymbol{z}] - \frac{1}{\sigma_t^2}\boldsymbol{z} = -\frac{1}{\tau^2\lambda_t^2 + \sigma_t^2}\boldsymbol{z} + \frac{\lambda_t}{\tau^2\lambda_t^2 + \sigma_t^2}\boldsymbol{A}\cdot\mathbb{E}[\boldsymbol{\theta}\mid\boldsymbol{z}].$$

We notice the equality in distribution $\boldsymbol{z}/\lambda_t \overset{d}{=} \boldsymbol{z}_* = \boldsymbol{A}\boldsymbol{\theta} + \bar{\boldsymbol{\varepsilon}}$ where $(\boldsymbol{\theta},\bar{\boldsymbol{\varepsilon}}) \sim \pi_0^m \otimes \mathcal{N}(\boldsymbol{0}, \bar{\tau}_t^2\mathbf{I}_d)$ (this $\boldsymbol{z}_*$ is as defined in Assumption 4). This implies

$$\boldsymbol{s}_t(\boldsymbol{z}) = -\frac{1}{\tau^2\lambda_t^2 + \sigma_t^2}\boldsymbol{z} + \frac{\lambda_t}{\tau^2\lambda_t^2 + \sigma_t^2}\boldsymbol{A}\cdot\boldsymbol{e}_t(\boldsymbol{z}/\lambda_t), \tag{60}$$

where $\boldsymbol{e}_t$ is as defined in Eq. (15).

Existence of a unique minimizer of the VI free energy

We analyze the VI free energy. We define

$$\mathcal{F}_t^{\rm sparse}(\boldsymbol{e};\boldsymbol{z}_*) := \sum_{i=1}^m \max_\lambda \Big[\lambda e_i - \log \mathbb{E}_{\beta\sim\pi_0}[e^{\lambda\beta-\beta^2\nu_t/2}]\Big] + \frac{1}{2\bar\tau_t^2}\|\boldsymbol{z}_* - \boldsymbol{A}\boldsymbol{e}\|_2^2 - \frac{1}{2}\langle \boldsymbol{e}, \boldsymbol{K}_t\boldsymbol{e}\rangle.$$

Let $G_t(\lambda) = \log \mathbb{E}_{\beta\sim\pi_0}[e^{\lambda\beta-\beta^2\nu_t/2}]$, and $\lambda_i = \arg\max_\lambda[\lambda e_i - G_t(\lambda)]$, then $e_i = G_t'(\lambda_i)$. Therefore,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}e_i}\left[\lambda_i e_i - G_t(\lambda_i)\right] = \lambda_i + \frac{e_i}{G_t''(\lambda_i)} - \frac{G_t'(\lambda_i)}{G_t''(\lambda_i)} = \lambda_i,\\
&\frac{\mathrm{d}^2}{\mathrm{d}^2 e_i}\left[\lambda_i e_i - G_t(\lambda_i)\right] = \frac{1}{G_t''(\lambda_i)}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\nabla_{\boldsymbol{e}}\mathcal{F}_t^{\rm sparse}(\boldsymbol{e};\boldsymbol{z}_*) &= (G_t')^{-1}(\boldsymbol{e}) - \frac{1}{\bar\tau_t^2}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}_* + \frac{1}{\bar\tau_t^2}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{e} - \boldsymbol{K}_t\boldsymbol{e},\\
\nabla_{\boldsymbol{e}}^2\mathcal{F}_t^{\rm sparse}(\boldsymbol{e};\boldsymbol{z}_*) &= \mathrm{diag}\{(G_t''(\lambda_i)^{-1})_{i\in[m]}\} + \frac{1}{\bar\tau_t^2}\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \boldsymbol{K}_t.
\end{aligned}$$

Note that $G_t''(\lambda_i) = \mathrm{Var}_{(\beta,z)\sim\pi_0\otimes\mathcal{N}(0,1)}[\,\beta \mid \beta + \nu_t^{-1/2}z = \lambda\nu_t^{-1}] \le \Pi^2$. In addition, note that $|G_t'(\lambda)| = |\mathbb{E}[\beta \mid \beta + \nu_t^{-1/2}z = \lambda\nu_t^{-1}]| \le \Pi$ for all λ. By assumption, $\|\bar\tau_t^{-2}\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \boldsymbol{K}_t\|_{\rm op} < \Pi^{-2}$, hence $\nabla_{\boldsymbol{e}}^2\mathcal{F}_t^{\rm sparse}(\boldsymbol{e};\boldsymbol{z}_*)$ is positive-definite and $\mathcal{F}_t^{\rm sparse}(\cdot;\boldsymbol{z}_*)$ as a function of $\boldsymbol{e}$ is strongly convex. That is to say, the equation

$$\boldsymbol{e} = G_t'\left((-\bar\tau_t^{-2}\boldsymbol{A}^\mathsf{T}\boldsymbol{A} + \boldsymbol{K}_t)\boldsymbol{e} + \bar\tau_t^{-2}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\right)$$

has a unique fixed point $\hat{\boldsymbol{e}}_t(\boldsymbol{z}_*)$.

Approximate the minimizer of the free energy via iterative algorithm

We denote by $f_t(\cdot)$ the function obtained from Lemma 5 that achieves ζ-uniform approximation to $G_t'(\cdot)$. By Lemma 6, we conclude that if we implement the following iteration

$$\tilde{\boldsymbol{e}}^0(\boldsymbol{z}_*) = \boldsymbol{0}, \qquad \tilde{\boldsymbol{e}}^{\ell+1}(\boldsymbol{z}_*) = f_t\left((-\bar\tau_t^{-2}\boldsymbol{A}^\mathsf{T}\boldsymbol{A} + \boldsymbol{K}_t)\tilde{\boldsymbol{e}}^\ell(\boldsymbol{z}_*) + \bar\tau_t^{-2}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\right), \tag{61}$$

then for all $\ell \in \mathbb{N}_+$, we have

$$\frac{1}{\sqrt{m}}\|\tilde{\boldsymbol{e}}^\ell(\boldsymbol{z}_*) - \hat{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2 \le \Pi\cdot(\Pi^2 A)^\ell + \frac{\zeta}{1-\Pi^2 A}. \tag{62}$$

Represent the iterative algorithm as a ResNet

We then show that $\boldsymbol{s}_t(\boldsymbol{z}) = (\lambda_t\boldsymbol{A}\boldsymbol{e}_t(\boldsymbol{z}/\lambda_t) - \boldsymbol{z})/(\tau^2\lambda_t^2 + \sigma_t^2)$ (c.f. Eq. (60)) can be expressed as a ResNet that takes input $\boldsymbol{z}$.

Lemma 13. *For all $t \in \{T - t_k\}_{0\le k\le N-1}$ and $\ell \in \mathbb{N}_+$, there exists $\boldsymbol{W} \in \mathcal{W}_{d,D,\ell,M,B}$, with*

$$\begin{aligned}
&D = 3m + d, \qquad M = (\lceil 2\Pi\zeta^{-1}\rceil + 3)m,\\
&B = \left(\lceil 2\Pi\zeta^{-1}\rceil - 1\right)\cdot\left(A + 1 + 2\Pi^2 + w_\zeta\right) + 2\Pi + 6 + (\|\boldsymbol{A}\|_{\rm op} + 1)/(1 - e^{-2\delta}) + \bar\tau_t^{-2}\lambda_t^{-1}\|\boldsymbol{A}\|_{\rm op} + \sqrt{m},
\end{aligned}$$

such that $(\lambda_t\boldsymbol{A}\tilde{\boldsymbol{e}}^\ell(\boldsymbol{z}/\lambda_t) - \boldsymbol{z})/(\tau^2\lambda_t^2 + \sigma_t^2) = \mathrm{ResN}_{\boldsymbol{W}}(\boldsymbol{z})$. Here, $\tilde{\boldsymbol{e}}^\ell$ is as defined in Eq. (61), and w_ζ is given by

$$w_\zeta = \sup_{t\in\{T-t_k\}_{0\le k\le N-1}} \inf\Big\{w : \textit{for all } \lambda_1 > \lambda_2 \ge w \textit{ or } \lambda_1 < \lambda_2 \le -w \textit{ we have } |G_t'(\lambda_1) - G_t'(\lambda_2)| < \zeta\Big\}.$$

Proof of Lemma 13. Recall that the ResNet is defined as (ResNet). Recall the definition of f_t as an approximation of G_t' as in Lemma 5. We shall choose the weight matrices appropriately, such that $\boldsymbol{u}^{(\ell)} = [\tilde{\boldsymbol{e}}^\ell(\boldsymbol{z}/\lambda_t); \bar\tau_t^{-2}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}/\lambda_t; \boldsymbol{1}_m; \boldsymbol{z}] \in \mathbb{R}^{3m+d}$. For $\ell = 0$, we set

$$\boldsymbol{W}_{\rm in} = \begin{bmatrix} \boldsymbol{0}_{d\times m} & \bar\tau_t^{-2}\lambda_t^{-1}\boldsymbol{A} & \boldsymbol{0}_{d\times m} & \mathbf{I}_d \\ \boldsymbol{0}_{1\times m} & \boldsymbol{0}_{1\times m} & \boldsymbol{1}_{1\times m} & \boldsymbol{0}_{1\times d}\end{bmatrix}^\mathsf{T} \in \mathbb{R}^{(3m+d)\times(d+1)}.$$

For $\ell \geq 1$, we set

$$\boldsymbol{W}_1^{(\ell)} = \begin{bmatrix} a_i\mathbf{I}_m & \cdots & a_{\lceil 2\Pi\zeta^{-1}\rceil-1}\mathbf{I}_m & -\mathbf{I}_m & \mathbf{I}_m & a_0\mathbf{I}_m & -a_0\mathbf{I}_m \\ \mathbf{0}_{m\times m} & \cdots & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} & \cdots & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \\ \mathbf{0}_{d\times m} & \cdots & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} \end{bmatrix} \in \mathbb{R}^{(3m+d)\times(\lceil 2\Pi\zeta^{-1}\rceil+3)m},$$

$$\boldsymbol{W}_2^{(\ell)} = \begin{bmatrix} -\bar{\tau}_t^{-2}\boldsymbol{A}^\top\boldsymbol{A}+\boldsymbol{K}_t & \cdots & -\bar{\tau}_t^{-2}\boldsymbol{A}^\top\boldsymbol{A}+\boldsymbol{K}_t & \mathbf{I}_m & -\mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \\ \mathbf{I}_m & \cdots & \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \\ -w_1\mathbf{I}_m & \cdots & -w_{\lceil 2\Pi\zeta^{-1}\rceil-1}\mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{I}_m & -\mathbf{I}_m \\ \mathbf{0}_{d\times m} & \cdots & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} & \mathbf{0}_{d\times m} \end{bmatrix}^\top \in \mathbb{R}^{(\lceil 2\Pi\zeta^{-1}\rceil+3)m\times(3m+d)}.$$

For the output layer, we let $\boldsymbol{W}_{\text{out}} = [\lambda_t\boldsymbol{A}/(\sigma_t^2+\tau^2\lambda_t^2), \mathbf{0}_{d\times m}, \mathbf{0}_{d\times m}, -(\tau^2\lambda_t^2+\sigma_t^2)^{-1}\mathbf{I}_d] \in \mathbb{R}^{d\times(3m+d)}$.

The following upper bounds are straightforward:

$$\|\boldsymbol{W}_{\text{in}}\|_{\text{op}} \leq \bar{\tau}_t^{-2}\lambda_t^{-1}\|\boldsymbol{A}\|_{\text{op}} + \sqrt{m}+1, \qquad \|\boldsymbol{W}_{\text{out}}\|_{\text{op}} \leq (\|\boldsymbol{A}\|_{\text{op}}+1)/(1-e^{-2\delta}),$$
$$\|\boldsymbol{W}_1^{(\ell)}\|_{\text{op}} \leq 2\Pi^2(\lceil 2\Pi\zeta^{-1}\rceil-1)+2\Pi+2, \qquad \|\boldsymbol{W}_2^{(\ell)}\|_{\text{op}} \leq \big(\lceil 2\Pi\zeta^{-1}\rceil-1\big)\cdot(A+1+w_\zeta)+4.$$

In summary, we have

$$\|\boldsymbol{W}\| \leq \big(\lceil 2\Pi\zeta^{-1}\rceil-1\big)\cdot\big(A+1+2\Pi^2+w_\zeta\big)+2\Pi+6+(\|\boldsymbol{A}\|_{\text{op}}+1)/(1-e^{-2\delta})+\bar{\tau}_t^{-2}\lambda_t^{-1}\|\boldsymbol{A}\|_{\text{op}}+\sqrt{m}.$$

This concludes the proof of Lemma 13. □

Proof of Theorem 4

Similar to the proof of Theorem 1, we obtain

$$\mathbb{E}[\|\hat{\boldsymbol{s}}_t(\boldsymbol{z})-\boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \leq \bar{\varepsilon}_{\text{app}}^2+\bar{\varepsilon}_{\text{gen}}^2, \tag{63}$$

where $\bar{\varepsilon}_{\text{app}}^2$ is the approximation error and $\bar{\varepsilon}_{\text{gen}}^2$ is the generalization error:

$$\bar{\varepsilon}_{\text{app}}^2 = \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\bar{\mathsf{P}}_t[\text{ResN}_{\boldsymbol{W}}](\boldsymbol{z})-\boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d,$$
$$\bar{\varepsilon}_{\text{gen}}^2 = 2\sup_{\boldsymbol{W}\in\mathcal{W}}\Big|\hat{\mathbb{E}}[\|\bar{\mathsf{P}}_t[\text{ResN}_{\boldsymbol{W}}](\boldsymbol{z})+\boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d-\mathbb{E}[\|\mathsf{P}_t\text{ResN}_{\boldsymbol{W}}(\boldsymbol{z})+\boldsymbol{\sigma}_t^{-1}\boldsymbol{g}\|_2^2]/d\Big|.$$

Applying Proposition 7 and taking $D = 3m+d$, we conclude that with probability at least $1-\eta$, simultaneously for any $t \in \{T-t_k\}_{0\leq k\leq N-1}$, when $n \geq \log(2/\eta)$, we have

$$\bar{\varepsilon}_{\text{gen}}^2 \lesssim \Big(\lambda_t^2\|\boldsymbol{A}\|_{\text{op}}^2\Pi^2(\tau^{-4}+1)\frac{m}{d}+\frac{\lambda_t^2}{\sigma_t^2}(1+\tau^2)\Big) \times\sqrt{\frac{(dD+LDM)\cdot[T+L\log(LB)+\log(nmT(\tau+1)(\|\boldsymbol{A}\|_{\text{op}}\Pi+1)\tau^{-1})]+\log(2N/\eta)}{n}}. \tag{64}$$

where we choose

$$B = M/m\cdot\big(A+1+2\Pi^2+w_\star\big)+2\Pi+6+(\|\boldsymbol{A}\|_{\text{op}}+1)/(1-e^{-2\delta})+\tau^{-2}\|\boldsymbol{A}\|_{\text{op}}+\sqrt{m},$$
$$w_\star = \sup_{t\in\{T-t_k\}_{0\leq k\leq N-1}}\inf\Big\{w:\ \text{for all } \lambda_1>\lambda_2\geq w \text{ or } \lambda_1<\lambda_2\leq -w,\ |G_t'(\lambda_1)-G_t'(\lambda_2)|<M/(6m\Pi)\Big\}. \tag{65}$$

We next upper bound $\bar{\varepsilon}_{\text{app}}^2$. Recall Eq. (60) and $\bar{\tau}_t^2 = \tau^2+\sigma_t^2/\lambda_t^2$, we have $\boldsymbol{s}_t(\boldsymbol{z}) = -\lambda_t^{-2}\bar{\tau}_t^{-2}\boldsymbol{z}+\lambda_t{}^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\bar{\boldsymbol{e}}_t(\boldsymbol{z}_*)$ (recall that $\boldsymbol{z}_* = \boldsymbol{z}/\lambda_t$) and recall $\bar{\mathsf{P}}_t[\text{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) = \text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\text{ResN}_{\boldsymbol{W}}(\boldsymbol{z})+\lambda_t^{-2}\bar{\tau}_t^{-2}\boldsymbol{z})-\lambda_t^{-2}\bar{\tau}_t^{-2}\boldsymbol{z}$. According to Lemma 13, recalling $\tilde{\boldsymbol{e}}^L$

as defined in Eq. (61), we have

$$\begin{aligned}
\bar{\varepsilon}_{\text{app}}^2 &= \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\bar{\mathsf{P}}_t[\text{ResN}_{\boldsymbol{W}}](\boldsymbol{z}) - \boldsymbol{s}_t(\boldsymbol{z})\|_2^2]/d \\
&= \inf_{\boldsymbol{W}\in\mathcal{W}} \mathbb{E}[\|\text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\text{ResN}_{\boldsymbol{W}}(\boldsymbol{z}) + \lambda_t^{-2}\bar{\tau}_t^{-2}\boldsymbol{z}) - \lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\bar{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2^2]/d \\
&\le \mathbb{E}[\|\text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\tilde{\boldsymbol{e}}^L(\boldsymbol{z}_*)) - \lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\bar{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2^2]/d \\
&\lesssim \mathbb{E}[\|\text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\tilde{\boldsymbol{e}}^L(\boldsymbol{z}_*)) - \text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\hat{\boldsymbol{e}}(\boldsymbol{z}_*))\|_2^2]/d \\
&\quad + \mathbb{E}[\|\text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\hat{\boldsymbol{e}}(\boldsymbol{z}_*)) - \lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\bar{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2^2]/d
\end{aligned} \tag{66}$$

where the last inequality is by the triangle inequality. By Eq. (62) and the 1-Lipschitzness of $\text{proj}(\cdot)$, we obtain that the first term in the right-hand side above is upper bounded by

$$\begin{aligned}
&\mathbb{E}[\|\text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\tilde{\boldsymbol{e}}^L(\boldsymbol{z}_*)) - \text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\hat{\boldsymbol{e}}(\boldsymbol{z}_*))\|_2^2]/d \\
&\lesssim \frac{m\|\boldsymbol{A}\|_{\text{op}}^2}{d\lambda_t^2\bar{\tau}_t^4}\cdot(\Pi^2\cdot(\Pi^2 A)^{2L} + \zeta^2(1-\Pi^2 A)^{-2}) \lesssim \frac{m\|\boldsymbol{A}\|_{\text{op}}^2}{d\lambda_t^2\bar{\tau}_t^4}\cdot\Big(\Pi^2\cdot(\Pi^2 A)^{2L} + \frac{m^2\Pi^2}{(1-\Pi^2 A)^2 M^2}\Big).
\end{aligned} \tag{67}$$

In the above display, the last inequality is by the fact that we can choose ζ such that $M = m\cdot(\lceil 2\Pi\zeta^{-1}\rceil + 3)$, which implies that $2m\Pi/M \le \zeta \le 6m\Pi/M$. Furthermore, by Assumption 4 and by the fact that $\|\hat{\boldsymbol{e}}(\boldsymbol{z}_*)\|_2 \le \sqrt{m}\,\Pi$, we obtain that the second quantity in the right-hand side of Eq. (66) is controlled by

$$\mathbb{E}[\|\text{proj}_{\sqrt{m}\|\boldsymbol{A}\|_{\text{op}}\Pi\lambda_t^{-1}\bar{\tau}_t^{-2}}(\lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\hat{\boldsymbol{e}}(\boldsymbol{z}_*)) - \lambda_t^{-1}\bar{\tau}_t^{-2}\boldsymbol{A}\bar{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2^2]/d \lesssim \frac{m\|\boldsymbol{A}\|_{\text{op}}^2}{d\lambda_t^2\bar{\tau}_t^4}\cdot\varepsilon_{\text{VI},t}^2(\boldsymbol{A}). \tag{68}$$

Finally, we combine Eq. (63), (64), (66), (67), (68). This completes the proof of Theorem 4.

F.4 Proof of Lemma 3

Consider the sparse coding problem $\boldsymbol{z}_* = \boldsymbol{A\theta} + \bar{\varepsilon} \in \mathbb{R}^d$ with dictionary $\boldsymbol{A} \in \mathbb{R}^{d\times m}$, sparse representation $\boldsymbol{\theta} \in \mathbb{R}^m$, and noise $\bar{\varepsilon} \in \mathbb{R}^d$. Assume that the model satisfies the following assumption.

Assumption 7 (Simplified version of Assumption 1 - 4 of Li et al. [2023b])**.** *Assume that $\boldsymbol{A} = \boldsymbol{QDO}^\mathsf{T}$ is the singular value decomposition of $\boldsymbol{A}$, where $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{O} \in \mathbb{R}^{m\times m}$ are orthogonal and $\boldsymbol{D} \in \mathbb{R}^{d\times m}$ is diagonal with diagonal elements $\{d_i\}_{i\in[\min\{d,m\}]}$. Assume that $\boldsymbol{Q}$, $\boldsymbol{D}$ are deterministic, $\boldsymbol{O}$, $\boldsymbol{\theta}$, ε are mutually independent, and $\boldsymbol{O} \sim \text{Haar}(\text{SO}(m))$ is uniformly distributed on the special orthogonal group. As $d, m \to \infty$, we assume $\mu_{\boldsymbol{D}} \stackrel{W}{\to} \mathsf{D}$ where $\mu_{\boldsymbol{D}}$ is the empirical distribution of coordinates of $\boldsymbol{D}$, D is a random variable with $\text{supp}\{\mathsf{D}^2\} \subseteq [d_-, d_+]$ and $0 < d_- < d_+ < \infty$, and $\stackrel{W}{\to}$ denotes Wasserstein-p convergence. Furthermore, $\min_i\{d_i^2\} \to d_-$ and $\max_i\{d_i^2\} \to d_+$. We further assume $\theta_i \sim_{iid} \pi_0$ with $\mathbb{E}_{\pi_0}[\theta] = 0$, $\mathbb{E}_{\pi_0}[\theta^2] > 0$, and π_0 is compactly supported. Finally, we have $\bar{\varepsilon}_i \sim_{iid} \mathcal{N}(0, \bar{\tau}^2)$.*

Denote the posterior mean of $\boldsymbol{\theta}$ given $(\boldsymbol{A}, \boldsymbol{z}_*)$ by $\boldsymbol{e}(\boldsymbol{z}_*) = \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{z}_*]$. Theorem 1.11 of Li et al. [2023b] proves the following.

Lemma 14 (Theorem 1.11 of Li et al. [2023b])**.** *Let Assumption 7 hold. There exists $\bar{\tau}_0^2$ that depends on $(\alpha, \pi_0, \mathsf{D})$, such that the following happens. For any $\bar{\tau}^2 \ge \bar{\tau}_0^2$, there exists $\nu_\star = (\alpha, \pi_0, \mathsf{D}, \bar{\tau}^2)$ that depends on $(\alpha, \pi_0, \mathsf{D}, \bar{\tau}^2)$ such that, taking $G(\lambda) = \log \mathbb{E}_{\beta\sim\pi_0}[e^{\lambda\beta - \beta^2\nu_\star/2}]$ we have almost surely*

$$\lim_{d,m\to\infty} \mathbb{E}_{\boldsymbol{z}_*}\Big[\Big\|\boldsymbol{e}(\boldsymbol{z}_*) - G'\big(-\bar{\tau}^{-2}\big((\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_\star\mathbf{I}_m)\boldsymbol{e}(\boldsymbol{z}_*) - \boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\big)\big)\Big\|_2^2\Big|\boldsymbol{A}\Big] = 0.$$

Furthermore, for any fixed $(\pi_0, \alpha, \mathsf{D})$, we have $\sup_{\bar{\tau}^2\ge\bar{\tau}_0^2}\nu_\star(\bar{\tau}^2) < \infty$.

We remark that Theorem 1.11 of Li et al. [2023b] assumes the fixed noise level $\bar{\tau}^2 = 1$. However, a simple rescaling argument could extend the result to general $\bar{\tau}^2$.

Given Lemma 14, we are now ready to prove Lemma 3. Taking $\bar{\tau}^2 = \bar{\tau}_t^2 = \tau^2 + \sigma_t^2/\lambda_t^2$, $\nu_t = \nu_\star(\bar{\tau}_t^2)$, $G_t = G$, and $\boldsymbol{K}_t = \bar{\tau}_t^{-2}\nu_\star(\bar{\tau}_t^2)$, we note that the minimizer of the VI free energy $\hat{\boldsymbol{e}}_t(\boldsymbol{z}_*) \in [-\Pi, \Pi]^m$ should satisfy

$$\hat{\boldsymbol{e}}_t(\boldsymbol{z}_*) = G_t'\left(-\bar{\tau}_t^{-2}\big((\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_t\mathbf{I}_m)\hat{\boldsymbol{e}}_t(\boldsymbol{z}_*) - \boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\big)\right).$$

For the posterior mean $\boldsymbol{e}_t(\boldsymbol{z}_*) \in [-\Pi, \Pi]^m$, we have

$$
\begin{aligned}
&\Big\|\boldsymbol{e}_t(\boldsymbol{z}_*) - G_t'\big(-\bar{\tau}_t^{-2}\big((\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_t \mathbf{I}_m)\boldsymbol{e}_t(\boldsymbol{z}_*) - \boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\big)\big)\Big\|_2 \\
\geq &\Big\|\boldsymbol{e}_t(\boldsymbol{z}_*) - \hat{\boldsymbol{e}}_t(\boldsymbol{z}_*)\Big\|_2 \\
&- \Big\|G_t'\big(-\bar{\tau}_t^{-2}\big((\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_t \mathbf{I}_m)\boldsymbol{e}_t(\boldsymbol{z}_*) - \boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\big)\big) - G_t'\big(-\bar{\tau}_t^{-2}\big((\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_t r \mathbf{I}_m)\hat{\boldsymbol{e}}_t(\boldsymbol{z}_*) - \boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\big)\Big\|_2 \\
\geq &\Big(1 - \Pi^2 \bar{\tau}_t^{-2}\|\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_t \mathbf{I}_m\|_{\mathrm{op}}\Big)\|\boldsymbol{e}_t(\boldsymbol{z}_*) - \hat{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2,
\end{aligned}
$$

where the last inequality used the fact that G_t' is Π^2-Lipschitz. Notice that by Lemma 14, $\sup_{\bar{\tau}^2 \geq \bar{\tau}_0^2} \nu_\star(\bar{\tau}^2) = \nu < \infty$, and $\|\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\|_{\mathrm{op}} = \max_i d_i^2$ bounded almost surely by some $D < \infty$ per Assumption 7. Therefore, when $\tau_0^2 \geq 2\Pi^2(D + \nu)$, we have $1 - \Pi^2\bar{\tau}_t^{-2}\|\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_\star \mathbf{I}_m\|_{\mathrm{op}} \geq 1/2$ for any $\tau^2 \geq \tau_0^2$ and any t. This gives

$$
\Big\|\boldsymbol{e}_t(\boldsymbol{z}_*) - G_t'\big(-\bar{\tau}_t^{-2}\big((\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_t \mathbf{I}_m)\boldsymbol{e}_t(\boldsymbol{z}_*) - \boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\big)\big)\Big\|_2 \geq \|\boldsymbol{e}_t(\boldsymbol{z}_*) - \hat{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2/2.
$$

Furthermore, by Lemma 14, the posterior mean $\boldsymbol{e}_t(\boldsymbol{z}_*)$ satisfies

$$
\lim_{d,m\to\infty} \mathbb{E}_{\boldsymbol{z}_*}\Big[\Big\|\boldsymbol{e}_t(\boldsymbol{z}_*) - G_t'\big(-\bar{\tau}_t^{-2}\big((\boldsymbol{A}^\mathsf{T}\boldsymbol{A} - \nu_t \mathbf{I}_m)\boldsymbol{e}_t(\boldsymbol{z}_*) - \boldsymbol{A}^\mathsf{T}\boldsymbol{z}_*\big)\big)\Big\|_2^2\Big|\boldsymbol{A}\Big] = 0.
$$

This implies that

$$
\lim_{d,m\to\infty} \mathbb{E}_{\boldsymbol{z}_*}[\|\boldsymbol{e}_t(\boldsymbol{z}_*) - \hat{\boldsymbol{e}}_t(\boldsymbol{z}_*)\|_2^2|\boldsymbol{A}] = 0,
$$

which concludes the proof of Lemma 3.