Taking a Big Step: Large Learning Rates in Denoising Score Matching Prevent Memorization

Yu-Han Wu YU-HAN.WU@ETU.SORBONNE-UNIVERSITE.FR
LPSM, Sorbonne Université

Pierre Marion PIERRE.MARION@EPFL.CH
Institute of Mathematics, EPFL

Gérard Biau GERARD.BIAU@SORBONNE-UNIVERSITE.FR
LPSM, Sorbonne Université, Institut universitaire de France

Claire Boyer CLAIRE.BOYER@UNIVERSITE-PARIS-SACLAY.FR
LMO, Université Paris-Saclay, Institut universitaire de France

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Denoising score matching plays a pivotal role in the performance of diffusion-based generative models. However, the empirical optimal score–the exact solution to the denoising score matching–leads to memorization, where generated samples replicate the training data. Yet, in practice, only a moderate degree of memorization is observed, even without explicit regularization. In this paper, we investigate this phenomenon by uncovering an implicit regularization mechanism driven by large learning rates. Specifically, we show that in the small-noise regime, the empirical optimal score exhibits high irregularity. We then prove that, when trained by stochastic gradient descent with a large enough learning rate, neural networks cannot stably converge to a local minimum with arbitrarily small excess risk. Consequently, the learned score cannot be arbitrarily close to the empirical optimal score, thereby mitigating memorization. To make the analysis tractable, we consider one-dimensional data and two-layer neural networks. Experiments validate the crucial role of the learning rate in preventing memorization, even beyond the one-dimensional setting.

Keywords: diffusion models, denoising score matching, implicit regularization, neural networks

1. Introduction

Diffusion models have achieved remarkable success in generative modeling across a wide range of tasks, including computer vision (Amit et al., 2021; Baranchuk et al., 2022), temporal data modeling (Chen et al., 2021; Alcaraz and Strodthoff, 2023), multimodal modeling (Ramesh et al., 2022; Rombach et al., 2022), and natural language processing (NLP, Austin et al., 2021; Savinov et al., 2022). Using diffusion models for generative modeling was first proposed by Sohl-Dickstein et al. (2015). Subsequently, denoising diffusions reached state-of-the-art performance (Song and Ermon, 2019; Ho et al., 2020) when trained efficiently with denoising score matching (Hyvärinen, 2005; Vincent, 2011). This training objective consists in learning to denoise artificially perturbed images from the training sample, which is mathematically equivalent to learning the gradient of the log-density, or *score*, of the noisy empirical data distribution. Once score matching has been performed, new observations can be generated by running a backward diffusion process that involves the learned score. Beyond applications in diffusion, denoising score matching is widely used in various tasks (see, for example, Milanfar and Delbracio, 2024), including image restoration (Venkatakrishnan et al., 2013; Teodoro et al., 2016) and nonlinear inverse problems (Wu et al., 2019).

However, despite the remarkable effectiveness of denoising score matching, the theoretical properties of this training procedure remain unclear. In particular, if the score matching objective were solved perfectly, meaning that the learned score is equal to the score of the empirical data distribution, then the distribution generated by the diffusion process would exactly coincide with the empirical distribution (Li et al., 2024a), a failure mode known as *memorization*. Avoiding this issue is crucial in terms of privacy, intellectual property rights (Vyas et al., 2023; Zhang et al., 2023), and ability of models to effectively generate new, unseen data.

Empirically, only a moderate amount of memorization is observed in practical settings (Carlini et al., 2023; Somepalli et al., 2023a; Gu et al., 2023; Kadkhodaie et al., 2024), even without explicitly regularizing the training objective (see Section 2 for details on explicit regularization). This suggests the existence of an *implicit regularization* mechanism that prevents exact solving of the denoising score matching problem and thus full memorization of the training data. However, the nature of this regularization is an open problem, as already highlighted in Biroli et al. (2024).

Contributions. In this paper, we approach the question through the lens of the (ir)regularity of the empirical optimal score, drawing a connection with the learning rate of stochastic gradient descent (SGD). We first show that the empirical optimal score is irregular in the sense that its derivative has a large (weighted) total variation (Section 5). Building upon a recent literature on the impact of the learning rate on the regularity of stable minima for SGD (Mulayoff et al., 2021; Qiao et al., 2024), we prove that the empirical optimal score cannot be stably learned by SGD unless the learning rate becomes vanishingly small (Section 6). Our main result can be informally stated as follows.

Theorem 1 *(informal) Consider the denoising score matching objective $\mathcal{R}_n$ over the class of two-layer neural networks for one-dimensional data. Then, for a sufficiently small level of noise σ and a learning rate $\eta \gtrsim \sigma^2$, the stochastic gradient descent on $\mathcal{R}_n$ cannot stably converge to a local minimum with arbitrarily small excess risk.*

The take-home message is that for σ small enough and η large enough, the learned score cannot be arbitrarily close to the empirical optimal one. To the best of our knowledge, this is the first demonstration of an implicit regularization mechanism in denoising score matching that prevents (full) memorization, thanks to the non-vanishing learning rate used in practice. Focusing on small values of σ is reasonable, as they correspond to the last steps of the backward diffusion, which are known to play a critical role in memorization (Raya and Ambrogioni, 2023; Biroli et al., 2024).

We state our result with SGD because the objective $\mathcal{R}_n$ writes as an expectation, hence requiring stochastic approximation. However, our proof also carries over to GD on the population risk—see Appendix D for further comments. Therefore, the non-convergence of SGD towards the global minimizer is not due to the lack of handling the variance of the gradient estimates, but rather to an (implicit) bias due to the large learning rate.

Our results are illustrated through experiments in Section 7, supporting the connection between the choice of learning rate and memorization, and suggesting that our findings extend beyond the one-dimensional case. Some open questions are discussed in Section 8.

2. Related work

Implicit bias of large learning rates and minima stability. Large learning rates are an essential implicit regularization mechanism in deep learning (see, e.g., Li et al., 2019; Andriushchenko et al., 2023). In particular, the learning rate provides an upper bound on the maximal eigenvalue of the

Hessian of the risk at a twice-differentiable stable minimum (Wu et al., 2018). This result can be extended to non-differentiable minima for underparameterized networks (Mulayoff et al., 2021). In addition, Qiao et al. (2024) prove a generalization bound for twice-differentiable stable minima in regression tasks, by relating the condition on the Hessian to the functions that can be represented by the neural network. In the present paper, we investigate the impact of large learning rates in the setting of denoising score matching. An important feature of our analysis is that the noise in the training objective of score matching has a regularizing effect, as highlighted by Lemma 5. This regularizing effect guarantees the necessary assumption that SGD reaches a twice-differentiable local minimum. This contrasts with the standard regression case, where this hypothesis is hardly met (indeed, in general, SGD with ReLU networks tends to align the kinks with the data points; see, e.g., Boursier and Flammarion 2023, resulting in lower regularity—C^1 instead of C^2). Importantly, our proof technique should be adaptable to other tasks with noise in the training objective, for example in robust learning or dropout regularization.

Memorization effect of diffusion models. Diffusion models were found to generate replicas of their training data (see, e.g., Carlini et al., 2023; Somepalli et al., 2023a,b), raising privacy and security concerns. Following these initial observations, a series of papers quantified this memorization phenomenon (Gu et al., 2023; Yoon et al., 2023; Kadkhodaie et al., 2024). These articles experimentally demonstrate a transition from memorization to generalization as the sample size increases, showing that with practical sample sizes, the extent of memorization is limited. Furthermore, Kadkhodaie et al. (2024) link the generalization ability of diffusion models to their adaptability to the underlying geometric structure of the data. Finally, Gu et al. (2023), Yi et al. (2023), and Li et al. (2024a) show that diffusion models with the empirical optimal score exhibit full memorization.

Regularization of denoising score matching. In practice, several methods can be used to mitigate memorization. Regularization techniques like weight decay, dropout, or data corruption, can help reduce the model's dependency on specific data points (Daras et al., 2023; Gu et al., 2023; Baptista et al., 2025). All these methods rely on explicitly regularizing the training process. On the contrary, the present paper studies the *implicit* regularization effect of the learning rate in denoising score matching, in order to explain the moderate amount of memorization observed in practice even without explicit regularization. The work by Zeno et al. (2024) is more closely aligned with our approach. They derive a closed-form formula for the minimum-norm interpolator of the 1d denoising problem and analyze its generalization properties. We adopt a complementary approach, focusing on SGD stability rather than interpolation and minimum-norm representation. Finally, our analysis supports experimental evidence by Li et al. (2024b), who observe that diffusion models capable of generalization tend to learn near-linear scores. Indeed, we show that the learning rate constrains the learned score's nonlinearity (via the total variation of its derivative), thus preventing full memorization.

3. Denoising score matching at a glance

In this section, we define the problem of denoising score matching and its connection with diffusion-based generative models.

Diffusion-based generative models. Let p_{true} be an unknown non-atomic distribution on $\mathbb{R}$ with finite variance. Diffusion-based generative models aim to generate new observations following p_{true},

given an i.i.d. sample of p_{true}. The principle is as follows. For $t \in [0, T]$, the forward diffusion

$$d\overrightarrow{X}_t = -\overrightarrow{X}_t dt + \sqrt{2} d\overrightarrow{B}_t, \quad \overrightarrow{X}_0 \sim p_{\text{true}}, \tag{1}$$

can be reversed in time using the backward diffusion

$$d\overleftarrow{X}_t = (\overleftarrow{X}_t + 2\nabla \log p_{T-t}(\overleftarrow{X}_t))dt + \sqrt{2} d\overleftarrow{B}_t, \quad \overleftarrow{X}_0 \sim p_T, \tag{2}$$

where p_t is the probability density of $\overrightarrow{X}_t$, and $\overrightarrow{B}_t$ (resp. $\overleftarrow{B}_t$) is a Brownian motion. Note that $\overrightarrow{X}_t$ has a density since it is convolved with Gaussian noise, and, according to Tweedie's formula (Robbins, 1956), the log-density of $\overrightarrow{X}_t$, i.e. $\log p_t$, is differentiable. Here and in the following, in a slight abuse of notation, the same notation refers to the distribution and its density. The time reversal means that $\overleftarrow{X}_{T-t}$ has the same distribution as $\overrightarrow{X}_t$. Thus, assuming that sampling from p_T is straightforward, the goal is to use the backward equation (2) to generate new observations. However, this requires learning the unknown *score function* $\nabla \log p_t$. To do so, a key observation is that

$$\overrightarrow{X}_t \overset{\mathcal{D}}{=} \mu(t)\overrightarrow{X}_0 + \sigma(t)\xi, \quad \xi \sim \mathcal{N}(0, 1), \tag{3}$$

where $\mu(t) = e^{-t}$ and $\sigma(t) = \sqrt{1 - e^{-2t}}$. Therefore, learning the score $\nabla \log p_t$ for every t is equivalent to learning the score of the convolution $(\mu(t)p_{\text{true}}) * \mathcal{N}(0, \sigma^2(t))$ for every t. An efficient method to do so is denoising score matching, which we introduce next.

Denoising score matching and empirical optimal score. In the following, we drop the time index t to cast the problem in the more general context of denoising. Let X be a real-valued random variable of unknown distribution, keeping in mind that in the diffusion model above, X corresponds to $\overrightarrow{X}_0$. Let $\mu, \sigma \in (0, 1)$ be two real numbers, and Y the random variable defined as $Y = \mu X + \sigma\xi$, where ξ is standard Gaussian noise independent of X. In particular, in the diffusion context (3), we have, at any given time t, $Y = \overrightarrow{X}_t$, $\mu = e^{-t}$, and $\sigma = \sqrt{1 - e^{-2t}}$. We let $p_{\mu,\sigma}$ be the density of Y.

The key to connect denoising, i.e, learning the conditional expectation function $\mathbb{E}[X|Y = y]$, and the score function $\nabla \log p_{\mu,\sigma}(y)$, is that, as shown by Robbins (1956) and Miyasawa (1961), $\mathbb{E}[X|Y] = \frac{1}{\mu}(Y + \sigma^2 \nabla \log p_{\mu,\sigma}(Y))$, and thus

$$\nabla \log p_{\mu,\sigma} \in \operatorname*{argmin}_{s \in L_2} \mathbb{E}\big[(s(Y) + \frac{1}{\sigma^2}(Y - \mu X))^2\big]. \tag{4}$$

This variational characterization, called *denoising score matching* (Vincent, 2011), is well-posed since $\nabla \log p_{\mu,\sigma}$ is in L_2 provided that p_{true} has finite variance (see, e.g., Benton et al., 2024, Lemma 6). Since the distributions of X and Y are unknown, this minimization problem is not directly solvable. Thus, given a sample $x_1, \ldots, x_n$ drawn from X, we instead consider the risk

$$\mathcal{R}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y \sim \mathcal{N}(\mu x_i, \sigma^2)}\big[(s(Y) + \frac{1}{\sigma^2}(Y - \mu x_i))^2\big], \tag{5}$$

where, for clarity, we use lowercase $x_1, \ldots, x_n$ to indicate that the expression is conditional on the sample. We emphasize that this risk is semi-empirical, as it retains an expectation with respect to the noise. This is in line with practice, where fresh noise is introduced at each step, ensuring that stochastic gradient descent indeed minimizes (5). Exploiting the convexity of $\mathcal{R}_n$ with respect to s,

one can show (Gu et al., 2023; Li et al., 2024a) that its minimizer over all measurable L_2 functions is

$$s^\star(y;\mu,\sigma) = \frac{\sum_{i=1}^n (\mu x_i - y) e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}}{\sigma^2 \sum_{i=1}^n e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}}, \quad y \in \mathbb{R}. \tag{6}$$

It is interesting to note the resemblance between $s^\star(y;\mu,\sigma)$ and a Nadaraya-Watson kernel estimator (Györfi et al., 2002, Chapter 7). Throughout, we refer to the function $s^\star$ as the *empirical optimal score*. We emphasize again its dependence on the sample $x_1, \dots, x_n$, which justifies the terminology *empirical*. This is not to be confused with the minimizer of the theoretical risk (4).

The memorization problem. The next logical step is to substitute the empirical optimal score, $s^\star$, for the unknown theoretical score, $\nabla \log p_t$, to generate new data from p_{true} by running the backward diffusion (2). However, Li et al. (2024a) showed that this procedure is undesirable as it leads to full memorization of the training data. More precisely, consider the backward diffusion

$$d\overleftarrow{X}_t = (\overleftarrow{X}_t + 2s^\star(\overleftarrow{X}_t; \mu(T-t), \sigma(T-t)))dt + \sqrt{2}d\overleftarrow{B}_t, \quad \overleftarrow{X}_0 \sim \mathcal{N}(0,1),$$

run from time $t = 0$ to time $t = T - \delta$. Then, according to Li et al. (2024a), the total variation distance between the distribution of $\overleftarrow{X}_{T-\delta}$ and a smoothed version of the empirical measure of the training sample can be made arbitrarily small as $T \to \infty$ and $\delta \to 0$. In practice, however, only a moderate degree of memorization is observed (see Section 2). The solution to this puzzle is that practitioners do not use the explicit form of $s^\star$, but perform stochastic gradient descent (SGD) to minimize the risk (5). Our goal is to substantiate this observation by showing that the score fitted with SGD with a large learning rate deviates from $s^\star$, thereby avoiding full memorization.

4. Network class, training algorithm, and stability

Model. In practice, the risk $\mathcal{R}_n$ is optimized over a class of parameterized functions. We consider in the present paper two-layer ReLU networks with m hidden neurons, i.e., a class $\mathcal{S}$ of the form

$$\mathcal{S} = \Big\{ s_\theta : \mathbb{R} \to \mathbb{R} : s_\theta(y) = \frac{1}{m} \sum_{\ell=1}^m w_\ell^{(2)} \phi(w_\ell^{(1)} y + b_\ell), w_\ell^{(1)} \in \{\pm 1\}, (w_\ell^{(2)}, b_\ell) \in [-A, A] \times \mathbb{R} \Big\},$$

where $\theta = (w_{1:m}^{(2)}, b_{1:m}) \in \mathbb{R}^{2m}$ and ϕ is the ReLU activation function. At training time, we consider a random initialization of the inner weights $w_\ell^{(1)}$ with values ± 1, keeping them fixed during training. This simplification is introduced for technical reasons, specifically to avoid non-differentiability issues when an inner weight vanishes. Due to the homogeneity of ReLU, this constraint does not affect the expressivity of $\mathcal{S}$. We also constrain the outer weights $w_\ell^{(2)}$ within a ball. This is a mild requirement insofar as A can be chosen arbitrarily large (provided it grows polynomially with $1/\sigma$). In particular, one can ensure by taking A large enough that the empirical optimal score $s^\star$ is approximated by functions in $\mathcal{S}$. More precisely, by Lemma 12 in Appendix C, one has $\int_{\mathbb{R}} |s^{\star\prime\prime}(y;\mu,\sigma)| dy \leqslant \frac{C_n}{2\sigma^6}$ for some explicit $C_n \geqslant 1$ depending on μ and on the training sample. Accordingly, Bach (2024, Section 9.3.3) shows that $s^\star$ may be approximated by $\mathcal{S}$ as soon as $A \geqslant \frac{C_n}{\sigma^6}$. To fix ideas, a safe choice is therefore $A = \frac{C_n}{\sigma^6}$. In addition, for simplicity, we denote the risk for s_θ as $\mathcal{R}_n(\theta)$ instead of $\mathcal{R}_n(s_\theta)$.

Training. Denoising score matching is performed by minimizing the objective $\mathcal{R}_n$ using SGD. Since $\mathcal{R}_n(\theta)$ is not directly computable, we use at each iteration j an unbiased estimator, given by

$$\hat{\mathcal{R}}_j(\theta) = \sum_{i \in \mathcal{B}_j} \big(s_\theta(Y_i) + \frac{1}{\sigma^2}(Y_i - \mu x_i)\big)^2,$$

where $\mathcal{B}_j$ is a random subset of $\{1, \dots, n\}$ and $Y_i \sim \mathcal{N}(\mu x_i, \sigma^2)$. Then, SGD updates are obtained as $\theta_{j+1} = \theta_j - m\eta \nabla \hat{\mathcal{R}}_j(\theta_j)$, where $\eta > 0$ denotes the learning rate. Note that the network output in the definition of $\mathcal{S}$ and the learning rate are rescaled depending on the width m. As shown in Chizat et al. (2019); Yang and Hu (2021), these are the correct normalization factors in the feature learning regime (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2022).

Linearly-stable minima. Our analysis is based on studying the stability of the sequence (θ_j) around a local minimum of the empirical risk $\mathcal{R}_n$, which allows us to link the Hessian of the risk with the learning rate. More precisely, following Mulayoff et al. (2021) and Qiao et al. (2024), the second-order Taylor expansion of $\hat{\mathcal{R}}_j$ around a twice-differentiable local minimum $\theta^\star$ of $\mathcal{R}_n$ is

$$\hat{\mathcal{R}}_j(\theta) \approx \hat{\mathcal{R}}_j(\theta^\star) + (\theta - \theta^\star)^\top \nabla \hat{\mathcal{R}}_j(\theta^\star) + \frac{1}{2}(\theta - \theta^\star)^\top \nabla^2 \hat{\mathcal{R}}_j(\theta^\star)(\theta - \theta^\star),$$

Therefore, for θ_j close to $\theta^\star$, this motivates considering the linearized SGD updates

$$\theta_{j+1} = \theta_j - m\eta(\nabla \hat{\mathcal{R}}_j(\theta^\star) + \nabla^2 \hat{\mathcal{R}}_j(\theta^\star)(\theta_j - \theta^\star)). \tag{7}$$

It is emphasized that this linearization is valid for $\theta^\star$ located in the interior of the constraint set, i.e., such that $w_\ell^{(2)} \in (-A, A)$ for $1 \leqslant \ell \leqslant m$, an assumption that is made throughout the paper. In this context, a local minimum $\theta^\star$ is said to be *linearly stable* if there exists some $\varepsilon > 0$ such that, for any θ_0 in the ε-ball $\mathcal{B}_\varepsilon(\theta^\star)$, the following condition holds:

$$\limsup_{j \to \infty} \mathbb{E}\|\theta_j - \theta^\star\|_2 \leqslant \varepsilon,$$

where θ_j follows the updates (7). The key property (Mulayoff et al., 2021, Lemma 1) we utilize is that if $\theta^\star$ is linearly stable, then

$$\lambda_{\max}(\nabla^2 \mathcal{R}_n(\theta^\star)) \leqslant \frac{2}{m\eta}, \tag{8}$$

where $\lambda_{\max}$ denotes the largest eigenvalue. This result connects the learning rate to the regularity of $\mathcal{R}_n$: the larger η, the flatter the empirical risk around stable minima for the linearized SGD (7).

5. Regularity of the empirical score function

In this section, we show that the empirical optimal score $s^\star$, defined in (6), becomes irregular for small σ. This analysis represents the first important insight into the memorization phenomenon.

Without loss of generality, we assume that the sample is ordered, i.e., $x_1 \leqslant \cdots \leqslant x_n$, and let $\Delta = \min_{2 \leqslant i \leqslant n}(x_i - x_{i-1})$ denote the minimum spacing between consecutive observations. Note that Δ is a random quantity depending on the sample. We further assume $\Delta > 0$, which is a.s. the case since p_{true} is non-atomic. As a typical example, if the x_i are sampled uniformly over an interval of length a, then Δ is of the order of a/n^2 (Molchanov and Reznikova, 1983; Nagaraja et al., 2015).

To quantify the regularity of $s^\star$, we first note by standard rules that this function is infinitely differentiable. This allows us to resort to the following weighted total variation of the derivative of $s^\star$ (the superscript (1) reminds us that this quantity is the TV of the *first* derivative of $s^\star$):

$$\mathrm{TV}_\pi^{(1)}(s^\star) = \int_{\mathbb{R}} |s^{\star\prime\prime}(y;\mu,\sigma)|\pi(y;\mu,\sigma)dy,$$

which, in this context, is interpreted as a measure of the regularity of $s^\star$: the larger the total variation $\mathrm{TV}_\pi^{(1)}(s^\star)$, the more the derivative of $s^\star$ fluctuates. The weight function π is defined as

$$\pi(y;\mu,\sigma) = \begin{cases} \mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}\big[\min\{\pi^+(y-\xi;\mu,\sigma),\pi^-(y-\xi;\mu,\sigma)\}\big], & \text{if } \mu x_1 \leqslant y \leqslant \mu x_n, \\ 0, & \text{otherwise,} \end{cases}$$

where, denoting by $U \sim \mathcal{U}(\{x_1,\dots,x_n\})$ a uniform draw from the dataset, for $y \in [\mu x_1, \mu x_n]$,

$$\pi^-(y;\mu,\sigma) = \mathbb{P}(\mu U < y)^2\mathbb{E}[y-\mu U \,|\, \mu U < y], \pi^+(y;\mu,\sigma) = \mathbb{P}(\mu U > y)^2\mathbb{E}[\mu U - y \,|\, \mu U > y].$$

Note that a similar, though distinct, weighting scheme was proposed by Mulayoff et al. (2021). As these authors highlight, π puts more weight towards the center of the support of the training data.

The next step consists in rewriting the score $s^\star$ from (6) into a more probabilistic manner. To do so, let, for $1 \leqslant i \leqslant n$, $\alpha_i(y;\mu,\sigma) = e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}/Z$, where Z normalizes the α_i's to sum to 1, and denote by $W(y;\mu,\sigma)$ a random variable taking values in $\{x_1,\dots,x_n\}$ such that the probability of picking x_i is $\alpha_i(y;\mu,\sigma)$. We arrive at the following identities:

$$s^\star(y;\mu,\sigma) = \frac{1}{\sigma^2}\big(-y+\mu\mathbb{E}[W(y;\mu,\sigma)]\big) \quad \text{and} \quad s^{\star\prime}(y;\mu,\sigma) = \frac{1}{\sigma^2}\big(-1+\frac{\mu^2}{\sigma^2}\mathbb{V}[W(y;\mu,\sigma)]\big). \tag{9}$$

The proof of the second identity is given in Lemma 11 in the Appendix. The appeal of this probabilistic formalism is that it connects the properties of $s^\star$ to the moments of W. The latter are the topic of the next proposition. All proofs are postponed to the Appendix (except for Theorem 4).

Proposition 2 *Let* $m_i = \frac{\mu(x_i+x_{i+1})}{2}$, $1 \leqslant i \leqslant n-1$. *Then*

$$\mathbb{V}[W(m_i;\mu,\sigma)] \geqslant \frac{1}{2n}(x_i - x_{i+1})^2.$$

If, in addition, $|y-\mu x_i| \leqslant \frac{\mu}{4}\Delta$ *and* $\Delta \geqslant 2\frac{\sigma}{\mu}$, *then*

$$|\mathbb{E}[W(y;\mu,\sigma)] - x_i| \leqslant n\Delta e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \quad \text{and} \quad \mathbb{V}[W(y;\mu,\sigma)] \leqslant 4n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}}.$$

Note that the lower bound for $\mathbb{V}[W(m_i;\mu,\sigma)]$ is independent of σ, whereas the upper bound of $\mathbb{V}[W(y;\mu,\sigma)]$ decreases to 0 as $\sigma \to 0$. This remark translates into a non-vacuous lower bound on the variation of $s^{\star\prime}$ in the following corollary.

Corollary 3 *If* $\Delta \geqslant 2\frac{\sigma}{\mu}$, *we have, for* $y \in \{x_i, x_{i+1}\}$,

$$|s^{\star\prime}(y;\mu,\sigma) - s^{\star\prime}(m_i;\mu,\sigma)| \geqslant \frac{\mu^2}{\sigma^4}\Big(\frac{(x_i-x_{i+1})^2}{2n} - 4n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}}\Big).$$

In addition,

$$\mathcal{R}_n(s^\star) \leqslant \frac{4\mu^2(x_n-x_1)^2}{\sigma^4}e^{-\frac{\mu^2\Delta^2}{32\sigma^2}}. \tag{10}$$

In particular, the upper bound of (10) *converges to 0 as* $\sigma \to 0$.

In the diffusion-based generative models (1)–(3), we have $\sigma = \sqrt{1-e^{-2t}}$. Thus the condition $\Delta \geqslant 2\frac{\sigma}{\mu}$ is satisfied when the diffusion time t is close to 0, that is, at the last steps of the backward diffusion. As discussed in Section 1, this part of the diffusion plays a key role in memorization. The condition $\Delta \geqslant 2\frac{\sigma}{\mu}$ can be interpreted as the fact that the noisy data points should remain far from each other. A similar small-noise setting has been previously explored in related work (Zeno et al., 2024, Assumption 1). In this context, the corollary implies that the loss tends to 0 as $t \to 0$.

Equipped with this foundation, we are now ready to establish a lower bound on $\mathrm{TV}_\pi^{(1)}(s^\star)$.

Theorem 4 *If* $16n^3 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \leqslant 1$, *then* $\mathrm{TV}_\pi^{(1)}(s^\star) \geqslant \dfrac{\mu^3 n\Delta^3}{2^{12}\sigma^4}$.

Proof The lower bound on $\mathrm{TV}_\pi^{(1)}(s^\star)$ is obtained by decomposing the integral over half-intervals between successive datapoints. Using the triangular inequality, we have

$$\begin{aligned}
\mathrm{TV}_\pi^{(1)}(s^\star) &= \int_{\mu x_1}^{\mu x_n} |s^{\star\prime\prime}(y;\mu,\sigma)|\pi(y;\mu,\sigma)dy \\
&= \sum_{i=1}^{n-1} \int_{\mu x_i}^{m_i} |s^{\star\prime\prime}(y;\mu,\sigma)|\pi(y;\mu,\sigma)dy + \int_{m_i}^{\mu x_{i+1}} |s^{\star\prime\prime}(y;\mu,\sigma)|\pi(y;\mu,\sigma)dy \\
&\geqslant \sum_{i=1}^{n-1} \Big(\min_{y\in[\mu x_i,\mu x_{i+1}]} \pi(y;\mu,\sigma)\Big)\Big(\int_{\mu x_i}^{m_i} |s^{\star\prime\prime}(y;\mu,\sigma)|dy + \int_{m_i}^{\mu x_{i+1}} |s^{\star\prime\prime}(y;\mu,\sigma)|dy\Big) \\
&\geqslant \sum_{i=1}^{n-1} \Big(\min_{y\in[\mu x_i,\mu x_{i+1}]} \pi(y;\mu,\sigma)\Big)\Big(\Big| \int_{\mu x_i}^{m_i} s^{\star\prime\prime}(y;\mu,\sigma)dy\Big| + \Big| \int_{m_i}^{\mu x_{i+1}} s^{\star\prime\prime}(y;\mu,\sigma)dy\Big|\Big).
\end{aligned}$$

Then, by the fundamental theorem of calculus,

$$\begin{aligned}
\mathrm{TV}_\pi^{(1)}(s^\star) &\geqslant \sum_{i=1}^{n-1} \Big(\min_{y\in[\mu x_i,\mu x_{i+1}]} \pi(y;\mu,\sigma)\Big)\Big(\big|s^{\star\prime}(m_i;\mu,\sigma) - s^{\star\prime}(\mu x_i;\mu,\sigma)\big| \\
&\qquad + \big|s^{\star\prime}(m_i;\mu,\sigma) - s^{\star\prime}(\mu x_{i+1};\mu,\sigma)\big|\Big) \\
&\geqslant \sum_{i=1}^{n-1} \frac{\mu}{n^2}\Big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{2\sigma^2}}\Big) \min\Big(\frac{i^2(i-1)}{2}, \frac{(n-i)^2(n-i+1)}{2}\Big)\Delta \\
&\qquad \times 2\frac{\mu^2}{\sigma^4}\Big(\frac{(x_i - x_{i+1})^2}{2n} - 4n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}}\Big),
\end{aligned}$$

where, in the last inequality, we used a lower bound on π given in Proposition 9 in Appendix A, combined with Corollary 3 (notice that the condition $16n^3 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \leqslant 1$ implies $\Delta \geqslant 2\frac{\sigma}{\mu}$). By considering only the i's that are between $\lceil n/4\rceil + 1$ and $\lfloor 3n/4\rfloor - 1$, we see that

$$\min\Big(\frac{i^2(i-1)}{2}, \frac{(n-i)^2(n-i+1)}{2}\Big) \geqslant \frac{n^3}{128}.$$

Then, we obtain that

$$\mathrm{TV}_\pi^{(1)}(s^\star) \geqslant \sum_{i=\lceil n/4\rceil+1}^{\lfloor 3n/4\rfloor-1} \frac{\mu^3 n\Delta}{64\sigma^4}\Big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{2\sigma^2}}\Big)\Big(\frac{\Delta^2}{2n} - 4n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}}\Big)$$

$$\geqslant \frac{\mu^3 n\Delta^3}{256\sigma^4}\Big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{2\sigma^2}}\Big)\Big(\frac{1}{2} - 4n^3 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}}\Big)$$
$$\geqslant \frac{\mu^3 n\Delta^3}{256\sigma^4}(\frac{1}{2} - \frac{1}{4})(\frac{1}{2} - \frac{1}{4}),$$

where the second inequality utilizes that there are at least $n/4$ points between $\lceil n/4\rceil + 1$ and $\lfloor 3n/4\rfloor - 1$ (for $n \geqslant 10$), and the last inequality unfolds from the assumption $16n^3 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \leqslant 1$. This concludes the proof. ■

The condition $16n^3 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \leqslant 1$ is equivalent to $\Delta^2 \geqslant 4(\sigma^2/\mu^2)\ln(16n^3)$. In other words, up to a log factor, the minimum spacing of sample points Δ is larger than the normalized standard deviation σ/μ. Consequently, for a fixed μ, in the small-noise regime $\sigma \to 0$, this condition is satisfied, and Theorem 4 shows that $s^\star$ becomes more and more irregular. This result can be recast in the diffusion framework (1)–(3). Indeed, in this situation where $\mu = e^{-t}$ and $\sigma = \sqrt{1 - e^{-2t}}$, the condition of the theorem is satisfied when t is close to 0. For small t, we have $\frac{\mu^3}{\sigma^4} = \frac{e^{-3t}}{(1-e^{-2t})^2} \geqslant \frac{1}{8t^2}$, and thus

$$\mathrm{TV}_\pi^{(1)}(s^\star) \geqslant \frac{Cn\Delta^3}{8t^2}.$$

The conclusion of this section is that $s^\star$ is highly irregular (at least in the sense of the $\mathrm{TV}_\pi^{(1)}$ measure) when σ is small. This suggests that gradient descent could struggle to learn it, as it is known to exhibit an inductive bias toward learning regular functions (see, e.g., Bach, 2024, Section 12.1). This provides a strong initial argument against the possibility of full memorization. We formalize this intuition in the next section within the context of two-layer neural networks.

6. Implicit regularization and memorization

We show that the mechanism of SGD protects the two-layer neural networks defined in Section 4 from memorization. This goal is achieved in Theorem 8 below by proving that SGD cannot converge to a local minimum $\theta^\star$ with low risk $\mathcal{R}_n(\theta^\star)$, unless the learning rate is small. To establish this result, we study the regularity of $s_{\theta^\star}$ as measured by $\mathrm{TV}_\pi^{(1)}(s_{\theta^\star})$ within the linear stability framework outlined in Section 4. This approach requires the risk $\mathcal{R}_n$ to be twice differentiable at $\theta^\star$. We start by showing that, in fact, it is twice differentiable everywhere.

Lemma 5 *For all $\theta = (w_{1:m}^{(2)}, b_{1:m}) \in \mathbb{R}^{2m}$, the risk $\mathcal{R}_n(\theta)$ is twice differentiable with respect to θ.*

The following proposition gives an upper bound on $\mathrm{TV}_\pi^{(1)}(s_{\theta^\star})$ when $\theta^\star$ is a linearly stable minimum, expressed in terms of the loss and the inverse of the learning rate. Note that, since s''_θ is a sum of Diracs, $\mathrm{TV}_\pi^{(1)}(s_\theta)$ is computed in the sense of the theory of distributions (this operation is possible since $\pi(y;\mu,\sigma)$ is a smooth function).

Proposition 6 *Let $\theta = (w_{1:m}^{(2)}, b_{1:m}) \in \mathbb{R}^{2m}$. Then*

$$\mathrm{TV}_\pi^{(1)}(s_\theta) \leqslant \frac{\lambda_{\max}(\nabla_\theta^2 \mathcal{R}_n(\theta))m}{4} + \frac{\sqrt{\mathcal{R}_n(\theta)}}{2} + \frac{A}{\sqrt{2\pi}\sigma}\max\Big(\sqrt{2n\mathcal{R}_n(\theta)}, \big(\sqrt{2\pi e}A\sigma n\mathcal{R}_n(\theta)\big)^{\frac{1}{3}}\Big).$$

In particular, if $\theta^\star$ is a linearly stable local minimum of $\mathcal{R}_n$, one has

$$\mathrm{TV}_\pi^{(1)}(s_{\theta^\star}) \leqslant \frac{1}{2\eta} + \frac{\sqrt{\mathcal{R}_n(\theta^\star)}}{2} + \frac{A}{\sqrt{2\pi}\sigma} \max\left(\sqrt{2n\mathcal{R}_n(\theta^\star)}, \left(\sqrt{2\pi e} A\sigma n \mathcal{R}_n(\theta^\star)\right)^{\frac{1}{3}}\right).$$

The proof of the first statement consists in carefully assessing the magnitude of the terms in the Hessian of the risk. A key step is to lower bound the largest eigenvalue of the neural tangent kernel term by $\mathrm{TV}_\pi^{(1)}(s_\theta)$. The second identity of the proposition then directly follows from (8).

Symmetrically, we next provide a lower bound on $\mathrm{TV}_\pi^{(1)}(s_\theta)$ for low-enough values of the risk.

Proposition 7 *If $\Delta \geqslant 8\frac{\sigma}{\mu}$, then for any θ such that $\mathcal{R}_n(\theta) \leqslant \frac{1}{16n\sigma^2}$, one has $\mathrm{TV}_\pi^{(1)}(s_\theta) \geqslant \frac{\mu n^2 \Delta}{2^{11}\sigma^2}$.*

This result is connected with Theorem 4. Indeed, the theorem gives a lower bound on $\mathrm{TV}_\pi^{(1)}(s_\theta)$ in the case where $s_\theta = s^\star$. Then, Proposition 7 relaxes this bound to all neural networks with small enough risk. Combining the above upper and lower bounds on the $\mathrm{TV}_\pi^{(1)}$ metric, we see that a low-risk *and* linearly stable minimum of $\mathcal{R}_n$ imposes a lower bound on $1/\eta$ of the order of $1/\sigma^2$. This observation, combined with Corollary 3 and elementary computations, leads to our main result.

Theorem 8 *Let $\theta^\star \in \mathbb{R}^{2m}$ be a linearly stable local minimum of $\mathcal{R}_n$. Then there exists $\sigma_0 > 0$, depending on μ and the training sample, such that if $\sigma \leqslant \sigma_0$ and $\eta > \frac{2^{12}\sigma^2}{\mu n^2 \Delta}$, one has*

$$\mathcal{R}_n(\theta^\star) - \mathcal{R}_n(s^\star) > \frac{\pi n^5 \mu^3 \Delta^3}{2^{36} e^{1/2} A^4 \sigma^4}.$$

The main message is that for a fixed (small enough) σ, if η is sufficiently large, then the excess risk $\mathcal{R}_n(\theta^\star) - \mathcal{R}_n(s^\star)$ cannot be made arbitrarily small. This is equivalent to stating that $s_{\theta^\star}$ cannot be arbitrarily close to $s^\star$, as can be seen by reformulating the theorem's conclusion as

$$\frac{\pi n^5 \mu^3 \Delta^3}{2^{36} e^{1/2} A^4 \sigma^4} < \mathcal{R}_n(\theta^\star) - \mathcal{R}_n(s^\star) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i, \sigma^2)}\left[(s_{\theta^\star}(Y) - s^\star(Y;\mu,\sigma))^2\right].$$

We refer to, e.g., Coste (2023) for a proof of this identity. The right-hand side can be interpreted as a weighted L_2 distance between $s_{\theta^\star}$ and $s^\star$, assigning greater weight around the noisy observations. In the small-noise regime, i.e., when $\sigma \to 0$, the two conditions of Theorem 8 are automatically satisfied. This is true in particular for diffusions (1)–(3) as $t \to 0$. Therefore, in this context, Theorem 8 suggests that setting a large learning rate prevents memorization of the training sample.

7. Experiments

In this section, we experimentally assess the closeness of the learned model $s_{\theta^\star}$ to the empirical optimal score $s^\star$, as well as the memorization effect, depending on the learning rate and the dimension of the data. Following our theoretical framework, we fix the model to be a 2-layer ReLU network of width 1000. In all experiments, the number of epochs scales inversely with the learning rate, to ensure comparable convergence across models. Experimental details are given in Appendix E.

Connection between the learning rate and the proximity of $s_{\theta^\star}$ to $s^\star$. For the first experiment, the training data $x_1, \dots, x_{20}$ are 20 i.i.d. samples of the one-dimensional standard Gaussian. We perform SGD on the score matching risk (5), for fixed values of μ and σ (taken so $\mu^2 + \sigma^2 = 1$). Figure 1 shows the graphs of the learned models $s_{\theta^\star}$ trained with different learning rates, together with the empirical optimal score $s^\star$. As expected from our theory (Theorem 8), we observe that a larger learning rate η or smaller noise variance σ prevents $s_{\theta^\star}$ from converging to $s^\star$. This leads to a larger excess risk, which is confirmed in Figure 2 (left). We present in Figure 2 (right) the result of the analogous experiment in dimension 10, highlighting the same pattern. This provides evidence that the model should not (fully) memorize the data, as we further investigate next. We also report the largest eigenvalue of the loss Hessian at the end of training for different learning rates (see Appendix E Figure 5 (left and middle)), which confirms that taking a larger learning rate leads to convergence to a flatter region of parameter space, which is key behind our analysis.

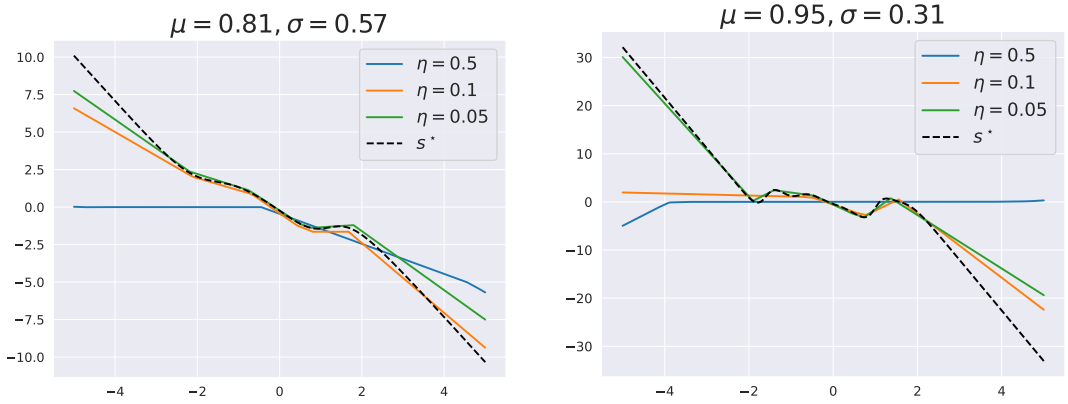

Figure 1: Graphs of the learned model $s_{\theta^\star}$ with different learning rates and of the empirical optimal score $s^\star$, for two pairs of (μ, σ). As the learning rate decreases, $s_{\theta^\star}$ approaches $s^\star$. When σ is smaller (right plot), $s^\star$ is more irregular, and a smaller learning rate is needed for $s_{\theta^\star}$ to approach $s^\star$.

Learning rate and memorization effect for denoising diffusions. In dimension $d = 2$, we sample 10 isotropic Gaussian observations, and aim to generate new ones using a diffusion model. We perform denoising score matching to learn the score $(t, x) \in \mathbb{R}^{d+1} \mapsto \nabla \log p_t(x) \in \mathbb{R}^d$ with a two-layer neural network $s_\theta(t, x)$ fitted on noisy observations (for various noise variances $\sigma(t)$). Running the diffusion with the learned score, we observe in Figure 3 (left and middle) that a small learning rate leads to generating observations close to the training data, indicating memorization. As expected, simulating the diffusion with $s^\star$ also induces memorization. In contrast, a larger learning rate leads to observations closer to the target distribution. This is further verified by measuring the maximum mean discrepancy (MMD) between generated and training data (Figure 3, right). This figure also suggests that a larger learning rate not only avoids memorization but also learns a Gaussian distribution fitted on the training sample. This is not too surprising since larger learning rates constraint the total variation of the derivative (Proposition 6); in the limit where $\mathrm{TV}_\pi^{(1)}(s_{\theta^\star}) \to 0$, the model can only implement a linear function, and the optimal linear score generates such a Gaussian distribution. This is also in line with findings of Li et al. (2024b)—see Section 2. In addition, we report the largest eigenvalue of the loss Hessian at thet end of training for different learning rates in Appendix E Figure 5 (right).

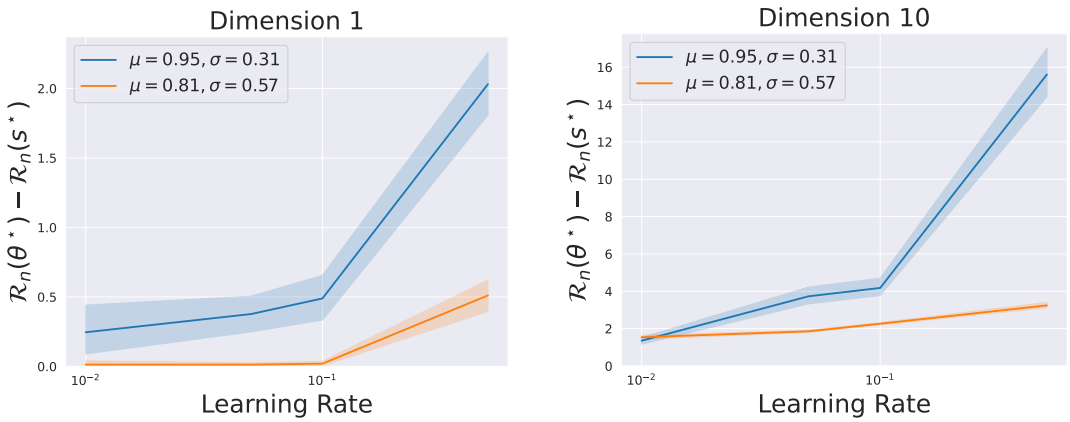

Figure 2: Excess risk of the learned model $s_{\theta^\star}$ trained with different learning rates, for two pairs of (μ, σ) and two dimensions of the data ($d = 1$, left, and $d = 10$, right). The x-axis is in logarithmic scale while the y-axis is in standard scale. Confidence intervals are computed with 30 simulations.

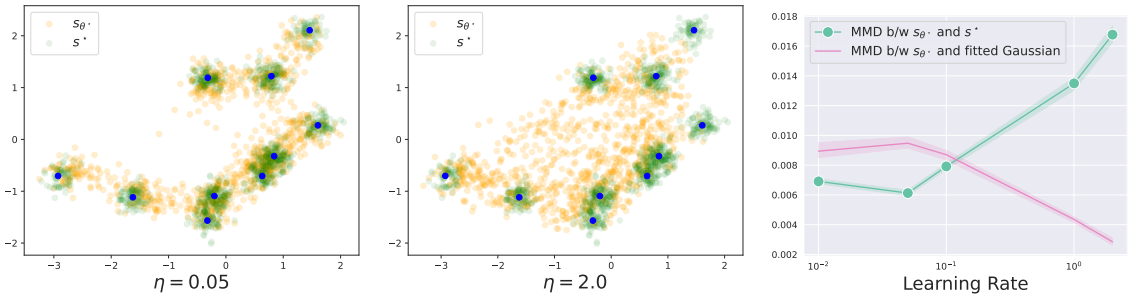

Figure 3: (left) Sample generated by $s^\star$ and $s_{\theta^\star}$ fitted with learning rate 0.05. The training data are the blue points. (middle) Same with $s_{\theta^\star}$ fitted with learning rate 2. (right) The green marked curve corresponds to the MMD between observations generated by $s^\star$ and observations generated by $s_{\theta^\star}$ (for different learning rates). The pink curve is the MMD between observations following the Gaussian distribution fitted on the training data and observations generated by $s_{\theta^\star}$.

Dimension and memorization effect for denoising diffusions. In this final experiment, we examine the effect of the dimension on memorization, while keeping the learning rate fixed and following the same experimental procedure as previously. Since our results depend on the minimum spacing between data points, which scales with dimensionality, memorization is expected to be less prominent in the high-dimensional regime—a common scenario in image generation. This is confirmed by Figure 4 (left and middle), which shows that, as the dimension increases, the fitted neural network $s_{\theta^\star}$ generates observations that are less similar to the training observations. Figure 4 (right) confirms this finding by measuring the MMD between observations generated by $s_{\theta^\star}$ and observations generated by $s^\star$. This indicates that avoiding memorization is easier in a high-dimensional setting. Further, in high dimension, the network seems to be learning a Gaussian distribution.

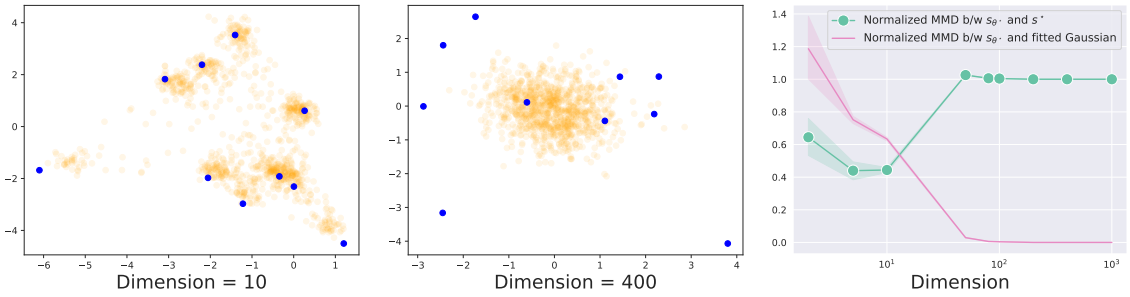

Figure 4: (left) Sample generated by $s_{\theta^\star}$ in dimension 10, projected on the first two axes. The training data are the blue points. (middle) Same in dimension 400. (right) The green marked curve corresponds to the MMD between observations generated by $s^\star$ and observations generated by $s_{\theta^\star}$, depending on the dimension. The pink curve is the MMD between observations following the Gaussian distribution fitted on the training data and observations generated by $s_{\theta^\star}$. Both distances are normalized by the MMD between observations generated by $s^\star$ and by the Gaussian distribution.

8. Conclusion

In this paper, we provide a theoretical explanation for why neural networks do not fully memorize the training data. Specifically, our main result, Theorem 8, shows that using a large learning rate acts as a regularization mechanism. This mechanism prevents the network from converging to the empirical optimal score, which becomes unstable under stochastic gradient descent.

Our approach can be extended in several directions. While our theoretical results focus on the one-dimensional case, it is our belief that a good understanding of the one-dimensional problem provides a solid basis for the more complex study of higher dimensional cases, for example by revealing the crucial role played by the data minimal spacing. On top of this, our experiments strongly suggest that similar results hold in a multivariate setting. Future work could explore this extension, for example by leveraging a multivariate version of the minimum stability framework (Nacson et al., 2023). Next, our results highlight the critical relationship between the learning rate η and the noise variance σ. However, the effects of the sample size n and the dimension d remain unclear. An interesting question is to analyze the role of these hyperparameters, connecting with the literature discussing the impact of n (see Section 2). Finally, even if our results indicate that η should not be too small to prevent memorization, they do not guarantee that a larger learning rate improves the quality of the generated data. Exploring trade-offs between memorization, generation quality, and training speed is a key avenue for future research.

Acknowledgments. Authors warmly thank Florentin Guth for insightful discussions that motivated them towards this line of research, as well as Peter Bartlett, Raphaël Berthier, and Serena Booth for helpful comments. P.M. is supported by a Google PhD Fellowship.

References

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2023.

Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. SegDiff: Image segmentation with diffusion probabilistic models. *arXiv:2112.00390*, 2021.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with large step sizes learns sparse features. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 903–925. PMLR, 2023.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17981–17993. Curran Associates, Inc., 2021.

Francis Bach. *Learning Theory from First Principles*. MIT press, Cambridge, Massachusetts, 2024.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B Kovachki, Assad Oberai, and Andrew M Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *The Tenth International Conference on Learning Representations*, 2022.

J. Benton, V. de Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 57795–57824. Curran Associates, Inc., 2023.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *The Ninth International Conference on Learning Representations*, 2021.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3040–3050. Curran Associates, Inc., 2018.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 2937–2947. Curran Associates, Inc., 2019.

Simon Coste. Diffusion models, 2023. URL https://scoste.fr/posts/diffusion/#denoising_score_matching. Accessed on 2025-01-20.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 288–313. Curran Associates, Inc., 2023.

Robert Donald Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12 (3):364–366, 1941.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv:2310.02664*, 2023.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model. *arXiv:2401.04856*, 2024a.

Xiang Li, Yixiang Dai, and Qing Qu. Understanding generalizability of diffusion models requires rethinking the hidden Gaussian structure. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2024b.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 11674–11685. Curran Associates, Inc., 2019.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.

Peyman Milanfar and Mauricio Delbracio. Denoising: A powerful building-block for imaging, inverse problems, and machine learning. *arXiv:2409.06219*, 2024.

Koichi Miyasawa. An empirical Bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38:181–188, 1961.

Stanislav A. Molchanov and A. Ya. Reznikova. Limit theorems for random partitions. *Theory of Probability & Its Applications*, 27:310–323, 1983.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17749–17761. Curran Associates, Inc., 2021.

Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Haikady Navada Nagaraja, Karthik Bharath, and Fangyuan Zhang. Spacings around an order statistic. *Annals of the Institute of Statistical Mathematics*, 67:515–540, 2015.

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. *arXiv:2406.06838*, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 66377–66389. Curran Associates, Inc., 2023.

Herbert Robbins. An empirical Bayes approach to statistics. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163, 1956.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the EEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75:1889–1935, 2022.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. Step-unrolled denoising autoencoders for text generation. In *The Tenth International Conference on Learning Representations*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265. PMLR, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47783–47803. Curran Associates, Inc., 2023b.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Afonso M. Teodoro, José M. Bioucas-Dias, and Mário A.T. Figueiredo. Image restoration and reconstruction using variable splitting and class-adapted image priors. In *2016 IEEE International Conference on Image Processing*, pages 3518–3522. IEEE, 2016.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23:1661–1674, 2011.

Nikhil Vyas, Sham M. Kakade, and Boaz Barak. On provable copyright protection for generative models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 35277–35299. PMLR, 2023.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Zihui Wu, Yu Sun, Jiaming Liu, and Ulugbek Kamilov. Online regularization by denoising with applications to phase retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 11727–11737. PMLR, 2021.

Mingyang Yi, Jiacheng Sun, and Zhenguo Li. On the generalization of diffusion model. *arXiv:2305.14712*, 2023.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Chen Zeno, Greg Ongie, Yaniv Blumenfeld, Nir Weinberger, and Daniel Soudry. How do minimum-norm shallow denoisers look in function space? In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2024.

Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, Haonan Wang, and Kenji Kawaguchi. On copyright risks of text-to-image diffusion models. *arXiv:2311.12803*, 2023.

Appendix

Organization of the Appendix. Appendix A presents two additional propositions of interest preliminary to the proofs of the results of the main text, which are given in Appendix B. Appendix C is dedicated to technical lemmas. Finally, Appendix E details our experimental setting.

Appendix A. Auxiliary propositions

The next proposition provides bounds on the weight function π defined in Section 5.

Proposition 9 *Let $1 \leqslant i \leqslant n-1$ and $y \in [\mu x_i, \mu x_{i+1}]$. Then*

$$\pi(y;\mu,\sigma) \geqslant \frac{\mu}{n^2}\Big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{2\sigma^2}}\Big)\min\Big(\frac{i^2(i-1)}{2}, \frac{(n-i)^2(n-i+1)}{2}\Big)\Delta.$$

On the other hand, for any $y \in \mathbb{R}$,

$$\pi(y;\mu,\sigma) \leqslant \mu(x_n - x_1).$$

Proof Recall that

$$\pi^-(y;\mu,\sigma) = \mathbb{P}(\mu U < y)^2 \mathbb{E}(y - \mu U | \mu U < y).$$

Lower bound. Let $x = \frac{y}{\mu}$. Then, clearly, $y \in [\mu x_i, \mu x_{i+1}]$ is equivalent to $x \in [x_i, x_{i+1}]$. Hence,

$$\mathbb{P}(\mu U < y)^2 = \mathbb{P}(U < x)^2 = \frac{i^2}{n^2}$$

and

$$\mathbb{E}[y - \mu U | \mu U < y] = \mu \mathbb{E}[x - U | U < x] = \mu\Big(x - \frac{1}{i}\sum_{i'=1}^{i} x_{i'}\Big).$$

Therefore,

$$\pi^-(\mu x;\mu,\sigma) = \frac{\mu i^2}{n^2}\Big(x - \frac{1}{i}\sum_{i'=1}^{i} x_{i'}\Big) \geqslant \frac{\mu i}{n^2}\sum_{i'<i}(x_i - x_{i'}),$$

and, similarly,

$$\pi^+(\mu x;\sigma,\mu) = \frac{\mu(n-i)^2}{n^2}\Big(\frac{1}{n-i}\Big(\sum_{i'=i+1}^{n} x_{i'}\Big) - x\Big) \geqslant \frac{\mu(n-i)}{n^2}\sum_{i'>i+1}(x_{i'} - x_{i+1}).$$

Recall that

$$\pi(\mu x;\mu,\sigma) = \mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)} \min(\pi^-(\mu x - \xi;\mu,\sigma), \pi^+(t, \mu x - \xi;\mu,\sigma)).$$

Since ξ in the expectation is Gaussian with mean 0, it is symmetric, and we may rewrite π as

$$\pi(\mu x;\mu,\sigma) = \mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)} \min(\pi^-(\mu x + \xi;\mu,\sigma), \pi^+(\mu x + \xi;\mu,\sigma)).$$

Using the previous bounds on π^+ and π^-, we obtain

$$\pi(\mu x;\mu,\sigma) \geqslant \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mu(x_i - x)}^{\mu(x_{i+1}-x)} \min(\pi^-(\mu x + z;\mu,\sigma), \pi^+(\mu x + z;\mu,\sigma))e^{-\frac{z^2}{2\sigma^2}}dz$$

$$\geqslant \frac{\mu}{n^2} \min\Big(i\sum_{i'<i}(x_i - x_{i'}), (n-i)\sum_{i'>i+1}(x_{i'} - x_{i+1})\Big) \times \mathbb{P}_{z\sim\mathcal{N}(0,1)}(\mu(x_i - x) < \sigma z < \mu(x_{i+1} - x)).$$

To bound the last term, we use the fact that $0 \in (\mu(x_i - x), \mu(x_{i+1} - x))$ and thus

$$\begin{aligned}
&\mathbb{P}(\mu(x_i - x) < \sigma z < \mu(x_{i+1} - x)) \\
&\quad = \mathbb{P}_{z\sim\mathcal{N}(0,1)}(\mu(x_i - x) < \sigma z < 0) + \mathbb{P}_{z\sim\mathcal{N}(0,1)}(0 < \sigma z < \mu(x_{i+1} - x)) \\
&\quad = \mathbb{P}_{z\sim\mathcal{N}(0,1)}(0 < \sigma z < \mu(x - x_i)) + \mathbb{P}_{z\sim\mathcal{N}(0,1)}(0 < \sigma z < \mu(x_{i+1} - x)) \\
&\quad \geqslant \mathbb{P}_{z\sim\mathcal{N}(0,1)}(\mu(x_{i+1} - x) < \sigma z < \mu(x_{i+1} - x) + \mu(x - x_i)) \\
&\qquad + \mathbb{P}_{z\sim\mathcal{N}(0,1)}(0 < \sigma z < \mu(x_{i+1} - x)), \\
&\quad = \mathbb{P}_{z\sim\mathcal{N}(0,1)}\Big(\sigma z \in (0, \mu(x_{i+1} - x)) \cup \big(\mu(x_{i+1} - x), \mu(x_{i+1} - x_i)\big)\Big) \\
&\quad = \mathbb{P}_{z\sim\mathcal{N}(0,1)}\big(\sigma z \in (0, \mu(x_{i+1} - x_i))\big).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\pi(\mu x; \mu, \sigma) &\geqslant \frac{\mu}{n^2} \min\Big(i\sum_{i'<i}(x_i - x_{i'}), (n-i)\sum_{i'>i+1}(x_{i'} - x_{i+1})\Big) \\
&\quad \times \mathbb{P}_{z\sim\mathcal{N}(0,1)}(0 < \sigma z < \mu(x_{i+1} - x_i)) \\
&\geqslant \frac{\mu}{n^2} \min\Big(i\sum_{i'<i}(x_i - x_{i'}), (n-i)\sum_{i'>i+1}(x_{i'} - x_{i+1})\Big)\Big(\frac{1}{2} - e^{-\frac{\mu^2(x_i - x_{i+1})^2}{2\sigma^2}}\Big).
\end{aligned}$$

In the last inequality, we used a tail bound of the Gaussian distribution for the last inequality, namely, $\mathbb{P}_{z\sim\mathcal{N}(0,\sigma^2)}(z \geqslant t) \leqslant e^{-\frac{t^2}{2\sigma^2}}$ (see, for instance, Gordon, 1941). To derive the lower bound of the Proposition, note that

$$\begin{aligned}
\sum_{i'<i}(x_i - x_{i'}) &\geqslant \sum_{i'=1}^{i-1}(i - i')\Delta = \frac{i(i-1)}{2}\Delta, \\
\sum_{i'>i}(x_{i'} - x_i) &\geqslant \sum_{i'=i+1}^{n}(i' - i)\Delta = \frac{(n-i)(n-i+1)}{2}\Delta.
\end{aligned}$$

Upper bound. Again, let $x = \frac{y}{\mu}$. Observe that, for $y \in [\mu x_1, \mu x_n]$, we have $x \in [x_1, x_n]$. Therefore,

$$\mathbb{E}[y - \mu U | \mu U < y] = \mu\mathbb{E}[x - U | U < x] \leqslant \mu(x_1 - x_n).$$

So,

$$\pi^-(y; \mu, \sigma) \leqslant \mu(x_n - x_1).$$

We may also upper bound $\pi^+(y; \mu, \sigma)$ with the same value. By taking the expectation, we have

$$\pi(y; \mu, \sigma) \leqslant \mu(x_n - x_1).$$

■

The next result is a key technical component of our analysis. It lower bounds the largest eigenvalue of the neural tangent kernel by $2/m$ times $\mathrm{TV}_\pi^{(1)}(s_\theta)$. The proof technique is inspired by Mulayoff et al. (2021, Lemma 4).

Proposition 10 *For any $s_\theta \in \mathcal{S}$,*

$$\lambda_{\max}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}[(\nabla_\theta s_\theta(Y))(\nabla_\theta s_\theta(Y))^\top] \geqslant \frac{2}{m}\int_{\mathbb{R}}|s_\theta''(y)|\pi(y;\mu,\sigma)dy.$$

Proof We start by rewriting the matrix on the left-hand side of the inequality. We have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[(\nabla_\theta s_\theta(Y))(\nabla_\theta s_\theta(Y))^\top\big]\\ &\quad=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}\big[(\nabla_\theta s_\theta(\xi+\mu x_i))(\nabla_\theta s_\theta(\xi+\mu x_i))^\top\big].\end{aligned}$$

Let $\Phi(\xi)=[\nabla_\theta s_\theta(\xi+\mu x_1),\ldots,\nabla_\theta s_\theta(\xi+\mu x_n)]\in\mathbb{R}^{3m\times n}$, and notice that the left-hand side is $\frac{1}{n}\mathbb{E}\Phi(\xi)\Phi(\xi)^\top$. Denoting $\mathbb{S}^{d-1}$ the sphere of $\mathbb{R}^d$, we may then deduce that

$$\begin{aligned}\lambda_{\max}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[(\nabla_\theta s_\theta(Y))(\nabla_\theta s_\theta(Y))^\top\big]\Big)&=\frac{1}{n}\max_{v\in\mathbb{S}^{2m-1}}\mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}[v^\top\Phi(\xi)\Phi(\xi)^\top v],\\&=\frac{1}{n}\max_{v\in\mathbb{S}^{2m-1}}\mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}[\|\Phi(\xi)^\top v\|^2],\\&=\frac{1}{n}\max_{u\in\mathbb{S}^{n-1}}\mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}[\|\Phi(\xi)u\|^2],\\&\geqslant\frac{1}{n}\mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}[\frac{1}{n}\|\Phi(\xi)\mathbf{1}\|^2],\end{aligned}$$

where $\mathbf{1}=(1,1,\ldots,1)^\top\in\mathbb{R}^n$, so that $\frac{1}{\sqrt{n}}\mathbf{1}$ is a unit vector. To this aim, let us lower bound $\frac{1}{n^2}\|\Phi(\xi)\mathbf{1}\|^2$ for a fixed $\xi\in\mathbb{R}^n$, and then take the expectation with respect to ξ. First, let $I_{i,\ell}=\delta(w_\ell^{(1)}(\xi+\mu x_i)+b_\ell>0)$, where $\delta(\cdot)$ equals 1 if the condition is satisfied otherwise equals 0, and $\mathbb{I}_i\in\mathbb{R}^m$ the vector whose ℓ-th entry is $I_{i,\ell}$. We may then calculate the gradient of s_θ as follows:

$$\nabla_\theta s_\theta(y)=\begin{pmatrix}\nabla_{\mathbf{w}^{(2)}}s_\theta(y)\\ \nabla_{\mathbf{b}}s_\theta(y)\end{pmatrix}=\begin{pmatrix}\frac{1}{m}\big(y\mathbf{w}^{(1)}+\mathbf{b}\big)\odot\mathbb{I}_i\\ \frac{1}{m}\mathbf{w}^{(2)}\odot\mathbb{I}_i\end{pmatrix}.$$

Then, by the inequality $a^2+b^2\geqslant 2ab$,

$$\begin{aligned}\frac{1}{n^2}\|\Phi(\xi)\mathbf{1}\|^2&=\frac{1}{m^2n^2}\sum_{\ell=1}^{m}\Big[\Big(\sum_{i=1}^{n}\phi\big(w_\ell^{(1)}(\xi+\mu x_i)+b_\ell\big)\Big)^2+\Big(\sum_{i=1}^{n}I_{i,\ell}w_\ell^{(2)}\Big)^2\Big]\\&\geqslant\frac{2}{m^2n^2}\sum_{\ell=1}^{m}\big|w_\ell^{(2)}\big|\times\Big|\sum_{i=1}^{n}\phi\big(w_\ell^{(1)}(\xi+\mu x_i)+b_\ell\big)\Big|\times\Big(\sum_{i=1}^{n}I_{i,\ell}\Big).\end{aligned}$$

Define $C_\ell = \{y \in \mathbb{R}, w_\ell^{(1)} y + b_\ell > 0\}$ and $n_\ell = \sum_{i=1}^n I_{i,\ell} = |C_\ell|$. Recall that U denotes a random variable drawn uniformly from the training data $\{x_j\}_{1 \leqslant j \leqslant n}$. Then, we have

$$\begin{aligned}\frac{1}{n^2}\|\Phi(\xi)\mathbf{1}\|^2 &\geqslant \frac{2}{m^2}\sum_{\ell=1}^m \Big(\frac{n_\ell}{n}\Big)^2 |w_\ell^{(2)}| \times \Big|\frac{1}{n_\ell}\sum_{(\xi+\mu U)\in C_\ell} w_\ell^{(1)}(\xi+\mu U) + b_\ell\Big| \\ &= \frac{2}{m^2}\sum_{\ell=1}^m \mathbb{P}((\xi+\mu U)\in C_\ell)^2 \times |w_\ell^{(2)}| \times \Big|\mathbb{E}\Big[w_\ell^{(1)}(\xi+\mu U) + b_\ell \mid (\xi+\mu U)\in C_\ell\Big]\Big|,\end{aligned}$$

where the probability and expectation are taken with respect to U. Next, define $\tau_\ell = -b_\ell / w_\ell^{(1)}$. Since $|w_\ell^{(1)}| = 1$, we may then rewrite

$$\mathbb{E}\big[w_\ell^{(1)}(\xi+\mu U) + b_\ell \big| (\xi+\mu U)\in C_\ell\big] = \mathbb{E}\big[|\xi+\mu U - \tau_\ell|\big|(\xi+\mu U)\in C_\ell\big].$$

Notice that the set C_ℓ is an interval with one end at $\pm\infty$ and another at τ_ℓ, depending on the sign of $w_\ell^{(1)}$. If $w_\ell^{(1)} = 1$, we have $C_\ell = (-b_\ell, \infty) = (\tau_\ell, \infty)$, and

$$\begin{aligned}&\mathbb{P}((\xi+\mu U)\in C_\ell)^2 \big|\mathbb{E}\big[\xi+\mu U - \tau_\ell \big| \xi+\mu U)\in C_\ell\big]\big| \\ &\qquad = \mathbb{P}(\mu U > \tau_\ell - \xi)^2 \times \mathbb{E}\big[\mu U - \tau_\ell + \xi \big| \mu U > \tau_\ell - \xi\big] \\ &\qquad = \pi^+(\tau_\ell - \xi),\end{aligned}$$

where we recall that π^+ is defined in Section 5. Likewise, if $w_\ell^{(1)} = -1$, we obtain

$$\mathbb{P}((\xi+\mu U)\in C_\ell)^2 \big|\mathbb{E}\big[\xi+\mu U - \tau_\ell \big| \xi+\mu U)\in C_\ell\big]\big| = \pi^-(\tau_\ell - \xi).$$

All in all, we get, for all $\ell \in [1, m]$,

$$\mathbb{P}((\xi+\mu U)\in C_\ell)^2 \times \big|\mathbb{E}\big[(\xi+\mu U) - \tau_\ell \big| (\xi+\mu U)\in C_\ell\big]\big| \geqslant \min\{\pi^+(\tau_\ell - \xi;\mu,\sigma), \pi^-(\tau_\ell - \xi;\mu,\sigma)\}.$$

Thus, since $s_\theta''(y) = \frac{1}{m}\sum_{\ell=1}^m w_\ell^{(1)} w_\ell^{(2)} \delta(y = \tau_\ell)$, we are led to

$$\begin{aligned}\frac{1}{n^2}\|\Phi(\xi)\mathbf{1}\|^2 &\geqslant \frac{2}{m^2}\sum_{\ell=1}^m |w_\ell^{(2)}| \min\{\pi^+(\tau_\ell - \xi;\mu,\sigma), \pi^-(\tau_\ell - \xi;\mu,\sigma)\}, \\ &\geqslant \frac{2}{m}\int_{\mathbb{R}} |s_\theta''(y)| \min\{\pi^+(y - \xi;\mu,\sigma), \pi^-(y - \xi;\mu,\sigma)\} dy.\end{aligned}$$

We may then take the expectation value on both side and conclude that

$$\lambda_{\max}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[(\nabla_\theta s_\theta(Y))(\nabla_\theta s_\theta(Y))^\top\big]\Big) \geqslant \frac{2}{m}\int_{\mathbb{R}} |s_\theta''(y)| \pi(y;\mu,\sigma) dy.$$

■

Appendix B. Proofs of the results of the main text

B.1. Proof of Proposition 2

We start by proving that

$$|\mathbb{E}[W(y;\mu,\sigma)] - x_i| \leqslant n\Delta e^{-\frac{\mu^2\Delta^2}{4\sigma^2}},$$

when $|y - \mu x_i| \leqslant \frac{\mu}{4}\Delta$. Let $x = \frac{y}{\mu}$, then $|y - \mu x_i| \leqslant \frac{\mu}{4}\Delta$ implies that $|x - x_i| \leqslant \frac{\Delta}{4}$. We show that

$$\mathbb{E}[W(\mu x;\mu,\sigma)] - x_i \leqslant n\Delta e^{-\frac{\mu^2\Delta^2}{4\sigma^2}}, \tag{11}$$

and the other side can be deduced similarly. Observe that

$$\begin{aligned}
\mathbb{E}[W(\mu x;\mu,\sigma)] - x_i &= \sum_{i'=1}^{n}(x_{i'} - x_i)\alpha_{i'}(\mu x;\mu,\sigma),\\
&\leqslant \frac{\sum_{i'=i+1}^{n}(x_{i'} - x_i)e^{-\frac{\mu^2(x_{i'}-x)^2}{2\sigma^2}}}{\sum_{i'=1}^{n} e^{-\frac{\mu^2(x_{i'}-x)^2}{2\sigma^2}}},\\
&= \frac{\sum_{i'=i+1}^{n}(x_{i'} - x_i)e^{-\frac{\mu^2\left((x_{i'}-x)^2-(x_i-x)^2\right)}{2\sigma^2}}}{\sum_{i'=1}^{n} e^{-\frac{\mu^2\left((x_{i'}-x)^2-(x_i-x)^2\right)}{2\sigma^2}}},\\
&\leqslant \sum_{i'=i+1}^{n}(x_{i'} - x_i)e^{-\frac{\mu^2(x_{i'}-x_i)(x_{i'}+x_i-2x)}{2\sigma^2}},
\end{aligned}$$

where, in the last inequality, we used the fact that when $i' = i$, $e^{-\frac{\mu^2\left((x_{i'}-x)^2-(x_i-x)^2\right)}{2\sigma^2}} = 1$, and so the denominator is larger than 1. Next, for $i' > i$, with the condition

$$-\frac{x_{i'} - x_i}{4} \leqslant -\frac{\Delta}{4} \leqslant x_i - x \leqslant 0,$$

which implies that $x_{i'} - x \geqslant \frac{3}{4}(x_{i'} - x_i)$, we have $x_{i'} + x_i - 2x \geqslant \frac{1}{2}(x_{i'} - x_i) \geqslant 0$. Therefore,

$$\begin{aligned}
\mathbb{E}[W(\mu x;\mu,\sigma)] - x_i &\leqslant \sum_{i'=i+1}^{n}(x_{i'} - x_i)e^{-\frac{\mu^2(x_i-x_{i'})^2}{4\sigma^2}}\\
&\leqslant n\Delta e^{-\frac{\mu^2\Delta^2}{4\sigma^2}},
\end{aligned} \tag{12}$$

where we apply Lemma 13 with $k = 1$ and $A = \frac{\mu^2}{2\sigma^2}$ the function $x \mapsto xe^{-\frac{\mu^2x^2}{4\sigma^2}}$ is strictly decreasing on $[\sqrt{2}\frac{\sigma}{\mu}, \infty)$, and $x_{i'} - x_i \geqslant \Delta \geqslant 2\frac{\sigma}{\mu} \geqslant \sqrt{2}\frac{\sigma}{\mu}$. Hence, we deduce (11). We now turn to the bounds of the variance of W.

Lower bounding $\mathbb{V}[W(m_i;\mu,\sigma)]$**.** We start by observing that

$$\alpha_i(m_i;\mu,\sigma) \geqslant \frac{1}{n} \quad \text{and } \alpha_{i+1}(m_i;\mu,\sigma) \geqslant \frac{1}{n},$$

since $\alpha_i(m_i;\mu,\sigma) = \frac{e^{-\frac{\mu^2(m_i-x_i)^2}{2\sigma^2}}}{\sum_{j=1}^n e^{-\frac{\mu^2(m_i-x_i)^2}{2\sigma^2}}}$ and, for $1 \leqslant j \leqslant n$,

$$e^{-\frac{\mu^2(m_i-x_j)^2}{2\sigma^2}} \leqslant e^{-\frac{\mu^2(m_i-x_i)^2}{2\sigma^2}}.$$

We may then lower bound $\mathbb{V}W(m_i)$ as follows:

$$\begin{aligned}
&\mathbb{V}[W(m_i;\mu,\sigma)] \\
&\quad\geqslant \alpha_i(m_i;\mu,\sigma)(x_i - \mathbb{E}[W(m_i;\mu,\sigma)])^2 + \alpha_{i+1}(m_i;\mu,\sigma)(x_{i+1} - \mathbb{E}[W(m_i;\mu,\sigma)])^2 \\
&\quad\geqslant \frac{1}{n}\big((x_i - \mathbb{E}[W(m_i;\mu,\sigma)])^2 + (x_{i+1} - \mathbb{E}[W(m_i;\mu,\sigma)])^2\big) \\
&\quad\geqslant \frac{1}{2n}(x_i - x_{i+1})^2,
\end{aligned}$$

where the last inequality is derived by applying the Cauchy-Schwarz inequality.

Upper bounding $\mathbb{V}[W(y;\mu,\sigma)]$**.** Let $x = \frac{y}{\mu}$. The condition $|y - \mu x_i| \leqslant \frac{\mu\Delta}{4}$ can be rewritten in terms of x by $|x - x_i| \leqslant \frac{\Delta}{4}$. For $1 \leqslant i' \leqslant n$, we have the following bound, which is a consequence of the Cauchy-Schwarz inequality and the previous paragraph:

$$(x_{i'} - \mathbb{E}[W(\mu x;\mu,\sigma)])^2 \leqslant 2(x_{i'} - x_i)^2 + 2(\mathbb{E}[W(\mu x;\mu,\sigma)] - x_i)^2 \leqslant 2(x_{i'} - x_i)^2 + 2n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{2\sigma^2}}.$$

Thus,

$$\begin{aligned}
&\mathbb{V}[W(\mu x;\mu,\sigma)] \\
&\quad= \alpha_i(\mu x;\mu,\sigma)(x_i - \mathbb{E}[W(\mu x;\mu,\sigma)])^2 + \sum_{i'\neq i} \alpha_{i'}(\mu x;\mu,\sigma)(x_{i'} - \mathbb{E}[W(\mu x;\mu,\sigma)])^2 \\
&\quad\leqslant \alpha_i(\mu x;\mu,\sigma) n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{2\sigma^2}} + 2\sum_{i'\neq i} \alpha_{i'}(\mu x;\mu,\sigma) n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{2\sigma^2}} \\
&\qquad + 2\sum_{i'\neq i}(x_{i'} - x_i)^2 \alpha_{i'}(\mu x;\mu,\sigma) \\
&\quad\leqslant 2n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{2\sigma^2}} + 2\sum_{i'\neq i}(x_{i'} - x_i)^2 e^{-\frac{\mu^2(x_{i'}-x_i)^2}{4\sigma^2}}
\end{aligned}$$

where we apply the same argument as in (12) to bound the last term. Next, applying Lemma 13 with $k = 2$ and $A = \frac{\mu^2}{2\sigma^2}$, we have that $x \mapsto x^2 e^{-\frac{\mu^2 x^2}{4\sigma^2}}$ is strictly decreasing on $[2\frac{\sigma}{\mu}, \infty)$. Since $x_{i'} - x_i \geqslant \Delta \geqslant 2\frac{\sigma}{\mu}$, we obtain

$$\mathbb{V}[W(\mu x;\mu,\sigma)] \leqslant 2n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{2\sigma^2}} + 2n\Delta^2 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \leqslant 4n^2\Delta^2 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}}.$$

This concludes the proof.

B.2. Proof of Corollary 3

The first inequality of the Corollary unfolds from Proposition 2 and equation (9). By applying Proposition 2 we have that

$$\left|s^\star(y;\mu,\sigma) - \frac{1}{\sigma^2}(\mu x_i - y)\right| \leqslant \frac{n\mu\Delta}{\sigma^2} e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \tag{13}$$

We now focus on upper bounding the loss of the empirical optimal score. We have

$$\begin{aligned}
\mathcal{R}_n(s^\star) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}[(s^\star(Y;\mu,\sigma) + \frac{1}{\sigma^2}(Y-\mu x_i))^2] \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\Big(\int_{\mu x_i - \frac{\mu}{4}\Delta}^{\mu x_i + \frac{\mu}{4}\Delta} + \int_{-\infty}^{\mu x_i - \frac{\mu}{4}\Delta} + \int_{\mu x_i + \frac{\mu}{4}\Delta}^{\infty}\Big)(s^\star(y;\mu,\sigma) + \frac{1}{\sigma^2}(y-\mu x_i))^2 e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}} dy \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \frac{n^2\mu^2\Delta^2}{\sigma^4} e^{-\frac{\mu^2\Delta^2}{2\sigma^2}} + \frac{1}{n}\sum_{i=1}^n \frac{\mu^2}{\sigma^4}(x_n - x_1)^2 \mathbb{P}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}(|Y-\mu x_i| > \frac{\mu}{4}\Delta),
\end{aligned}$$

where for the first term in the last inequality we use (13) and for the second term, we use the fact that

$$|s^\star(y;\mu,\sigma) + \frac{1}{\sigma^2}(y-\mu x_i)| = \frac{\mu}{\sigma^2}|x_i - \mathbb{E}[W(y;\mu,\sigma)]|.$$

Since both $W(y;\mu,\sigma)$ and x_i only take value between $[x_1, x_n]$, we have that

$$|s^\star(y;\mu,\sigma) + \frac{1}{\sigma^2}(y-\mu x_i)| \leqslant \frac{\mu}{\sigma^2}(x_n - x_1).$$

We then obtain by applying a tail bound of the Gaussian distribution (Gordon, 1941) that

$$\mathcal{R}_n(s^\star) \leqslant \frac{n^2\mu^2\Delta^2}{\sigma^4} e^{-\frac{\mu^2\Delta^2}{2\sigma^2}} + \frac{2\mu^2(x_n-x_1)^2}{\sigma^4} e^{-\frac{\mu^2\Delta^2}{32\sigma^2}}.$$

Observe that $n\Delta \leqslant 2(x_n - x_1)$. Thus we get

$$\mathcal{R}_n(s^\star) \leqslant \frac{4\mu^2(x_n-x_1)^2}{\sigma^4} e^{-\frac{\mu^2\Delta^2}{32\sigma^2}},$$

which concludes the proof.

B.3. Proof of Lemma 5

Our approach to proving that $\mathcal{R}_n(\theta)$ is twice differentiable involves explicitly computing its Hessian, as the resulting expression will be instrumental in the subsequent proof. First note, since ϕ is differentiable almost everywhere, that by the dominated convergence theorem,

$$\nabla_\theta \mathcal{R}_n(\theta) = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)\nabla_\theta s_\theta(Y)\big],$$

with

$$\nabla_\theta s_\theta(y) = \begin{pmatrix} \nabla_{\mathbf{w}^{(2)}} s_\theta(y) \\ \nabla_{\mathbf{b}} s_\theta(y) \end{pmatrix} = \begin{pmatrix} \frac{1}{m}\phi\big(y\mathbf{w}^{(1)} + \mathbf{b}\big) \\ \frac{1}{m}\mathbf{w}^{(2)} \odot \mathbf{1}_{y\mathbf{w}^{(1)}+\mathbf{b}\geqslant 0} \end{pmatrix}. \tag{14}$$

In the expression above and throughout the remainder of this proof, bold symbols represent vectors in $\mathbb{R}^m$, where each entry corresponds to a neuron. For example, $\mathbf{b} = (b_1, \dots, b_m)$. We will also make use of the notation $\mathrm{diag}(v)$ to denote the square matrix whose diagonal is the vector v. Next, using again that ϕ is differentiable almost everywhere to differentiate a second time the first part of the expression, we obtain

$$\nabla_\theta^2 \mathcal{R}_n(\theta) = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[(\nabla_\theta s_\theta(Y))(\nabla_\theta s_\theta(Y))^\top\big] + \frac{2}{n}\sum_{i=1}^n \nabla_\nu \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)(\nabla_\nu s_\nu(Y))^\top\big]\Big|_{\nu=\theta}.$$

The notation in the second sum means that we are only considering the derivative with respect to the parameters appearing in the gradient term $\nabla_\nu s_\nu(Y)$, and not in the loss term $s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)$. Fixing $i \in \{1,\dots,n\}$, and denoting by $M \in \mathbb{R}^{2m\times 2m}$ the matrix inside the second sum, we observe that M is a block matrix, each block corresponding to differentiating with respect to either $\mathbf{w}$ or $\mathbf{b}$. More precisely,

$$M = \begin{pmatrix} M_{\mathbf{ww}} & M_{\mathbf{wb}} \\ M_{\mathbf{bw}} & M_{\mathbf{bb}} \end{pmatrix},$$

where three blocks $M_{\mathbf{ww}}$, $M_{\mathbf{wb}}$, and $M_{\mathbf{bw}}$ are straightforward to compute, i.e.,

$$M_{\mathbf{ww}} = \mathbf{0}_{m\times m},$$
$$M_{\mathbf{wb}} = M_{\mathbf{bw}} = \frac{1}{m}\,\mathrm{diag}\,\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)\mathbf{1}_{Y\mathbf{w}^{(1)}+\mathbf{b}\geqslant 0}\big],$$

where $\mathbf{0}_{m\times m}$ is the null matrix in $\mathbb{R}^{m\times m}$. Computing the last block $M_{\mathbf{bb}}$ is more delicate, because of the term $\mathbf{1}_{y\mathbf{w}^{(1)}+\mathbf{b}\geqslant 0}$ appearing in the gradient (14) of $\mathcal{R}_n$ with respect to $\mathbf{b}$. Observe that $M_{\mathbf{bb}}$ is a diagonal matrix, whose ℓ-th diagonal element is

$$(M_{\mathbf{bb}})_{\ell\ell} = \partial_b \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)\frac{1}{m}w_\ell^{(2)}\mathbf{1}_{Yw_\ell^{(1)}+b\geqslant 0}\big]\Big|_{b=b_\ell}.$$

Here, this notation indicates once again that we take the derivative only with respect to the term b appearing in the indicator function, and not with respect to $s_\theta(Y)$. To compute this quantity, we consider two cases based on the value of $w_\ell^{(1)} = \pm 1$. If $w_\ell^{(1)} = 1$, we have

$$\begin{aligned}(M_{\mathbf{bb}})_{\ell\ell} &= \frac{1}{m}w_\ell^{(2)}\Big(\partial_b \int_{-b}^{\infty}\big(s_\theta(y) + \frac{1}{\sigma^2}(y-\mu x_i)\big)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}dy\Big)\Big|_{b=b_\ell} \\ &= -\frac{1}{\sqrt{2\pi}m\sigma}w_\ell^{(2)}\big(s_\theta(-b_\ell) + \frac{1}{\sigma^2}(-b_\ell-\mu x_i)\big)e^{-\frac{(-b_\ell-\mu x_i)^2}{2\sigma^2}}.\end{aligned}$$

A similar computation shows that, if $w_\ell^{(1)} = -1$,

$$(M_{\mathbf{bb}})_{\ell\ell} = \frac{1}{\sqrt{2\pi}m\sigma}w_\ell^{(2)}\big(s_\theta(b_\ell) + \frac{1}{\sigma^2}(b_\ell-\mu x_i)\big)e^{-\frac{(b_\ell-\mu x_i)^2}{2\sigma^2}}.$$

Letting $\tau_\ell = -b_\ell/w_\ell^{(1)}$, we can summarize both cases in a single formula:

$$(M_{\mathbf{bb}})_{\ell\ell} = -\frac{1}{\sqrt{2\pi}m\sigma}w_\ell^{(2)}w_\ell^{(1)}\big(s_\theta(\tau_\ell) + \frac{1}{\sigma^2}(\tau_\ell-\mu x_i)\big)e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}}.$$

All in all, we are led to

$$\begin{aligned}\nabla^2_\theta \mathcal{R}_n(\theta) = &\frac{2}{n}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[(\nabla_\theta s_\theta(Y))(\nabla_\theta s_\theta(Y))^\top\big] \\ &+ \frac{2}{mn}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)H(Y)\big] \\ &- \frac{\sqrt{2}}{\sqrt{\pi}mn\sigma}\sum_{i=1}^n \operatorname{diag}(\mathbf{0}_m, \mathbf{w^{(2)}w^{(1)}}\big(s_\theta(\boldsymbol{\tau}) + \frac{1}{\sigma^2}(\boldsymbol{\tau}-\mu x_i)\big)e^{-\frac{(\boldsymbol{\tau}-\mu x_i)^2}{2\sigma^2}}), \end{aligned} \tag{15}$$

where $\mathbf{0}_m$ denotes the null vector in $\mathbb{R}^m$ and

$$H(y) = \begin{pmatrix} \mathbf{0}_{m\times m} & \operatorname{diag}\mathbf{1}_{y\mathbf{w}^{(1)}+\mathbf{b}\geqslant 0} \\ \operatorname{diag}\mathbf{1}_{y\mathbf{w}^{(1)}+\mathbf{b}\geqslant 0} & \mathbf{0}_{m\times m}\end{pmatrix}. \tag{16}$$

This concludes the proof.

B.4. Proof of Proposition 6

We start from the formula for the Hessian of the loss provided by (15). Let D be the diagonal matrix in the third term of the Hessian, and let v be a unit eigenvector of the first term with respect to its largest eigenvalue. Recalling that, for any matrix M, $\lambda_{\max}(M) \geqslant v^\top M v$ with equality if v is an eigenvector of M with eigenvalue $\lambda_{\max}(M)$, we then have

$$\begin{aligned}\lambda_{\max}(\nabla^2_\theta \mathcal{R}_n(\theta)) &\geqslant v^\top \nabla^2_\theta \mathcal{R}_n(\theta) v \\ &= \lambda_{\max}\Big(\frac{2}{mn}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\Big[(\nabla_\theta s_\theta(Y))(\nabla_\theta s_\theta(Y))^\top\Big]\Big) \\ &\quad + \frac{2}{mn}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\Big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)v^\top H(Y) v\Big] \\ &\quad + v^\top D v.\end{aligned}$$

The first term is lower bounded using Proposition 10. Thus, rearranging the terms, we obtain

$$\begin{aligned}&\lambda_{\max}(\nabla^2_\theta \mathcal{R}_n(\theta)) + \Big|\frac{2}{mn}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\Big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)v^\top H(Y) v\Big]\Big| + |v^\top D v| \\ &\geqslant \frac{4}{m}\mathrm{TV}^{(1)}_\pi(s_\theta).\end{aligned} \tag{17}$$

We now bound the second and third term on the left-hand side of the inequality above. As for the second term, using the Cauchy-Schwarz inequality twice,

$$\Big|\frac{2}{mn}\sum_{i=1}^n \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\Big[\big(s_\theta(Y) + \frac{1}{\sigma^2}(Y-\mu x_i)\big)v^\top H(Y) v\Big]\Big|$$

$$
\begin{aligned}
&\leqslant \frac{2}{mn}\sum_{i=1}^{n}\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\Big[\big(s_\theta(Y)+\frac{1}{\sigma^2}(Y-\mu x_i)\big)^2\Big]}\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big(v^\top H(Y)v\big)^2}\\
&\leqslant \frac{2}{m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\Big[\big(s_\theta(Y)+\frac{1}{\sigma^2}(Y-\mu x_i)\big)^2\Big]}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big(v^\top H(Y)v\big)^2}\\
&\leqslant \frac{2}{m}\sqrt{\mathcal{R}_n(\theta)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[\lambda_{\max}(H(Y))^2\big]}.
\end{aligned}
$$

By inspecting formula (16) for $H(y)$, we easily see that, for any $y\in\mathbb{R}$, $\lambda_{\max}(H(y))\leqslant 1$. So,

$$
\Big|\frac{2}{mn}\sum_{i=1}^{n}\mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\Big[\big(s_\theta(Y)+\frac{1}{\sigma^2}(Y-\mu x_i)\big)v^\top H(Y)v\Big]\Big|\leqslant\frac{2}{m}\sqrt{\mathcal{R}_n(\theta)}.
$$

We now proceed to bound the term $|v^\top Dv|$ in (17), where we recall that $\tau_\ell=-b_\ell/w_\ell^{(1)}$. Since D is a diagonal matrix, we have

$$
\begin{aligned}
|v^\top Dv| &\leqslant \max_{1\leqslant\ell\leqslant m}|D_{\ell\ell}|\\
&= \max_{1\leqslant\ell\leqslant m}\frac{\sqrt{2}}{\sqrt{\pi}mn\sigma}\Big|\sum_{i=1}^{n}w_\ell^{(2)}w_\ell^{(1)}\big(s_\theta(\tau_\ell)+\frac{1}{\sigma^2}(\tau_\ell-\mu x_i)\big)e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}}\Big|\\
&\leqslant \max_{1\leqslant\ell\leqslant m}\frac{\sqrt{2}}{\sqrt{\pi}mn\sigma}\sum_{i=1}^{n}|w_\ell^{(2)}|\times\Big|s_\theta(\tau_\ell)+\frac{1}{\sigma^2}(\tau_\ell-\mu x_i)\Big|e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}}\\
&\leqslant \max_{1\leqslant\ell\leqslant m,1\leqslant i\leqslant n}\frac{\sqrt{2}A}{\sqrt{\pi}m\sigma}\Big|s_\theta(\tau_\ell)+\frac{1}{\sigma^2}(\tau_\ell-\mu x_i)\Big|e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}}.
\end{aligned}
$$

Thus, so far, we have proved that

$$
\begin{aligned}
\frac{4}{m}\mathrm{TV}_\pi^{(1)}(s_\theta)\leqslant\lambda_{\max}(\nabla_\theta^2\mathcal{R}_n(\theta))+\frac{2}{m}\sqrt{\mathcal{R}_n(\theta)}\\
+\frac{\sqrt{2}A}{\sqrt{\pi}m\sigma}\max_{1\leqslant\ell\leqslant m,1\leqslant i\leqslant n}\Big|s_\theta(\tau_\ell)+\frac{1}{\sigma^2}(\tau_\ell-\mu x_i)\Big|e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}}.
\end{aligned}
\tag{18}
$$

Now, let $f_\theta(y)=s_\theta(y)+\frac{1}{\sigma^2}(y-\mu x_i)$. Our proof strategy consists in deriving a bound on $|f_\theta(\tau_\ell)|e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}}$ depending on the value of the risk $\mathcal{R}_n(\theta)$, and valid for all $\ell\in\{1,\dots,m\}$ and all $i\in\{1,\dots,n\}$. To this aim, first note that

$$
n\mathcal{R}_n(\theta)\geqslant\mathbb{E}_{Y\sim\mathcal{N}(\mu xi,\sigma^2)}\Big[\big(s_\theta(Y)+\frac{1}{\sigma^2}(Y-\mu x_i)\big)^2\Big]=\mathbb{E}_{Y\sim\mathcal{N}(\mu xi,\sigma^2)}\big[f_\theta(Y)^2\big].
$$

We observe that $f_\theta(y)$ is Lipschitz continuous with Lipschitz constant at most $A+\frac{1}{\sigma^2}\leqslant 2A$, where the inequality holds since we assumed that $A\geqslant C_n/\sigma^6\geqslant 1/\sigma^2$. Thus, for $y\in\mathbb{R}$,

$$
|f_\theta(y)|\geqslant|f_\theta(\tau_\ell)|-2A|y-\tau_\ell|.
$$

So, if $|y - \tau_\ell| \leqslant \frac{|f_\theta(\tau_\ell)|}{4A}$,

$$|f_\theta(y)| \geqslant \frac{|f_\theta(\tau_\ell)|}{2}.$$

Therefore,

$$\begin{aligned}
n\mathcal{R}_n(\theta) &\geqslant \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} f_\theta(y)^2 e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}} dy \\
&\geqslant \frac{1}{\sqrt{2\pi}\sigma} \int_{\tau_\ell - |f_\theta(\tau_\ell)|/4A}^{\tau_\ell + |f_\theta(\tau_\ell)|/4A} f_\theta(y)^2 e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}} dy \\
&\geqslant \frac{1}{\sqrt{2\pi}\sigma} \int_{\tau_\ell - |f_\theta(\tau_\ell)|/4A}^{\tau_\ell + |f_\theta(\tau_\ell)|/4A} \frac{f_\theta(\tau_\ell)^2}{4} e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}} dy \\
&= \frac{f_\theta(\tau_\ell)^2}{4\sqrt{2\pi}\sigma} \int_{\tau_\ell - \mu x_i - |f_\theta(\tau_\ell)|/4A}^{\tau_\ell - \mu x_i + |f_\theta(\tau_\ell)|/4A} e^{-\frac{y^2}{2\sigma^2}} dy \\
&= \frac{f_\theta(\tau_\ell)^2}{4\sqrt{2\pi}\sigma} \int_{|\tau_\ell - \mu x_i| - |f_\theta(\tau_\ell)|/4A}^{|\tau_\ell - \mu x_i| + |f_\theta(\tau_\ell)|/4A} e^{-\frac{y^2}{2\sigma^2}} dy,
\end{aligned}$$

where the last step follows from the symmetry of the Gaussian distribution around 0. Denote by I the last integral and D its integration domain. To lower bound I, two cases are considered:

Case 1. $[-\sigma, \sigma]$ **is included in** D**.** In this case,

$$\frac{1}{\sqrt{2\pi}\sigma} I \geqslant \frac{1}{\sqrt{2\pi}\sigma} \int_{-\sigma}^{\sigma} e^{-\frac{y^2}{2\sigma^2}} dy = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{y^2}{2}} dy \geqslant \frac{1}{2},$$

and thus $n\mathcal{R}_n(\theta) \geqslant \frac{f_\theta(\tau_\ell)^2}{8}$. We conclude, when $[-\sigma, \sigma]$ is included in D, that

$$|f_\theta(\tau_\ell)| e^{-\frac{(\tau_\ell - \mu x_i)^2}{2\sigma^2}} \leqslant |f_\theta(\tau_\ell)| \leqslant 2\sqrt{2n\mathcal{R}_n(\theta)}.$$

Case 2. $[-\sigma, \sigma]$ **is not included in** D**.** Since the absolute value of the upper endpoint of D is larger than the absolute value of its lower endpoint, the condition implies that the lower endpoint of D is larger than $-\sigma$. Therefore, we have $D \subset \big[-\sigma, |\tau_\ell - \mu x_i|\big]$. Hence, for all $y \in D$,

$$e^{-\frac{y^2}{2\sigma^2}} \geqslant \frac{1}{\sqrt{e}} e^{-\frac{(\tau_\ell - \mu x_i)^2}{2\sigma^2}}.$$

To see this, notice that, if $y \in D$ and $y < 0$, one has $y \geqslant -\sigma$ and so $e^{-\frac{y^2}{2\sigma^2}} \geqslant e^{-\frac{1}{2}}$. On the other hand, if $y \geqslant 0$, then $y \leqslant |\tau_\ell - \mu x_i|$ gives $e^{-\frac{y^2}{2\sigma^2}} \geqslant e^{-\frac{(\tau_\ell - \mu x_i)^2}{2\sigma^2}}$. We are led to

$$I \geqslant \frac{|f_\theta(\tau_\ell)|}{2A} \frac{1}{\sqrt{e}} e^{-\frac{(\tau_\ell - \mu x_i)^2}{2\sigma^2}}.$$

Then

$$n\mathcal{R}_n(\theta) \geqslant \frac{|f_\theta(\tau_\ell)|^3}{8\sqrt{2\pi e} A\sigma} e^{-\frac{(\tau_\ell - \mu x_i)^2}{2\sigma^2}}.$$

We deduce, still in the case where $[-\sigma, \sigma]$ is not included in D, that

$$\begin{aligned}|f_\theta(\tau_\ell)|e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}} &\leqslant |f_\theta(\tau_\ell)|e^{-\frac{(\tau_\ell-\mu x_i)^2}{6\sigma^2}} \\ &= \big(|f_\theta(\tau_\ell)|^3 e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}}\big)^{\frac{1}{3}} \\ &\leqslant \big(8\sqrt{2\pi e}A\sigma n\mathcal{R}_n(\theta)\big)^{\frac{1}{3}} \\ &= 2\big(\sqrt{2\pi e}A\sigma n\mathcal{R}_n(\theta)\big)^{\frac{1}{3}}.\end{aligned}$$

Putting both cases together, we obtain

$$|f_\theta(\tau_\ell)|e^{-\frac{(\tau_\ell-\mu x_i)^2}{2\sigma^2}} \leqslant 2\max\Big(\sqrt{2n\mathcal{R}_n(\theta)}, \big(\sqrt{2\pi e}A\sigma n\mathcal{R}_n(\theta)\big)^{\frac{1}{3}}\Big).$$

We conclude, coming back to (18), that

$$\begin{aligned}&\frac{4}{m}\mathrm{TV}_\pi^{(1)}(s_\theta) \\ &\qquad\leqslant \lambda_{\max}(\nabla_\theta^2\mathcal{R}_n(\theta)) + \frac{2}{m}\sqrt{\mathcal{R}_n(\theta)} + \frac{2\sqrt{2}A}{\sqrt{\pi}m\sigma}\max\Big(\sqrt{2n\mathcal{R}_n(\theta)}, \big(\sqrt{2\pi e}A\sigma n\mathcal{R}_n(\theta)\big)^{\frac{1}{3}}\Big).\end{aligned}$$

This shows the first statement of the proposition. Finally, if $\theta = \theta^\star$ is a linearly stable minimum of $\mathcal{R}_n$, we apply (8) to get the second inequality.

B.5. Proof of Proposition 7

We start by showing by contradiction that for any $1 \leqslant i \leqslant n$, there exists $a_i \in [\mu x_i - \frac{\mu\Delta}{2}, \mu x_i]$ and $b_i \in [\mu x_i, \mu x_i + \frac{\mu\Delta}{2}]$ such that $s_{\theta^\star}(a_i) > 0$ and $s_{\theta^\star}(b_i) < 0$. If, for all $y \in [\mu x_i - \frac{\mu\Delta}{2}, \mu x_i]$, one has $s_{\theta^\star}(y) \leqslant 0$, then

$$\begin{aligned}n\mathcal{R}_n(\theta^\star) &\geqslant \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[(s_{\theta^\star}(Y) - \frac{1}{\sigma^2}(\mu x_i - Y))^2\big] \\ &\geqslant \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mu x_i - \frac{\mu\Delta}{2}}^{\mu x_i}\big(s_{\theta^\star}(y) - \frac{1}{\sigma^2}(\mu x_i - y)\big)^2 e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}dy \\ &\geqslant \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mu x_i - \frac{\mu\Delta}{2}}^{\mu x_i}\frac{1}{\sigma^4}(\mu x_i - y)^2 e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}dy \\ &= \frac{1}{\sigma^4\sqrt{2\pi\sigma^2}}\int_{-\frac{\mu\Delta}{2}}^{0} y^2 e^{-\frac{y^2}{2\sigma^2}}dy \\ &= \frac{1}{\sigma^2\sqrt{2\pi\sigma^2}}\Big(\frac{\mu\Delta}{2}e^{-\frac{\mu^2\Delta^2}{8\sigma^2}} + \int_{-\frac{\mu\Delta}{2}}^{0} e^{-\frac{y^2}{2\sigma^2}}dy\Big) \\ &\geqslant \frac{1}{\sigma^2\sqrt{2\pi\sigma^2}}\frac{\mu\Delta}{2}e^{-\frac{\mu^2\Delta^2}{8\sigma^2}} + \frac{1}{\sigma^2}\big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{8\sigma^2}}\big),\end{aligned}$$

where we integrate by parts to derive the second last equation, and then use a tail bound of the Gaussian distribution (Gordon, 1941). Since $\Delta \geqslant 8\frac{\sigma}{\mu}$, we get

$$n\mathcal{R}_n(\theta^\star) \geqslant \frac{1}{\sigma^2}\big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{8\sigma^2}}\big) \geqslant \frac{1}{\sigma^2}\big(\frac{1}{2} - e^{-8}\big) \geqslant \frac{1}{4\sigma^2},$$

which is a contradiction with $\mathcal{R}_n(\theta^\star) \leqslant \frac{1}{16n\sigma^2}$. Thus, there must exist $a_i \in \big[\mu x_i - \frac{\mu\Delta}{2}, \mu x_i\big]$ such that $s_{\theta^\star}(a_i) \geqslant 0$. A similar argument proves the existence of b_i. Hence, for every $1 \leqslant i \leqslant n-1$, there exists $c_i \in [b_i, a_{i+1}] \subset [\mu x_i, \mu x_{i+1}]$ such that $s'_{\theta^\star}(c_i) \geqslant 0$.

Assume now that for all $y \in [-\frac{\mu\Delta}{2} + \mu x_i, \frac{\mu\Delta}{2} + \mu x_i]$, we have

$$s'_{\theta^\star}(y) > -\frac{1}{\sigma^2} + \sqrt{\frac{2n\mathcal{R}_n(\theta^\star)}{\sigma^2}},$$

and aim again at reaching a contradiction. By applying Lemma 14 with $f(x) = s_{\theta^\star}(x + \mu x_i)$, $\beta = \frac{1}{\sigma^2}$, $\gamma = \frac{1}{\sigma^2} - \sqrt{\frac{2n\mathcal{R}_n(\theta^\star)}{\sigma^2}}$ and $\delta = \frac{\mu\Delta}{2}$ we have

$$\begin{aligned} n\mathcal{R}_n(\theta^\star) &> \mathbb{E}_{y\sim\mathcal{N}(\mu x_i,\sigma^2)}[(s_{\theta^\star}(y) + \frac{1}{\sigma^2}(y - \mu x_i))^2] \\ &= \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)}[(f(z) + \frac{1}{\sigma^2}z)^2] \\ &\geqslant \sigma^2 \frac{2n\mathcal{R}_n(\theta^\star)}{\sigma^2}\Big(1 - 2\Big(\frac{\mu\Delta}{2\sqrt{2\pi\sigma^2}} + 1\Big)e^{-\frac{\mu^2\Delta^2}{8\sigma^2}}\Big). \end{aligned}$$

Applying Lemma 13 to $x \mapsto xe^{-\frac{x^2}{2}}$ at $x = \frac{\mu\Delta}{4\sigma} \geqslant 2$, we obtain that $\frac{\mu\Delta}{4\sigma}e^{-\frac{\mu^2\Delta^2}{8\sigma^2}} \leqslant e^{-2}$, and thus

$$1 - 2\Big(\frac{\mu\Delta}{2\sqrt{2\pi\sigma^2}} + 1\Big)e^{-\frac{\mu^2\Delta^2}{8\sigma^2}} \geqslant 1 - \frac{2\sqrt{2}}{\sqrt{\pi}}e^{-2} - 2e^{-8} \geqslant \frac{1}{2}.$$

We thus obtain $n\mathcal{R}_n(\theta^\star) > n\mathcal{R}_n(\theta^\star)$, which is a contradiction. So, there must exist $y_i \in [-\frac{\mu\Delta}{2} + \mu x_i, \frac{\mu\Delta}{2} + \mu x_i]$ such that

$$s'_{\theta^\star}(y_i) \leqslant -\frac{1}{\sigma^2} + \sqrt{\frac{2n\mathcal{R}_n(\theta^\star)}{\sigma^2}}.$$

Note that $[y_i, c_i] \cap [y_{i+2}, c_{i+2}] = \emptyset$. Let $x_{(n/4)}$ be the smallest x_i such that $i > n/4$ and let $x_{(3n/4)}$ be the largest x_i such that $i < 3n/4$. Then, using arguments similar to those employed in the proof of Theorem 4,

$$\begin{aligned} \mathrm{TV}^{(1)}_\pi(s_{\theta^\star}) &\geqslant \sum_{\frac{n}{4}<2i<\frac{3n}{4}} \Big(\min_{y\in[x_{(n/4)},x_{(3n/4)}]} \pi(y;\mu,\sigma)\Big)\Big|\int_{y_{2i}}^{c_{2i}} s''_{\theta^\star}(y)dy\Big| \\ &\geqslant \sum_{\frac{n}{4}<2i<\frac{3n}{4}} \frac{\mu n\Delta}{128}\Big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{2\sigma^2}}\Big)\Big(\frac{1}{\sigma^2} - \sqrt{\frac{2n\mathcal{R}_n(\theta^\star)}{\sigma^2}}\Big) \\ &\geqslant \frac{\mu n^2\Delta}{1024}\Big(\frac{1}{2} - e^{-\frac{\mu^2\Delta^2}{2\sigma^2}}\Big)\Big(\frac{1}{\sigma^2} - \sqrt{\frac{2n\mathcal{R}_n(\theta^\star)}{\sigma^2}}\Big), \end{aligned}$$

where the last inequality utilizes that there are at least $n/4$ points between $\lceil n/4\rceil + 1$ and $\lfloor 3n/4\rfloor - 1$ (for $n \geqslant 10$), so we sum over at least $n/8$ points given that we consider one point out of two. Then, using our assumption $\frac{\mu\Delta}{\sigma} \geqslant 8$ and $\mathcal{R}_n(\theta^\star) \leqslant \frac{1}{16n\sigma^2}$, we obtain

$$\mathrm{TV}^{(1)}_\pi(s_{\theta^\star}) \geqslant \frac{\mu n^2\Delta}{1024}(\frac{1}{2} - e^{-32})\Big(\frac{1}{\sigma^2} - \frac{1}{2\sqrt{2}\sigma^2}\Big) \geqslant \frac{\mu n^2\Delta}{2^{11}\sigma^2}.$$

B.6. Proof of Theorem 8

We reason by contraposition, that is, we assume that

$$\mathcal{R}_n(\theta^\star) - \mathcal{R}_n(s^\star) \leqslant \frac{\pi n^5 \mu^3 \Delta^3}{2^{36} e^{1/2} A^4 \sigma^4},$$

and show that it implies that $\eta \leqslant \frac{2^{12}\sigma^2}{\mu n^2 \Delta}$. For $\sigma \leqslant \sigma_1 := \frac{\mu\Delta}{8}$, we can apply Corollary 3, which gives

$$\mathcal{R}_n(s^\star) \leqslant \frac{4\mu^2 (x_n - x_1)^2}{\sigma^4} e^{-\frac{\mu^2\Delta^2}{32\sigma^2}}.$$

Thus

$$\mathcal{R}_n(\theta^\star) \leqslant \frac{\pi n^5 \mu^3 \Delta^3}{2^{36} e^{1/2} A^4 \sigma^4} + \frac{4\mu^2 (x_n - x_1)^2}{\sigma^4} e^{-\frac{\mu^2\Delta^2}{32\sigma^2}}.$$

Let us show that this implies

$$\mathcal{R}_n(\theta^\star) \leqslant \frac{\pi n^5 \mu^3 \Delta^3}{2^{35} e^{1/2} A^4 \sigma^4}. \tag{19}$$

By rearranging terms, one can see that this holds as soon as

$$e^{-\frac{\mu^2\Delta^2}{32\sigma^2}} \leqslant \frac{\pi n^5 \mu}{2^{38} e^{1/2} (x_n - x_1)^2 A^4}.$$

Recalling that A grows polynomially fast with $1/\sigma$, we observe that the left-hand side of the previous inequality decays exponentially fast when $\sigma \to 0$, while the right-hand side decays polynomially fast. This implies the existence of some σ_2 depending on the training data and on μ such that this inequality holds true for $\sigma \leqslant \sigma_2$.

Next, observe that (19) implies in particular that, for $\sigma \leqslant \sigma_3 := \frac{1}{n}$,

$$\mathcal{R}_n(\theta^\star) \leqslant \frac{1}{16 n \sigma^2}. \tag{20}$$

This enables us to apply Proposition 7 to $\theta = \theta^\star$, which entails that, for $\sigma \leqslant \sigma_1$,

$$\mathrm{TV}_\pi^{(1)}(s_{\theta^\star}) \geqslant \frac{\mu n^2 \Delta}{2^{11}\sigma^2}$$

Combining this lower bound with the upper bound of Proposition 6, we obtain that

$$\frac{1}{2\eta} + \frac{\sqrt{\mathcal{R}_n(\theta^\star)}}{2} + \frac{A}{\sqrt{2\pi}\sigma} \max\left(\sqrt{2n\mathcal{R}_n(\theta^\star)}, \left(\sqrt{2\pi e} A\sigma n \mathcal{R}_n(\theta^\star)\right)^{\frac{1}{3}}\right) \geqslant \frac{\mu n^2 \Delta}{2^{11}\sigma^2}.$$

Note that

$$\begin{aligned}\sqrt{2n\mathcal{R}_n(\theta^\star)} \leqslant \left(\sqrt{2\pi e} A\sigma n \mathcal{R}_n(\theta^\star)\right)^{\frac{1}{3}} &\Leftrightarrow 8n^3 \mathcal{R}_n(\theta^\star)^3 \leqslant 2\pi e A^2 \sigma^2 n^2 \mathcal{R}_n(\theta^\star)^2 \\ &\Leftrightarrow \mathcal{R}_n(\theta^\star) \leqslant \frac{\pi e}{4} \frac{A^2\sigma^2}{n},\end{aligned}$$

which holds true by (20). Hence, we obtain

$$\frac{1}{2\eta} + \frac{\sqrt{\mathcal{R}_n(\theta^\star)}}{2} + \frac{A}{\sqrt{2\pi}\sigma}\big(\sqrt{2\pi e}A\sigma n\mathcal{R}_n(\theta^\star)\big)^{\frac{1}{3}} \geqslant \frac{\mu n^2\Delta}{2^{11}\sigma^2}.$$

Rewriting the third term, we have

$$\frac{1}{2\eta} + \frac{\sqrt{\mathcal{R}_n(\theta^\star)}}{2} + \frac{e^{1/6}A^{4/3}n^{1/3}\mathcal{R}_n(\theta^\star)^{1/3}}{(2\pi)^{1/3}\sigma^{2/3}} \geqslant \frac{\mu n^2\Delta}{2^{11}\sigma^2}.$$

By (19), recalling again that $A \geqslant \frac{C_n}{\sigma^6}$, observe that there exists σ_4 such that, for $\sigma \leqslant \sigma_4$,

$$\frac{\sqrt{\mathcal{R}_n(\theta^\star)}}{2} \leqslant \frac{\mu n^2\Delta}{2^{13}\sigma^2}.$$

Thus, using again (19), we get

$$\frac{1}{2\eta} + \frac{\mu n^2\Delta}{2^{13}\sigma^2} + \frac{\mu n^2\Delta}{2^{12}\sigma^2} \geqslant \frac{\mu n^2\Delta}{2^{11}\sigma^2},$$

which implies that $\eta \leqslant \frac{2^{12}\sigma^2}{\mu n^2\Delta}$, thereby concluding the proof with $\sigma_0 = \min(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$.

Appendix C. Technical lemmas

The first lemma relates the derivatives of $s^\star$ with the moments of W.

Lemma 11 *Let $s^\star$ and W be defined as in Sections 3 and 5. Then we have*

$$s^{\star\prime}(y;\mu,\sigma) = \frac{1}{\sigma^2}\Big(-1 + \frac{\mu^2}{\sigma^2}\mathbb{V}[W(y;\mu,\sigma)]\Big),$$

and

$$s^{\star\prime\prime}(y;\mu,\sigma) = \frac{\mu^3}{\sigma^6}\mathbb{E}[(W(y;\mu,\sigma) - \mathbb{E}[W(y;\mu,\sigma)])^3].$$

Proof We start by calculating the derivatives of α_j with respect to y. Observe that

$$\begin{aligned}
\alpha_j'(y;\mu,\sigma) &= -\frac{e^{-\frac{(y-\mu x_j)^2}{2\sigma^2}}\sum_{i=1}^n(-y+\mu x_i)e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}}{\sigma^2\big(\sum_{i=1}^n e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}\big)^2} + \frac{(-y+\mu x_j)e^{-\frac{(y-\mu x_j)^2}{2\sigma^2}}}{\sigma^2\sum_{i=1}^n e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}},\\
&= \frac{1}{\sigma^2}\Big(y\alpha_j(y;\mu,\sigma) - \mu\alpha_j(y;\mu,\sigma)\sum_{i=1}^n x_i\alpha_i(y;\mu,\sigma) - y\alpha_j(y;\mu,\sigma) + \mu x_j\alpha_j(y;\mu,\sigma)\Big),\\
&= \frac{\alpha_j(y;\mu,\sigma)}{\sigma^2}\Big(\mu x_j - \mu\sum_{i=1}^n x_i\alpha_i(y;\mu,\sigma)\Big),\\
&= \frac{\mu\alpha_j(y;\mu,\sigma)}{\sigma^2}\big(x_j - \mathbb{E}[W(y;\mu,\sigma)]\big).
\end{aligned}$$

In addition, we may also compute $\alpha_j''(y;\mu,\sigma)$ as follows

$$\begin{aligned}
\alpha_j''(y;\mu,\sigma) &= \frac{d}{dy}\Big(\frac{\mu\alpha_j(y;\mu,\sigma)}{\sigma^2}\big(x_j - \sum_{i=1}^n x_i\alpha_i(y;\mu,\sigma)\big)\Big),\\
&= \frac{\mu\alpha_j'(y;\mu,\sigma)}{\sigma^2}\Big(x_j - \sum_{i=1}^n x_i\alpha_i(y;\mu,\sigma)\Big) - \frac{\mu\alpha_j(y;\mu,\sigma)}{\sigma^2}\sum_{i=1}^n x_i\alpha_i'(y;\mu,\sigma),\\
&= \frac{\mu^2\alpha_j(y;\mu,\sigma)}{\sigma^4}\Big(x_j - \sum_{i=1}^n x_i\alpha_i(y;\mu,\sigma)\Big)^2\\
&\quad - \frac{\mu^2\alpha_j(y;\mu,\sigma)}{\sigma^2}\sum_{i=1}^n x_i\frac{\alpha_i(y;\mu,\sigma)}{\sigma^2}\Big(x_i - \sum_{i'=1}^n x_{i'}\alpha_{i'}(y;\mu,\sigma)\Big),\\
&= \frac{\mu^2\alpha_j(y;\mu,\sigma)}{\sigma^2}\Big(x_j - \mathbb{E}[W(y;\mu,\sigma)]\Big)^2\\
&\quad - \frac{\mu^2\alpha_j(y;\mu,\sigma)}{\sigma^4}\Big(\mathbb{E}[W^2(y;\mu,\sigma)] - \mathbb{E}^2[W(y;\mu,\sigma)]\Big),\\
&= \frac{\mu^2\alpha_j(y;\mu,\sigma)}{\sigma^4}\big(x_j^2 - 2x_j\mathbb{E}[W(y;\mu,\sigma)] - \mathbb{E}[W^2(y;\mu,\sigma)] + 2\mathbb{E}^2[W(y;\mu,\sigma)]\big)
\end{aligned}$$

Consequently,

$$\begin{aligned}
s^{\star\prime}(y;\mu,\sigma) &= \frac{1}{\sigma^2}\Big(-1 + \mu\sum_{i=1}^n x_i\alpha_i'(y;\mu,\sigma)\Big)\\
&= \frac{1}{\sigma^2}\Big(-1 + \frac{\mu^2}{\sigma^2}\Big(\sum_{i=1}^n \alpha_i(y;\mu,\sigma)x_i^2 - \mathbb{E}[W(y;\mu,\sigma)]\sum_{i=1}^n \alpha_i(y;\mu,\sigma)x_i\Big)\Big)\\
&= \frac{1}{\sigma^2}\Big(-1 + \frac{\mu^2}{\sigma^2}\big(\mathbb{E}[W^2(y;\mu,\sigma)] - \mathbb{E}^2[W(y;\mu,\sigma)]\big)\Big)\\
&= \frac{1}{\sigma^2}\Big(-1 + \frac{\mu^2}{\sigma^2}\mathbb{V}[W(y;\mu,\sigma)]\Big),
\end{aligned}$$

and

$$\begin{aligned}
s^{\star\prime\prime}(y;\mu,\sigma) &= \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\alpha_i''(y;\mu,\sigma),\\
&= \frac{\mu}{\sigma^2}\sum_{i=1}^n \frac{\mu^2 x_i\alpha_i(y;\mu,\sigma)}{\sigma^4}\big(x_i^2 - 2x_i\mathbb{E}[W(y;\mu,\sigma)] - \mathbb{E}[W^2(y;\mu,\sigma)] + 2\mathbb{E}^2[W(y;\mu,\sigma)]\big),\\
&= \frac{\mu^3}{\sigma^6}(\mathbb{E}[W^3(y;\mu,\sigma)] - 2\mathbb{E}[W^2(y;\mu,\sigma)]\mathbb{E}[W(y;\mu,\sigma)])\\
&\quad + \frac{\mu^3}{\sigma^6}(-\mathbb{E}[W^2(y;\mu,\sigma)]\mathbb{E}[W(y;\mu,\sigma)] + 2\mathbb{E}^3[W(y;\mu,\sigma)]),\\
&= \frac{\mu^3}{\sigma^6}(\mathbb{E}[W^3(y;\mu,\sigma)] - 3\mathbb{E}[W^2(y;\mu,\sigma)]\mathbb{E}[W(y;\mu,\sigma)] + 2\mathbb{E}^3[W(y;\mu,\sigma)]),\\
&= \frac{\mu^3}{\sigma^6}\mathbb{E}[(W(y;\mu,\sigma) - \mathbb{E}[W(y;\mu,\sigma)])^3].
\end{aligned}$$

This concludes the proof. ■

The next lemma bounds the total variation of the derivative of $s^\star$.

Lemma 12 *Let* $(x_n)_+ = \max(0, x_n)$ *and* $(x_1)_- = \min(0, x_1)$. *Then*

$$\int_{\mathbb{R}} |s^{\star\prime\prime}(y;\mu,\sigma)|dy \leqslant \frac{4\mu^2(x_n - x_1)^3}{\sigma^6}\Big(\mu^2((x_n)_+ - (x_1)_-) + \frac{2(n-1)\sigma^2}{\Delta}\Big).$$

Proof We start by proving that, for $1 \leqslant i \leqslant n-1$ and $y \geqslant 2\mu(x_n)_+$, one has

$$\alpha_i(y;\mu,\sigma) \leqslant e^{-\frac{y\mu\Delta}{2\sigma^2}} \quad \text{and} \quad |s^{\star\prime\prime}(y;\mu,\sigma)| \leqslant \frac{2\mu^3(n-1)(x_n - x_1)^3}{\sigma^6} e^{-\frac{y\mu\Delta}{2\sigma^2}}.$$

Observe that

$$\alpha_i(y;\mu,\sigma) = \frac{e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}}{\sum_{i'=1}^n e^{-\frac{(y-\mu x_{i'})^2}{2\sigma^2}}} \leqslant \frac{e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}}{e^{-\frac{(y-\mu x_n)^2}{2\sigma^2}}} = e^{\frac{-2y\mu(x_n - x_i) + \mu^2(x_n^2 - x_i^2)}{2\sigma^2}}.$$

Since $y \geqslant 2\mu(x_n)_+ \geqslant 2\mu x_n$ implies $y \geqslant \mu(x_n + x_i)$, noticing that $x_n - x_i \geqslant \Delta$, we have

$$-2y\mu(x_n - x_i) + \mu^2(x_n^2 - x_i^2) = \mu(x_n - x_i)(-2y + \mu(x_n + x_i)) \leqslant -y\mu(x_n - x_i) \leqslant -y\mu\Delta,$$

where we also used the fact that $y \geqslant 0$ in the last inequality. It follows that

$$\alpha_i(y;\mu,\sigma) \leqslant e^{\frac{-2y\mu(x_n - x_i) + \mu^2(x_n^2 - x_i^2)}{2\sigma^2}} \leqslant e^{-\frac{y\mu\Delta}{2\sigma^2}}. \tag{21}$$

To upper bound $|s^{\star\prime\prime}(y;\mu,\sigma)|$, we first remark that $W(y;\mu,\sigma)$ takes value in $\{x_1, \dots, x_n\}$. Hence, for all $1 \leqslant i \leqslant n$, $|x_i - \mathbb{E}[W(y;\mu,\sigma)]| \leqslant x_n - x_1$. Applying Lemma 11, we are led to

$$\begin{aligned}
|s^{\star\prime\prime}(y;\mu,\sigma)| &\leqslant \frac{\mu^3}{\sigma^6}\mathbb{E}\big[|W(y;\mu,\sigma) - \mathbb{E}[W(y;\mu,\sigma)]|^3\big] \qquad (22)\\
&= \frac{\mu^3}{\sigma^6}\sum_{i=1}^{n-1}\big|x_i - \mathbb{E}[W(y;\mu,\sigma)]\big|^3\alpha_i(y;\mu,\sigma) + \frac{\mu^3}{\sigma^6}\big|x_n - \mathbb{E}[W(y;\mu,\sigma)]\big|^3\alpha_n(y;\mu,\sigma)\\
&\leqslant \frac{\mu^3}{\sigma^6}\sum_{i=1}^{n-1}(x_n - x_1)^3 e^{-\frac{y\mu\Delta}{2\sigma^2}} + \frac{\mu^3}{\sigma^6}\Big|x_n - \sum_{i=1}^n x_i\alpha_i(y;\mu,\sigma)\Big|^3\alpha_n(y;\mu,\sigma)\\
&\leqslant \frac{\mu^3}{\sigma^6}(n-1)(x_n - x_1)^3 e^{-\frac{y\mu\Delta}{2\sigma^2}} + \frac{\mu^3}{\sigma^6}\Big|\sum_{i=1}^{n-1}(x_n - x_i)\alpha_i(y;\mu,\sigma)\Big|^3\\
&\leqslant \frac{\mu^3}{\sigma^6}(n-1)(x_n - x_1)^3 e^{-\frac{y\mu\Delta}{2\sigma^2}} + \frac{\mu^3}{\sigma^6}\sum_{i=1}^{n-1}(x_n - x_i)^3\alpha_i(y;\mu,\sigma)\\
&\leqslant \frac{2\mu^3(n-1)(x_n - x_1)^3}{\sigma^6} e^{-\frac{y\mu\Delta}{2\sigma^2}},
\end{aligned}$$

where, in the second to last line we use the fact that $x \mapsto x^3$ is convex on $\mathbb{R}_+$ and $x_n - x_i > 0$ for all $1 \leqslant i \leqslant n-1$, and the last inequality follows from (21).

A similar argument apply for $y < 2\mu(x_1)_-$ and $2 \leqslant i \leqslant n$. In this case,

$$\alpha_i(y;\mu,\sigma) \leqslant e^{\frac{y\mu\Delta}{2\sigma^2}} \quad \text{and} \quad |s^{\star\prime\prime}(y;\mu,\sigma)| \leqslant \frac{2\mu^3(n-1)(x_n-x_1)^3}{\sigma^6} e^{\frac{y\mu\Delta}{2\sigma^2}}.$$

We can now proceed to bounding $\int_{\mathbb{R}} |s^{\star\prime\prime}(y;\mu,\sigma)|dy$. To this aim, we first split the integral as follows

$$\begin{aligned}&\int_{\mathbb{R}} |s^{\star\prime\prime}(y;\mu,\sigma)|dy \\ &= \int_{2\mu(x_1)_-}^{2\mu(x_n)_+} |s^{\star\prime\prime}(y;\mu,\sigma)|dy + \int_{-\infty}^{2\mu(x_1)_-} |s^{\star\prime\prime}(y;\mu,\sigma)|dy + \int_{2\mu(x_n)_+}^{\infty} |s^{\star\prime\prime}(y;\mu,\sigma)|dy.\end{aligned} \tag{23}$$

Similar to the argument in (22), we have $|s^{\star\prime\prime}(y;\mu,\sigma)| \leqslant \frac{\mu^3}{\sigma^6}\mathbb{E}[|W(y;\mu,\sigma) - \mathbb{E}[W(y;\mu,\sigma)]|^3] \leqslant \frac{\mu^3}{\sigma^6}(x_n - x_1)^3$. Therefore,

$$\int_{2\mu(x_1)_-}^{2\mu(x_n)_+} |s^{\star\prime\prime}(y;\mu,\sigma)|dy \leqslant \frac{4\mu^4(x_n-x_1)^3((x_n)_+ - (x_1)_-)}{\sigma^6}.$$

To bound the last two terms on the right-hand side of (23), we use the previous derivations, and see that

$$\begin{aligned}\int_{-\infty}^{2\mu(x_1)_-} |s^{\star\prime\prime}(y;\mu,\sigma)|dy &\leqslant \frac{2\mu^3(n-1)(x_n-x_1)^3}{\sigma^6} \int_{-\infty}^{2\mu(x_1)_-} e^{\frac{y\mu\Delta}{2\sigma^2}} dy \\ &= \frac{2\mu^3(n-1)(x_n-x_1)^3}{\sigma^6} \frac{2\sigma^2}{\mu\Delta} e^{\frac{\mu^2\Delta(x_1)_-}{\sigma^2}} \\ &\leqslant \frac{4\mu^2(n-1)(x_n-x_1)^3}{\sigma^4\Delta},\end{aligned}$$

since $(x_1)_- \leqslant 0$ implies $e^{\frac{\mu^2\Delta(x_1)_-}{\sigma^2}} \leqslant 1$. Similarly,

$$\int_{2\mu(x_n)_+}^{\infty} |s^{\star\prime\prime}(y;\mu,\sigma)|dy \leqslant \frac{4\mu^2(n-1)(x_n-x_1)^3}{\sigma^4\Delta}.$$

Putting everything together, we have

$$\begin{aligned}\int_{\mathbb{R}} |s^{\star\prime\prime}(y;\mu,\sigma)|dy &\leqslant \frac{4\mu^4(x_n-x_1)^3((x_n)_+ - (x_1)_-)}{\sigma^6} + \frac{8\mu^2(n-1)(x_n-x_1)^3}{\sigma^4\Delta} \\ &\leqslant \frac{4\mu^2(x_n-x_1)^3}{\sigma^6}\Big(\mu^2((x_n)_+ - (x_1)_-) + \frac{2(n-1)\sigma^2}{\Delta}\Big),\end{aligned}$$

which is the desired result. ■

The final two lemmas are technical properties of moments of Gaussian distributions.

Lemma 13 *Let k be a positive integer. Then the function $f_k : \mathbb{R} \to \mathbb{R}$ defined by $f_k(x) = x^k e^{-\frac{Ax^2}{2}}$ is strictly increasing on $(0, \sqrt{\frac{k}{A}})$ and strictly decreasing on $(\sqrt{\frac{k}{A}}, \infty)$.*

Proof Consider the derivative of f_k, which is given by

$$f_k'(x) = kx^{k-1}e^{-\frac{Ax^2}{2}} - Ax^{k+1}e^{-\frac{Ax^2}{2}} = (k - Ax^2)x^{k-1}e^{-\frac{Ax^2}{2}}.$$

Clearly, $f_k' < 0$ on $(\sqrt{\frac{k}{A}}, \infty)$ and $f_k' > 0$ on $(0, \sqrt{\frac{k}{A}})$. ■

Lemma 14 *Let β, γ, δ be positive numbers such that $\beta > \gamma$. Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function satisfying $f'(x) \geqslant -\gamma$ for $x \in [-\delta, \delta]$. Then*

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}(0,\sigma^2)}\big[(f(x) + \beta x)^2\big] &\geqslant \frac{(\beta-\gamma)^2}{\sqrt{2\pi\sigma^2}}\int_{-\delta}^{\delta} x^2 e^{-\frac{x^2}{2\sigma^2}}dx \\ &\geqslant \sigma^2(\beta-\gamma)^2\Big(1 - 2\Big(\frac{\delta}{\sqrt{2\pi\sigma^2}} + 1\Big)e^{-\frac{\delta^2}{2\sigma^2}}\Big).\end{aligned}$$

Proof Notice that

$$\begin{aligned}\mathbb{E}_{Z\sim\mathcal{N}(0,\sigma^2)}\big[(f(x) + \beta x)^2\big] &\geqslant \frac{1}{\sqrt{2\pi\sigma^2}}\int_{[-\delta,\delta]} (f(y) + \beta y)^2 e^{-\frac{y^2}{2\sigma^2}}dy \\ &\geqslant \frac{1}{\sqrt{2\pi\sigma^2}}\int_{[-\delta,\delta]} (-\gamma y + f(0) + \beta y)^2 e^{-\frac{y^2}{2\sigma^2}}dy.\end{aligned}$$

The last term reads

$$\begin{aligned}&\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\delta}^{\delta}\big((\beta-\gamma)^2y^2 + 2(\beta-\gamma)f(0)y + f^2(0)\big)e^{-\frac{y^2}{2\sigma^2}}dy \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\delta}^{\delta}\big((\beta-\gamma)^2y^2 + f^2(0)\big)e^{-\frac{y^2}{2\sigma^2}}dy \\ &\geqslant \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\delta}^{\delta}(\beta-\gamma)^2y^2e^{-\frac{y^2}{2\sigma^2}}dy. \\ &= \frac{-\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\delta}^{\delta}(\beta-\gamma)^2y\frac{de^{-\frac{y^2}{2\sigma^2}}}{dy}dy, \\ &= \Big[\frac{-\sigma^2}{\sqrt{2\pi\sigma^2}}(\beta-\gamma)^2ye^{-\frac{y^2}{2\sigma^2}}\Big]_{-\delta}^{\delta} + \frac{\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\delta}^{\delta}(\beta-\gamma)^2e^{-\frac{y^2}{2\sigma^2}}dy \\ &= \sigma^2(\beta-\gamma)^2\Big(\frac{-2\delta}{\sqrt{2\pi\sigma^2}}e^{-\frac{\delta^2}{2\sigma^2}} + \mathbb{P}_{\xi\in\mathcal{N}(0,1)}(|\sigma\xi| \leqslant \delta)\Big) \\ &\geqslant \sigma^2(\beta-\gamma)^2\Big(\frac{-2\delta}{\sqrt{2\pi\sigma^2}}e^{-\frac{\delta^2}{2\sigma^2}} + 1 - 2e^{-\frac{-\delta^2}{2\sigma^2}}\Big)\end{aligned}$$

where in the fifth line we use integration by part and in the last line we use a Gaussian tail bound (Gordon, 1941). ■

Appendix D. Comments on gradient descent

The choice of SGD is motivated by the fact that the loss is defined as an expectation (5), thus requiring a stochastic approximation. However our analysis also applies to an idealized scenario: GD performed directly on the population risk $\mathcal{R}_n$ (5).

For $j \in \mathbb{N}$, the standard GD update is $\theta_{j+1} = \theta_j - m\eta\nabla\mathcal{R}_n(\theta_j)$. Similar to the argument in Section 4, our analysis primarily relies on the linearized dynamics around the linearly stable global minimizer $\theta^\star$, where $\nabla\mathcal{R}_n(\theta_j) \approx \nabla\mathcal{R}_n(\theta^\star) + \nabla^2\mathcal{R}_n(\theta^\star)(\theta_j - \theta^\star)$. Since $\nabla\mathcal{R}_n(\theta^\star) = 0$, the linearized GD updates may be written as follows:

$$\theta_{j+1} = \theta_j - m\eta\nabla^2\mathcal{R}_n(\theta^\star)(\theta_j - \theta^\star).$$

Therefore, the central property (8), which relates the learning rate to the Hessian at the optimum $\nabla^2\mathcal{R}_n(\theta^\star)$, still holds in this deterministic setting. Consequently, our proof carries over and shows a lower bound on the learning rate above which GD cannot converge to the global minimizer. This demonstrates that our proof of the main result does not fundamentally depend on the stochasticity of the optimization algorithm. The idealized scenario of GD on the population risk effectively represents the case where the variance of the gradient estimates is zero. In other words, the non-convergence of SGD towards the global minimizer is not due to the lack of handling the variance of the gradient estimates, but rather to an (implicit) bias due to the large learning rate.

Furthermore, our proof does not necessarily require constant learning rate. If the learning rate varies per iteration as η_j, as long as there exists $\eta_{\min} > 0$ such that $\eta_j \geq \eta_{\min}$ for every $j \in \mathbb{N}$, it suffices to replace the η in (8) with $\eta_{\min}$. Specifically, the condition becomes:

$$\lambda_{\max}(\nabla^2\mathcal{R}_n(\theta^\star)) \leq \frac{2}{m\eta_{\min}}.$$

By reaplacing the η with $\eta_{\min}$ in subsequent propositions and bounds, the proof remains valid.

Appendix E. Experimental details and additional results

Our code is available at
https://github.com/pojoowu/Prevent-Memorization-via-implicit-regularization.

Model. For all the experiments, we fix the model to be a 2-layer ReLU network with a hidden width $m = 1000$. We initialize outer weights as standard Gaussian random variables, inner weights as Gaussian random variables of variance $1/d$, and the inner bias to be 0. Note that this is the standard initialization scheme of 2-layer networks (in the feature learning regime), and differs from our theoretical setup from Section 4 where the inner weights are set to ± 1.

Figure 1. We use a set of learning rates in $\{.5, .1, .05\}$ and with number of epochs

$$\{5,000, 25,000, 50,000\}$$

respectively. The batch size is set to 50. For each pair of (μ, σ) we generate 20 training data by sampling the standard Gaussian distribution, and keep the training data to be the same for every learning rate.

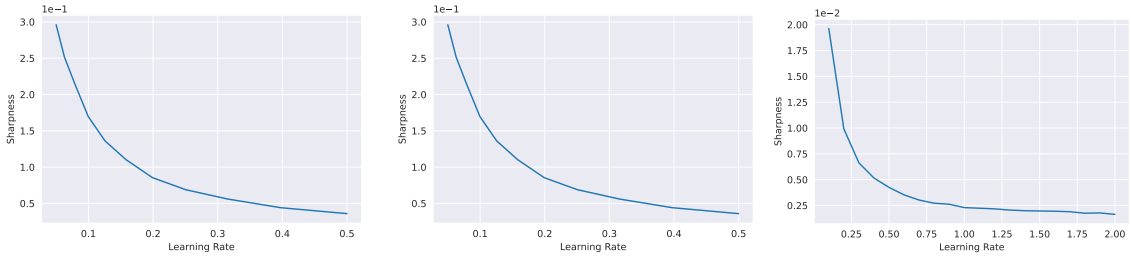

Figure 5: Largest eigenvalue of the loss Hessian (or sharpness) at the end of training, as a function of the learning rate. (left) and (middle) for the experiment of Figures 1 and 2, for $d = 1$ and $d = 10$ respectively, with $(\mu, \sigma) = (0.81, 0.57)$. (right) for the experiment of Figure 3.

Figure 2. The model is trained with different learning rates with 30 simulations. The set of learning rates is the same as previously, with additionally the learning rate .01 (and $200,000$ epochs). In each simulation, we generate 20 training data with the standard Gaussian law and use a batch size of 50. In addition, to estimate the excess risk

$$\mathcal{R}_n(\theta^\star) - \mathcal{R}_n(s^\star) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\left[(s_{\theta^\star}(Y) - s^\star(Y;\mu,\sigma))^2\right],$$

we generate 5000 Gaussian noises for each training data to simulate the expectation.

Figure 3. The training data is sampled from the isotropic Gaussian distribution of standard deviation 2, and we keep the dataset the same for training with each learning rate. We minimize by SGD over $r_\theta : \mathbb{R}^{d+1} \to \mathbb{R}^d$ the risk

$$\int_\delta^T \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z\sim\mathcal{N}(0,I)}\left[(r_\theta(t,\mu(t)Y + \sigma(t)Z) - Z)^2\right],$$

where $\mu(t) = e^{-t}$ and $\sigma(t) = \sqrt{1 - e^{-2t}}$. The integral over T is discretized over 100 equally spaces times. The batch size is set to $5,000$. For each batch element, the time t is sampled among the 100 discretization points, with a probability proportional to $\sigma(t)$. We use a set of learning rate in $\{2., 0.05\}$ and number of epochs to be $\{2 \times 10^4, 10^6\}$. Note that the risk differs from the one we analyze (c.f. (5)) by an affine transform. This is standard in practice for numerical stability reasons. Accordingly, we take the score to be $s_\theta(t,x) = -\frac{1}{\sigma(t)} r_\theta(t,x)$. We then generate new samples using the backward Ornstein-Uhlenbeck process starting from $T = 1$ and ending at $\delta = .01$. The MMD distance is calculated with the Gaussian kernel with bandwidth equal to 1. We also checked that other metrics give qualitatively similar conclusions (MMD with other bandwidth, Wasserstein distance).

Figure 4. We fix the learning rate to be .5 and the number of epochs to be $50,000$. We use a set of dimensions $\{2, 5, 10, 50, 80, 100, 200, 400, 1000\}$. For each dimension, we train the model with 3 simulations and in each simulation generate the training data by sampling the isotropic Gaussian of standard deviation 2. The score matching objective and generation procedure are the same as previously. We generate 5 sets of data with each simulations, and the MMD is computed with the Gaussian kernel with bandwidth set to 1.