# DIFFUSION MODELS ARE MINIMAX OPTIMAL DISTRIBUTION ESTIMATORS

**Kazusato Oko**
The University of Tokyo
RIKEN Center for Advanced Intelligence Project
oko-kazusato@g.ecc.u-tokyo.ac.jp

**Shunta Akiyama**
The University of Tokyo
shunta_akiyama@mist.i.u-tokyo.ac.jp

**Taiji Suzuki**
The University of Tokyo
RIKEN Center for Advanced Intelligence Project
taiji@mist.i.u-tokyo.ac.jp

## ABSTRACT

This paper provides the first rigorous analysis of estimation error bounds of diffusion modeling, trained with a finite sample, for well-known function spaces. The highlight of this paper is that when the true density function belongs to the Besov space and the empirical score matching loss is properly minimized, the generated data distribution achieves the nearly minimax optimal estimation rates in the total variation distance and in the Wasserstein distance of order one. Furthermore, we extend our theory to demonstrate how diffusion models adapt to low-dimensional data distributions. We expect these results advance theoretical understandings of diffusion modeling and its ability to generate verisimilar outputs.

## 1 INTRODUCTION

Diffusion modeling, in particular, score-based generative modeling (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2020; Ho et al., 2020; Vahdat et al., 2021), requires the gradient of the logarithmic density of the (diffused) data distribution, called the score. In practice, however, we have only access to the true distribution through a finite sample from that, and therefore the true score is replaced by a neural network (score network). We train the score network based on the score of the diffusion process from the empirical distribution, using the score matching technique (Hyvärinen & Dayan, 2005; Vincent, 2011). This replacement causes the difference between the true data distribution and the distribution of the outputs generated by diffusion modeling, which motivates the analysis of this error as a distribution estimation problem. In other words, *is diffusion modeling a good distribution estimator?*

Existing literature has analyzed the estimation error given the score approximation error bound as an assumption. (i) Under the $L^2$-bound on the score approximation accuracy, Song et al. (2020) showed the bound in the Kullback–Leibler (KL) divergence via Girsanov theorem, for continuous-time dynamics. Recently, the polynomial bound has appeared in discrete-time in the total variation distance (TV) (Lee et al., 2022b). Lee et al. (2022b) assumed the log-Sobolev inequality (LSI) for the true density, which was later eliminated by Chen et al. (2022) and Lee et al. (2022a). (ii) Concurrently, with the $L^\infty$-bound of the approximation error, De Bortoli et al. (2021) (also with dissipativily) and De Bortoli (2022) (under the manifold hypothesis) derived non-polynomial bounds in TV and in the Wasserstein distance of order one ($W_1$), respectively.

However, the important problem has been unaddressed, that is, whether the score can be appropriately approximated with a finite number of sample via score matching. As the only exception, De Bortoli (2022) derived the $n^{-1/d}$ bound in $W_1$ for $n$ data and a $d$-dimensional distribution. However, in their analysis, the neural network is assumed to almost perfectly fit the empirical score and the estimation bound depends on the convergence rate of the empirical distribution to the true one (Weed & Bach, 2019). Because of the same lower bound for the convergence of empirical measures

(Dudley, 1969), their $n^{-1/d}$ bound is essentially unimprovable with any structural assumption on the data distribution. Therefore, it is impossible to extend their result to formal density estimation problems, where the faster convergence rates depending on the smoothness of the true density are expected. We also mention generalization error analysis mainly on each one discretized step by Block et al. (2020), but they do not explicitly state the final estimation error and their intermediate bounds depend on the unknown Rademacher complexity which should be sufficiently large so that the hypothesis class well approximates the true score.

In summary, the fundamental question on the performance of diffusion models as a distribution learner largely remains open.

### 1.1 Our contributions

This work establishes a statistical learning theory for diffusion modeling. The convergence rate of the estimation error is derived assuming that the true density belongs to well-known function spaces and deep neural network is employed as an estimator. Surprisingly, we find that diffusion modeling can achieve the nearly minimax estimation rates. The contributions are detailed as follows:

- (i) We give the explicit form of approximation of the score with a neural network and derive the error bound in $L^2(p_t)$ at each $t$, where the initial density is supported in $[-1,1]^d$, in the Besov space $B^s_{p,q}([-1,1]^d)$, and smooth in the boundary.
- (ii) We then convert the approximation error analysis into the estimation error bounds. We derive the bound of $n^{-\frac{s}{d+2s}}$ in TV. Moreover, the rate of $n^{-\frac{s+1-\delta}{d+2s}}$ in $W_1$ is derived for an arbitrary fixed $\delta > 0$, with modified score matching. Notably, both of them are nearly minimax optimal.
- (iii) We extend our theory to demonstrate that diffusion models can avoid the curse of dimensionality under the manifold hypothesis, considering when the true data is distributed over a low-dimensional plane. This is a special case of De Bortoli (2022) but by far tighter.

Appendix A provides further related works about estimation problems for those less familiar with statistical learning theory and also introduces existing literature regarding the manifold hypothesis.

## 2 Problem settings

**Diffusion modeling** We basically follow the setting of Song et al. (2020) and the notation of De Bortoli (2022). $(B_t)_{[0,\overline{T}]}$ denote $d$-dimensional Brownian motion. We use $p_t$ for the distribution of $X_t$, and therefore $p_0$ is the data distribution. As a forward process $(X_t)_{[0,\overline{T}]}$ in $\mathbb{R}^d$, we consider the following Ornstein–Ulhenbeck (OU) process:

$$\mathrm{d}X_t = -\beta_t X_t \mathrm{d}t + \sqrt{2\beta_t}\mathrm{d}B_t, \quad X_0 \sim p_0.$$

Then we have that $X_t|X_0 \sim \mathcal{N}(m_t X_0, \sigma_t)$, where $m_t = \exp(-\int_0^t \beta_s \mathrm{d}s)$ and $\sigma_t^2 = 1 - \exp(-2\int_0^t \beta_s \mathrm{d}s)$. Under mild assumptions on $p_0$, that are easily verified for our setting (Haussmann & Pardoux, 1986), the backward process $(Y_t)_{[0,T]}$ with $Y_t = X_{\overline{T}-t}$ satisfies

$$\mathrm{d}Y_t = \beta_t(Y_t + 2\nabla \log p_t(Y_t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t, Y_0 \sim p_{\overline{T}}.$$

$\nabla \log p_t(x)$ is called the score, which is replaced by the score network $\hat{s}(x,t)$ trained with the finite sample. Also, because $p_t$ approaches $\mathcal{N}(0, I_d)$, we take $\overline{T} = \tilde{\mathcal{O}}(1)$ and replace the initial noise distribution of $Y_0$ by $\mathcal{N}(0, I_d)$. Then the modified backward process $(\hat{Y}_t)_{[0,\overline{T}]}$ is defined as

$$\mathrm{d}\hat{Y}_t = \beta_t(\hat{Y}_t + 2\hat{s}(\hat{Y}_t, t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t, \hat{Y}_0 \sim \mathcal{N}(0, I_d).$$

**Class of neural networks** As usual in approximation with neural networks (Yarotsky, 2017; Liang, 2017), the score network is selected from a class of deep neural network with the ReLU activation $\mathrm{ReLU}(x) = \max\{0, x\}$ (operated element-wise for a vector) (Nair & Hinton, 2010; Glorot et al., 2011) with a sparsity constraint (on the number of non-zero parameters). The score network is a function from $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$ to $y \in \mathbb{R}^d$.

**Definition 2.1.** A class of neural networks $\Phi(L, W, S, B)$ with height $L$, width $W$, sparsity constraint $S$, and norm constraint $B$ is defined as $\Phi(L, W, S, B) := \{(A^{(L)}\mathrm{ReLU}(\cdot) + b^{(L)}) \circ \cdots \circ (A^{(1)}x + b^{(1)}) | \, A^{(i)} \in \mathbb{R}^{W_i \times W_{i+1}}, b^{(i)} \in \mathbb{R}^{W_{i+1}}, \sum_{i=1}^{l}(\|A^{(i)}\|_0 + \|b^{(i)}\|_0) \leq S, \max_i \|A^{(i)}\|_\infty \vee \|b^{(i)}\|_\infty \leq B\}$.

**Score matching** Score matching with finite data $\{x_{0,i}\}_{i=1}^n$ selects the score network $\hat{s}$ from the hypothesis $\mathcal{S}$ so that $\hat{s}$ minimizes the empirical score matching loss:

$$\hat{s} \in \arg\min_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} \int_{t=\underline{T}}^{\overline{T}} \mathop{\mathbb{E}}_{x_t \sim p_t(x_t|x_{0,i})} [\|s(x_t, t) - \nabla \log p_t(x_t|x_{0,i})\|^2] \mathrm{d}t, \tag{1}$$

where $x_{0,i} \overset{\text{i.i.d.}}{\sim} p_0$ is assumed. We clip the integral interval by $\underline{T} > 0$ because generally the score blows up as $t \to 0$ and (1) gets $\infty$ for any neural network. We can use finite sample of $t$ and $x_t$, instead of taking expectation, which is explained in Appendix H.

### 2.1 Assumptions

We evaluate $\mathrm{TV}(X_0, \hat{Y}_{\overline{T}-\underline{T}})$ and $W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}})$ under the following assumptions. Let $d$ be a dimenision of the space, $n$ be the sample size, and $0 < p, q \leq \infty, s > 0$ with $s > (1/p - 1/2)_+$ be parameters of the Besov space. The besov spaces include the Sobolev Hölder spaces, and can contain not continuous functions (see Appendix B for details). Our main assumption is as follows.

**Assumption 2.2.** The true density $p_0$ is supported on $[-1, 1]^d$, where it is upper and lower bounded by $C_f$ and $C_f^{-1}$, respectively. Also, $p_0$ belongs to $U(B_{p,q}^s([-1, 1]^d); C)$ for some constant $C$.

$U(\cdot; C)$ means the ball of radius $C$ and we sometimes write it as $U(\cdot)$. We additionally make two technical assumptions. One is the smoothness of $\beta_t$.

**Assumption 2.3.** $\beta_\cdot : [0, \overline{T}] \to \mathbb{R}_+$ satisfies $0 < \underline{\beta} \leq \beta_t \leq \overline{\beta}$ for all $t$ and $\beta_\cdot \in U(\mathcal{C}^\infty([0, \overline{T}]), 1)$ as a function of $t$.

The other is the smoothness of the true density $p_0$ on the boundary region. Let $a_0$ be a sufficiently small value defined later.

**Assumption 2.4.** $p_0$ also belongs to $U(\mathcal{C}^\infty([-1, 1]^d \setminus [-1 + a_0, 1 - a_0]^d); 1)$.

This is used in the region where $p_t$ is not lower bounded. This is necessarily because in density estimation lower boundedness is typically assumed (Tsybakov, 2009) and without lower boundedness the minimax optimal rate gets worse than otherwise (Niles-Weed & Berthet, 2022). This assumption can be replaced by sufficiently slow decay of the density, such as LSI, used in Lee et al. (2022b).

## 3 Main results

Throughout this section, we fix $\delta > 0$ as a constant. We assume $n \gg 1$ and let $N = n^{\frac{d}{d+2s}}, \underline{T} = \mathrm{poly}(n^{-1})$, and $\overline{T} \simeq \log n$. Here $N$ is the parameter that controls the score network size. We take $a_0 = n^{-\frac{1-\delta}{d+2s}} = N^{-\frac{1-\delta}{d}}$ in Assumption 2.4.

### 3.1 Approximation of the true score

First, we consider approximating the true score $\nabla \log p_t$ via a deep neural network.

**Theorem 3.1.** *There exists a neural network $\phi_{\mathrm{score}} \in \Phi(L, W, S, B)$ that satisfies, for $t \in [\underline{T}, \overline{T}]$,*

$$\int_x p_t(x) \|\phi_{\mathrm{score}}(x, t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d}} \log(N)}{\sigma_t^2}.$$

*Here, $L, W, S$ and $B$ are evaluated as $L = \mathcal{O}(\log^4 N), \|W\|_\infty = \mathcal{O}(N \log^6 N), S = \mathcal{O}(N \log^8 N)$, and $B = \exp(\mathcal{O}(\log^4 N))$. Moreover, we can take $\phi_{\mathrm{score}}$ so that $\|\phi_{\mathrm{score}}(\cdot, t)\|_\infty = \mathcal{O}(\sigma_t^{-1} \log^{\frac{1}{2}} N)$ holds.*

**Proof overview** In order to obtain this result, the approximation should be constructed in the following ways. (i) It should reflect the structure of $p_0(x)$, especially the fact of $p_0(x) \in U(B^s_{p,q})$. (ii) It should serve as a good score approximation for different timepoints simultaneously, as a function of both $x$ and $t$. To address these issues, we construct a novel basis decomposition in the space of $\mathbb{R}^d \times [\underline{T}, \overline{T}]$, specially designed for score approximation. Moreover, as usual in approximation theory (e.g., Yarotsky (2017)) each basis can be realized by a neural network very efficiently, meaning that a polylogarithmic-sized network suffices with respect to the permissible error.

The basis decomposition goes as follows. First remind the B-spline basis decomposition of the Besov functions (DeVore & Popov, 1988; Dũng, 2011; Suzuki, 2018). Let $\mathcal{N}_l(x)$ be *cardinal B-spline* of order $l$, and for $k \in \mathbb{N}^d$ and $j \in \mathbb{Z}^d$, take the *tensor product B-spline basis* as $M^d_{k,j}(x) = \prod_{i=1}^d \mathcal{N}(2^{k_i}x - j_i)$. This is the basis function in $\mathbb{R}^d$ and a function $f$ in the Besov space is approximated by a super-position of $M^d_{k,j}(x)$ as $f_N = \sum_{i=1}^N \alpha_i M^d_{k_i,j_i}(x)$.

We decompose $p_0$ as $p_0(x) \approx \sum_{i=1}^N \alpha_i M^d_{k_i,j_i}(x)$. Defining the transition kernel $K_t(x|y) = \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp(-\frac{\|x - m_t y\|^2}{2\sigma_t^2})$, we have that $p_t(x) = \int p_0(y) K_t(x|y)\mathrm{d}y$. Now, $p_t(x)$ is approximated as $p_t(x) \approx \sum_{i=1}^N \alpha_i \int M^d_{k_i,j_i}(y) K(x|y)\mathrm{d}y$. Moreover, $E_{k,j}(x,t) = \int M^d_{k,j}(y)K(x|y)\mathrm{d}y$ is further decomposed as $E_{k,j}(x,t) = \prod_{i=1}^d \int \frac{\mathcal{N}(2^{k_i}y_i - j_i)}{\sigma_t\sqrt{2\pi}} \exp(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2})\mathrm{d}y_i$. We name $\mathcal{D}_{k_i,j_i}(x_i,t) = \int \frac{\mathcal{N}(2^{k_i}y_i - j_i)}{\sigma_t\sqrt{2\pi}} \exp(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2})\mathrm{d}x_i$ as the *diffused B-spline basis* and $E_{k,j}$ as the *tensor product diffused B-spline basis*. We show that there exists a neural network that approximates $\mathcal{D}_{k,j}$ and $E_{k,j}$. Then we obtain an efficient approximation of $p_t(x)$. In the same way, we can approximate $\nabla p_t(x)$. Based on these, we finally obtain the approximation of the score $\nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)}$.

The complete proof can be found in Appendix D. For more detailed proof sketch, see Appendix D.1.

### 3.2 Generalization of the score network

We then consider the generalization error of the score network. Before stating the bound, we limit the hypothesis $\Phi$ given in Theorem 3.1 into $\mathcal{S}$, which consists of a network $\phi$ satisfying $\|\phi(\cdot,t)\|_\infty = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} n)$ because we can take $\phi_{\mathrm{score}}$ so that $\|\phi_{\mathrm{score}}(\cdot,t)\|_\infty = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} N)$ holds according to Theorem 3.1. Then we let $\mathcal{L} = \{\ell_s \colon [-1,1]^d \to \mathbb{R}_+ |\ s \in \mathcal{S}\}$, where $\ell_s$ is defined by $\ell_s(x) = \int_{t=\underline{T}}^{\overline{T}} \int \|s(x_t,t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x)\mathrm{d}x_t\mathrm{d}t$. Note that the empirical score matching loss (1) is written as $\frac{1}{n}\sum_{i=1}^n \ell_{\hat{s}}(x_i)$. For $\mathcal{L}$, let $\mathcal{N} = \mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty([-1,1]^d)}, \varepsilon)$ be the $\varepsilon$-covering number of $\mathcal{L}$ with the $L^\infty$ norm. Based on this, we can bound the generalization error of the score network selected in the empirical score matching.

**Theorem 3.2.** *For sufficiently small $\varepsilon > 0$, the minimizer $\hat{s}$ of the empirical score matching loss* (1) *over $\mathcal{S}$ satisfies that*

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{x_t\sim p_t}[\|\hat{s}(x_t,t) - \nabla \log p_t(x_t)\|^2]\mathrm{d}t\right] \tag{2}$$

$$\lesssim \inf_{s\in\mathcal{S}} \int_{\underline{T}}^{\overline{T}} \mathbb{E}_{x_t\sim p_t}[\|s(x_t,t) - \nabla \log p_t(x_t)\|_2^2]\mathrm{d}t + \frac{\sup_{s\in\mathcal{S}}\|\ell_s\|_{L^\infty([-1,1]^d)}\log(\mathcal{N})}{n} + \varepsilon. \tag{3}$$

The proof is inspired by Schmidt-Hieber (2020); Hayakawa & Suzuki (2020). See Appendix E.4.

The first term can be bounded by $N^{-2s/d}\log N(\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T}))$, according to Corollary D.13, which is obtained from Theorem 3.1. The second term is bounded by $\lesssim N\log^2(n)(\log^{16}(N) + \log^{12}(N)\log(\varepsilon^{-1}))$, because Appendix E.2 gives $\sup_{s\in\mathcal{S}}\|\ell_s\|_{L^\infty([-1,1]^d)} \lesssim \log^2 n$ and Appendix E.3 gives $\log(\mathcal{N}) \lesssim N(\log^{16} N + \log^{12} N \log(\varepsilon^{-1}))$. Now, we apply $N = n^{d/(2s+d)}$ and set $\varepsilon = n^{-2s/(2s+d)}$ to obtain (2) $\lesssim n^{-\frac{2s}{d+2s}}\log^{18}(n)$. For more detailed sketch, see Appendix E.1.

### 3.3 Estimation error analysis

Here we finally obtain the estimation error bounds. As a small modification, if $\|\hat{Y}_{\overline{T}-\underline{T}}\|_\infty \geq 2$, then we reset it to $\hat{Y}_{\overline{T}-\underline{T}} = 0$. First, the estimation error in the total variation distance is presented.

**Theorem 3.3.** *Let $\underline{T} = n^{-\mathcal{O}(1)}$ and $\overline{T} = \frac{s \log n}{\underline{\beta}(d+2s)}$. Then,*

$$\mathbb{E}[\mathrm{TV}(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-s/(2s+d)} \log^9 n.$$

On the other hand, we can show the following lower bound exists.

**Proposition 3.4.** *For $0 < p, q \leq \infty$, $s > 0$, and $s > \max\{d(\frac{1}{p} - \frac{1}{2}), 0\}$, we have that*

$$\inf_{\hat{\mu}} \sup_{p \in B^s_{p,q}} \mathbb{E}[\mathrm{TV}(\hat{\mu}, p)] \gtrsim n^{-s/(2s+d)},$$

*where $\hat{\mu}$ runs over all estimators based on $n$ observations.*

We have proven that diffusion modeling achieves the minimax estimation rate for the Besov space $B^s_{p,q}$ in the total variation distance up to the logarithmic factor. Appendix F.1 provides the proofs.

Moreover, we also have the following bound in the Wesserstein distance of order one.

**Theorem 3.5.** *We can train the score network with $n$ sample and with that we have*

$$\mathbb{E}[W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-(s+1-\delta)/(d+2s)}. \tag{4}$$

The minimax rate in $W_1$ is $n^{-\frac{s+1}{2s+d}}$ (Niles-Weed & Berthet, 2022), and thus (4) is also nearly minimax optimal up to $\delta$. For Theorem 3.5, we switch the score networks during the backward process, where each network is adjusted to a different time interval, as a technical modification. This is explained in Appendix F.2.

Also, the bound on the time discretization error is discussed in Appendix I.

## 4 Error analysis under the manifold hypothesis

When the true data is distributed over a $d'$-dimensional plane with $d' < d$, we can replace $d$ in (4) by $d'$, obtaining the improved bound in $W_1$. See Appendix A.2 for motivations and related works.

We assume that the true density $p_0$ is a probability measure that is absolutely continuous with respect to the Lebesgue measure on the plane. Its probability density, as a function on the canonical coordinate system of the plane, is assumed to satisfy Assumptions 2.2 and 2.4, with $d$ replaced by $d'$. We also assume Assumption 2.3 as well. Then, we obtain the following bound.

**Theorem 4.1.** *We can train the score network with $n$ sample so that the estimation error in the Wasserstein distance of order one is bounded by*

$$\mathbb{E}[W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-\frac{s+1-\delta}{d'+2s}}.$$

Contrary to Theorem 3.3, the upper bound here depends on $d'$ (not on $d$). Thus, we conclude that the diffusion models can avoid the curse of dimensionality.

In Appendix G, Appendix G.1 states the formal settings, Appendix G.2 provides the proof overview, and Appendix G.3 gives the complete proof. Simply put, we decompose the score function into two parts. One is determined by the diffusion process on the $d'$-dimensional plane, which the difficulty in approximation mostly depends on. The other part corresponds to the diffusion process on the orthocomplement and is easy to be approximated.

## 5 Conclusion

We showed that diffusion modeling can achieve nearly minimax estimation rates in both TV and $W_1$. To approximate the score, the novel basis is introduced, which we call the diffused B-spline basis. We also demonstrated that diffusion models can avoid the curse of dimensionality under the manifold hypothesis. In summary, we analyzed diffusion models from the statistical learning theory and provided theoretical supports for the real-world success of diffusion models.

ACKNOWLEDGMENTS

KO was partially supported by Fujitsu Ltd. SA was partially supported by JSPS KAKENHI (22J13388). TS was partially supported by JSPS KAKENHI (20H00576) and JST CREST.

## REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

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

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Nicolas Boullé, Yuji Nakatsukasa, and Alex Townsend. Rational neural networks. *Advances in Neural Information Processing Systems*, 33:14243–14253, 2020.

Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

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

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=MhK5aXo3gB`.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Ronald A DeVore and Vasil A Popov. Interpolation of besov spaces. *Transactions of the American Mathematical Society*, 305(1):397–414, 1988.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Dinh Dũng. Optimal adaptive sampling recovery. *Advances in Computational Mathematics*, 34(1): 1–41, 2011.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

U. G. Haussmann and E. Pardoux. Time Reversal of Diffusions. *The Annals of Probability*, 14 (4):1188–1205, 1986. doi: 10.1214/aop/1176992362. URL `https://doi.org/10.1214/aop/1176992362`.

Satoshi Hayakawa and Taiji Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

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

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

I. A. Ibragimov and Rafail Z. Khas'minskii. More on the estimation of distribution densities. *Journal of Soviet Mathematics*, 25:1155–1165, 1984.

Ioannis Karatzas, Ioannis Karatzas, Steven Shreve, and Steven E Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022a.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022b.

Jing Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

Tengyuan Liang. How well can generative adversarial networks learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

Lukas Meier, Sara Van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Hrushikesh N Mhaskar and Charles A Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied mathematics*, 13(3):350–373, 1992.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21(174):1–38, 2020.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=AiNrnIrDfD9`.

Johannes Schmidt-Hieber. Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pp. 1051–1071. PMLR, 2021.

Shashank Singh and Barnabás Póczos. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2018.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. *Advances in Neural Information Processing Systems*, 34: 3609–3621, 2021.

Matus Telgarsky. Neural networks and rational functions. In *International Conference on Machine Learning*, pp. 3387–3393. PMLR, 2017.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Hans Triebel. *Theory of Function Spaces*. Monographs in Mathematics. Birkhäuser Verlag, 1983. ISBN 9783764313814.

Hans Triebel. Entropy numbers in function spaces with mixed integrability. *Revista matemática complutense*, 24(1):169–188, 2011.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794. URL https://doi.org/10.1007/b13794.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pp. 1564–1599, 1999.

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

## TABLE OF CONTENTS

# A Additional related works

## A.1 Distribution and function estimation

Recently, minimax estimation rates in the Wasserstein distance have been investigated by several works (empirical distribution (Weed & Bach, 2019; Singh & Póczos, 2018; Lei, 2020); smooth density (Liang, 2017; Singh et al., 2018; Schreuder et al., 2021)); Besov space (Niles-Weed & Berthet, 2022)). Niles-Weed & Berthet (2022) utilized the wavelet basis for the Besov space, while Liang (2017) used neural networks as an estimator motivated by Generative Adversarial Networks (GAN) (Goodfellow et al., 2020).

We would like to emphasize that our work is not replacement of wavelet expansion of Niles-Weed & Berthet (2022) with neural networks. In diffusion modeling, we first minimize the squared-error-like score matching loss, and then consider the estimation error. This makes existing sharp bounds in $W_1$ unavailable. Contrary to the analysis of GAN, where the minimax problem of the final goal directly relates to $W_1$, analysis of diffusion models requires conversion of the score approximation error to the estimation error.

What we are built on is rather the theory of function estimation with deep neural networks in $L^p$ norms (Barron, 1993; Yarotsky, 2017; Petersen & Voigtlaender, 2018; Suzuki, 2018; Schmidt-Hieber, 2020; Hayakawa & Suzuki, 2020). Our approximation result can be seen as an extension of the B-spline basis expansion used in Suzuki (2018). On the other hand, our generalization bound relies on Schmidt-Hieber (2020); Hayakawa & Suzuki (2020).

## A.2 Analysis under the manifold hypothesis

Although the obtained rates in Theorem 3.3 is minimax optimal, it still suffers from the *curse of dimensionality*: the exponent of the convergence rate depends on $n$. One approach to avoid this curse of dimensionality in statistics is to assume mixed or anisotropic smoothness (Ibragimov & Khas'minskii, 1984; Meier et al., 2009; Suzuki, 2018; Suzuki & Nitanda, 2021), and our theory directly applies to them. On the other hand, the *manifold hypothesis*, that the distributions of real-world data lie in low dimensional manifolds, has been proposed (Tenenbaum et al., 2000; Fefferman et al., 2016), and this is another assumption that convergence rates dependent not on the dimension $d$ of the space itself but on the manifold's dimension can be obtained Nakada & Imaizumi (2020); Schmidt-Hieber (2019).

Recently, the convergence of diffusion modeling under the manifold hypothesis has begun to be investigated. The bound by Pidstrigach (2022), however, is not quantitative and does not consider the estimation rate. De Bortoli (2022) considered the estimation rates, but the approximation error should be exponentially small with respect to the desired estimation rate. Therefore, none of the literature has shown that diffusion models can ease the curse of dimensionality. This is what we work on in Appendix G, defining the specific class of density function with intrinsic dimensionality.

## B The Besov space

As a class of the true density, we used the Besov space, because this allows us to discuss many well-known function classes in a unified manner. Here we give formal definition and some properties of the Besov spaces. We first introduce the modulus of smoothness. We assume that $\Omega$ be a cube in $\mathbb{R}^d$.

**Definition B.1.** For a function $f \in L^p(\Omega)$ for some $p \in (0, \infty]$, the $r$-th modulus of smoothness of $f$ is defined by

$$w_{r,p}(f,t) = \sup_{\|h\|_2 \le t} \|\Delta_h^r(f)\|_p,$$

$$\text{where } \Delta_h^r(f)(x) = \begin{cases} \sum_{j=0}^r \binom{r}{j}(-1)^{r-j} f(x+jh) & (\text{if } x+jh \in \Omega \text{ for all } j) \\ 0 & (\text{otherwise}). \end{cases}$$

**Definition B.2** (Besov space $B^s_{p,q}(\Omega)$)**.** For $0 < p, q \le \infty, s > 0, r := \lfloor s \rfloor + 1$, let the seminorm $|\cdot|_{B^s_{p,q}}$ be

$$|f|_{B^s_{p,q}} = \begin{cases} \left(\int_0^\infty (t^{-s} w_{r,p}(f,t))^q \frac{dt}{t}\right)^{\frac{1}{q}} & (q < \infty), \\ \sup_{t>0} t^{-s} w_{r,p}(f,t) & (q = \infty). \end{cases}$$

The norm of the Besov space $B^s_{p,q}$ is defined by $\|f\|_{B^s_{p,q}} = \|f\|_p + |f|_{B^s_{p,q}}$, and we have $B^s_{p,q} = \{f \in L^p(\Omega) | \, \|f\|_{B^s_{p,q}} < \infty\}$.

Let us take several examples of function classes that can be embedded in the Besov spaces. For $\alpha \in \mathbb{Z}^d_+$, let $\partial^\alpha = \frac{\partial^{|\alpha|} f}{\partial^{\alpha_1}_{x_1} \cdots \partial^{\alpha_d}_{x_d}}(x)$. The Hölder space for $s \in \mathbb{R}_{>0} \setminus \mathbb{Z}_+$ is a set of $\lfloor s \rfloor$ times differentiable functions $\mathcal{C}^s(\Omega) = \{f \colon \Omega \to \mathbb{R} | \, \|f\|_{\mathcal{C}^s} := \max_{|\alpha| \le s} \|D^\alpha\|_\infty + \max_{m=\lfloor s \rfloor} \sup_{x,y \in \Omega} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|^{s-\lfloor s \rfloor}} < \infty\}$ for $s \in \mathbb{R}_{>0} \setminus \mathbb{Z}_+$. The Sobolev space for $s \in \mathbb{N}, 1 \le p \le \infty$ is a set of $s$ times differentiable functions $W^s_p(\Omega) := \{f \colon \Omega \to \mathbb{R} | \, \|f\|_{W^s_p} := (\sum_{|\alpha| \le s} \|\partial^\alpha f\|_p^p)^{\frac{1}{p}} < \infty\}$. Then the following relationships are due to Triebel (1983):

- For $s \in \mathbb{N}$, $B^s_{p,1}(\Omega) \hookrightarrow W^s_p(\Omega) \hookrightarrow B^s_{p,\infty}(\Omega)$.
- $B^s_{2,2}(\Omega) = W^s_2(\Omega)$.
- For $s \in \mathbb{R}_{>0} \setminus \mathbb{Z}_+$, $\mathcal{C}^s(\Omega) = B^s_{\infty,\infty}(\Omega)$.

If $s > d/p$, $B^s_{p,q}(\Omega)$ is continuously embedded in the set of the continuous functions. Otherwise, the elements in the space is no longer continuous. Our result is valid for $B^s_{p,q}(\Omega)$ with $s > d(1/p - 1/2)_+$, and thus can include not continuous functions, unlike existing bounds assuming smoothness or Lipschitzness (Lee et al., 2022b;a; Chen et al., 2022).

## C Several high-probability bounds on the backward paths

One of the difficulties in the analysis is the unboundedness of the space and the value of the score. This subsection aims to provide several treatments for such issues, before going into the main part of the proofs. These inequalities allow us to focus on the score approximation within the bounded region. We note that, however, some of the following bounds still depend on the time $t$, and therefore the level of difficulty for approximation and estimation of the score differs with respect to $t$.

In the following, we define several constants $C_{\mathrm{a},i}$. Other than in this section, we simply denote them as $C_{\mathrm{a}}$ for simplicity.

### C.1 Bounds on $\|Y_t\|$ and $\|\Delta Y_t\|$ with high probability

We first provide several high-probability bounds, which guarantee that most of the paths travel within some bounded region.

**Lemma C.1** (Bounds on $\|Y_t\|$ and $\|\Delta Y_t\|$ with high probability)**.** *There exists a constant $C_{\mathrm{a},1}$ such that*

$$\mathbb{P}\left[\|Y_t\|_\infty \leq m_{\overline{T}-t} + C_{\mathrm{a},1}\sigma_{\overline{T}-t}\sqrt{\log(\varepsilon^{-1}\underline{T}^{-1}\overline{T})} \textit{ for all } t \in [0, \overline{T}-\underline{T}]\right] \geq 1-\varepsilon.$$

*Moreover, for an arbitrarily fixed $0 < \tau \leq 1$,*

$$\mathbb{P}\left[\|Y_t - Y_{t+\tau}\|_\infty \leq C_{\mathrm{a},1}\sqrt{\tau\log(\varepsilon^{-1}\tau^{-1}\overline{T})} \textit{ for all } t \in [0, \overline{T}-\tau]\right] \geq 1-\varepsilon.$$

*Proof.* Remind that $Y_t = X_{\overline{T}-t}$. Thus we discuss bounding $X_t$ in the following.

We begin with the first assertion. Let $t_1, t_2, \cdots, t_K$ be time steps satisfying $\underline{T} = t_1 < t_2 < \cdots < t_K = \overline{T}$ with $t_i - t_{i-1} = \Delta t$ that is some scaler value specified later. We first show the following for some constant $C_1$:

$$\mathbb{P}\left[\|X_t\|_\infty \leq m_t + C_1\sigma_t\sqrt{\log\varepsilon^{-1}} \text{ for all } t = t_i \ (i = 1, 2, \cdots, K)\right] \geq 1 - \varepsilon K. \tag{5}$$

Remind that $X_t|X_0$ follows $\mathcal{N}(m_t X_0, \sigma_t^2)$ and $\|X_0\|_\infty \leq 1$. Lemma J.14 yields that

$$\mathbb{P}\left[\|X\|_\infty \leq m_t + C_1\sigma_t\sqrt{\log\varepsilon^{-1}} \text{ for some fixed } t = t_i\right] \geq 1-\varepsilon,$$

which immediately yields (5).

Then we consider how far each particle $X_t$ moves from $t = t_{i-1}$ to $t_i$. Equivalently, we consider $X_t$ and decompose it into

$$X_t = \exp\left(-\int_{s=t_{i-1}}^{t_i}\beta_s \mathrm{d}s\right)X_{t_{i-1}} + B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s)}, \tag{6}$$

where $B_s$ denotes a $d$-dimensional Brownian motion. This is obtained by considering the Ornstein-Uhlenbeck process starting from $t = t_{i-1}$. By Lemma J.15, with probability at least $\varepsilon$, the following holds uniformly over $t \in [t_{i-1}, t_i]$:

$$\begin{aligned}\|X_t\|_\infty &\leq \exp\left(-\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s\right)\|X_{t_{i-1}}\|_\infty + \sqrt{1-\exp(-2\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s)}\cdot 2\sqrt{\overline{\beta}2\log d\varepsilon^{-1}} \\ &\leq \exp\left(-\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s\right)\|X_{t_{i-1}}\|_\infty + \sqrt{2\underline{\beta}\Delta t}\cdot 2\sqrt{\overline{\beta}2\log d\varepsilon^{-1}}.\end{aligned}$$

If $\|X_{t_{i-1}}\|_\infty \leq m_{t_{i-1}} + C_1\sigma_{t_{i-1}}\sqrt{\log\varepsilon^{-1}}$, this is further bounded by

$$\|X_t\|_\infty \leq m_{t_{i-1}} + C_1\sigma_{t_{i-1}}\sqrt{\log\varepsilon^{-1}} + \sqrt{\Delta t}\cdot 4\sqrt{\overline{\beta}\underline{\beta}\log d\varepsilon^{-1}}.$$

Because we can check that $\sigma_t \simeq \sqrt{t}\wedge 1 \geq \sqrt{\underline{T}}$ holds, if we take $\Delta \leq \underline{T}$, then we have that

$$C_1\sigma_{t_{i-1}}\sqrt{\log\varepsilon^{-1}} + \sqrt{\Delta t}\cdot 4\sqrt{\overline{\beta}\underline{\beta}\log d\varepsilon^{-1}} \lesssim C_2\sigma_{t_{i-1}}\sqrt{\log\varepsilon^{-1}} \tag{7}$$

for all $t \in [t_{i-1}, t_i]$, with some constant $C_2$.

Therefore, with probability $1 - 2K\varepsilon$ we have (5), and (7) for all $i$. We need to take $K = \mathcal{O}(\overline{T}/\underline{T})$ to satisfy $\Delta \leq \underline{T}$. We reset $\frac{\varepsilon}{K}$ as a new $\varepsilon$ and adjust $C_2$ accordingly. Now the first assertion is proved.

Next, we consider the second assertion. Let us consider a different time discretization $t_0 = 0, t_1 = \tau, t_2 = 2\tau, \cdots, t_K = K\tau$ with $K = \min\{i \in \mathbb{N} | K\tau \geq \overline{T}\}$. Then, from the first argument, we have that $\|X_t\|_\infty \le m_t + C_2\sigma_t\sqrt{\log(\varepsilon^{-1}\tau^{-1}\overline{T})}$ holds with probability at least $1-\varepsilon$, for all $t = t_0, t_1, \cdots, t_K$. We condition the event conditioned by this. By (6), we have that, for $t \ge t_{i-1}$,

$$X_t - X_{t_{i-1}} = \left[\exp\left(-\int_{s=t_{i-1}}^{t_i} \beta_s \mathrm{d}s\right) - 1\right] X_{t_{i-1}} + B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i} \beta_s \mathrm{d}s)},$$

which yields that

$$\begin{aligned}\|X_t - X_{t_{i-1}}\|_\infty &\le \left|\exp\left(-\int_{s=t_{i-1}}^{t_i} \beta_s \mathrm{d}s\right) - 1\right| \|X_{t_{i-1}}\|_\infty + \left\|B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i} \beta_s \mathrm{d}s)}\right\|_\infty \\ &\le \tau\overline{\beta}(m_{t_{i-1}} + C_2\sigma_{t_{i-1}}\sqrt{\log(\varepsilon^{-1}\tau^{-1}\overline{T})}) + \left\|B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i} \beta_s \mathrm{d}s)}\right\|_\infty\end{aligned}$$

We bound the last term over $t \in [t_{i-1}, t_i]$. With probability at least $1 - \frac{\varepsilon}{K}$, that is bounded by $\sqrt{2\underline{\beta}\tau} \cdot 2\sqrt{\overline{\beta}2\log dK\varepsilon^{-1}}$ according to Lemma J.15. To summarize, with probability at least $1-2\varepsilon$,

$$\sup_{t\in[t_{i-1},t_i]} \|X_t - X_{t_{i-1}}\|_\infty \le \tau\overline{\beta}(m_{t_{i-1}} + C_2\sigma_{t_{i-1}}\sqrt{\log(\varepsilon^{-1}\tau^{-1}\overline{T})}) + \sqrt{2\underline{\beta}\tau} \cdot 2\sqrt{\overline{\beta}2\log dK\varepsilon^{-1}}$$

holds for all $i = 0, 1, \cdots, K-1$. RHS is bounded by $C_3\sqrt{\tau\log\varepsilon^{-1}\tau^{-1}\overline{T}}$ with some sufficiently large constant $C_3$.

Then, for any $t$, there exists $i$ such that $t \le t_i \le t+\tau$. Thus, with probability $1-2\varepsilon$, $\|X_t - X_{t+\tau}\|_\infty \le \|X_t - X_{t_{i-1}}\|_\infty + \|X_{t_i} - X_{t_{i-1}}\|_\infty + \|X_{t+\tau} - X_{t_i}\|_\infty$ is bounded by $3C_3\sqrt{\tau\log\varepsilon^{-1}\tau^{-1}\overline{T}}$ for all $t$. Setting $2\varepsilon$ to $\varepsilon$ yields the second assertion. □

### C.2 Bounds on $p_t(x)$

We then give upper and lower bounds on $p_t(x)$.

**Lemma C.2** (Upper and lower bounds on the density $p_t(x)$)**.** *The following upper and lower bounds on $p_t(x)$ holds for a constant $C_{\mathrm{a},2}$ depending on $C_f$ and $d$:*

$$C_{\mathrm{a},2}^{-1}\exp\left(-\frac{d(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\right) \le p_t(x) \le C_{\mathrm{a},2}\exp\left(-\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2}\right). \quad \textit{(for all } t.)$$

*Proof.* We first consider the case when $x \in [-m_t, m_t]^d$. The upper bound is relatively easy. $f(y) \le C_f \mathbb{1}[y \in [-1,1]^d]$ means

$$\begin{aligned}p_t(x) &= \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y \quad (8)\\ &\le \int \frac{C_f\mathbb{1}[y\in[-1,1]^d]}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y = \frac{2^dC_f}{\sigma_t^d(2\pi)^{\frac{d}{2}}}.\end{aligned}$$

At the same time, we have that

$$p_t(x) \le \int \frac{C_f}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y = \frac{C_f}{m_t^d}. \quad (9)$$

Thus, according to (8) and (9), $p_t(x)$ is bounded by $\min\left\{\frac{2^dC_f}{\sigma_t^d(2\pi)^{\frac{d}{2}}}, \frac{C_f}{m_t^d}\right\}$. This is further bounded by a constant that depends only on $C_f$ and $d$, because $m_t^2 + \sigma_t^2 = 1$ holds for all $t$.

The lower bound can be understood as follows. We have

$$\begin{aligned}p_t(x) &= \int \frac{C_f^{-1}}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y\\ &\ge \frac{1}{(2\pi)^{\frac{d}{2}}}\int f(x/m_t - \sigma_t y)\exp\left(-\frac{\|m_ty\|^2}{2}\right)\mathrm{d}y \quad \text{(by letting } (x-m_ty)/\sigma_t \mapsto m_ty).\end{aligned} \quad (10)$$

Since $x \in [-m_t, m_t]^d$, we have $x/m_t \in [-1,1]^d$. Thus, $|\{y \in [-1,1]^d |\, x/m_t - \sigma_t y \in [-1,1]\}| \geq 1$. Moreover, $\exp\left(-\frac{\|m_t y\|^2}{2}\right) \geq \exp(-d^2/2)$ in $y \in [-1,1]^d$. Therefore, the integral (10) is lower bounded by $\exp(-d^2/2)$.

We then consider the case when $x \notin [-m_t, m_t]^d$. For such $x$, let $r = (\|x\|_\infty - m_t)/\sigma_t$ and choose $i^*$ from $\{1, 2, \cdots, d\}$ such that $|x_{i^*}| = \|x\|_\infty = m_t + r/\sigma_t$ holds. Then, we have the upper bound of $p_t(x)$ as

$$
\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&\leq C_f \prod_{i=1}^d \int \frac{\mathbb{1}[-1 \leq y_i \leq 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i \\
&\lesssim C_f \int_{y_{i^*} \in [-1,1]} \frac{1}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_{i^*} - m_t y_{i^*})^2}{2\sigma_t^2}\right) \mathrm{d}y \qquad (11) \\
&\Bigg(\text{because} \int \frac{\mathbb{1}[-1 \leq y_i \leq 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i \text{ for } i \neq i^* \text{ is bounded by } \mathcal{O}(1), \\
&\qquad\qquad \text{as } p_t(x) \text{ for } x \in [-m_t, m_t]^d.\Bigg) \\
&\leq \frac{C_f}{m_t} \int_{a = r/\sqrt{2}}^\infty \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}a \qquad (\text{by } a = x_{i*} - m_t y_{i_*}/\sqrt{2}\sigma_t) \\
&\leq \frac{C_f}{m_t} \exp\left(-r^2/2\right) = \frac{C_f}{m_t} \exp\left(-\frac{(\|x\|_\infty - m_t)^2}{2\sigma_t^2}\right)
\end{aligned}
$$

where we used $\int_z^\infty e^{-a^2} \mathrm{d}a \leq e^{-z^2}$ (see, e.g. Chang et al. (2011)) for the last inequality. Also, (11) is alternatively bounded by $\frac{2C_f}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(\|x\|_\infty - m_t)^2}{2\sigma_t^2}\right)$. Because $m_t^2 + \sigma_t^2 = 1$ means that $\min\{m_t, \sigma_t\} \gtrsim 1$, it holds that $p_t(x) \lesssim C_f \exp\left(-\frac{(\|x\|_\infty - m_t)^2}{2\sigma_t^2}\right)$.

On the other hand,

$$
\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&\geq C_f^{-1} \prod_{i=1}^d \underbrace{\int_{y_i \in [-1,1]} \frac{1}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y}_{\text{(a)}} \\
&= C_f^{-1} \left(\int_{y_{i^*} \in [-1,1]} \frac{1}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_{i^*} - m_t y_{i^*})^2}{2\sigma_t^2}\right) \mathrm{d}y\right)^d \\
&\qquad\qquad (\text{because (a) is minimized when } i = i_*) \\
&\geq \frac{C_f^{-1}}{m_t^d} \left(\int_{a = r/\sqrt{2}}^{r/\sqrt{2} + \sqrt{2} m_t/\sigma_t} \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}y\right)^d \qquad (\text{by } (x_{i^*} - m_t y_{i^*})/\sqrt{2}\sigma_t) \\
&\geq \frac{C_f^{-1}}{m_t^d} \left(\int_{a = r/\sqrt{2}}^{r/\sqrt{2} + \sqrt{2} m_t} \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}y\right)^d
\end{aligned}
$$

$$\geq \frac{C_f^{-1}}{m_t^d}\left(\frac{\sqrt{2}m_t}{\sqrt{\pi}}\exp\left(-(r/\sqrt{2}+\sqrt{2}m_t)^2\right)\right)^d$$

(by lower bounding $\exp(-a^2)$ in the integral interval and just multiplying the width of the interval)

$$\geq \frac{C_f^{-1}}{m_t^d}\left(\frac{\sqrt{2}m_t}{\sqrt{\pi}}\exp\left(-r^2-4\right)\mathrm{d}a\right)^d$$

$$\geq \frac{C_f^{-1}2^{d/2}}{e^{4d}\pi^{d/2}}\exp\left(-dr^2\right),$$

which gives the lower bound on $p_t(x)$. □

### C.3 Bounds on the derivatives of $p_t(x)$ and the score

This subsection evaluates the derivatives of $p_t(x)$ and the score. On the one hand, straightforward argument yields that the derivatives of $p_t(x)$ is bounded by $\partial^k p_t(x) = \mathcal{O}(1/\sigma_t^k) = \mathcal{O}(t^{-k/2})$. On the other hand, as for the score, $\sup_{x\in\mathbb{R}^d}\|\nabla\log p_t(x)\| = \infty$ holds in general, which prevents us to construct an approximation of the score with neural networks. This is because $\nabla\log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)}$ and $p_t(x)$ can be arbitrarily small as $\|x\|\to\infty$. Nevertheless, using Lemma C.2, we can show the bounds on the score dependent on $x$ and $t$, in the next Lemma C.3. In Lemma C.4, Lemma C.3 is used to show that the decay of $p_t$ is so fast that the approximation error in the region with small $p_t(x)$ (that can be $\gg 1$ in some $x$) does not much affects the $L^2(p_t)$ approximation error bound; We can show that $\|\nabla\log p_t(x)\| = \tilde{\mathcal{O}}(1/\sigma_t) = \tilde{\mathcal{O}}(1\vee 1/\sqrt{t})$ with high probability (when $x\sim p_t$).

**Lemma C.3** (Boundedness of derivatives). *For $k\in\mathbb{Z}_+$, there exists a constant $C_{\mathrm{a},3}$ depending only on $k$, $d$, and $C_f$ such that*

$$|\partial_{x_{i_1}}\partial_{x_{i_2}}\cdots\partial_{x_{i_k}}p_t(x)| \leq \frac{C_{\mathrm{a},3}}{\sigma_t^k}. \tag{12}$$

*Moreover, we have that*

$$\|\nabla\log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t}\cdot\left(\frac{(\|x\|_\infty - m_t)_+}{\sigma_t}\vee 1\right), \tag{13}$$

*and that for $i\in\{1,2,\cdots,d\}$,*

$$\|\partial_{x_i}\nabla\log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t^2}\left(\frac{(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\vee 1\right). \tag{14}$$

*and that*

$$\|\partial_t\nabla\log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t^3}\left[|\partial_t\sigma_t| + |\partial_t m_t|\right]\left(\frac{(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\vee 1\right)^{\frac{3}{2}}. \tag{15}$$

*Proof.* First, we consider (12). Let $g_1(x) = p_t(x) = \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y$. For $s\in\mathbb{Z}_+^d$, we abbreviate the notation as $g_1^{(s)}(x) = \partial_{x_1}^{s_1}\partial_{x_2}^{s_2}\cdots\partial_{x_d}^{s_d}g_1(x)$. For $s\in\mathbb{Z}_+^d$, we define $B_s = \{s'\in\mathbb{Z}_+^d | s'_i\leq s_i\ (i=1,\cdots,d)\}$ and a constant $c_s$ such that $\partial_{x_1}^{s_1}\partial_{x_2}^{s_2}\cdots\partial_{x_d}^{s_d}e^{-\|x\|^2/2} = \sum_{s'\in B_s}c_{s'}x_1^{s'_1}x_2^{s'_2}\cdots x_d^{s'_d}e^{-\|x\|^2/2}$ holds. Then, because of $\partial_{x_i} = \frac{1}{\sigma}\partial_{\frac{x_i}{\sigma}}$, we can write $g_1^{(s)}(x)$ as

$$g_1^{(s)}(x) = \frac{\sum_{s'\in B_s}c_{s'}}{\sigma_t^{\sum_{i=1}^d s_i}}\underbrace{\int\prod_{i=1}^d\left(\frac{x_i-my_i}{\sigma_t}\right)^{s'_i}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y}_{\text{(a)}}. \tag{16}$$

Note that $\max_{s:\ \sum s_i\leq k}\{\sum_{s'\in B_s}c_{s'}\}$ is bounded by a constant that only depends on $k$. Thus we focus on the evaluation of (a). When $t\leq 1$, (a) in (16) can be bounded by $\mathcal{O}(1/m_t^d)\simeq\mathcal{O}(1)$ (we hide dependency on $\sum_{i=1}^d s_i' \le k$ and $C_f$). This is because $m_t \simeq 1$ and $f(y) \le C_f$. On the other hand, when $t \ge 1$, $\sigma_t \gtrsim 1$ holds, we can bound (a) by $\mathcal{O}(1)$ by noting that $f(y) \neq 0$ only for $y \in [-1,1]^d$. Now, the first statement (12) has been proven.

We then consider $\nabla \log p_t(x)$ and its derivatives. We can focus on $[\nabla \log p_t(x)]_1$, and all the other coordinates of the score are bounded in the same way. Let $g_2(x) = \sigma_t[\nabla p_t(x)]_1 = -\int \frac{x_1 - m_t y_1}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y$, and define $g_2^{(s)}$ in the same way as that for $g_1^{(s)}$.

We can see that

$$[\nabla \log p_t(x)]_1 = \frac{1}{\sigma_t} \cdot \frac{g_2(x)}{g_1(x)}, \quad [\partial_{x_i} \nabla \log p_t(x)]_1 = \frac{1}{\sigma_t} \cdot \frac{\partial_{x_i} g_2(x)}{g_1(x)} - \frac{1}{\sigma_t} \cdot \frac{g_2(x)(\partial_{x_i} g_1(x))}{g_1^2(x)}. \tag{17}$$

Moreover,

$$\frac{g_2(x)}{g_1(x)} = \frac{-\int \frac{x_1 - m_t y_1}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}, \tag{18}$$

$$\frac{\partial_{x_i} g_1(x)}{g_1(x)} = \frac{1}{\sigma_t} \cdot \frac{-\int \frac{x_i - m_t y_i}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}, \tag{19}$$

$$\frac{\partial_{x_i} g_2(x)}{g_1(x)} = -\frac{1}{\sigma_t} \cdot \frac{\int \frac{\mathbb{1}[i=1] - \frac{x_1 - m_t y_1}{\sigma_t} \frac{x_i - m_t y_i}{\sigma_t}}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}. \tag{20}$$

In order to bound them, we consider the following quantity with $\sum_{i=1}^d s_i \le 2$. Also, let $\varepsilon$ be a scaler value specified later, with which we assume $p_t(x) \ge \varepsilon$ holds for the moment.

$$\frac{\int \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{\sigma_t^2}\right) \mathrm{d}y} \tag{21}$$

According to Lemma J.10, we have that

$$\left| \int_{A^x} \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma_t^2}\right) \mathrm{d}y \right.$$
$$\left. - \int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma_t^2}\right) \mathrm{d}y \right| \le \frac{\varepsilon}{2}.$$

where $A^x = \prod_{i=1}^d a_i^x$ with $a_i^x = [\frac{x_1}{m_t} - \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log 2\varepsilon^{-1}}, \frac{x_1}{m_t} + \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log 2\varepsilon^{-1}}]$. Note that $C_{\mathrm{f}}$ only depends on $\sum_{i=1}^d s_i$, $d$, and $C_f$.

Therefore, when $p_t(x) = g_1(x) \ge \varepsilon$,

$$(21) \le \frac{2\int \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int_{A^x} \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{\sigma_t^2}\right) \mathrm{d}y}$$

$$\leq \frac{2 \int_{A^x} \prod_{i=1}^d \left( \frac{x_i - m_t y_i}{\sigma_t} \right)^{s_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y}{\int_{A^x} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left( -\frac{\|x - m_t y\|^2}{\sigma_t^2} \right) \mathrm{d}y} + \frac{2 \cdot \frac{\varepsilon}{2}}{\varepsilon}$$

(note that the denominator is larger than $\varepsilon$)

$$\leq 2 \max_{y \in A_x} \left[ \prod_{i=1}^d \left( \frac{x_i - m_t y_i}{\sigma_t} \right)^{s_i} \right] + 1$$

$$\leq 2 \left( C_{\mathrm{f}}^2 \log \varepsilon^{-1} \right)^{(\sum_{i=1}^d s_i)/2} + 1. \tag{22}$$

Applying this bound to (18), (19), and (20), $\frac{g_2(x)}{g_1(x)}$, $\frac{\partial_{x_i} g_1(x)}{g_1(x)}$, and $\frac{\partial_{x_i} g_2(x)}{g_1(x)}$ are bounded by

$$\log^{1/2} \varepsilon^{-1}, \frac{\log^{1/2} \varepsilon^{-1}}{\sigma_t}, \text{ and } \frac{\log \varepsilon^{-1}}{\sigma_t},$$

up to constant factors, respectively. Finally, we apply this to (17) and obtain that

$$\|\nabla \log p_t(x)\| \lesssim \frac{\log^{1/2} \varepsilon^{-1}}{\sigma_t} \text{ and, } \|\partial_{x_i} \nabla \log p_t(x)\| \lesssim \frac{\log \varepsilon^{-1}}{\sigma_t^2}.$$

Now we replace $\varepsilon$ with a specific value. Remember that $\varepsilon$ should satisfy $\varepsilon \leq p_t(x)$. According to Lemma C.2, we have $C_{\mathrm{a},2}^{-1} \exp\left( -\frac{d(\|x\|_\infty - m_t)_+^2}{\sigma_t^2} \right) \leq p_t(x)$, which yields that

$$\|\nabla \log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t} \cdot \frac{(\|x\|_\infty - m_t)_+}{\sigma_t} \vee 1, \text{ and } \quad \|\partial_{x_i} \nabla \log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t^2} \left( \frac{(\|x\|_\infty - m_t)_+^2}{\sigma_t^2} \vee 1 \right),$$

with $C_{\mathrm{a},3}$ depending on $k$, $d$ and $C_f$. Thus, we obtain (13) and (14).

Finally, we consider $\partial_t \nabla \log p_t(x)$.

$$\partial_t [\nabla \log p_t(x)]_1 = \partial_t \left( \frac{1}{\sigma_t} \cdot \frac{g_2(x)}{g_1(x)} \right) = \left( \partial_t \frac{1}{\sigma_t} \right) \frac{g_2(x)}{g_1(x)} - \frac{1}{\sigma_t} \cdot \frac{(\partial_t g_1(x))}{g_1(x)} \cdot \frac{g_2(x)}{g_1(x)} + \frac{1}{\sigma_t} \cdot \frac{\partial_t g_2(x)}{g_1(x)}$$

$$= \frac{(-\partial_t \sigma_t)}{\sigma_t} [\nabla \log p_t(x)]_1$$

$$- \frac{1}{\sigma_t} \cdot \frac{\int A_1(y) f(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y}{\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y} \cdot [\nabla \log p_t(x)]_1$$

$$+ \frac{1}{\sigma_t} \cdot \frac{\int A_2(y) f(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y}{\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y}, \tag{23}$$

where

$$A_1(y) = \frac{-d(\partial_t \sigma_t) \sigma_t^{-1} + \|x - m_t y\|^2 (\partial_t \sigma_t) \sigma_t^{-3} - (\partial_t m_t) y^\top (m_t y - x) \sigma_t^{-2}}{\sigma_t^d (2\pi)^{\frac{d}{2}}},$$

$$A_2(y) = \frac{(\partial_t m_t) y_1 + (x_1 - m_t y_1)((d+1)(\partial_t \sigma_t) \sigma_t^{-1} - \|x - m_t y\|^2 (\nabla_t \sigma_t) \sigma_t^{-3} + (\partial_t m_t) y^\top (m_t y - x) \sigma_t^{-2})}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}}.$$

By carefully decomposing (23) into the sum of (21), and then applying (22) and Lemma C.2, we have the final bound (15). □

Now, based on Lemma C.3 we show that we only need to approximate $\nabla \log p_t(x)$ on some bounded region and on $x$ where $p_t(x)$ is not too small.

**Lemma C.4** (Error bounds due to clipping operations)**.** *Let $t \geq \underline{T}$. There exists a constant $C_{\mathrm{a},4}$ depending on $\mathrm{d}$ and $C_f$, we have*

$$\int_{\|x\|_\infty \geq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}\underline{T}^{-1}}} p_t(x)\|\nabla \log p_t(x)\|^2 \mathrm{d}x \leq \varepsilon, \tag{24}$$

$$\int_{\|x\|_\infty \geq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}\underline{T}^{-1}}} p_t(x) \mathrm{d}x \leq \varepsilon \tag{25}$$

*for all $t \geq \underline{T}$.*

*Moreover, there exists a constant $C_{\mathrm{a},5}$ depending on $\mathrm{d}$ and $C_f$ and, for $x$ such that $\|x\|_\infty \leq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}}$, we have*

$$\|\nabla \log p_t(x)\| \leq \frac{C_{\mathrm{a},5}}{\sigma_t}\sqrt{\log \varepsilon^{-1}}.$$

*Therefore,*

$$\int_{\|x\|_\infty \leq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}\underline{T}^{-1}}} p_t(x)\mathbb{1}[p_t(x) \leq \varepsilon]\|\nabla \log p_t(x)\|^2 \mathrm{d}x \leq \frac{C_{\mathrm{a},5}\varepsilon}{\sigma_t^2} \cdot \log^{\frac{d+2}{2}}(\varepsilon^{-1}\underline{T}^{-1}), \tag{26}$$

$$\int_{\|x\|_\infty \leq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}\underline{T}^{-1}}} p_t(x)\mathbb{1}[p_t(x) \leq \varepsilon] \mathrm{d}x \leq C_{\mathrm{a},5}\varepsilon \cdot \log^{\frac{d}{2}}(\varepsilon^{-1}\underline{T}^{-1}). \tag{27}$$

*Proof.* According to Lemma C.2 and Lemma C.3,

$$\begin{aligned} p_t(x)\|\nabla \log p_t(x)\|^2 &\leq C_{\mathrm{a},2}\exp\left(-\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2}\right) \cdot \frac{C_{\mathrm{a},3}^2}{\sigma_t^2}\frac{(\|x\|_\infty - m_t)_+^2}{\sigma_t^2} \\ &\leq \frac{C_{\mathrm{a},2}C_{\mathrm{a},3}^2}{\sigma_t^2}\exp\left(-\frac{r^2}{2}\right)r^2, \end{aligned}$$

where we let $r := (\|x\|_\infty - m_t)_+/\sigma_t$. Then,

$$\begin{aligned} &\int_{\|x\|_\infty \geq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}}} p_t(x)\|\nabla \log p_t(x)\|^2 \mathrm{d}x \\ &\leq \int_{C_{\mathrm{a},4}\sqrt{\log \varepsilon^{-1}}}^\infty \frac{C_{\mathrm{a},2}C_{\mathrm{a},3}^2}{\sigma_t}\exp\left(-\frac{r^2}{2}\right)r^2(d-1)(\sigma_t r + m_t)^{d-1}\mathrm{d}r \\ &\lesssim \frac{1}{\sigma_t}\varepsilon\log^{d/2}\varepsilon^{-1}. \end{aligned}$$

We can make sure the final inequality by integration by parts. Because $\sigma_t \gtrsim \sqrt{\underline{T}}$, if we take $\varepsilon' = \sqrt{\underline{T}} \cdot \varepsilon^2$ then we have that $\frac{1}{\sigma_t}\varepsilon'\log^{d/2}((\varepsilon')^{-1}) \lesssim \varepsilon$. Therefore, replacing $\varepsilon$ with $\varepsilon'$ and adjusting $C_{\mathrm{a},4}$ yield the bound (24).

In the same way,

$$\begin{aligned} \int_{\|x\|_\infty \geq m + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}}} p_t(x)\mathrm{d}x &\leq \int_{C_{\mathrm{a},4}\sqrt{\log \varepsilon^{-1}}}^\infty C_{\mathrm{a},2}\sigma_t\exp\left(-\frac{r^2}{2}\right)(d-1)(\sigma_t r + m)^{d-1}\mathrm{d}r \\ &\lesssim \sigma_t\varepsilon\log^{(d-2)/2}\varepsilon^{-1}, \end{aligned}$$

which yields (25).

We then consider the second part of the lemma. Eq. (25) is a direct corollary of Lemma C.3: for $x$ with $\|x\|_\infty \leq m_t + C_{\mathrm{a},5}\sigma_t\sqrt{\log \varepsilon^{-1}}$

$$\|\nabla \log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t} \cdot C_{\mathrm{a},4}\sqrt{\log \varepsilon^{-1}} \leq \frac{C_{\mathrm{a},5}}{\sigma_t}\sqrt{\log \varepsilon^{-1}}. \quad \text{(by taking } C_{\mathrm{a},5} \text{ larger than } C_{\mathrm{a},3}C_{\mathrm{a},4}.\text{)}$$

Using this, we have

$$\begin{aligned} &\int_{\|x\|_\infty \leq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}}} p_t(x)\mathbb{1}[p_t(x) \leq \varepsilon]\|\nabla \log p_t(x)\|^2 \mathrm{d}x \\ &\lesssim \varepsilon \cdot \frac{C_{\mathrm{a},4}^2}{\sigma_t^2}\log \varepsilon^{-1} \cdot (m_t + C_{\mathrm{a},5}\sigma_t\sqrt{\log \varepsilon^{-1}})^d. \end{aligned}$$

Adjusting $C_{\mathrm{a},4}, C_{\mathrm{a},5}$ and resetting $\varepsilon$ yields (26). Eq. (27) follows in the same way. □

# D APPROXIMATION OF THE SCORE FUNCTION

This section corresponds to Section 3.1. In this section, we analyze approximation error for the (ideal) score matching loss minimization. We construct a neural network that approximates $\nabla \log p_t(x)$ and bound the approximation error over different time $t$. Throughout this section, we take a sufficiently large $N$ as a parameter that determines the size of the neural network, and $\underline{T} = \mathrm{poly}(N^{-1})$ and $\overline{T} = \mathcal{O}(\log N)$.

## D.1 DETAILED PROOF SKETCH

Here we provide detailed proof sketch of Theorem 3.1.

**Approximation via the diffused B-spline Basis** We consider the approximation for $t \ll 1$. First remind the B-spline basis decomposition of the Besov functions (DeVore & Popov, 1988; Suzuki, 2018). Let $\mathcal{N}(x) = 1\ (x \in [0,1]), 0$ (otherwise). The *cardinal B-spline of order* $l$ is defined by $\mathcal{N}_l(x) = \underbrace{\mathcal{N} * \mathcal{N} * \cdots * \mathcal{N}}_{l+1 \text{ times convolution}}(x)$, where $(f * g)(x) = \int f(x-t)g(t)\mathrm{d}t$. Then, the *tensor product B-spline basis* in $\mathbb{R}^d$ is defined for $k \in \mathbb{N}^d$ and $j \in \mathbb{Z}^d$ as $M^d_{k,j}(x) = \prod_{i=1}^d \mathcal{N}(2^{k_i}x - j_i)$. It is known that a function $f$ in the Besov space is approximated by a super-position of $M^d_{k,j}(x)$ as $f_N = \sum_{(k,j)} \alpha_{(k,j)} M^d_{k,j}(x)$.

**Lemma D.1** (Informal version of Lemma J.13; Suzuki (2018))**.** *For any $p_0 \in U(B^s_{p,q})$, there exists a super-position $f_N$ of $N$ tensor-product B-spline bases satisfying*

$$\|p_0 - f_N\|_{L^2} \lesssim N^{-s/d}\|f\|_{B^s_{p,q}}.$$

Inspired by this, we introduce our basis decomposition. Because of $X_t|X_0 \sim \mathcal{N}(m_t X_0, \sigma_t)$, we can write $p_t$ as

$$p_t(x) = \int p_0(y) \underbrace{\frac{1}{\sigma^d(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)}_{=:K_t(x|y)} \mathrm{d}y.$$

Because the transition kernel $K_t(x|y)$ linearly applies to $p_0$ and $p_0$ is approximated by $f_N = \sum_{(k,j)} \alpha_{(k,j)} M^d_{k,j}(x)$, we come up with the following approximation of $p_t$:

$$p_t(x) \approx \sum_{(k,j)} \alpha_{(k,j)} \underbrace{\int M^d_{k,j}(y) K(x|y) \mathrm{d}y}_{=:E_{k,j}(x,t)}.$$

Moreover, $E_{k,j}$ is further decomposed as

$$E_{k,j}(x,t) = \prod_{i=1}^d \underbrace{\int \frac{\mathcal{N}(2^{k_i}x_i - j_i)}{\sigma_t\sqrt{2\pi}} \exp(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}) \mathrm{d}x_i}_{=:\mathcal{D}_{k,j}(x_i,t)}.$$

We name $\mathcal{D}_{k,j}$ as the *diffused B-spline basis* and $E_{k,j}$ as the *tensor product diffused B-spline basis*. We show that there exists a neural network that approximates $\mathcal{D}_{k,j}$ and $E_{k,j}$ very efficiently. Our construction then goes as follows. We construct networks approximating $m_t$ and $\sigma_t$.

**Lemma D.2** (See also Lemma D.6)**.** *Under Assumption 2.4, there exists neural networks $\phi_m(t), \phi_\sigma(t) \in \Phi(L, W, B, S)$ that approximates $m_t$ and $\sigma_t$ up to $\varepsilon$ for all $t \geq 0$, where $L = \mathcal{O}(\log^2(\varepsilon^{-1})), \|W\|_\infty = \mathcal{O}(\log^3(\varepsilon^{-1})), S = \mathcal{O}(\log^4(\varepsilon^{-1}))$, and $B = \exp(\mathcal{O}(\log^2(\varepsilon^{-1})))$.*

Next we clip the integral interval of $\mathcal{D}_{k,j}$ and approximate the integrand by a rational function of $(x, m_t, \sigma_t)$. Then the following is obtained as an informal version of Lemma D.8.

**Lemma D.3.** *For $\varepsilon > 0$, there exists a neural network $\phi_{\mathrm{TDB}}\colon \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$ that satisfies $\|\phi_{\mathrm{TDB}}(x,t) - E_{k,j}(x,t)\|_\infty \le \varepsilon$. Here, $\phi_{\mathrm{TDB}} \in \Phi(L,W,S,B)$ with $L = \mathcal{O}(\log^4(\varepsilon^{-1}))$, $\|W\|_\infty = \mathcal{O}(\log^6(\varepsilon^{-1}))$, $S = \mathcal{O}(\log^8(\varepsilon^{-1}))$, $B = \mathcal{O}(\exp(\mathcal{O}(\log^4(\varepsilon^{-1}))))$.*

Here $\phi_{\mathrm{TDB}}$ approximates $E_{k,j}(x,t)$ given $(x, m_t, \sigma_t)$. Then we use $\phi_{\mathrm{TDB}}(x, \phi_m(t), \phi_\sigma(t))$ as the approximation of $E_{k,j}(x,t)$, and $p_t(x)$ is finally approximated. Similar approximation can also be made for $\nabla p_t(x)$, and the score is finally approximated together with $\nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)}$ and we obtain the bound as in Theorem 3.1.

We remark that the bounds on the network class parameters given above are slightly larger than that for the B-spline basis (Suzuki (2018)) because approximating integrals and exponential functions (Appendix J.3) and rational functions (Appendix J.2) is more difficult than realizing the B-spline basis via polynomials. Especially, $B = \exp(O(\log^4 N))$ comes from approximation of exponential functions. Because $B$ affects the generalization error only in a $\log B$ term (see Lemma E.2), this super-polynomial scaling does not much affects the the final estimation errors.

We also remark that, in this construction, the approximation error for $\nabla p_t(x)$ is amplified in the area where $p_t(x) \ll 1$. This is why we need the higher-order smoothness of $p_0$ in the area with distance less than $\tilde{\mathcal{O}}(\sqrt{t})$ from the edge of the support (Assumption 2.4). This approach is used during $t \in [\underline{T}, 3N^{-\frac{2-\delta}{d}}]$, and it suffices to set $a_0$ to $a_0 = N^{-\frac{1-\delta}{d}}$.

**Utilizing the smoothness induced by the noise** The above approach enables approximation of the score in $t \ll 1$, when the score is highly non-smooth, by using the structure of $p_0$. On the other hand, after a certain period of time, the shape of $p_t$ gets almost like a Gaussian, very smooth and easy to be approximated. This paragraph extends the previous approach and gives an alternative approximation based on the smoothness induced by the noise, yielding a tighter bound.

We begin with evaluating the derivatives of $p_t$ w.r.t. $t$.

**Lemma D.4.** *For any $k \in \mathbb{Z}_+$, there exists a constant $C_{\mathrm{a}}$ depending only on $k$, $d$, and $C_f$ such that*

$$\left|\partial_{x_{i_1}} \partial_{x_{i_2}} \cdots \partial_{x_{i_k}} p_t(x)\right| \le \frac{C_{\mathrm{a}}}{\sigma_t^k}.$$

We have that $\|p_{t_*}\|_{W_p^k} = \mathcal{O}({t_*}^{-\frac{k}{2}})$ for $t_* > 0$ from this, and that $W_p^k \hookrightarrow B_{p,\infty}^k$. For $t > t_*$, consider $p_t$ as the diffused distribution from $p_{t_*}$, instead of $p_0$. We can show that $\nabla \log p_t$ can be approximated with a neural network with the size $N'$, with an $L^2$ error of $\mathcal{O}\left(\frac{N'^{-2k/d}}{\sigma_t^2} \cdot t_*^{-k}\right)$. If $N'$ and $k$ are sufficiently large, this is tighter than the previous bound of $\frac{N^{-\frac{2s}{d}}}{\sigma_t^2}$. This argument is formalized as follows. In Appendix D, this is presented as Lemma D.12.

**Lemma D.5.** *Let $N \gg 1$ and $N' \ge t_*^{-d/2} N^{\delta/2}$. Suppose $t_* \ge N^{-(2-\delta)/d}$. Then there exists a neural network $\phi'_{\mathrm{score}} \in \Phi(L,W,S,B)$ that satisfies*

$$\int_x p_t(x) \|\phi'_{\mathrm{score}}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^2}$$

*for $t \in [2t_*, \overline{T}]$. Specifically, $L = \mathcal{O}(\log^4(N))$, $\|W\|_\infty = \mathcal{O}(N)$, $S = \mathcal{O}(N')$, and $B = \exp(\mathcal{O}(\log^4 N))$.*

Setting $t_* = N^{-\frac{2-\delta}{d}}$ and $N' = N$ in this lemma, we obtain the bound in Theorem 3.1 after $t \gtrsim t_*$, without Assumption 2.4. Moreover, further exploiting this lemma later plays an important role for achieving the minimax optimal estimation rate in the $W_1$ distance.

### D.2 Approximation of $m_t$ and $\sigma_t$

We begin with construction of sub-networks that approximate $m_t$ and $\sigma_t$. In addition to the true data distribution $p_0(x)$, the score $\nabla \log p_t(x)$ also depends on $m_t$ and $\sigma_t$. Indeed, in our construction, each diffused B-spline basis is approximated as a rational function of $x$, $m_t$ and $\sigma_t$. Here, $m_t$ and

$\sigma_t$ are as important as $x$, because we use exponentiation of $m_t$ and $\sigma_t$, as well as that of $x$, while exact values of $m_t$ and $\sigma_t$ are unavailable. In other words, because approximation errors of $m_t$ and $\sigma_t$ are amplified via such exponentiation, approximating $m_t$ and $\sigma_t$ with high accuracy is necessary for obtaining tight bounds. Therefore, in this subsection, we construct sub-networks for efficient approximation of $m_t$ and $\sigma_t$. The following is the formal version of Lemma D.2.

**Lemma D.6.** *Let $0 < \varepsilon < \frac{1}{2}$. Then, there exists a neural network $\phi_m(t) \in \Phi(L, W, B, S)$ that approximates $m_t$ for all $t \geq 0$, within the additive error of $\varepsilon$, where $L = \mathcal{O}(\log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log \varepsilon^{-1}), S = \mathcal{O}(\log^2 \varepsilon^{-1})$, and $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$.*

*Also, there exists a neural network $\phi_\sigma(t) \in \Phi(L, W, B, S)$ that approximates $\sigma_t$ for all $t \geq \varepsilon$, within the additive error of $\varepsilon$, where $L \leq \mathcal{O}(\log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1}), S = \mathcal{O}(\log^4 \varepsilon^{-1})$, and $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$.*

*Proof.* First we consider $m_t = \exp(-\int_0^t \beta_s \mathrm{d}s)$. Since $\beta \geq \underline{\beta}$, $\int_0^t \beta_s \mathrm{d}s \geq \log 4\varepsilon^{-1}$ for all $t \geq A := \log 4\varepsilon^{-1}/\underline{\beta}$. We limit ourselves within $[0, A]$. Then, from Assumption 2.3, we can expand $\beta_s$ as $\beta_s = \sum_{i=0}^{k-1} \frac{\beta^{(i)}}{i!} s^i + \frac{\beta^{(k)}}{k!}(\theta s)^k$ with $|\beta^{(i)}| \leq 1$ and $0 < \theta < 1$, and therefore we obtain that

$$\left| \int_0^t \beta_s \mathrm{d}s - \int_0^t \sum_{i=1}^{k-1} \frac{\beta^{(i)}}{i!} s^i \mathrm{d}s \right| \leq \frac{|\beta^{(k)}| A^{k+1}}{(k+1)!} \leq \frac{A^{k+1}}{(k+1)!}.$$

We take $k = \max\{2eA, \lceil \log_2 4\varepsilon^{-1} \rceil\} - 1$ so that we have $\frac{A^{k+1}}{(k+1)!} \leq \left(\frac{eA}{k+1}\right)^{k+1} \leq \frac{\varepsilon}{4}$. $\int_0^t \sum_{i=1}^{k-1} \frac{\beta^{(i)}}{i!} s^i = \sum_{i=1}^{k-1} \frac{\beta^{(i)}}{(i+1)!} t^{i+1}$ can be realized with an additive error up to $\frac{\varepsilon}{4}$ by the neural network with $L = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}) = \mathcal{O}(\log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(A + \log \varepsilon^{-1}) = \mathcal{O}(\log \varepsilon^{-1}), S = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}) = \mathcal{O}(\log^2 \varepsilon^{-1}), B = \exp(\log^2 \mathcal{O}(A + \log \varepsilon^{-1})) = \mathcal{O}(\log^2 \varepsilon^{-1})$, using Lemmas J.3 and J.6. From the definition of $A$, we can easily check that $e^{-A} \leq \frac{\varepsilon}{4}$ holds. We clip the input with $[0, A]$ to obtain the neural network $\phi_1$, which approximates $\int_0^t \beta_s \mathrm{d}s$ with an additive error of $\frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}$ for $x \in [0, A]$, and satisfies $|\phi_1(x)| = |\phi_1(A)|$ for all $x \geq A$.

Then we apply Lemma J.12 with $\varepsilon = \frac{\varepsilon}{4}$. Then we obtain the neural network $\phi_m$ of the desired size, which approximates $m_t = \exp(-\int_0^t \beta_s \mathrm{d}s)$ with an additive error of $\frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \frac{3\varepsilon}{4}$ for $x \in [0, A]$ and $|\phi_m(x) - e^{-x}| \leq |\phi_m(x) - \phi_m(A)| + |\phi_m(A) - e^{-A}| + |e^{-A} - e^{-x}| \leq 0 + \frac{3\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon$ for $x \geq A$.

Similarly, we can approximate $\sigma^2 = 1 - \exp(-2\int_0^t \beta_s \mathrm{d}s)$ with an additive error of $\mathcal{O}(\varepsilon^{1.5})$ using a neural network with $L = \mathcal{O}(\log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log \varepsilon^{-1}), S = \mathcal{O}(\log^2 \varepsilon^{-1}), B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$. Since $t \geq \varepsilon$, we have $\sigma_t^2 = 1 - \exp(-2\int_0^t \beta_s \mathrm{d}s) \geq c\varepsilon$ for some constant $c$ depending on $\underline{\beta}$. Then, we apply Lemma J.9 with $\varepsilon = c\varepsilon$ and finally obtain a neural network $\phi_\sigma(t)$ that approximates $\sigma_t$ with an additive error of $c\varepsilon + \frac{\varepsilon^{1.5}}{\sqrt{c\varepsilon}} = \mathcal{O}(\varepsilon)$, with $L = \mathcal{O}(\log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1}), S = \mathcal{O}(\log^4 \varepsilon^{-1})$, and $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$. Adjusting hidden constants can make the approximation error smaller than $\varepsilon$, and concludes the proof. □

### D.3 Approximation via the diffused B-spline basis

This subsection introduces the approximation via the *diffused B-spline basis* and the *tensor-product diffused B-spline basis*, which enable us to approximate the score $\nabla \log p_t(x)$ in the space of $\mathbb{R}^d \times [\underline{T}, \overline{T}]$. Although we consider the function approximation in a $(d+1)$-dimensional space, the obtained rate (Theorem 3.1) is the typical one for a $d$-dimensional space. This is because our basis decomposition can reflect the structure of $p_0$ for $t > 0$. Before beginning the formal proof, we provide extended proof outline about the approximation via the diffusion B-spline basis and tensor-product diffused B-spline basis, which is more detailed than that in Section 3.

Remind that the cardinal B-spline basis of order $l$ can be written as

$$\mathcal{N}_m(x) = \frac{1}{l!} \mathbb{1}[0 \leq x \leq l+1] \sum_{l'=0}^{l} (-1)^j {}_{l+1}\mathrm{C}_{l'} (x - l')_+^l$$

(see Eq. (4.28) of Mhaskar & Micchelli (1992) for example) and the function in the Besov space can be approximated by a sum of $M_{k,j}^d(x)$

$$M_{k,j}^d(x) = \prod_{i=1}^{d} \mathcal{N}_m(2^{k_i}x_i - j_i)$$

where $k \in \mathbb{Z}_+^d$ and $j \in \mathbb{Z}^d$.

Therefore, the denominator and numerator of the score

$$\nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)} = -\frac{1}{\sigma_t} \cdot \frac{\int \frac{x-m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}$$

are decomposed into the sum of

$$E_{k,j}^{(1)}(x,t) := \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} \mathbb{1}[\|y\|_\infty \le C_{\mathrm{b},1}] M_{k,j}^d(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \tag{28}$$

and

$$E_{k,j}^{(2)}(x,t) := \int \frac{x-m_t y}{\sigma^{d+1}(2\pi)^{\frac{d}{2}}} \mathbb{1}[\|y\|_\infty \le C_{\mathrm{b},1}] M_{k,j}^d(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y, \tag{29}$$

respectively. This corresponds to what we called the tensor-product diffused B-spline basis in Section 3. Here $E_{k,j}^{(1)}(x,t)$ is the same as $E_{k,j}(x,t)$ in Section 3, except for the term of $\mathbb{1}[\|y\|_\infty \le C_{\mathrm{b},1}]$. Note that $C_{\mathrm{b},1}$ be a scaler value adjusted later. We then approximate each of the denominator and numerator of $\nabla \log p_t(x)$ combining sub-networks that approximates each $E_{k,j}^{(1)}(x,t)$ or $E_{k,j}^{(2)}(x,t)$.

Here we briefly remark why $\mathbb{1}[\|y\|_\infty \le C_{\mathrm{b},1}]$ appears. Let us assume $C_{\mathrm{b},1} = 1$ and approximate $p_t(x)$ based on basis decomposition of $p_0(x)$, although later we need to consider other situations. If we use basis decomposition as $p_0(x) \approx f_N(x) = \sum M_{k,j}^d(x)$, existing results such as Lemma J.13 only assure that the approximation is valid within $[-1,1]^d$ and do not guarantee anything outside the region. This might harm the approximation accuracy when we integrate the approximation of $p_t(x)$ over all $\mathbb{R}^d$. Therefore, we need to force $f_N(x) = 0$ if $\|x\|_\infty > 1$ by the indicator function.

From now, we realize the (modified) tensor-product diffused B-spline basis with neural networks. We take $E_{k,j}^{(1)}$ as an example, and the procedures for $E_{k,j}^{(2)}$ is essentially the same. Remind that in Section 3 we decomposed $E_{k,j}$ into the product of the diffused B-spline basis:

$$\mathcal{D}_{k,j}(x_i,t) = \int \frac{\mathcal{N}(2^k x_i - j_i)}{\sigma_t\sqrt{2\pi}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}x_i.$$

Although the way we proceed is essentially the same as that in Section 3, here, more formally, we first truncate the integral intervals. We clip the integral interval as

$$\begin{aligned}
E_{k,j}^{(1)}(x,t) &\fallingdotseq \int_{y\in A^{x,t}} \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} \mathbb{1}[\|y\|_\infty \le C_{\mathrm{b},1}] M_{k,j}^d(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&= \prod_{i=1}^{d}\left(\sum_{l'=0}^{l+1} \frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!} \int_{y_i\in a_i^x} \frac{1}{\sigma_t(2\pi)^{\frac{1}{2}}} \mathbb{1}[|y_i| \le C_{\mathrm{b},1}]\mathbb{1}[0 \le 2^{k_i}y_i - j_i \le l+1]\right. \\
&\qquad\qquad \left. \times (2^k y_i - l' - j_i)_+^l \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i\right),
\end{aligned} \tag{30}$$

where $A^{x,t} = \prod_{i=1}^d a_i^{x,t}$ with $a_i^{x,t} = [\frac{x_i}{m_t} - \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}}, \frac{x_i}{m_t} + \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}}]$, $C_{\mathrm{f}} = \mathcal{O}(1)$, and $0 < \varepsilon < 1$. This clipping causes the error at most $\mathcal{O}(\varepsilon)$ according to Lemma J.10 and the observation $\mathbb{1}[\|y\|_\infty \le C_{\mathrm{b},1}] M_{k,j}^d(y) \le \left((l+1)^{l+1}2^{l+1}\right)^d$. In summary, owing to the fact that $M_{k,j}^d(x)$ is a product of univariate functions of $x_i$ $(i = 1, 2, \cdots, d)$, the integral over $\mathbb{R}^d$ is now decomposed into the integral with respect to only one variable over the bounded region, which is a truncated version of the diffused B-spline basis $\mathcal{D}_{k,j}$ introduced in Section 3.

We now begin the formal proof with the following lemma. We approximate

$$\int_{y_i \in a_i^{x,t}} \frac{1}{\sigma_t (2\pi)^{\frac{1}{2}}} \mathbb{1}[|y_i| \leq C_{\mathrm{b},1}] \mathbb{1}[0 \leq 2^k y_i - j_i \leq l+1](2^{k_i} y_i - l' - j_i)_+^l \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i \tag{31}$$

(remind (30)). Note that $\mathbb{1}[|y_i| \leq C_{\mathrm{b},1}]\mathbb{1}[0 \leq 2^k y_i - j_i \leq l+1] \equiv 0$ or $= \mathbb{1}[a \leq 2^k y_i \leq b]$ holds with $a, b$ satisfying

$$-C2^k - l \leq \min_i j_i \leq j_i \leq a < b \leq j_i + l + 1 \leq \max_i j_i + l + 1 \leq C2^k + l + 1, \tag{32}$$

if we assume $\mathrm{supp}(p_0) = [-C, C]^d$ (see Lemma J.13). Based on (32), (31) (if $\mathbb{1}[|y_i| \leq C_{\mathrm{b},1}]\mathbb{1}[0 \leq 2^k y_i - j_i \leq l+1](2^k y_i - l' - j_i)_+^l \not\equiv 0$) can alternatively written as

$$\int_{y_i \in a_i^{x,t}} \frac{1}{\sigma_t (2\pi)^{\frac{1}{2}}} \mathbb{1}[\underline{j} \leq 2^k y \leq \overline{j}](2^k y_i - j')^l \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i, \tag{33}$$
$$\text{with } \underline{j}, \overline{j}, j' \in \mathbb{R}, \quad \overline{j} - l - 1 \leq j' \leq \underline{j} \leq \overline{j}, \quad -C2^k - l \leq j', \underline{j}, \overline{j} \leq C2^k + l + 1.$$

In the following lemma, we consider the approximation of (33). We omit the subscript $i$ for the coordinates, for simple presentation. Also, $j'$ in (33) is denoted by $j$, because $j \in \mathbb{R}^d$ will not be used in the following lemma.

**Lemma D.7** (Approximation of the diffused B-spline basis)**.** *Let $j, k, l \in \mathbb{Z}$, $\underline{j}, \overline{j} \in \mathbb{R}$ satisfy $\overline{j} - l - 1 \leq j \leq \underline{j} \leq \overline{j}$, $-C2^k - l \leq j, \underline{j}, \overline{j} \leq C2^k + l + 1$, and $k, l \geq 0$. Assume that $|\sigma' - \sigma_t|, |m' - m_t| \leq \varepsilon_{\mathrm{error}}$, and take $\varepsilon$ from $0 < \varepsilon < \frac{1}{2}$ and $C > 0$ arbitrarily. Then, there exists a neural network $\phi_{\mathrm{dif},1}^{j,\overline{j},\underline{j},k} \in \Phi(L, W, S, B)$ with*

$$\begin{aligned} L &= \mathcal{O}(\log^4 \varepsilon^{-1} + \log^2 C + k), \\ \|W\|_\infty &= \mathcal{O}(\log^6 \varepsilon^{-1}), \\ S &= \mathcal{O}(\log^8 \varepsilon^{-1} + \log^2 C + k), \\ B &= \mathcal{O}(C^l 2^{kl}) + \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}. \end{aligned}$$

*such that*

$$\left| \phi_{\mathrm{dif},1}^{j,\overline{j},\underline{j},k}(x, \sigma', m') - \int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}} \frac{1}{\sqrt{2\pi}\sigma_t} \mathbb{1}[\underline{j} \leq 2^k y \leq \overline{j}](2^k y - j)^l \exp\left(-\frac{(x - m_t y)^2}{2\sigma_t^2}\right) \mathrm{d}y \right|$$
$$\leq \tilde{\mathcal{O}}(\varepsilon) + \varepsilon_{\mathrm{error}} C^{4l} 2^{k(4l+1)} \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}.$$

*holds for all $x$ in $-C \leq x \leq C$ and for all $t \geq \varepsilon$.*

*Also, with the same conditions, there exists a neural network $\phi_{\mathrm{dif},2}^{j,\overline{j},\underline{j},k} \in \Phi(L, W, S, B)$ with the same bounds on $L, \|W\|_\infty, S, B$ as above such that*

$$\left| \phi_{\mathrm{dif},2}^{j,\overline{j},\underline{j},k}(x, \sigma', m') - \int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}} \frac{[x - m_t y]_i}{\sqrt{2\pi}\sigma_t^2} \mathbb{1}[\underline{j} \leq 2^k y \leq \overline{j}](2^k y - j)^l \exp\left(-\frac{(x - m_t y)^2}{2\sigma_t^2}\right) \mathrm{d}y \right|$$
$$\leq \tilde{\mathcal{O}}(\varepsilon) + \varepsilon_{\mathrm{error}} C^{4l} 2^{k(4l+1)} \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}.$$

*holds for all $x$ in $-C \leq x \leq C$ and for all $t \geq \varepsilon$.*

*Furthermore, we can take these networks so that $\|\phi_{\mathrm{dif},1}^{j,\overline{j},\underline{j},k}\|_\infty$, $\|\phi_{\mathrm{dif},2}^{j,\overline{j},\underline{j},k}\|_\infty = \mathcal{O}(1)$ hold.*

*Proof.* Here we only consider $\phi_{\mathrm{dif},1}^{j,\overline{j},\underline{j},k}$, because the assertion for $\phi_{\mathrm{dif},2}^{j,\overline{j},\underline{j},k}$ essentially follows the argument for $\phi_{\mathrm{dif},1}^{j,\overline{j},\underline{j},k}$.

First, we approximate the exponential function within the closed interval, using polynomials of degree at most $\mathcal{O}(\log \varepsilon^{-1})$. Note that $\mathbb{1}[\underline{j} \le 2^k y \le \overline{j}](2^k y - j)^l$ is bounded by $(l+1)^l$, from the assumption of $\overline{j} - l - 1 \le j \le \underline{j} \le \overline{j}$. Therefore, according to Lemma J.11, there exists $S = \mathcal{O}(\log \varepsilon^{-1})$ and we have that

$$\left| \exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right) - \sum_{s=0}^{S-1} \frac{(-1)^s}{s!} \frac{(x-m_t y)^{2s}}{2^s \sigma_t^{2s}} \right| \le \varepsilon^2$$

for all $y \in [-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + x, \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + x]$. Then, we have that

$$\begin{aligned}
&\left| \int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}} \frac{1}{\sqrt{2\pi}\sigma_t} \mathbb{1}[\underline{j} \le 2^k y \le \overline{j}](2^k y - j)^l \exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right) \mathrm{d}y \right. \\
&\quad \left. - \int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}} \frac{1}{\sqrt{2\pi}\sigma_t} \mathbb{1}[\underline{j} \le 2^k y \le \overline{j}](2^k y - j)^l \left( \sum_{s=0}^{S-1} \frac{(-1)^s}{s!} \frac{(x-m_t y)^{2s}}{2^s \sigma_t^{2s}} \right) \mathrm{d}y \right| \\
&\le \max\left\{ \frac{2\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}}, (l+1) \right\} \cdot \frac{1}{\sqrt{2\pi\sigma_t^2}} (l+1)^l \cdot \varepsilon \lesssim \varepsilon \log^{\frac{1}{2}} \varepsilon^{-1}.
\end{aligned}$$

Here, $\frac{2\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}}$ comes from the length of the integral interval and $l+1$ comes from the interval where $\mathbb{1}[\underline{j} \le 2^k y \le \overline{j}] = 1$ holds.

Now all we need is to approximate the integral of polynomials over the closed interval:

$$\begin{aligned}
&\sum_{s=0}^{S-1} \int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}} \frac{1}{\sqrt{2\pi}\sigma_t} \mathbb{1}[\underline{j} \le 2^k y \le \overline{j}](2^k y - j)^l \cdot \frac{(-1)^s}{s!} \frac{(x-m_t y)^{2s}}{2^s \sigma_t^{2s}} \mathrm{d}y \\
&= \sum_{s=0}^{S-1} \sum_{l'=0}^{l} \frac{-(-1)^{s+l}}{\sqrt{2\pi} m_t^{l+1} s! 2^s} \\
&\cdot \left[ {}_l\mathrm{C}_{l'} (2^k \sigma_t)^{l'} (j m_t - 2^k x)^{l-l'} \int_{-C_{\mathrm{f}}\sqrt{\log \varepsilon^{-1}}}^{C_{\mathrm{f}}\sqrt{\log \varepsilon^{-1}}} \mathbb{1}\left[ \frac{x - m_t 2^{-k}\overline{j}}{\sigma_t} \le y \le \frac{x - m_t 2^{-k}\underline{j}}{\sigma_t} \right] y^{l'+2s} \mathrm{d}y \right] \\
&\qquad \left( \text{by resetting } y \leftarrow \frac{x - m_t y}{\sigma_t} \right) \\
&= \sum_{s=0}^{S-1} \sum_{l'=0}^{l} \frac{-(-1)^{s+l} {}_l\mathrm{C}_{l'} 2^{kl'} \sigma^{l'} (j m_t - 2^k x)^{l-l'}}{\sqrt{2\pi} m_t^{l+1} s! 2^s (l'+2s+1)} \\
&\quad \cdot \left[ \left( \min\left\{ C_{\mathrm{f}}\sqrt{\log \varepsilon^{-1}}, \max\left\{ \frac{x - m_t 2^{-k}\underline{j}}{\sigma_t}, -C_{\mathrm{f}}\sqrt{\log \varepsilon^{-1}} \right\} \right\} \right)^{l'+2s+1} \right. \\
&\qquad \left. - \left( \min\left\{ C_{\mathrm{f}}\sqrt{\log \varepsilon^{-1}}, \max\left\{ \frac{x - m_t 2^{-k}\overline{j}}{\sigma_t}, -C_{\mathrm{f}}\sqrt{\log \varepsilon^{-1}} \right\} \right\} \right)^{l'+2s+1} \right]. \quad (34)
\end{aligned}$$

We decompose (34) into the following sub-modules for convenience. We let

$$\begin{aligned}
f_1^{l',s}(x,\sigma,m) &= (\min\{C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1}), \max\{\frac{x - m 2^{-k}\underline{j}}{\sigma}, -C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1})\}\})^{l'+2s+1}, \\
f_2^{l',s}(x,\sigma,m) &= (\min\{C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1}), \max\{\frac{x - m 2^{-k}\overline{j}}{\sigma}, -C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1})\}\})^{l'+2s+1}, \\
f_3^{l',s}(x,\sigma,m) &= f_1^{l',s}(x,\sigma,m) - f_2^{l',s}(x,\sigma,m) \\
f_4^{l'}(x,m) &= (jm - 2^k x)^{l-l'}, \\
f_5^{l'}(\sigma) &= \sigma^{l'}, \\
f_6(m) &= m^{-(l+1)}, \\
f_7^{l',s}(x,\sigma,m) &= f_3^{l',s}(x,\sigma,m) f_4^{l'}(x,m) f_5^{l'}(\sigma) f_6(m).
\end{aligned}$$

They also depends on $j, \underline{j}, \overline{j}, k$, and $l$, but we omit the dependency on these variables for simple presentation. We take some $\varepsilon_1 > 0$, which is adjusted at the final part of the proof.

We first consider approximation of $f_1^{l',s}(x,\sigma,m)$. We realize this as

$$\begin{aligned}
&f_1^{l',s}(x,\sigma,m) \fallingdotseq \phi_1^{l',s}(x,\sigma,m)\\
&:= \phi_{\mathrm{mult}}(\cdot; l'+2s+1) \circ \phi_{\mathrm{clip}}(\cdot; -C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1}), -C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1})) \circ (\phi_{\mathrm{mult}}(x - m2^{-k}\underline{j}, \phi_{\mathrm{rec}}(\sigma))).
\end{aligned}$$

by setting $\varepsilon = \min\{\sigma_\varepsilon, \varepsilon_1\}$ in Corollary J.8 for $\phi_{\mathrm{rec}}$, $\varepsilon = \varepsilon_1, C = \max\{2C + l + 1, \sigma_\varepsilon^{-1}\} \geq \max\{|x| + m2^{-k}\underline{j}, \sigma_\varepsilon^{-1}\}$ in Lemma J.6 for the first $\phi_{\mathrm{mult}}$, $a = -C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1}), b = C_{\mathrm{f}} \log^{\frac{1}{2}}(\varepsilon^{-1})$ in Lemma J.4 for $\phi_{\mathrm{clip}}$, and $\varepsilon = \varepsilon_1, C = C_{\mathrm{f}} \log^{\frac{1}{2}}(2\varepsilon^{-1})$ in Lemma J.6 for the second $\phi_{\mathrm{mult}}$. Note that $\sigma_\varepsilon \simeq \sqrt{\varepsilon}$. Then, using Lemmas J.1, J.4, J.6 and J.7 the size of the network is at most

$$\begin{aligned}
&L = \mathcal{O}(\log^2 \varepsilon_1^{-1} + \log^2 \varepsilon^{-1} + \log^2 C),\\
&\|W\|_\infty = \mathcal{O}(\log^3 \varepsilon_1^{-1} + \log^3 \varepsilon^{-1}),\\
&S = \mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^4 \varepsilon^{-1} + \log^2 C),\\
&B = \mathcal{O}(\varepsilon_1^{-2} + C^2) + \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}.
\end{aligned} \tag{35}$$

Approximation error between $f_1^{l',s}(x,\sigma_t,m_t)$ and $\phi_1^{l',s}(x,\sigma',m')$ is bounded by

$$\begin{aligned}
&\varepsilon_1 + \mathcal{O}(\log \varepsilon^{-1})(C_{\mathrm{f}} \log^{\frac{1}{2}} \varepsilon^{-1})^{\mathcal{O}(\log \varepsilon^{-1})} \cdot (\varepsilon_1 + \max\{C + l + 2, \sigma_\varepsilon^{-1}\}^2 \cdot (\varepsilon_1 + \varepsilon_{\mathrm{error}}(\varepsilon_1^{-2} + \varepsilon^{-2})))\\
&= (\varepsilon_1 + \varepsilon_{\mathrm{error}}) \left(\log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1} + C^2\right).
\end{aligned}$$

$f_2^{l',s}(x,\sigma_t,m_t)$ is also approximated in the same way, and therefore aggregating $f_1^{l',s}(x,\sigma_t,m_t)$ and $f_2^{l',s}(x,\sigma_t,m_t)$ (by using Lemma J.3) yields that $f_3^{l',s}(x,\sigma_t,m_t)$ is approximated by $\phi_3^{l',s}(x,\sigma',m')$ with the error up to an additive error of $(\varepsilon_1 + \varepsilon_{\mathrm{error}}) \left(\log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1} + C^2\right)$ using a neural network with the same size as that of (35).

Next, we consider $f_4^{l'}(x, m_t)$. Since $2^k x = \mathcal{O}(C2^k)$ and $|jm_t - jm'| \leq \mathcal{O}(C2^k \varepsilon_{\mathrm{error}})$, we approximate $f_4^{l'}(x, m_t)$ with a neural network $\phi_4^{l'}(x, m') \in \Phi(L, W, S, B)$, where $L, \|W\|_\infty, S, B$ are evaluated by Lemmas J.1 and J.6 (setting $\varepsilon = \varepsilon_1, C = \mathcal{O}(C2^k)$) as

$$L = \mathcal{O}(\log \varepsilon_1^{-1} + k \log C), \quad W = \mathcal{O}(1), \quad S = \mathcal{O}(\log \varepsilon_1^{-1} + k \log C), \quad B = \mathcal{O}(C^l 2^{kl}).$$

Approximation error between $f_4^{l'}(x, m_t)$ and $\phi_4^{l'}(x, m')$ is bounded as $\varepsilon_1 + \mathcal{O}(C^l 2^{kl})\varepsilon_{\mathrm{error}}$, using Lemma J.6.

The arguments for $f_5^{l'}(\sigma)$ and $f_6(m)$ are just setting appropriate parameters in Lemma J.6 and Corollary J.8, respectively. For $f_5^{l'}(\sigma_t)$, there exists a neural network $\phi_5^{l'}(\sigma')$ with $L = \mathcal{O}(\log \varepsilon_1^{-1}), \|W\|_\infty = 48l, S = \mathcal{O}(\log \varepsilon_1^{-1}), B = 1$ and the approximation error between $f_5^{l'}(\sigma)$ and $\phi_5^{l'}(\sigma')$ is bounded by $\varepsilon_1 + l\varepsilon_{\mathrm{error}}$, by setting $d = l'(\leq l), \varepsilon = \varepsilon_1$ in Lemma J.6. For $f_6(m_t)$, there exists a neural network $\phi_6(m')$ with $L = \mathcal{O}(\log^2 \varepsilon_1^{-1} + \log^2 m_\varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log^3 \varepsilon_1^{-1} + \log^3 m_\varepsilon^{-1}), S = \mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^4 m_\varepsilon^{-1}), B = \mathcal{O}(\varepsilon_1^{-l-1} + \underline{m}^{-l-1})$ and the approximation error between $f_6(m_t)$ and $\phi_6(m')$ is bounded by $\varepsilon_1 + (l+1)\varepsilon_1^{-l-2}\varepsilon_{\mathrm{error}} + (l+1)m_\varepsilon^{-l-2}\varepsilon_{\mathrm{error}}$, by setting $d = l+1, \varepsilon = \min\{\varepsilon_1, m_\varepsilon\}$ in Corollary J.8. Note that $m_\varepsilon \gtrsim 1$.

Therefore, Lemma J.6 with $\varepsilon = \varepsilon_1$ yields that there exists a neural network $\phi_7^{l',s}(x,m,\sigma)$ such that

$$\begin{aligned}
L &= \mathcal{O}(\log^2 \varepsilon_1^{-1} + \log^2 \varepsilon^{-1} + \log^2 C + k),\\
\|W\|_\infty &= \mathcal{O}(\log^3 \varepsilon_1^{-1} + \log^3 \varepsilon^{-1}),\\
S &= \mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^4 \varepsilon^{-1} + \log^2 C + k),\\
B &= \mathcal{O}(\varepsilon_1^{-2} + C^2) + \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1} + C^l 2^{kl}.
\end{aligned}$$

where approximation error between $f_7^{l',s}(x,m_t,\sigma_t)$ and $\phi_7^{l',s}(x,m',\sigma')$ is bounded as

$$\left|f_7^{l',s}(x,\sigma,m) - \phi_7^{l',s}(x,m',\sigma')\right| \leq (\varepsilon_1 + \varepsilon_{\mathrm{error}}(\varepsilon_1^{-l-2} + C^{4l}2^{4kl})) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}.$$

Finally, we sum up $\phi_7^{l',s}(x,m',\sigma')$ multiplied $\frac{-(-1)^{s+l}{}_l\mathrm{C}_{l'}2^{kl'}}{\sqrt{2\pi}s!2^s(l'+2s+2)}$ over $(l',s)$, according to (34) and using Lemma J.3. Here, the coefficient is bounded by $2^{(k+1)l}$ and the total number of possible combinations $(l',s)$ is bounded by $\mathcal{O}(lS)=\mathcal{O}(\log\varepsilon^{-1})$. Then, approximation error for (34) is bounded as

$$2^{(k+1)l}(\varepsilon_1+\varepsilon_{\mathrm{error}}(\varepsilon_1^{-l-2}+C^{4l}2^{4kl}))\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}.$$

In order to bound the terms related to $\varepsilon_1$ by $\mathcal{O}(\varepsilon)$, we take $\varepsilon_1=\mathcal{O}(2^{-(k+1)l}\log^{-\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1})$. Then, the total approximation error is bounded by $\tilde{\mathcal{O}}(\varepsilon)+\varepsilon_{\mathrm{error}}C^{4l}2^{k(4l+1)}\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}$ and this is achieved by a neural network with

$$\begin{aligned}L&=\mathcal{O}(\log^4\varepsilon^{-1}+\log^2C+k),\\ \|W\|_\infty&=\mathcal{O}(\log^6\varepsilon^{-1}),\\ S&=\mathcal{O}(\log^8\varepsilon^{-1}+\log^2C+k),\\ B&=\mathcal{O}(C^l2^{kl})+\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}.\end{aligned}$$

Finally, because

$$\begin{aligned}&\left|\int_{-\frac{\sigma_tC_{\mathrm{f},1}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m}}^{\frac{\sigma_tC_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}\frac{1}{\sqrt{2\pi}\sigma_t}\mathbb{1}[\underline{j}\le 2^ky\le\bar{j}](2^ky-j)^l\exp\left(-\frac{(x-m_ty)^2}{2\sigma_t^2}\right)\mathrm{d}y\right|\\ &\le\int\frac{1}{\sqrt{2\pi}\sigma_t}\mathbb{1}[\underline{j}\le 2^ky\le\bar{j}](l+1)^l\exp\left(-\frac{(x-m_ty)^2}{2\sigma_t^2}\right)\mathrm{d}y\lesssim C_f,\end{aligned}$$

we can clip $\phi_{\mathrm{dif},1}^{\underline{j},\bar{j},j,k}$ so that it is bounded by $\mathcal{O}(1)$. □

We now approximate the (modified) tensor product diffused B-spline basis. The following is the formal version of Lemma D.3. Without the term of $\mathbb{1}[\|y\|_\infty\le C_{\mathrm{b},1}]$, the statement matches that of Lemma D.3. This network $\phi_{\mathrm{dif},3}$ corresponds to $\phi_{\mathrm{TDB}}$ in Lemma D.3.

**Lemma D.8** (Approximation of the tensor-product diffused B-spline bases)**.** *Let $k\in\mathbb{Z}_+, j\in\mathbb{Z}^d, l\in\mathbb{Z}_+$ with $-C2^k-l\le j_i\le C2^k$ $(i=1,2,\cdots,d)$, $\varepsilon$ $(0<\varepsilon<\frac{1}{2})$ and $C>0$. There exists a neural network $\phi_{\mathrm{dif},3}(x,t)\in\Phi(L,W,S,B)$ with*

$$\begin{aligned}L&=\mathcal{O}(\log^4\varepsilon^{-1}+\log^2C+k^2),\\ \|W\|_\infty&=\mathcal{O}(\log^6\varepsilon^{-1}+\log^3C+k^3),\\ S&=\mathcal{O}(\log^8\varepsilon^{-1}+\log^4C+k^4),\\ B&=\exp\left(\log^4\varepsilon^{-1}+\log C+k\right),\end{aligned}$$

*such that*

$$\left|\phi_{\mathrm{dif},3}^{k,j}(x,t)-\int_{\mathbb{R}^d}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\mathbb{1}[\|y\|_\infty\le C_{\mathrm{b},1}]M_{k,j}^d(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y\right|\le\varepsilon$$

*holds for all $x\in[-C,C]^d$.*

*Also, with the same conditions, there exists a neural network $\phi_{\mathrm{dif},4}\in\Phi(L,W,S,B)$ with the same bounds on $L,\|W\|_\infty,S,B$ as above such that*

$$\left\|\phi_{\mathrm{dif},4}^{k,j}(x,\sigma',m')-\int_{\mathbb{R}^d}\frac{x-m_ty}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}}\mathbb{1}[\|y\|_\infty\le C_{\mathrm{b},1}]M_{k,j}^d(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y\right\|\le\varepsilon.$$

*holds for all $x\in[-C,C]^d$.*

*Furthermore, we can choose these networks so that $\|\phi_{\mathrm{dif},3}^{k,j}\|_\infty,\ \|\phi_{\mathrm{dif},4}^{k,j}\|_\infty=\mathcal{O}(1)$ hold.*

*Proof.* Here we only prove the first part, because the second part follows in the same way. We assume $|\sigma' - \sigma_t|, |m' - m_t| \le \varepsilon_{\rm error}$.

From the discussion (30), we approximate

$$\prod_{i=1}^d \left( \sum_{l'=0}^{l+1} \frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!} \int_{y_i \in a_i^x} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \mathbb{1}[|y_i| \le C_{\rm b,1}] \mathbb{1}[0 \le 2^k y_i - j_i \le l+1] \right.$$
$$\left. \times (2^{k_i} y_i - l' - j_i)_+^l \exp\left( - \frac{(x_i - m y_i)^2}{2\sigma^2} \right) \mathrm{d}y_i \right), \qquad (36)$$

which is equal to $D^d_{k,j}(x)$ within an additive error of $\mathcal{O}(\varepsilon)$, so we approximate (36). Here $a_i^x = [\frac{x_i}{m_t} - \frac{\sigma_t C_{\rm f}}{m_t}\sqrt{\log \varepsilon^{-1}}, \frac{x_i}{m_t} + \frac{\sigma_t C_{\rm f}}{m_t}\sqrt{\log \varepsilon^{-1}}]$.

We let $f_i(y_i; j_i, k, l') := \mathbb{1}[|y_i| \le C_{\rm b,1}] \mathbb{1}[0 \le 2^k y_i - j_i \le l+1](2^k y_i - l' - j_i)_+^l \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i$. First, $\sum_{l'=0}^{l+1} \frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!} f_i(y_i; j_i, k, l')$ is approximated by $\sum_{l'=0}^{l+1} \frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!} \phi_{\rm dif,1}^{j_i - l', \bar{j}_{l'}, \underline{j}_{l'}, k}(y_i, \sigma', m')$ (see Lemma J.3 for aggregation of the networks). Here, $\bar{j}_{l'}$ and $\underline{j}_{l'}$ are defined so that $\mathbb{1}[\underline{j}_{l'} \le 2^k y \le \bar{j}_{l'}] = \mathbb{1}[|y_i| \le C_{\rm b,1}]\mathbb{1}[0 \le 2^k y_i - j_i \le l+1]$ holds.

Now we multiply $\sum_{l'=0}^{l+1} \frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!} \phi_{\rm dif,1}^{j_i, \bar{j}_{l'}, \underline{j}_{l'}, k}(y_i, \sigma', m')$ over $i = 1, 2, \cdots, d$ using $\phi_{\rm mult}$ to obtain the desired network $\phi^{k,j}_{\rm dif,3}$. According to Lemma D.7 with $\varepsilon = \varepsilon$ and Lemma J.6 with $\varepsilon = \varepsilon$ and $C = \mathcal{O}(1)$ (because $\|\phi_{\rm dif,1}^{j_i, \bar{j}_{l'}, \underline{j}_{l'}, k}\|_\infty = \mathcal{O}(1)$), there exists a neural network $\phi_1(x, m', \sigma') \in \Phi(L, W, S, B)$ with

$$\begin{aligned} L &= \mathcal{O}(\log^4 \varepsilon^{-1} + \log^2 C + k), \\ \|W\|_\infty &= \mathcal{O}(\log^6 \varepsilon^{-1}), \\ S &= \mathcal{O}(\log^8 \varepsilon^{-1} + \log^2 C + k), \\ B &= \mathcal{O}(C^l 2^{kl}) + \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1} \end{aligned}$$

and we can bound the approximation error between $\phi_1(x, m', \sigma')$ and (36) with

$$\tilde{\mathcal{O}}(\varepsilon) + \varepsilon_{\rm error} C^{4l} 2^{k(4l+1)} \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}. \qquad (37)$$

Now, we consider $\phi_{\rm dif,3} = \phi_1(x, \phi_m(t), \phi_\sigma(t))$. We apply Lemma D.6 with $\varepsilon = C^{-4l} 2^{-k(4l+1)} \log^{-\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}$, so that $\varepsilon_{\rm error}$ gets small enough and (37) is bounded by $\tilde{\mathcal{O}}(\varepsilon)$. Then, the size of $\phi_{\rm dif,3}$ is bounded by

$$\begin{aligned} L &= \mathcal{O}(\log^4 \varepsilon^{-1} + \log^2 C + k^2), \\ \|W\|_\infty &= \mathcal{O}(\log^6 \varepsilon^{-1} + \log^3 C + k^3), \\ S &= \mathcal{O}(\log^8 \varepsilon^{-1} + \log^4 C + k^4), \\ B &= \exp\left(\log^4 \varepsilon^{-1} + \log C + k\right). \end{aligned}$$

Now, adjusting $\varepsilon$ to replace $\tilde{\mathcal{O}}(\varepsilon)$ by $\varepsilon$ yields the first assertion.

We can make $\|\phi^{k,j}_{\rm dif,3}\|_\infty$ hold, because $\int_{\mathbb{R}^d} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \mathbb{1}[\|y\|_\infty \le C_{\rm b,1}] M^d_{k,j}(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y = \mathcal{O}(1)$.

□

### D.4 APPROXIMATION ERROR BOUND: BASED ON $p_0$

Now we put it all together and derive Theorem 3.1. Throughout this and the next subsections, we take $N \gg 1$, $T_1 = \underline{T} = \mathrm{poly}(N^{-1})$ and $T_5 = \overline{T} = \mathcal{O}(\log N)$. Moreover, we let $T_2 = N^{-(2-\delta)/d}$, $T_3 = 2T_2$, $T_4 = 3T_2$. This subsection considers the approximation for $t \in [T_1, T_4]$.

We begin with the following lemma, which gives the basis decompositon of the Besov functions.

**Lemma D.9** (Basis decomposition). *Under $N \gg 1$, Assumptions 2.2, 2.3, 2.4 with $a_0 = N^{-(1-\delta)/d}$, there exists $f_N$ that satisfies*

$$\|p_0 - f_N\|_{L^2([-1,1]^d)} \lesssim N^{-s/d},$$
$$\|p_0 - f_N\|_{L^2([-1,1]^d \setminus [-1+N^{-(1-\delta)/d}, 1-N^{-(1-\delta)/d}]^d)} \lesssim N^{-(3s+2)/d},$$

*and $f_N(x) = 0$ for all $x$ with $\|x\|_\infty \geq 1$, and has the following form:*

$$f_N(x) = \sum_{i=1}^{N} \alpha_i \mathbb{1}[\|x\|_\infty \leq 1] M^d_{k,j_i}(x) + \sum_{i=N+1}^{3N} \alpha_i \mathbb{1}[\|x\|_\infty \leq 1 - N^{-(1-\delta)/d}] M^d_{k,j_i}(x), \quad (38)$$

*where $-2^{(k)_m} - l \leq (j_i)_m \leq 2^{(k)_m}$ $(i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$, $|k| \leq K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ for $\delta = d(1/p - 1/r)_+$ and $\nu = (2s - \delta)/(2\delta)$. Moreover, $|\alpha_i| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)_+}$.*

*Proof.* Because $p_0 \in \mathcal{C}^{3s+2}([-1,1]^d \setminus [-1 + N^{-(1-\delta)/d}, 1 - N^{-(1-\delta)/d}]^d)$, according to Lemma J.13, we have $f_1$ such that

$$\|p_0 - f_1\|_{L^2([-1,1]^d \setminus [-1+N^{-(1-\delta)/d}, 1-N^{-(1-\delta)/d}]^d)} \lesssim N^{-(3s+2)/d}.$$

and has the following form:

$$f_1(x) = \sum_{i=1}^{N} \alpha_i M^d_{k,j_i}(x),$$

where $-2^{(k)_m} - l \leq (j_i)_m \leq 2^{(k)_m}$ $(i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$, $|k| \leq K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ for $\delta = d(1/p - 1/r)_+$ and $\nu = (2s - \delta)/(2\delta)$. Moreover, $|\alpha_{1,i}| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-2s)_+}$.

Next let us approximate $f$ in $[-1,1]^d$. Because $\|p_0\|_{B^s_{p,q}} \lesssim 1$, we have $f_2$ such that

$$\|p_0 - f_2\|_{L^2([-1,1]^d)} \lesssim N^{-s/d}.$$

and has the following form:

$$f_2(x) = \sum_{i=N+1}^{2N} \alpha_i M^d_{k,j_i}(x),$$

where $-2^{(k)_j} - l \leq (j_i)_j \leq 2^{(k)_j}$ $(i = 1, 2, \cdots, N,\ j = 1, 2, \cdots, d)$, $|k| \leq K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ for $\delta = d(1/p - 1/r)_+$ and $\nu = (s - \delta)/(2\delta)$. Moreover, $|\alpha_{2,i}| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)}$.

Therefore,

$$\begin{aligned}
&\mathbb{1}[\|x\|_\infty \leq 1] f_1(x) - \mathbb{1}[\|x\|_\infty \leq 1 - N^{-(1-\delta)/d}] f_1(x) + \mathbb{1}[\|x\|_\infty \leq 1 - N^{-(1-\delta)/d}] f_2(x) \\
&= \sum_{i=1}^{N} \alpha_i M^d_{k_i,j_i}(x) - \sum_{i=1}^{N} \alpha_i \mathbb{1}[\|x\|_\infty \leq 1 - N^{-(1-\delta)/d}] M^d_{k_i,j_i}(x) \\
&\quad + \sum_{i=N+1}^{2N} \alpha_i \mathbb{1}[\|x\|_\infty \leq 1 - N^{-(1-\delta)/d}] M^d_{k_i,j_i}(x)
\end{aligned}$$

holds and reindexing the bases gives the result. □

The following lemma gives neural network that approximates $\nabla \log p_t(x)$ in $[T_1, T_4]$.

**Lemma D.10** (Approximation of score function for $T_1 \leq t \leq T_4$). *There exists a neural network $\phi_{\text{score},1} \in \Phi(L, W, S, B)$ that satisfies*

$$\int p_t(x) \|\phi_{\text{score},1}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t \lesssim \frac{N^{-2s/d} \log N}{\sigma_t^2} \quad (39)$$

*Here, $L, \|W\|_\infty, S, B$ is evaluated as*

$$L = \mathcal{O}(\log^4 N), \quad \|W\|_\infty = \mathcal{O}(N \log^6 N), \quad S = \mathcal{O}(N \log^8 N), \quad \text{and } B = \exp(\mathcal{O}(\log^4 N)).$$

*Proof.* Before we proceed to the main part of the proof, we limit the discussion into the bounded region. According to Lemma C.4, we have that

$$\int_{\|x\|_\infty \geq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x)\|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{\underline{T}}{N^{(2s+1)/d}} \left(1 + \|s(\cdot,t)\|_\infty^2\right), \quad (40)$$

with a sifficiently large hidden constant in $\mathcal{O}(1)$. Because $\|\nabla \log p_t(x)\|$ is bounded with $\frac{\log^{\frac{1}{2}} N}{\sigma_t}$ in $\|x\|_\infty \geq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}$ due to Lemma C.3, $s$ can be taken so that $\|s(\cdot,t)\|_\infty \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}$ and therefore (40) is bounded by $\frac{\underline{T}}{N^{(2s+1)}} \cdot \frac{\log N}{\underline{T}} = N^{-(2s+1)/d}\log N$, which is smaller than the upper bound of (39). Thus, we can focus on the approximation of the score $\nabla \log p_t(x)$ within $\|x\|_\infty \leq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$. Moreover, we can also exclude the case where $p_t(x) \leq N^{-(2s+1)/d}$, because Lemma C.4 can bound the error

$$\begin{aligned}
&\int_{\|x\|_\infty \leq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x)\mathbb{1}[p_t(x) \leq \varepsilon]\|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \\
&\quad \lesssim \frac{\varepsilon}{\sigma_t^2} \log^{\frac{d+2}{2}}(\varepsilon^{-1}\underline{T}^{-1}) + \varepsilon\|s(x,t)\| \\
&\lesssim \frac{\varepsilon}{\sigma_t^2} \log^{\frac{d+2}{2}}(\varepsilon^{-1}\underline{T}^{-1}) + \frac{\varepsilon}{\sigma_t^2}\log N, \qquad (41)
\end{aligned}$$

and setting $\varepsilon = N^{-(2s+1)/d}$ makes (41) smaller than the bound (39).

Thus, in the following, we consider $x$ such that $\|x\|_\infty \leq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$ and $p_t(x) \geq N^{-(2s+1)/d}$ holds. In this case, we have $\|\nabla \log p_t(x)\| \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}$.

The construction is straightforward. Based on (38) of Lemma D.9, we let

$$\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&\risingdotseq \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y = \sum_{i=1}^N \alpha_i E^{(1)}_{k_i,j_i}(x,t) =: \tilde{f}_1(x,t), \\
f_1(x,t) &:= \tilde{f}_1(x,t) \vee N^{-(2s+1)/d},
\end{aligned}$$

and

$$\begin{aligned}
\sigma_t \nabla p_t(x) &= \int \frac{x - m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&\risingdotseq \int \frac{x - m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y = \sum_{i=1}^N \alpha_i E^{(2)}_{k_i,j_i}(x,t) =: f_2(x,t), \\
f_3(x,t) &:= \frac{f_2(x,t)}{f_1(x,t)} \mathbb{1}\left[\left\|\frac{f_2(x,t)}{f_1(x,t)}\right\| \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}\right]
\end{aligned}$$

so that $\alpha_i$, $E^{(1)}_{k_i,j_i}(x,t)$ and $E^{(2)}_{k_i,j_i}(x,t)$ correspond to the basis decomposition in Lemma D.9. Thus, $|\alpha_i| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)_+}$ and $|k_i| = \mathcal{O}(\log N)$. We remark that $C_{\mathrm{b},1}$ is set to be $1$ or $1 - N^{-(1-\delta)/d}$ in (28) and (29). We approximate $E^{(1)}_{k,j_i}$ and $E^{(2)}_{k,j_i}$ by $\phi^{k_i,j_i}_{\mathrm{dif},3}$ and $\phi^{k_i,j_i}_{\mathrm{dif},4}$ in Lemma D.8, by setting $\varepsilon = \varepsilon_1$ and $C = m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$ (because $\sigma_t \leq \sigma_{T_2} \lesssim \log^{-\frac{1}{2}} N$), where $\varepsilon_1 = \mathrm{poly}(N^{-1})$ is a scaler value adjusted below. Then we sum up these sub-networks using Lemma J.3 and obtain neural networks $\phi_{\mathrm{dif},5}(x,t)$ and $\phi_{\mathrm{dif},6}(x,t)$ that approximate $f_1(x,t)$ and $f_2(x,t)$, respectively.

Because we can decompose the error as

$$\int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x)\mathbb{1}[p_t(x) \ge N^{-\frac{2s+1}{d}}]\|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x$$

$$\lesssim \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{1}[p_t(x) \ge N^{-\frac{2s+1}{d}}]p_t(x) \left\|\phi_{\text{score},1}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\|^2 \mathrm{d}x \tag{42}$$

$$+ \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{1}[p_t(x) \ge N^{-\frac{2s+1}{d}}]p_t(x) \left\|\frac{f_3(x,t)}{\sigma_t} - \nabla \log p_t(x)\right\|^2 \mathrm{d}x, \tag{43}$$

we consider the approximation of $\frac{f_3(x,t)}{\sigma_t}$ for the moment, instead of $\nabla \log p_t(x) = \frac{\nabla p_t(x,t)}{f_1(x,t)}$, and bound (42). From the construction of the networks, we have the following bounds:

$$|f_1(x,t) - \phi_{\text{dif},5}(x,t)| \lesssim N \cdot \max|\alpha_i| \cdot \varepsilon_1, \tag{44}$$

$$\|f_2(x,t) - \phi_{\text{dif},6}(x,t)\| \lesssim N \cdot \max|\alpha_i| \cdot \varepsilon_1. \tag{45}$$

for all $x$ with $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$. Note that $\max|\alpha_i|$ is bounded by $N^{(\nu^{-1}+d^{-1})(d/p-s)_+}$. Thus, we take $\varepsilon_1 \lesssim N^{-1} \cdot N^{-(\nu^{-1}+d^{-1})(d/p-s)_+} \cdot N^{-\frac{9s+3}{d}}$ so that (44) and (45) are bounded by $N^{-\frac{9s+3}{d}}$ in Lemma J.6.

Then we define $\phi_{\text{dif},7}$ as

$$[\phi_{\text{dif},7}(x,t)]_i := \phi_{\text{clip}}(\phi_{\text{mult}} \\ (\phi_{\text{rec}}(\phi_{\text{clip}}(\phi_{\text{dif},5}(x,t); N^{-(2s+1)/d}, \mathcal{O}(1)))), [\phi_{\text{dif},6}(x,t)]_i); -\mathcal{O}(\log^{\frac{1}{2}} N), \mathcal{O}(\log^{\frac{1}{2}} N)).$$

to approximate $\sigma_t \nabla \log p_t(x)$. Here we used the boundedness of $p_t(x)$ with $[N^{-(2s+1)/d}, \mathcal{O}(1)]$ to clip $\phi_{\text{dif},5}(x,t)$ and the boundedness of $\sigma_t\nabla \log p_t(x)$ with $[-\mathcal{O}(\log^{\frac{1}{2}} N), \mathcal{O}(\log^{\frac{1}{2}} N)]$ to clip the whole output. For $\phi_{\text{rec}}$ we let $\varepsilon = N^{-(3s+1)/d}$ in Lemma J.7 and for $\phi_{\text{mult}}$ we let $\varepsilon = N^{-s/d}$ and $C = N^{(2s+1)/d}$. Then,

$$\|\phi_{\text{dif},7}(x,t) - f_3(x,t)\| = \left\|\phi_{\text{dif},7}(x,t) - \frac{f_2(x,t)}{f_1(x,t)}\mathbb{1}\left[\left\|\frac{f_2(x,t)}{f_1(x,t)}\right\| \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}\right]\right\|$$

$$\lesssim N^{-s/d}$$

$$+ N^{(2s+1)/d} \cdot (N^{-(3s+1)/d} + N^{2(3s+1)/d}|f_1(x,t) - \phi_{\text{dif},5}(x,t)| + \|f_2(x,t) - \phi_{\text{dif},6}(x,t)\|)$$

$$\lesssim N^{-s/d} + N^{(8s+3)/d}|f_1(x,t) - \phi_{\text{dif},5}(x,t)| + N^{(2s+1)/d}\|f_2(x,t) - \phi_{\text{dif},6}(x,t)\|. \tag{46}$$

Applying (44)$\le N^{-\frac{9s+3}{d}}$ and (45)$\le N^{-\frac{9s+3}{d}}$ yields that (46)$\le N^{-\frac{s}{d}}$.

Finally, we let

$$\phi_{\text{score},1}(x,t) := \phi_{\text{mult}}(\phi_{\text{dif},7}(x,t), \phi_\sigma(t)).$$

By setting $\varepsilon = N^{-s/d}$ and $C \simeq \max\{\log^{\frac{1}{2}} N, \sigma_{\underline{T}}\} \lesssim \text{poly}(N)$ in Lemma J.6 for $\phi_{\text{mult}}$ and $\varepsilon = N^{-s/d}/\text{poly}(N)$ in Lemma D.6 for $\phi_\sigma$. Then,

$$\left\|\phi_{\text{score},1}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\| \lesssim N^{-s/d} + \text{poly}(N) \cdot N^{-s/d}/\text{poly}(N) \lesssim N^{-s/d},$$

which yields

$$(42) = \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{1}[p_t(x) \ge N^{-\frac{2s+1}{d}}]p_t(x) \left\|\phi_{\text{score},1}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\|^2 \mathrm{d}x$$

$$\lesssim N^{-2s/d}.$$

The structure of $\phi_{\text{dif},7}$ and $\phi_{\text{score},1}$ are evaluated as

$$L = \mathcal{O}(\log^4 N),\ \|W\|_\infty = \mathcal{O}(N\log^6 N),\ S = \mathcal{O}(N\log^8 N),\ \text{and}\ B = \exp\left(\log^4 N\right).$$

Here we used $|k_i| = \mathcal{O}(\log N)$ and $C = \mathcal{O}(1)$.

We move to the error analysis between $\frac{f_3(x,t)}{\sigma_t}$ and $\nabla \log p_t(x)$ to bound (43). Remind that we consider $x$ such that $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$ and $p_t(x) \ge N^{-(2s+1)/d}$ holds. In this case, we have $\|\nabla \log p_t(x)\| \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}$. First, we consider the case $x \in [-m_t, m_t]^d$. Since $p_t(x)$ is lower bounded by $C_{\mathrm{a}}^{-1}$ according to Lemma C.2, as long as $|f_1(x,t) - p_t(x)| \le C_{\mathrm{a}}^{-1}/2$, we can say that the approximation error is bounded by $\lesssim |f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|$. On the other hand, if $|f_1(x,t) - p_t(x)| \ge C_{\mathrm{a}}^{-1}/2$, we no longer have such bound, but this time we can use the fact that $\frac{f_2(x,t)}{f_1(x,t)}$ and $\sigma_t \frac{\sigma_t \nabla p_t(x)}{p_t(x)}$ is bounded by $\log^{\frac{1}{2}} N$. Therefore, when $x \in [-m_t, m_t]^d$, we can bound the approximation error as

$$\begin{aligned}&\left\| f_3(x,t) - \sigma_t \frac{\nabla p_t(x)}{p_t(x)} \right\| \le \left\| \frac{f_2(x,t)}{f_1(x,t)} - \sigma_t \frac{\nabla p_t(x)}{p_t(x)} \right\| \\ &\lesssim \log^{\frac{1}{2}} N (|f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|).\end{aligned}$$

Next, we consider the case when $x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}]^d \setminus [-m_t, m_t]^d$. Then, we have that

$$\begin{aligned}&\left\| f_3(x,t) - \sigma_t \frac{\nabla p_t(x)}{p_t(x)} \right\| \le \left\| \frac{f_2(x,t)}{f_1(x,t)} - \sigma_t \frac{\nabla p_t(x)}{p_t(x)} \right\| \\ &\lesssim \frac{\|f_2(x,t) - \sigma_t \nabla p_t(x)\|}{f_1(x,t)} + \|\sigma_t \nabla p_t(x)\| \left| \frac{1}{f_1(x,t)} - \frac{1}{p_t(x)} \right|. \end{aligned} \tag{47}$$

The first term is bounded by $N^{(2s+1)/d}\|f_2(x,t)(x,t) - \sigma_t \nabla p_t(x)\|$ because we focus on the case $p_t(x) \ge N^{-(2s+1)/d}$. For the second term, because $\|\nabla \log p_t(x)\| = \left\|\sigma_t \frac{\nabla p_t(x)}{p_t(x)}\right\| \lesssim \frac{\log^{\frac{1}{2}}}{\sigma_t}$, we have $\|\sigma_t \nabla p_t(x)\| \lesssim p_t(x) \log^{\frac{1}{2}} N$. By using this, we can bound the second term as

$$\begin{aligned}\|\sigma_t \nabla p_t(x)\| \left| \frac{1}{f_1(x,t)} - \frac{1}{p_t(x)} \right| &\lesssim \log^{\frac{1}{2}} N p_t(x) \left| \frac{1}{f_1(x,t)} - \frac{1}{p_t(x)} \right| \\ &\lesssim \log^{\frac{1}{2}} N \frac{|p_t(x) - f_1(x,t)|}{f_1(x,t)} \\ &\lesssim N^{\frac{2s+1}{d}} \log^{\frac{1}{2}} N \, |p_t(x) - f_1(x,t)| ,\end{aligned}$$

where we used $f_1(x,t) \ge N^{-(2s+1)/d}$. Thus, for $x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}]^d \setminus [-m_t, m_t]^d$ and $p_t(x) \ge N^{-\frac{2s+1}{d}}$, (47) is bounded by

$$\left\| \phi_{\mathrm{dif},7}(x,t) - \frac{\sigma_t \nabla p_t(x)}{p_t(x)} \right\| \lesssim N^{\frac{2s+1}{d}} \log^{\frac{1}{2}} N (|\phi_{\mathrm{dif},5}(x,t) - p_t(x)| + \|\phi_{\mathrm{dif},6}(x,t) - \sigma_t \nabla p_t(x)\|).$$

Therefore, we have that

$$\begin{aligned}&\left\| \frac{f_2(x,t)}{\sigma_t f_1(x,t)} - \frac{\nabla p_t(x)}{p_t(x)} \right\| \\ &\lesssim \begin{cases} \log^{\frac{1}{2}} N (|f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|)/\sigma_t \quad (\|x\|_\infty \le m_t) \\ N^{\frac{2s+1}{d}} \log^{\frac{1}{2}} N (|f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|)/\sigma_t \\ \qquad (x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}]^d \setminus [-m_t, m_t]^d). \end{cases}\end{aligned} \tag{48}$$

We consider the $L^2(p_t)$ loss of (48). First, we consider the case of $\|x\|_\infty \le m_t$.

$$
\begin{aligned}
&\int_{\|x\|_\infty \le m_t} p_t(x) \left\| \frac{f_2(x,t)}{\sigma_t f_1(x,t)} - \frac{\nabla p_t(x)}{p_t(x)} \right\|^2 \mathrm{d}x \\
&\lesssim \int_{\|x\|_\infty \le m_t} (|f_1(x,t) - p_t(x)|^2 + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|^2) \log N / \sigma_t^2 \mathrm{d}x \\
&\quad \text{(we used(48) and } p_t(x) = \mathcal{O}(1) \text{ by Lemma C.2.)} \\
&\lesssim \int_{\|x\|_\infty \le m_t} \log N / \sigma_t^2 \mathrm{d}x \\
&\left( \left| \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right|^2 + \right. \\
&\left. \left\| \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right\|^2 \right) \\
&\lesssim \log N / \sigma_t^2 \cdot \int_{\|x\|_\infty \le m_t} \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x \\
&\quad + \log N / \sigma_t^2 \cdot \int_{\|x\|_\infty \le m_t} \int \frac{|x - m_t y|}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x \\
&= \log N / \sigma_t^2 \cdot \int \int_{\|x\|_\infty \le m_t} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}x \mathrm{d}y \\
&\quad + \log N / \sigma_t^2 \cdot \int \int_{\|x\|_\infty \le m_t} \frac{|x - m_t y|}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}x \mathrm{d}y \\
&\lesssim \log N / \sigma_t^2 \cdot \int |p_0(y) - f_N(y)|^2 \mathrm{d}y + \log N / \sigma_t^2 \cdot \int |p_0(y) - f_N(y)|^2 \mathrm{d}y \lesssim \log N / \sigma_t^2 \cdot N^{-2s/d}.
\end{aligned}
$$

For the third inequality, we used Jensen's inequality. For the second last inequality, we used the construction of $f_N$ and Lemma D.9.

We then consider the case of $x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}]^d \setminus [-m_t, m_t]^d$. Most of the part is the same as previously.

$$
\begin{aligned}
&\int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x) \mathbb{1}[p_t(x) \ge N^{-\frac{2s+1}{d}}] \left\| \frac{f_2(x,t)}{\sigma_t f_1(x,t)} - \frac{\nabla p_t(x)}{p_t(x)} \right\|^2 \mathrm{d}x \\
&\lesssim \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} (|f_1(x,t) - p_t(x)|^2 + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|^2) N^{\frac{4s+2}{d}} \log N / \sigma_t^2 \mathrm{d}x \\
&\lesssim \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} N^{\frac{4s+2}{d}} \log N / \sigma_t^2 \mathrm{d}x \\
&\quad \cdot \left( \left| \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right|^2 \right. \\
&\quad \left. + \left\| \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right\|^2 \right) \\
&\lesssim N^{\frac{4s+2}{d}} \log N / \sigma_t^2 \cdot \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x \\
&\quad + N^{\frac{4s+2}{d}} \log N / \sigma_t^2 \\
&\quad \cdot \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \int \frac{|x - m_t y|^2}{\sigma_t^{d+2} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x
\end{aligned}
$$

$$\lesssim N^{\frac{4s+2}{d}} \log N/\sigma_t^2 \cdot \Bigg[\int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}}$$
$$\Bigg[\int_{\|\frac{x}{m_t}-y\|_\infty \le \mathcal{O}(1)\sigma_t\sqrt{\log N}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2 \mathrm{d}y + N^{-\frac{6s+2}{d}}\Bigg]\mathrm{d}x$$
$$+\int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}}$$
$$\Bigg[\int_{\|\frac{x}{m_t}-y\|_\infty \le \mathcal{O}(1)\sigma_t\sqrt{\log N}} \frac{|x-m_t y|^2}{\sigma_t^{d+2} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2 \mathrm{d}y + N^{-\frac{6s+2}{d}}\Bigg]\mathrm{d}x\Bigg]$$

(we used Lemma J.10.)

$$\lesssim N^{\frac{4s+2}{d}} \log N/\sigma_t^2 \cdot \Bigg[N^{-\frac{6s+2}{d}} + \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathrm{d}x$$
$$\Bigg[\int_{\|\frac{x}{m_t}-y\|_\infty \le \mathcal{O}(1)\sigma_t\sqrt{\log N}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2 \mathrm{d}y\Bigg]$$
$$+\int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathrm{d}x$$
$$\Bigg[\int_{\|\frac{x}{m_t}-y\|_\infty \le \mathcal{O}(1)\sigma_t\sqrt{\log N}} \frac{\log N}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2 \mathrm{d}y\Bigg]\Bigg]$$
$$\lesssim N^{\frac{4s+2}{d}} \log^2 N/\sigma_t^2 \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \int_{\|\frac{x}{m_t}-y\|_\infty \le \mathcal{O}(1)\sigma_t\sqrt{\log N}}$$
$$\frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2 \mathrm{d}x\mathrm{d}y + N^{-\frac{2s}{d}} \log N/\sigma_t^2 \tag{49}$$

For the third inequality, we used Jensen's inequality. Here, we note that if $(x, y)$ satisfies $m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$ and $\|\frac{x}{m_t} - y\|_\infty \le \mathcal{O}(1)\sigma_t\sqrt{\log N}$, then we have that $1-\mathcal{O}(1)\sigma_t\sqrt{\log N} \le \|y\|_\infty \le 1+\mathcal{O}(1)\frac{\sigma_t}{m_t}\sqrt{\log N}$ and that $1-\mathcal{O}(1)\sqrt{t} \le \|y\|_\infty \le 1+\mathcal{O}(1)\sqrt{t}$. Because we are considering the time $t \le T_4 = 3N^{-\frac{2-\delta}{d}}$, $\mathcal{O}(1)\sqrt{t} \lesssim N^{-\frac{1-\delta}{d}}$ holds for sufficiently large $N$. Therefore, (49) is further bounded by

$$\begin{aligned}
&(49)\\
&\lesssim N^{\frac{4s+2}{d}} \log^2 N/\sigma_t^2 \cdot \\
&\quad \int_x \int_{1-N^{-\frac{1-\delta}{d}} \le \|y\|_\infty \le 1+N^{-\frac{1-\delta}{d}}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2 \mathrm{d}x\mathrm{d}y \\
&\quad + N^{-\frac{2s}{d}} \log N/\sigma_t^2 \\
&= N^{\frac{4s+2}{d}} \log^2 N/\sigma_t^2 \cdot \\
&\quad \int_{1-N^{-\frac{1-\delta}{d}} \le \|y\|_\infty \le 1+N^{-\frac{1-\delta}{d}}} \int_x \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2 \mathrm{d}y\mathrm{d}x \\
&\quad + N^{-\frac{2s}{d}} \log N/\sigma_t^2 \\
&\lesssim N^{\frac{4s+2}{d}} \log^2 N/\sigma_t^2 \cdot N^{-\frac{6s+4}{d}} + N^{-\frac{2s}{d}} \log N/\sigma_t^2 \lesssim N^{-\frac{2s}{d}} \log N/\sigma_t^2,
\end{aligned}$$

where we used the construction of $f_N$ and Lemma D.9 for the second last inequality. Now we successfully bounded (43) and the conclusion follows. □

### D.5 Approximation error bound: using the induced smoothness

We then consider the approximation for $t \gtrsim T_2 = N^{-(2-\delta)/d}$. This can be proved by considering diffusion process starting at $t = t_* > 0$. We begin with the following lemma.

**Lemma D.11** (Basis decomposition of $p_t$ at $t = t_*$). *If $N, N' \gg 1$ and $N' \geq t_*^{-\frac{d}{2}} N^{\frac{\delta}{2}}$, there exists $f_{N'}$ such that*

$$\|p_{t_*} - f_{N'}\|_{L^2(\mathbb{R}^d)} \lesssim N^{-(3s+5)/d},$$

*$f_{N'}(x) = 0$ for $x$ with $\|x\|_\infty \gtrsim \mathcal{O}(\sqrt{\log N})$, and has the following form:*

$$f_N(x) = \sum_{i=1}^{N'} \mathbb{1}[\|x\|_\infty \lesssim \mathcal{O}(\sqrt{\log N})] M^d_{k_i,j_i}(x),$$

*where $-\sqrt{\log N} 2^{(k_i)_m} - l \lesssim (j_i)_l \lesssim \sqrt{\log N} 2^{(k_i)_l}$ $(i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$, $\|k_i\|_\infty \leq K = \mathcal{O}(d^{-1} \log N)$ and $|\alpha_i| \lesssim N^{\frac{(3s+6)(2-\delta)}{\delta}}$.*

*Proof.* Let $\alpha = \frac{2(3s+6)}{\delta} + 1$. According to Lemma C.3, for any $x$, we have

$$\|\partial_{x_{i_1}} \partial_{x_{i_2}} \cdots \partial_{x_{i_k}} p_{T_2}(x)\| \leq \frac{C_{\mathrm{a}}}{\sigma^k_{t_*}}.$$

Because all derivatives up to order $\alpha$ is bounded by $\sigma_{t_*}^{-\alpha} \lesssim t_*^{-\frac{\alpha}{2}} \vee 1$, $\frac{p_{t_*}(x)}{t_*^{-\frac{\alpha}{2}} \vee a}$ belongs to $W^\alpha_\infty$ and its norm in $W^\alpha_\infty$ is bounded by a constant depending on $\alpha$, and hence to $B^\alpha_{\infty,\infty}$. Therefore, according to Lemma J.13, there exists a basis decomposition with the order of the B-spline basis $l = \alpha + 2$:

$$f_{N'}(x) = (t_*^{-\frac{\alpha}{2}} \vee 1) \sum_{i=1}^{N'} \alpha_i M^d_{k_i,j_i}(x).$$

such that

$$\begin{aligned} &\|p_{t_*} - f_{N'}\|_{L^2([-\mathcal{O}(\sqrt{\log N}), \mathcal{O}(\sqrt{\log N})]^d)} \lesssim (\sqrt{\log N})^\alpha {N'}^{-\alpha/d} t_*^{-\frac{\alpha}{2}} \\ &= (\sqrt{\log N})^\alpha N^{\alpha\delta/2d} = (\sqrt{\log N})^\alpha N^{-(3s+6)/d} \lesssim N^{-(3s+5)/d}, \end{aligned}$$

where $-\sqrt{\log N} 2^{(k_i)_m} - l \lesssim (j_i)_l \lesssim \sqrt{\log N} 2^{(k_i)_l}$ $(i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$, $\|k_i\|_\infty \leq K = \mathcal{O}(d^{-1} \log N)$, and $|\alpha_i| \lesssim 1$. Also, Lemma C.4 with $\varepsilon = N^{-\frac{6s+10}{d}}$ and $m_{t_*} + \mathcal{O}(1)\sigma_{t_*}\sqrt{\log N} \lesssim \sqrt{\log N}$ guarantees that $\|p_{T_2} - f_N\|_{L^2(\mathbb{R}^d \subseteq [-\mathcal{O}(\sqrt{\log N}), \mathcal{O}(\sqrt{\log N})]^d)} \lesssim N^{-(3s+5)/d}$. Therefore, by resetting $\alpha_i \leftarrow (t_*^{-\frac{\alpha}{2}} \vee 1)\alpha_i$, the assertion holds. ($\alpha_i$ is then bounded by $T_2^{-\frac{\alpha}{2}}$.) □

Lemma D.11 gives a concrete construction of the neural network for $T_3 \leq t \leq T_5$.

**Lemma D.12** (Approximation of score function for $T_3 \leq t \leq T_5$; Lemma D.5). *Let $N \gg 1$ and $N' \geq t_*^{-d/2} N^{\delta/2}$. Suppose $t_* \geq N^{-(2-\delta)/d}$. Then there exists a neural network $\phi_{\mathrm{score},2} \in \Phi(L, W, S, B)$ that satisfies*

$$\int_x p_t(x) \|\phi_{\mathrm{score},2}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^2}$$

*for $t \in [2t_*, \overline{T}]$. Specifically, $L = \mathcal{O}(\log^4(N)), \|W\|_\infty = \mathcal{O}(N), S = \mathcal{O}(N')$, and $B = \exp(\mathcal{O}(\log^4 N))$. Moreover, we can take $\phi_{\mathrm{score},2}$ satisfying $\|\phi_{\mathrm{score},2}\|_\infty = \mathcal{O}(\sigma_t^{-1} \log^{\frac{1}{2}} N)$.*

*Proof.* The proof is essentially the same as that of Lemma D.10. Here, the slight differences are that (i) $p_t$, $\phi_{\mathrm{dif},8}$, and $f_1$ are lower bounded by $N^{-(2s+3)/d}$, not by $N^{-(2s+1)/d}$, that (ii) $L^2(p_t)$ error should be bounded by $\frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^2}$, not by $\frac{N^{-\frac{2s}{d}}}{\sigma_t^2}$, and that (iii) $p_{t_*}$ is supported on $\mathbb{R}^d$, not on $[-1,1]^d$. Bounding the difference between Observe that $t_* \geq T_1 = N^{-\frac{2-\delta}{d}}$ holds, which is necessary to apply the argument of Lemma D.10.

Let us reset the time $t \leftarrow t - t_*$ in the following proof and consider the diffusion process from $p_0$ (in the new definition), for simplicity. We have $t \geq t_* \gtrsim \mathrm{poly}(N^{-1})$ in the new definition. According to Lemma C.4, we have that

$$\int_{\|x\|_\infty \geq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x)\|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{t_*}{N^{(2s+2)/d}}\left(1 + \|s(\cdot,t)\|_\infty^2\right), \quad (50)$$

with a sifficiently large hidden constant in $\mathcal{O}(1)$. We limit the domain of $x$ into $\|x\|_\infty \leq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(\sqrt{\log N})$. In this region, Lemma C.3 yields $\|\nabla \log p_t(x)\| \lesssim \frac{\sqrt{\log N}}{\sigma_t}$, and therefore we can take $s$ such that $\|s(\cdot,t)\|_\infty \leq \frac{\sqrt{\log N}}{\sigma_t} \lesssim \frac{\sqrt{\log N}}{\sqrt{t_*}\wedge 1}$ holds. Then, (50) is bounded by $N^{-2(s+1)/d}$. Moreover,

$$\begin{aligned}
&\int_{\|x\|_\infty \leq m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x)\mathbb{1}[p_t(x) \leq N^{-(2s+3)/d}]\|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \\
&\lesssim \frac{\varepsilon}{\sigma_t^2}\log^{\frac{d+2}{2}}(N) + \varepsilon\|s(x,t)\| \\
&\lesssim \left(\frac{N^{-(2s+3)/d}}{\sigma_t^2}\log^{\frac{d+2}{2}}(N) + \frac{N^{-(2s+3)/d}}{\sigma_t^2}\log N\right)\log^{\frac{d}{2}} N \lesssim N^{-2(s+1)/d}.
\end{aligned}$$

This means that we only need to consider $x$ with $p_t(x) \geq N^{-(2s+3)/d}$.

Using the basis decomposition in the previous lemma, we let

$$\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} p_0(y)\exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y \\
&\fallingdotseq \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f_N(y)\exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y = \sum_{i=1}^{N'} \alpha_i E^{(1)}_{k_i,j_i}(x,t) =: \tilde{f}_1(x,t), \\
f_1(x,t) &:= \tilde{f}_1(x,t) \vee N^{-(2s+3)/d},
\end{aligned}$$

and

$$\begin{aligned}
\sigma_t\nabla p_t(x) &= \int \frac{x - m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} p_0(y)\exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y \\
&\fallingdotseq \int \frac{x - m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f_N(y)\exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y = \sum_{i=1}^{N'} \alpha_i E^{(2)}_{k_i,j_i}(x,t) =: f_2(x,t), \\
f_3(x,t) &:= \frac{f_2(x,t)}{f_1(x,t)}\mathbb{1}\left[\left\|\frac{f_2(x,t)}{f_1(x,t)}\right\| \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}\right]
\end{aligned}$$

(exactly the same definitions as that in Lemma D.10, except for $f_1(x,t) := \tilde{f}_1(x,t) \vee N^{-(2s+3)/d}$). Then we approximate each $\alpha_i E^{(1)}_{k_i,j_i}(x,t)$ and $\alpha_i E^{(2)}_{k_i,j_i}(x,t)$ using Lemma D.8 with $\varepsilon \lesssim N'^{-1} \cdot N^{\frac{(3s+6)(2-\delta)}{\delta}} \cdot N^{-\frac{9s+10}{d}}$ and $C = m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(\sqrt{\log N})$ and aggregate them by Lemma J.3 to obtain $\phi_{\mathrm{dif},8}(x,t)$ and $\phi_{\mathrm{dif},9}(x,t)$, that approximate $f_1$ and $f_2$, respectively, and satisfy

$$|f_1(x,t) - \phi_{\mathrm{dif},8}(x,t)| \lesssim N^{-\frac{9s+3}{d}}, \quad \|f_2(x,t) - \phi_{\mathrm{dif},9}(x,t)\| \lesssim N^{-\frac{9s+10}{d}}.$$

for all $x$ with $\|x\|_\infty = \mathcal{O}(\sqrt{\log N})$. Now, we define $\phi_{\mathrm{dif},7}$ as

$$\begin{aligned}
&[\phi_{\mathrm{dif},10}(x,t)]_i := \phi_{\mathrm{clip}}(\phi_{\mathrm{mult}} \\
&\quad(\phi_{\mathrm{rec}}(\phi_{\mathrm{clip}}(\phi_{\mathrm{dif},8}(x,t); N^{-(2s+3)/d}, \mathcal{O}(1)))), [\phi_{\mathrm{dif},9}(x,t)]_i); -\mathcal{O}(\log^{\frac{1}{2}} N), \mathcal{O}(\log^{\frac{1}{2}} N)),
\end{aligned}$$

where we let $\varepsilon = N^{-(3s+4)/d}$ in Lemma J.7 for $\phi_{\mathrm{rec}}$ and we let $\varepsilon = N^{-(s+1)/d}$ and $C = N^{(2s+3)/d}$ for $\phi_{\mathrm{mult}}$ in Lemma J.6. Finally, we let

$$\phi_{\mathrm{score},2}(x,t) := \phi_{\mathrm{mult}}(\phi_{\mathrm{dif},10}(x,t), \phi_\sigma(t)).$$

where $\varepsilon = N^{-(s+1)/d}$ and $C \simeq \max\{\log^{\frac{1}{2}} N, \sigma_{\underline{T}}\} \lesssim \mathrm{poly}(N)$ in Lemma J.6 for $\phi_{\mathrm{mult}}$ and $\varepsilon = N^{-(s+1)/d}/\mathrm{poly}(N)$ in Lemma D.6 for $\phi_\sigma$. In summary, we can check that

$$\left\|\phi_{\mathrm{score},2}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\| \lesssim N^{-(s+1)/d}$$

holds for all $x$ with $\|x\|_\infty \lesssim \sqrt{\log N}$ and therefore

$$\int_{\|x\|_\infty \lesssim \sqrt{\log N}} p_t(x) \left\|\phi_{\mathrm{score},2}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\|^2 \lesssim N^{-(s+1)/d}. \tag{51}$$

Moreover, the size of $\phi_{\mathrm{score},2}$ is bounded by

$$L = \mathcal{O}(\log^4 N),\ \|W\|_\infty = \mathcal{O}(N' \log^6 N) \lesssim \mathcal{O}(N),\ S = \mathcal{O}(N' \log^8 N),\ \text{and}\ B = \exp\left(\log^4 N\right). \tag{52}$$

Now, we consider the difference between $f_3(x,t)/\sigma_t$ and $\nabla \log p_t(x)$. Its $L^2$ error in $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}$ is bounded as previously, and we finally get

$$\begin{aligned}
&\int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{1}[p_t(x) \ge N^{-\frac{2s+3}{d}}] p_t(x) \left\|\frac{f_3(x,t)}{\sigma_t} - \frac{\nabla p_t(x)}{p_t(x)}\right\|^2 \mathrm{d}x \\
&\lesssim N^{\frac{4s+6}{d}} \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} (|f_1(x,t) - p_t(x)|^2 + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|^2) \log N/\sigma_t^2 \mathrm{d}x \\
&\lesssim N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \left|\int_y \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)(p_0(y) - f_N(y))\mathrm{d}y\right|^2 \mathrm{d}x \\
&\quad + N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \left|\int_y \frac{x - m_t y}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)(p_0(y) - f_N(y))\mathrm{d}y\right|^2 \mathrm{d}x \\
&\lesssim N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \int_y \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)|p_0(y) - f_N(y)|^2 \mathrm{d}y\mathrm{d}x \\
&\quad + N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \int_y \frac{|x - m_t y|}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)|p_0(y) - f_N(y)|^2 \mathrm{d}y\mathrm{d}x \\
&\lesssim N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \int_y \int_x \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)|p_0(y) - f_N(y)|^2 \mathrm{d}x\mathrm{d}y \\
&\quad + N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \int_y \int_x \frac{|x - m_t y|}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)|p_0(y) - f_N(y)|^2 \mathrm{d}x\mathrm{d}y \\
&\lesssim N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \int_y |p_0(y) - f_N(y)|^2 \mathrm{d}y \lesssim N^{\frac{4s+6}{d}} \log N/\sigma_t^2 \cdot N^{-\frac{6s+10}{d}} \lesssim N^{-\frac{2(s+1)}{d}}/\sigma_t^2.
\end{aligned} \tag{53}$$

Here we used the result of the previous lemma for the last inequality. Eqs. (51) and (52), (53) yield the conclusion.

□

Combining Lemmas D.10 and D.12, where we use Lemma D.10 for $T_1 \le t \le T_4$ and Lemma D.12 for $T_3 \le t \le T_5$, we immediately obtain Theorem 3.1.

*Proof of Theorem 3.1.* Note that we can set $N' = N$ and $t_* = N^{-(2-\delta)/d}$ in Lemma D.12. According to Lemmas D.10 and D.12, we have two neural networks $\phi_{\mathrm{score},1}(x,t)$ and $\phi_{\mathrm{score},2}(x,t)$, that approximate the score function in $[T_1, T_4]$ and $[T_3, T_5]$. Therefore, letting $\bar{t}_1 = T_4$ and $\underline{t}_2 = T_3$ in Lemma J.5, $\phi_{\mathrm{score}}(x,t) = \phi^1_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1)\phi_{\mathrm{score},1}(x,t) + \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1)\phi_{\mathrm{score},2}(x,t)$ approximates the approximation error in $L^2(p_t)$ with an additive error of $\frac{N^{-2s/d}\log N}{\sigma_t^2}$. Realization of the multiplications ($\phi^1_{\mathrm{swit}}\phi_{\mathrm{score},1}$ and $\phi^2_{\mathrm{swit}}\phi_{\mathrm{score},2}$ and aggregation $\phi^1_{\mathrm{swit}}\phi_{\mathrm{score},1} + \phi^2_{\mathrm{swit}}\phi_{\mathrm{score},2}$ is trivial. Finally, according to Lemmas D.10 and D.12, the size of the network is bounded by

$$L = \mathcal{O}(\log^4(N)), \|W\|_\infty = \mathcal{O}(N\log^6 N), S = \mathcal{O}(N \log^8 N), \quad \text{and } B = \exp(\mathcal{O}(\log^4 N)),$$

which concludes the proof. □

We also prepare an integral form of the approximation theorems.

**Corollary D.13** (Approximation theorem)**.** *Suppose Assumptions 2.2, 2.3, 2.4 with* $a_0 = N^{-(1-\delta)/d}$*,* $N \gg 1$*,* $\underline{T} = \mathrm{poly}(N^{-1})$*, and* $\overline{T} \simeq \log N$*. Then there exists a neural network* $\phi_{\mathrm{score}} \in \Phi(L, W, S, B)$ *that satisfies*

$$\int_{t=\underline{T}}^{\overline{T}} \int_x p_t(x) \|\phi_{\mathrm{score}}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \mathrm{d}t \lesssim N^{-2s/d} \log N (\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T})).$$

*Here,* $L, \|W\|_\infty, S, B$ *is evaluated as*

$$L = \mathcal{O}(\log^4 N), \quad \|W\|_\infty = \mathcal{O}(N), \quad S = \mathcal{O}(N), \quad \text{and } B = \exp(\mathcal{O}(\log^4 N)).$$

*Moreover, suppose* $N' \geq t_*^{-d/2} N^{\delta/2}$*,* $t_* \geq N^{-(2-\delta)/d}$*, and* $\underline{T} \geq 2t_*$*, then there exists a neural network* $\phi_{\mathrm{score}} \in \Phi(L, W, S, B)$ *that satisfies*

$$\int_{t=\underline{T}}^{\overline{T}} \int_x p_t(x) \|\phi_{\mathrm{score}}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \mathrm{d}t \lesssim N^{-\frac{2(s+1)}{d}} (\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T})).$$

*Specifically,* $L = \mathcal{O}(\log^4(N)), \|W\|_\infty = \mathcal{O}(N), S = \mathcal{O}(N')$*, and* $B = \exp(\mathcal{O}(\log^4 N))$*.*

*Proof.* We only show the first part; the second part comes from Lemma D.12 in the same way. According to Theorem 3.1, there exists a network $\phi_{\mathrm{score}}$ with the desired size that satisfies

$$\int_x p_t(x) \|\phi_{\mathrm{score}}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d}} \log(N)}{\sigma_t^2}.$$

Note that $\sigma_t \gtrsim t \wedge 1$. Therefore,

$$\int_{t=\underline{T}}^{\overline{T}} \frac{N^{-\frac{2s}{d}} \log(N)}{\sigma_t^2} \mathrm{d}t \lesssim \int_{t=\underline{T}}^{\overline{T}} N^{-\frac{2s}{d}} \log(N)(1 \vee 1/t) \mathrm{d}t \leq N^{-\frac{2s}{d}} \log(N) (\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T})),$$

which gives the first part of the theorem. □

# E Generalization of the score network

## E.1 Detailed proof sketch

This section corresponds to Section 3.2. Here we provide detailed proof sketch of Theorem 3.2. We begin with the following fact (Lemma E.5; Vincent (2011)).

**Lemma E.1.** *The following holds for all* $s(x,t)$ *and* $t > 0$*:*

$$\begin{aligned} &\int_x \int_y \|s(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y) p_0(y) \mathrm{d}y \mathrm{d}x \qquad (54) \\ &= \int_x \|s(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}x + C_t. \end{aligned}$$

Here $C_t$ is a constant depending on $p_t$. According to this, minimizing the population score matching loss (54) is equivalent to minimizing the difference between the network and the score in $L^2(p_t)$.

Let us define

$$\ell_s(x) = \int_{t=\underline{T}}^{\overline{T}} \int \|s(x_t, t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x_t \mathrm{d}t,$$

so that the expected score matching loss (54) and the empirical score matching loss (1) are written as $\mathbb{E}_{x \sim p_0}[\ell(x)]$ and $\frac{1}{n}\sum_{i=1}^n \hat{\ell}(x_i)$, respectively. For the hypothesis $\mathcal{S}$ which we specify later, we define $\mathcal{L} = \{\ell_s |\ s \in \mathcal{S}\}$. Define the empirical loss minimizer $\hat{s} \in \arg\min_{s \in \mathcal{S}} \frac{1}{n} \sum_i \ell_s(x_{0,i})$. Then we can evaluate the difference between the empirical loss $\frac{1}{n}\sum_{i=1}^n \hat{\ell}(x_i)$ and the polutation loss $\mathbb{E}_{x \sim p_0}[\ell(x)]$ for $\hat{s}$, which yields Theorem 3.2.

The first term of (3) in Theorem 3.2 can be bounded by $N^{\frac{-2s/d}{\log}} N(\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T}))$, according to Corollary D.13, which is obtained from Theorem 3.1. In order to evaluate the second term in Theorem 3.2, we need to bound (i) $\|\ell\|_\infty$ uniformly over $\mathcal{L}$ and (ii) the covering number of $\mathcal{L}$.

**(i) Bounding sup-norm** According to Theorem 3.1, $\hat{s}(x,t)$ can be taken so that $\|\hat{s}(\cdot,t)\|_\infty \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}$. Thus we limit $\Phi(L,W,S,B)$ of Theorem 3.1 into

$$\mathcal{S} := \{\phi \in \Phi(L,W,S,B) | \, \|\phi(\cdot,t)\|_\infty \lesssim \frac{\log^{\frac{1}{2}} n}{\sigma_t}\}.$$

Then Appendix E.2 shows that,

$$\sup_{s\in\mathcal{S}} \sup_{x_0\in[-1,1]^d} \ell_s(x_0) \lesssim \log^2 n.$$

**(ii) Covering number evaluation** By Lemma 3 of Suzuki (2018) and the fact that $\|\ell_s\|_\infty$ is bounded by $\|s\|_\infty$ up to $\mathrm{poly}(n)$, we obtain the following.

**Lemma E.2.** *The covering number of $\mathcal{L}$ is evaluated by*

$$\log \mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty([-1,1]^d)}, \delta) \lesssim SL \log(\delta^{-1} L \|W\|_\infty Bn).$$

The proof is found in Appendix E.3. Applying this to the specified values of $L$, $\|W\|_\infty$, $S$, and $B$ in Theorem 3.1, the covering number is bounded by $\log \mathcal{N} \lesssim N(\log^{16} N + \log^{12} N \log \varepsilon^{-1})$.

Putting it all together, the second term of (3) in Theorem 3.2 can be bounded by $\lesssim N\log^2(n)(\log^{16}(N) + \log^{12}(N)\log(\varepsilon^{-1}))$. Now, (2) is bounded by

$$(2) \lesssim N^{-2s/d} \log N(\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T})) + N\log^2(n)(\log^{16}(N) + \log^{12}(N)\log(\varepsilon^{-1})) + \varepsilon.$$

Applying $N = n^{\frac{d}{d+2s}}$, $\underline{T} = \mathrm{poly}(n^{-1})$, and $\overline{T} \simeq \log n$ and setting $\varepsilon = n^{-\frac{2s}{d+2s}}$ yield

$$(2) \lesssim n^{-\frac{2s}{d+2s}} \log^2(n) + n^{-\frac{2s}{d+2s}} \log^{18}(n) + n^{-\frac{2s}{d+2s}} \lesssim n^{-\frac{2s}{d+2s}} \log^{18}(n).$$

In the following, we will first consider (i) (see Appendix E.2)and (ii) (see Appendix E.3), and then we will give the proof of Theorem 3.2 in Appendix E.4.

## E.2 BOUNDING SUP-NORM

**Lemma E.3.** *Suppose that $\|s(\cdot,t)\|_\infty = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} n)$, $\underline{T} = \mathrm{poly}(n^{-1})$ and $\overline{T} \simeq \log n$. Then, we have that*

$$\int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t \mathrm{d}t \lesssim \log^2 n.$$

*Proof.* The evaluation is mostly straightforward.

$$\begin{aligned}
&\int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t \mathrm{d}t \\
&\le 2\int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t|x_0) \mathrm{d}x \mathrm{d}t + 2\int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|\log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t \mathrm{d}t \\
&\lesssim \int_{t=\underline{T}}^{\overline{T}} \frac{\log n}{\sigma_t^2} \mathrm{d}t + \int_{t=\underline{T}}^{\overline{T}} \frac{1}{\sigma_t^2} \mathrm{d}t \\
&\lesssim \int_{t=\underline{T}}^{\overline{T}} \frac{\log n}{t \wedge 1} \mathrm{d}t \le (\log n) \cdot (\log \underline{T}^{-1} + \overline{T}) \lesssim \log^2 n
\end{aligned}$$

For the evaluation of $\int_{x_t} \|\log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t$, we used the fact that $p_t(x_t|x_0)$ is the density function of $\mathcal{N}(m_t x_0, \sigma_t^2)$. Also, we used that $\underline{T} = \mathrm{poly}(n^{-1})$ and $\overline{T} \simeq \log n$ for the last inequality. □

### E.3 COVERING NUMBER EVALUATION

**Lemma E.4** (Covering number of $\mathcal{L}$)**.** *For a neural network $s \cdot \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$, we define $\ell \cdot \mathbb{R}^d \to \mathbb{R}$ as*

$$\ell_s(x) = \int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x \mathrm{d}t.$$

*For the hypothesis network class $\mathcal{S} \in \Phi(L,W,S,B)$, we define a function class $\mathcal{L} = \{\ell_s|\ s \in \mathcal{S}\}$. If the corresponding $s$ is obvious for some $\ell_s$, we sometimes abbreviate $\ell_s$ as $\ell$.*

*Assume that $s(x,t)$ is bounded by $\|\|s(\cdot,t)\|_2\|_{L^\infty} = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} n)$ uniformly over all $s \in \mathcal{S}$ and $C \geq 1$. Then the covering number of $\mathcal{S}$ is evaluated by*

$$\log \mathcal{N}(\mathcal{S}, \|\|\cdot\|_2\|_{L^\infty([-C,C]^{d+1})}, \varepsilon) \lesssim 2SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1) C), \tag{55}$$

*and based on this, the covering number of $\mathcal{L}$ is evaluated by*

$$\log \mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty([-1,1]^d)}, \varepsilon) \lesssim SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1) n) \tag{56}$$

*when $\varepsilon^{-1}, \underline{T}^{-1}, \overline{T}, N = \mathrm{poly}(n)$.*

*Proof.* The first bound (55) is directly obtained from Suzuki (2018), with a slight modification of the input region. By following their proof, we can see that their $\varepsilon$-net for the $L^\infty([0,1]^d)$-norm serves as the $C\varepsilon$-net for the $L^\infty([-C,C]^d)$-norm. Therefore, we simply set $\varepsilon \leftarrow C^{-1}\varepsilon$ in their bound to obtain (55).

We next consider (56). First we clip the integral interval in the definition of $\ell$.

$$\begin{aligned}
&\left| \ell_s(x) - \int_{t=\underline{T}}^{\overline{T}} \int_{\|x_t\|_\infty \leq \mathcal{O}(\sqrt{\log n})} \|s(x_t,t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x_t \mathrm{d}t \right| \\
&\leq \int_{t=\underline{T}}^{\overline{T}} \int_{\|x_t\|_\infty \geq \mathcal{O}(\sqrt{\log n})} \|s(x_t,t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x_t \mathrm{d}t \mathrm{d}t \\
&\leq \|\|s(\cdot,\cdot)\|_2\|_{L^\infty}^2 \int_{t=\underline{T}}^{\overline{T}} \int_{\|x_t\|_\infty \geq \mathcal{O}(\sqrt{\log n})} p_t(x_t|x) \mathrm{d}x_t \mathrm{d}t \\
&\quad + \int_{t=\underline{T}}^{\overline{T}} \int_{\|x_t\|_\infty \geq \mathcal{O}(\sqrt{\log n})} \|\nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x_t \mathrm{d}t.
\end{aligned} \tag{57}$$

Because $p_t(x_t|x)$ is the density function of $\mathcal{N}(m_t x|\sigma_t^2)$, we can show that $\int_{\|x_t\|_\infty \geq \mathcal{O}(\sqrt{\log n})} p_t(x_t|x) \mathrm{d}x_t$ and $\int_{\|x_t\|_\infty \geq \mathcal{O}(\sqrt{\log n})} \|\nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x_t$ are bounded by $\frac{\varepsilon}{3\overline{T}(\|\|s(\cdot,\cdot)\|_2\|_{L^\infty}^2 \vee 1)}$ if $\varepsilon^{-1}, \underline{T}^{-1}, \overline{T}, N = \mathrm{poly}(n)$ and the hidden constant in $\mathcal{O}(\sqrt{\log n})$ is sufficiently large (see Lemma J.14). Therefore, (57) is bounded by

$$\|\|s(\cdot,\cdot)\|_2\|_{L^\infty}(\overline{T} - \underline{T}) \cdot \frac{\varepsilon}{3\overline{T}\|\|s(\cdot,\cdot)\|_2\|_{L^\infty}} + (\overline{T} - \underline{T}) \cdot \frac{\varepsilon}{3\overline{T}} \leq \frac{2}{3}\varepsilon. \tag{58}$$

We then take $C = \mathrm{poly}(n) \gtrsim \sqrt{\log n}$ and construct $\frac{\varepsilon}{3}$-net for a set of

$$\ell'(x) := \int_{t=\underline{T}}^{\overline{T}} \int_{\|x_t\|_\infty \leq C} \|s(x_t,t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x_t \mathrm{d}t \tag{59}$$

over all $s \in \mathcal{S}$. For this, we take $\frac{\varepsilon}{n^{\mathcal{O}(1)}}$-net of $\mathcal{S}$ with the $L^\infty([-C,C]^{d+1})$-norm. According to (55), the covering number is evaluated as

$$\log \mathcal{N}\left(\mathcal{S}, \|\|\cdot\|_2\|_{L^\infty([-C,C]^{d+1})}, \frac{\varepsilon}{n^{\mathcal{O}(1)}}\right) \lesssim 2SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1) n).$$

For different $s$ and $s'$, because $\|\nabla \log p_t(x_t|x)\| \lesssim \frac{C}{\sigma_t^2}$ for $\|x_t\|_\infty \leq C$, we have that

$$
\begin{aligned}
&|\|s(x_t,t) - \nabla \log p_t(x_t|x)\|^2 - \|s'(x_t,t) - \nabla \log p_t(x_t|x)\|^2| \qquad (60)\\
&\leq (\|s(x_t,t) - \nabla \log p_t(x_t|x)\| + \|s'(x_t,t) - \nabla \log p_t(x_t|x)\|^2)\\
&\quad \cdot |\|s(x_t,t) - \nabla \log p_t(x_t|x)\| - \|s'(x_t,t) - \nabla \log p_t(x_t|x)\||\\
&\leq (\|\|s(\cdot,\cdot)\|_2\|_{L^\infty} + \|\|s'(\cdot,\cdot)\|_2\|_{L^\infty} + 2C/\sigma_t^2) \cdot \frac{\varepsilon}{n^{\mathcal{O}(1)}}. \qquad (61)
\end{aligned}
$$

By taking the hidden constant in $\frac{\varepsilon}{n^{\mathcal{O}(1)}}$ sufficiently large, this is further bounded by $\frac{\varepsilon}{3\overline{T}(2C)^d}$ when $C, \underline{T}^{-1}, \overline{T} = \mathrm{poly}(n)$. Integrating (60) and (61) over $\int_{t=\underline{T}}^{\overline{T}} \int_{\|x_t\|_\infty \leq C} \mathrm{d}x_t \mathrm{d}t$ yields that this $\frac{\varepsilon}{n^{\mathcal{O}(1)}}$-net of $\mathcal{S}$ actually gives the $\frac{\varepsilon}{3}$-net for the set of (59); finally, we have obtained the $\varepsilon$-net for $\mathcal{L}$ together with (58). □

### E.4 Generalization error bound on the score matching loss

This subsection gives the complete proof of Theorem 3.2. First, the following relationship is useful. This shows the equivalence of explicit score matching and denoising score matching, and can be used to show that the minimizer of the empirical denoising score matching also approximately minimizes the explicit score matching loss.

**Lemma E.5** (Equivalence of explicit score matching and denoising score matching (Vincent (2011))). *The following equality holds for all $s(x_t,t)$ and $t > 0$:*

$$
\begin{aligned}
&\int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t)\|^2 p_t(x_t) \mathrm{d}x_t \\
&\quad = \int_{x_0}\int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) p_0(x_0) \mathrm{d}x_0 \mathrm{d}x_0 + C,
\end{aligned}
$$

*where* $C = \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t) \mathrm{d}x_t - \int_{x_0}\int_{x_t} \|\nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) p_0(x_0) \mathrm{d}x_t \mathrm{d}x_0.$

*Proof.* The proof follows Vincent (2011).

$$
\begin{aligned}
&\int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_t} p_t(x_t)s(x_t,t)^\top \nabla \log p_t(x_t)\mathrm{d}x \\
&\quad + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x \\
&= -2\int_{x_t} s(x_t,t)^\top \nabla p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x \\
&= -2\int_{x_t} s(x_t,t)^\top \nabla \left(\int_{x_0} p_t(x_t|x_0)p_0(x_0)\mathrm{d}x_0\right)\mathrm{d}x_t \\
&\quad + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_t} s(x_t,t)^\top \left(\int_{x_0} p_0(x_0)\nabla p_t(x_t|x_0)\mathrm{d}x_0\right)\mathrm{d}x_t \\
&\quad + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_t} p_t(x_t|y)p_0(x_0)s(x_t,t)^\top \left(\int_{x_0} \nabla \log p_t(x_t|x_0)\mathrm{d}x_0\right)\mathrm{d}x_t \\
&\quad + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)s(x_t,t)^\top \nabla \log p_t(x_t|x_0)\mathrm{d}x_t\mathrm{d}x_0 \\
&\quad + \int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)\|s(x_t,t)\|^2\mathrm{d}x_t\mathrm{d}x_0 + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= \int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)\|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2\mathrm{d}x_t\mathrm{d}x_0 + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&\quad - \int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)\|\nabla \log p_t(x_t|x_0)\|^2\mathrm{d}x_t\mathrm{d}x_0,
\end{aligned}
$$

where we used $\nabla \log p_t(x_t) = (\nabla p_t(x_t))/p_t(x_t)$ for the second, $p_t(x_t) = \int_{x_0} p_t(x_t|x_0)p_0(x_0)\mathrm{d}x_0$ for the third, $\nabla \log p_t(x_t|x_0) = (\nabla p_t(x_t|x_0))/p_t(x_t|x_0)$ for the fifth equalities.

□

Now, we evaluate the generalization error and the following theorem is a formal version of Theorem 3.2.

**Theorem E.6** (Generalization error bound based on the covering number)**.** *Let $\hat{s}$ be the minimizer of*

$$\frac{1}{n}\sum_{i=1}^{n}\int_{t=\underline{T}}^{\overline{T}}\int_x \|s(x,t) - \nabla \log p_t(x|x_i)\|_2^2 p_t(x|x_{0,i})\mathrm{d}x\mathrm{d}t, \tag{62}$$

*taking values in $\mathcal{S} \subset L^2(\mathbb{R}^d \times [\underline{T},\overline{T}])$. For each $s \in \mathcal{S}$, let $\ell(x) = \int_{t=\underline{T}}^{\overline{T}}\int_x \|s(x,t) - \nabla \log p_t(y|x)\|_2^2 p_t(y|x)\mathrm{d}y\mathrm{d}t$ and $\mathcal{L}$ be a set of $\ell$ corresponding to each $s \in \mathcal{S}$. Suppose every element $\ell \in \mathcal{L}$ satisfies $\|\ell\|_{L^\infty([-1,1]^d)} \leq C_\ell$ for some fixed $0 < C_\ell$. For an arbitrary $\varepsilon > 0$, if*

$\mathcal{N} := \mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty([-1,1]^d)}, \varepsilon) \geq 3$, *then we have that*

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x\right]$$
$$\leq 2 \inf_{s\in\mathcal{S}} \int_x \int_{\underline{T}}^{\overline{T}} \|s(x,t) - \nabla \log p_t(x)\|_2^2 p_t(x) \mathrm{d}x\mathrm{d}t + \frac{2C_\ell}{n}\left(\frac{37}{9}\log\mathcal{N} + 32\right) + 3\varepsilon.$$

*Proof.* In the following proof, $x_{0,i}$ is denoted as $x_i$ for simplicity. (62) is written as $\frac{1}{n}\sum_{i=1}^n \ell(x_i)$. Also, with $s^\circ(x,t) = \nabla \log p_t(x)$, we write

$$\begin{aligned}
R(\hat{\ell}, \ell^\circ) &:= \int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x \\
&= \int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x - \underbrace{\int_x \int_{t=\underline{T}}^{\overline{T}} \|s^\circ(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x}_{=0} \\
&= \int_y \int_{t=\underline{T}}^{\overline{T}} \int_x \|s(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y) p_0(x) \mathrm{d}y\mathrm{d}t\mathrm{d}x + C(\overline{T} - \underline{T}) \\
&\quad - \int_y \int_{t=\underline{T}}^{\overline{T}} \int_x \|s^\circ(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y) p_0(x) \mathrm{d}y\mathrm{d}t\mathrm{d}x - C(\overline{T} - \underline{T}) \\
&= \mathbb{E}_{\{x_i'\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\ell}(x_i') - \ell^\circ(x_i'))\right] \qquad (63)
\end{aligned}$$

with $\{x_i'\}_{i=1}^n$, that is an i.i.d. sample from $p_0$ and independent of $\{x_i\}_{i=1}^n$. For the second equality, we used Lemma E.5.

First, we evaluate the value of

$$D := \left|\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\ell}(x_i) - \ell^\circ(x_i))\right] - R(\hat{\ell}, \ell^\circ)\right|.$$

Using (63), we obtain

$$\begin{aligned}
D &= \left|\mathbb{E}_{x_i,x_i'}\left[\frac{1}{n}\sum_{i=1}^n ((\hat{\ell}(x_i) - \ell^\circ(x_i)) - (\hat{\ell}(x_i') - \ell^\circ(x_i')))\right]\right| \\
&\leq \frac{1}{n}\mathbb{E}_{x_i,x_i'}\left[\left|\sum_{i=1}^n ((\hat{\ell}(x_i) - \ell^\circ(x_i)) - (\hat{\ell}(x_i') - \ell^\circ(x_i')))\right|\right].
\end{aligned}$$

Let $\mathcal{L}_d = \{\ell_1, \ell_2, \cdots, \ell_N\}$ be a $\varepsilon$-covering of $\mathcal{L}$ with the minimum cardinality in the $L^\infty([-1,1]^d)$ metric. From the assumption of $N(\mathcal{L}, \|\cdot\|_\infty, \varepsilon) \geq 3$, we have $\log N \geq 1$. We define $g_j(x,x') = (\ell_j(x) - \ell^\circ(x)) - (\ell_j(x') - \ell^\circ(x'))$ and a random variable $J$ taking values in $\{1, 2, \cdots, N\}$ such that $\|\hat{\ell} - f_J\|_\infty \leq \varepsilon$, so that we have

$$\begin{aligned}
D &\leq \frac{1}{n}\mathbb{E}_{x_i,x_i'}\left[\left|\sum_{i=1}^n g_J(x_i,x_i')\right|\right] + \|(\hat{\ell}_j(x) - \ell_J(x)) - (\hat{\ell}_j(x') - \ell_J(x')\|_\infty \\
&\leq \frac{1}{n}\mathbb{E}_{x_i,x_i'}\left[\left|\sum_{i=1}^n g_J(x_i,x_i')\right|\right] + \varepsilon. \qquad (64)
\end{aligned}$$

Then we define $r_j := \max\{A, \sqrt{\mathbb{E}_{x'}[\ell_j(x') - \ell^\circ(x')]}\}$ $(j = 1, 2, \cdots, \mathcal{N})$ and a random variable

$$G := \max_{1\leq j\leq N}\left|\sum_{i=1}^n \frac{g_j(x_i,x_i')}{r_j}\right|,$$

where $A > 0$ is a constant adjusted later. Then we further evaluate (64) as

$$D \leq \frac{1}{n}\mathbb{E}_{x_i,x_i'}[r_J G] + \varepsilon \leq \frac{1}{n}\sqrt{\mathbb{E}_{x_i,x_i'}[r_J^2]\mathbb{E}_{x_i,x_i'}[G^2]} + \varepsilon \leq \frac{1}{2}\mathbb{E}_{x_i,x_i'}[r_J^2] + \frac{1}{2n^2}\mathbb{E}_{x_i,x_i'}[G^2] + \varepsilon, \tag{65}$$

by the Cauthy-Schwarz inequality and the AM-GM inequality. The definition of $J$ yields that

$$\mathbb{E}_{x_i,x_i'}[r_J^2] \leq A^2 + \mathbb{E}_{x'}[\ell_J(x') - \ell^\circ(x')] \leq A^2 + \mathbb{E}_{x'}[\hat{\ell}(x') - \ell^\circ(x')] + \varepsilon = R(\hat{\ell}, \ell^\circ) + A^2 + \varepsilon. \tag{66}$$

Because of the independence of $x_i$ and $x_i'$, we have that

$$\begin{aligned}
\mathbb{E}_{x_i,x_i'}\left[\left(\sum_{i=1}^n \frac{g_j(x_i,x_i')}{r_j}\right)^2\right] &\leq \sum_{i=1}^n \mathbb{E}_{x_i,x_i'}\left[\left(\frac{g_j(x_i,x_i')}{r_j}\right)^2\right] \\
&= \sum_{i=1}^n \left(\mathbb{E}_{x_i,x_i'}\left[\frac{(\ell_j(x_i) - \ell^\circ(x_i))^2}{r_j^2}\right] + \mathbb{E}_{x_i,x_i'}\left[\frac{(\ell_j(x_i') - \ell^\circ(x_i'))^2}{r_j^2}\right]\right) \\
&\leq 2C_\ell n
\end{aligned} \tag{67}$$

holds, where we used the fact that $g_j(x_i, x_i')$ is centered and $|\ell_j(x) - \ell^\circ(x)|$ is bounded by $C_\ell$. Also, $\frac{g_j(x_i,x_i')}{r_j}$ is bounded with $C_\ell/A$. Then, using Bernstein's inequality, we have that

$$\mathbb{P}[G^2 \geq t] = \mathbb{P}[G \geq \sqrt{t}] \leq 2\mathcal{N}\exp\left(-\frac{t}{2C_\ell(2n + \frac{\sqrt{t}}{3A})}\right),$$

for any $t \geq 0$. This gives evaluation of $\mathbb{E}_{x_i,x_i'}[G^2]$. For any $t_0 > 0$, we have that

$$\begin{aligned}
\mathbb{E}_{x_i,x_i'}[G^2] &= \int_0^\infty \mathbb{P}[G^2 \geq t]\mathrm{d}t \\
&\leq t_0 + \int_{t_0}^\infty \mathbb{P}[G^2 \geq t]\mathrm{d}t \\
&\leq t_0 + 2\mathcal{N}\int_{t_0}^\infty \exp\left(-\frac{t}{8C_\ell n}\right)\mathrm{d}t + 2\mathcal{N}\int_{t_0}^\infty \exp\left(-\frac{3A\sqrt{t}}{4C_\ell}\right)\mathrm{d}t.
\end{aligned}$$

These two integrals are computed as

$$\int_{t_0}^\infty \exp\left(-\frac{t}{8C_\ell n}\right)\mathrm{d}t = \left[-8C_\ell n\exp\left(-\frac{t}{8C_\ell n}\right)\right]_{t_0}^\infty = 8C_\ell n\exp\left(-\frac{t_0}{8C_\ell n}\right)$$

$$\begin{aligned}
\int_{t_0}^\infty \exp\left(-\frac{3A\sqrt{t}}{4C_\ell}\right)\mathrm{d}t &= \int_{t_0}^\infty \exp\left(-a\sqrt{t}\right)\mathrm{d}t \qquad (a := 3A/4C_\ell) \\
&= \left[-\frac{2(a\sqrt{t}+1)}{a^2}\exp(-a\sqrt{t})\right]_{t_0}^\infty \\
&= \frac{8C_\ell\sqrt{t_0}}{3A}\exp\left(-\frac{3A\sqrt{t_0}}{4C_\ell}\right) + \frac{32C_\ell}{9A^2}\exp\left(-\frac{3A\sqrt{t_0}}{4C_\ell}\right).
\end{aligned}$$

We take $A = \sqrt{t_0}6n$ so that

$$\begin{aligned}
\mathbb{E}_{x_i,x_i'}[G^2] &\leq t_0 + 2\mathcal{N}\left(8C_\ell n + 16C_\ell n + \frac{128C_\ell n^2}{t_0}\right)\exp\left(-\frac{t_0}{8C_\ell n}\right) \\
&\leq t_0 + 16\mathcal{N}C_\ell n(3 + 16n/t_0)\exp\left(-\frac{t_0}{8C_\ell n}\right)
\end{aligned}$$

holds. Furthermore, we take $t_0 = 8C_\ell n\log\mathcal{N}$, and then it holds that

$$\mathbb{E}_{x_i,x_i'}[G^2] \leq 8C_\ell n\left(\log\mathcal{N} + 6 + \frac{2}{C_\ell\log\mathcal{N}}\right). \tag{68}$$

Now, we combine (65), (66), (68), and $A^2 = \frac{2C_\ell \log N}{9n}$ to obtain

$$D \le \left(\frac{1}{2}R(\hat{\ell},\ell^\circ) + \frac{1}{2}A^2 + \frac{1}{2}\varepsilon\right) + \frac{4C_\ell}{n}\left(\log\mathcal{N} + 6 + \frac{2}{C_\ell \log\mathcal{N}}\right) + \varepsilon$$
$$\le \frac{1}{2}R(\hat{\ell},\ell^\circ) + \frac{C_\ell}{n}\left(\frac{37}{9}\log\mathcal{N} + 32\right) + \frac{3}{2}\varepsilon,$$

where we have used that $\log N \ge 1$. Therefore, we obtain

$$R(\hat{\ell},\ell^\circ) \le 2\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n(\hat{\ell}(x_i) - \ell^\circ(x_i))\right] + \frac{2C_\ell}{n}\left(\frac{37}{9}\log\mathcal{N} + 32\right) + 3\varepsilon. \quad (69)$$

For any fixed $\ell \in \mathcal{L}$,

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n(\hat{\ell}(x_i) - \ell^\circ(x_i))\right] \le \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n(\ell(x_i) - \ell^\circ(x_i))\right] = \mathbb{E}_x[\ell(x) - \ell^\circ(x)].$$

RHS is minimized as $\inf_{\ell\in\mathcal{L}} \mathbb{E}_x[\ell(x) - \ell^\circ(x)]$. Finally, combining this with (69), we obtain

$$R(\hat{\ell},\ell^\circ) \le 2\inf_{\ell\in\mathcal{L}} \mathbb{E}_x[\ell(x) - \ell^\circ(x)] + \frac{2C_\ell}{n}\left(\frac{37}{9}\log\mathcal{N} + 32\right) + 3\varepsilon.$$

According to Lemma E.5, we have

$$R(\hat{\ell},\ell^\circ) \le 2\inf_{s\in\mathcal{S}}\int_{\underline{T}}^{\overline{T}}\int_x \|s(x,t) - \nabla\log p_t(x)\|_2^2 p_t(x)\mathrm{d}x\mathrm{d}t + \frac{2C_\ell}{n}\left(\frac{37}{9}\log\mathcal{N} + 32\right) + 3\varepsilon.$$

□

# F ESTIMATION ERROR ANALYSIS

This section corresponds to Appendix F.

Let us define $(\bar{Y}_t)_{t=0}^{\overline{T}-\underline{T}}$, that replaces $\hat{Y}_0 \sim \mathcal{N}(0, I_d)$ in the definition of $(\hat{Y}_t)_{t=0}^{\overline{T}-\underline{T}}$ by $\bar{Y}_0 \sim p_{\underline{T}}$.

The following Girsanov theorem is useful when converting the error of the score matching to the estimation error.

**Proposition F.1** (Girsanov's Theorem (Karatzas et al., 1991))**.** *Let $p_0$ be any probability distribution, and let $Z = (Z_t)_{t\in[0,T]}, Z' = (Z'_t)_{t\in[0,T]}$ be two different processes satisfying*

$$\mathrm{d}Z_t = b(Z_t,t)\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \quad Z_0 \sim p_0,$$
$$\mathrm{d}Z'_t = b'(Z'_t,t)\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \quad Z'_0 \sim p_0.$$

*We define the distributions of $Z_t$ and $Z'_t$ as $p_t$ and $p'_t$, and the path measures of $Z$ and $Z'$ as $\mathbb{P}$ and $\mathbb{P}'$, respectively.*

*Suppose the following Novikov's condition:*

$$\mathbb{E}_{\mathbb{P}}\left[\exp\left(\int_0^T \frac{1}{2}\int_x \sigma^{-2}(t)\|(b-b')(x,t)\|^2\mathrm{d}x\mathrm{d}t\right)\right] < \infty. \quad (70)$$

*Then, the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{P}'$ is*

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}'}(Z) = \exp\left\{-\frac{1}{2}\int_0^T \sigma(t)^{-2}\|(b-b')(Z_t,t)\|^2\mathrm{d}t - \int_0^T \sigma(t)^{-1}(b-b')(Z_t,t)\mathrm{d}B_t\right\},$$

*and therefore we have that*

$$\mathrm{KL}(p_T|p'_T) \le \mathrm{KL}(\mathbb{P}|\mathbb{P}') = \int_0^T \frac{1}{2}\int_x p_t(x)\sigma(t)^{-2}\|(b-b')(x,t)\|^2\mathrm{d}x\mathrm{d}t.$$

*Moreover, Chen et al. (2022) showed that if $\int_x p_t(x)\sigma^{-2}(t)\|(b-b')(x,t)\|^2\mathrm{d}x \le C$ holds for some consant $C$ over all $t$, we have that*

$$\mathrm{KL}(p_T|p'_T) \le \int_0^T \frac{1}{2}\int_x p_t(x)\sigma(t)^2\|(b-b')(x,t)\|^2\mathrm{d}x\mathrm{d}t,$$

*even if the Novikov's condition (70) is not satisfied.*

### F.1 ESTIMATION BOUNDS IN THE TV DISTANCE

We show the upper and lower estimation rates in the total variation distance in this subsection. Let $\bar{Y}$ be $\hat{Y}$ with replacing $\hat{Y}_0 \sim \mathcal{N}(0, I_d)$ by $\bar{Y}_0 \sim p_t$. First notice that

$$\begin{aligned}\mathbb{E}[\mathrm{TV}(\mathrm{X}_0, \hat{\mathrm{Y}}_{\overline{\mathrm{T}}-\underline{\mathrm{T}}})] &\lesssim \mathbb{E}[\mathrm{TV}(\mathrm{Y}_{\overline{\mathrm{T}}}, \mathrm{Y}_{\overline{\mathrm{T}}-\underline{\mathrm{T}}})] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\ &\lesssim \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{\mathrm{T}}}) + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0)] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\ &= \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{\mathrm{T}}}) + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d))] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \end{aligned} \tag{71}$$

Here, $\mathbb{E}[\mathrm{TV}(\mathrm{Y}_{\overline{\mathrm{T}}}, \mathrm{Y}_{\overline{\mathrm{T}}-\underline{\mathrm{T}}})] = \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{\mathrm{T}}}) + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0)]$ follows from the correspondence between the forward and backward processes, and $\mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] \le \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0)]$ follows from the definitions of $\hat{Y}$ and $\bar{Y}$ (the only difference is the initial distribution.). We then bound the three terms in (71) in a row. We begin with the first term.

**Theorem F.2.** *We have that*

$$\mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{\mathrm{T}}}) \lesssim \sqrt{\underline{T}} n^{\mathcal{O}(1)}$$

*for* $\underline{T} \lesssim n^{-\mathcal{O}(1)}$. *Therefore, by taking* $\underline{T} \lesssim n^{-\mathcal{O}(1)}$, *we have that* $\mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{\mathrm{T}}}) \lesssim n^{-s/(d+2s)}$.

*Proof.* We need to evaluate $\|p_0 - p_{\underline{T}}\|_{L_1}$. When $p_0$ is Lipschitz continous, an intuitive proof strategy is as follows: For small $t$, $p_t(x)$ is an average of $p_0(y)$ nearby $x$. Because of the Lipshitzness, $p_0(x)$ and $p_0(y)$ with $|x - y| \ll 1$ are close, and therefore $p_0(x)$ and $p_t(x)$ are close. However, our setting also includes the not continous functions. To consider these cases in a uniform manner, we approximate $p_0$ with the B-spline basis decomposition because each B-spline basis is a Lipschitz function.

Remember that $p_0$ is decomposed as

$$f_N(x) = \sum_{i=1}^{N} \alpha_i \mathbb{1}[\|x\|_\infty \le 1] M^d_{k_i, j_i}(x)$$

in Lemma J.13, where $\|k\|_\infty \le K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ for $\delta = d(1/p - 1)_+$ and $\nu = (2s - \delta)/(2\delta)$, and $\|p_0 - f_N\|_{L^1([-1,1]^d)} \lesssim N^{-s/d} \simeq n^{-s/(2s+d)}$ holds. Because we take $N = n^{d/(2s+d)} = n^{\mathcal{O}(1)}$, we can say that each $M^d_{k_i, j_i}(x)$ is $n^{\mathcal{O}(1)}$-Lipschitz. Moreover, $|\alpha_i| \lesssim N^{(\nu^{-1} + d^{-1})(d/p - s)} = n^{\mathcal{O}(1)}$. Therefore, $f_N$ is $n^{\mathcal{O}(1)}$-Lipschitz.

We decompose $p_0$ as $p_0 = f_N + (p_0 - f_N)$ using the above $f_N$. Then we have that

$$\begin{aligned} &\left| p_{\underline{T}}(x) - \int \frac{f_N(y)}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \right| \\ &= \left| \int \frac{(p_0(y) - f_N(y))}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \right| \\ &\le \int \frac{|p_0(y) - f_N(y)|}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y. \end{aligned} \tag{72}$$

Integrating this over all $x$ yields that

$$\begin{aligned} &\int \left| p_{\underline{T}}(x) - \int \frac{f_N(y)}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \right| \mathrm{d}x \\ &\le \int \int \frac{|p_0(y) - f_N(y)|}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \mathrm{d}x \\ &= \int |p_0(y) - f_N(y)| \int \frac{1}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left( -\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}x \mathrm{d}y \\ &\le \int |p_0(y) - f_N(y)| \, \mathrm{d}y = \|p_0 - f_N\|_{L^1([-1,1]^d)}. \end{aligned}$$

Thus, $\|p_0 - p_{\underline{T}}\|_{L_1}$ is upper bounded by

$$\|p_0 - f_N\|_{L^1([-1,1]^d)} + \underbrace{\int \left| f_N(x) - \int \frac{f_N(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \right| \mathrm{d}x}_{\text{if } f_N \text{ is replaced by } p_0\text{, this is equal to } \|p_0 - p_t\|_{L_1}}$$
$$+ \underbrace{\|p_0 - f_N\|_{L^1([-1,1]^d)}}_{(72)}. \tag{73}$$

Because $\|p_0 - f_N\|_{L^1([-1,1]^d)}$ is bounded by $n^{-s/(2s+d)}$, we focus on the second term.

Note that at each $x$,

$$\left| \int \frac{f_N(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - \int_{A^x} \frac{f_N(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \right| \lesssim n^{-s/(d+2s)}, \tag{74}$$

where $A^x = \prod_{i=1}^d a_i^x$ with $a_i^x = [\frac{x_i}{m_{\underline{T}}} - \frac{\sigma_{\underline{T}} \mathcal{O}(1)}{m_{\underline{T}}} \sqrt{\log n}, \frac{x_i}{m_{\underline{T}}} + \frac{\sigma_{\underline{T}} \mathcal{O}(1)}{m_{\underline{T}}} \sqrt{\log n}]$, according to Lemma J.10. Because $\sigma_{\underline{T}} = \mathcal{O}(\sqrt{\underline{T}})$ and $m_{\underline{T}} = \mathcal{O}(1)$ for sufficiently small $\underline{T}$, the value of $p_{\underline{T}}(x)$ is almost determined by the value from points that is only $\mathcal{O}(\sqrt{\underline{T} \log n})$ away from $x$. Because of the Lipschitzness of $p_0$, for each $x \in [-m_{\underline{T}} - \mathcal{O}(\sqrt{\underline{T} \log n}), m_{\underline{T}} + \mathcal{O}(\sqrt{\underline{T} \log n})]^d$,

$$\left| \int_{A^x} \frac{f_N(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - \int_{A^x} \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \right| \tag{75}$$
$$\leq n^{\mathcal{O}(1)} \cdot \sqrt{\underline{T} \log n}.$$

where we used the Lipshitzness of $f_N$. By taking $\underline{T}$ polynomially small w.r.t. $n$, we have that (75) $\lesssim n^{-s/(d+2s)}$. Moreover,

$$\begin{aligned} &\left| \int_{A^x} \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - f_N(x) \right| \\ &= \left| \int_{A^x} \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - \int \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \right| \\ &\lesssim n^{-s/(d+2s)}, \end{aligned} \tag{76}$$

again with Lemma J.10.

Therefore, combining (73), (74), (75), and (76), we obtain that

$$\|p_0 - p_{\underline{T}}\|_{L_1} \lesssim \sqrt{\underline{T}} n^{\mathcal{O}(1)} \lesssim n^{-s/(d+2s)}.$$

for $\underline{T} = n^{-\mathcal{O}(1)}$. □

We next consider the second term.

**Lemma F.3.** *We can bound* $\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d))$ *as follows.*

$$\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d)) \lesssim \exp(-\underline{\beta}\overline{T}).$$

*Proof.* Exponential convergence of the Ornstein–Ulhenbeck process (Bakry et al., 2014) yields that

$$\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d)) \lesssim \sqrt{\mathrm{KL}(p_{\overline{T}} \| \mathcal{N}(0, I_d))} \leq \exp(-\underline{\beta}\overline{T}) \sqrt{\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d))} \lesssim \exp(-\underline{\beta}\overline{T}),$$

because $C_f^{-1} \leq p_0 \leq C_f$ holds and the density of $\mathcal{N}(0, I_d)$ is lower bounded by $\gtrsim 1$ in $\mathrm{supp}(p_0) = [-1, 1]^d$, which means that $\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d)) = \mathcal{O}(1)$. □

Therefore, by setting $\overline{T} = \frac{s \log n}{\underline{\beta}(d+2s)}$, the second term is bounded by $n^{-s/(d+2s)}$.

The third term $\mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})]$ in (71) is bounded by Girsanov's theorem Proposition F.1 and the generalization error bound from Section 3.2:

$$\begin{aligned}\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}}) &\lesssim \mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \sqrt{\int_{t=\underline{T}}^{\overline{T}} p_t(x) \beta_t^{-2} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \mathrm{d}t} \\ &\lesssim \sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \int_{t=\underline{T}}^{\overline{T}} p_t(x) \beta_t^{-2} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \mathrm{d}t} \\ &\lesssim \sqrt{n^{-\frac{2s}{d+2s}} \log^{18} n} \\ &\lesssim n^{-\frac{s}{d+2s}} \log^9 n.\end{aligned}$$

Therefore, all three terms in (71) are bounded as above and the first part of Theorem 3.3 follows. We also show the lower bound as follows. This is the rephrased version of Proposition 3.4.

**Proposition F.4.** *Assume that* $0 < p, q \leq \infty$, $s > 0$, *and*

$$s > \left\{ d\left(\frac{1}{p} - \frac{1}{2}\right), d\left(\frac{1}{p} - 1\right), 0 \right\}$$

*holds. Then, we have that*

$$\inf_{\hat{\mu}} \sup_{p \in B^s_{p,q}([-1,1]^d)} \mathbb{E}[\mathrm{TV}(\hat{\mu}, p)] \gtrsim n^{-s/(d+2s)},$$

*where the expectation is with respect to the sample, and the infimum is taken over all estimators based on* $n$ *observations.*

*Proof.* Theorem 10 of Triebel (2011) showed that, for a bounded domain $\Omega \subset \mathbb{R}^d$,

$$\log N(U(B^s_{p,q}(\Omega)), \|\cdot\|_r, \varepsilon) \simeq \varepsilon^{-d/s}, \tag{77}$$

for $0 < p, q \leq \infty, 1 \leq r < \infty$, and $s > 0$ that satisfy

$$s > \max\left\{ d\left(\frac{1}{p} - \frac{1}{r}\right), d\left(\frac{1}{p} - 1\right), 0 \right\}.$$

Although they considered all Besov functions that does not satisfy $\int f \mathrm{d}\mu = 1$, we can check by following their proof that bounding the functions does not harm the order of the entropy number. Now we use Theorem 4 of Yang & Barron (1999). Note that the equivalence of the covering number and the entropy holds because $\|\cdot\|_r$ is a distance, and therefore (77) is transferred to the entropy. The condition 2 of the theorem is checked directly from (77). Moreover, the condition 3 holds if we take $f_*(x) = 1/2^d$ $(x \in [-1,1]^d), 0$ (otherwise) for all $\alpha \in (0,1)$. Finally, if $s > \left\{ d(\frac{1}{p} - \frac{1}{2}), d(\frac{1}{p} - 1), 0 \right\}$, $\log N(U(B^s_{p,q}(\Omega)), \|\cdot\|_2, \varepsilon) \simeq \log N(U(B^s_{p,q}(\Omega)), \|\cdot\|_1, \varepsilon)$ holds. Therefore, Theorem 4 (i) of Yang & Barron (1999) is applied, and we get

$$\min_{\hat{\mu}} \max_{p \in B^s_{p,q}} \mathbb{E}[\|\hat{\mu} - p\|_1] \simeq \varepsilon_n,$$

where $\varepsilon_n$ is chosen as $\log N(U(B^s_{p,q}(\Omega)), \|\cdot\|_r, \varepsilon_n) = n\varepsilon_n^2$ holds. Together with (77), we obtain the assertion. ☐

### F.2 Estimation rate in the $W_1$ distance

**Switching score networks** First, let us explain our proof sketch. Theorem 3.3 directly yields the convergence rate of $n^{-s/(2s+d)} \log^9 n$. However, it is known from Niles-Weed & Berthet (2022) that the minimax estimation rate in $W_1$ is faster than this. Thus, this approach yields the sub-optimal rate. To overcome this issue, let us carefully consider where we lose the estimation rate, going back to the approximation error analysis in the previous subsection. Although we used Theorem 3.1 for all $\underline{T} \le t \le \overline{T}$, Lemma D.5 tells us that if $t \gtrsim N^{-\frac{2-\delta}{d}} \simeq n^{-\frac{2-\delta}{2s+d}}$, we can make the approximation error smaller than $\frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^{-2}} = \frac{n^{-\frac{2(s+1)}{d+2s}}}{\sigma_t^{-2}}$ with a smaller network of size $N' \le N$. This means that we have used a sub-optimal network for $t \gtrsim n^{-\frac{2-\delta}{d+2s}}$ in terms of both approximation and generalization errors.

Based on this discussion, we divide the time into $t_0 = \underline{T} < t_1 = 2n^{-\frac{2-\delta}{d+2s}} < \cdots < t_{K_*} = \overline{T} - \underline{T}$ with $t_{i+1}/t_i = \text{const.} \le 2$ $(i \ge 1)$. The number of intervals amounts to $K_* = \mathcal{O}(\log n)$. We consider to train a tailored network for each time interval $[t_i, t_{i+1}]$ and to switch them for different intervals. Lemma D.5 yields that for $i \ge 1$ these exists a network $s_i \in \Phi(L_i, W_i, S_i, W_i)$ such that

$$\mathbb{E}_{x\sim p_t}[\|s_i(x,t) - \nabla \log p_t(x)\|^2] \lesssim \frac{n^{-\frac{2(s+1)}{d+2s}}}{\sigma_t^2} \ (t \in [t_i, t_{i+1}]),$$

with $L = \mathcal{O}(\log^4(N))$, $\|W\|_\infty = \mathcal{O}(N)$, $S = \mathcal{O}(t_i^{-d/2} N^{\delta/2})$, and $B = \exp(\mathcal{O}(\log^4 N))$. Therefore, we choose a sequence of score networks $\hat{s}_i$ so that $\hat{s}_i$ minimizes the score matching loss restricted to $[t_i, t_{i+1}]$:

$$\frac{1}{n}\sum_{j=1}^{n} \mathop{\mathbb{E}}_{\substack{t\sim \mathrm{Unif}[t_i, t_{j+1}] \\ x_t \sim p_t(x_t|x_{0,j})}} [\|s(x_t, t) - \nabla \log p_t(x_t|x_{0,j})\|^2].$$

In other words, we let $\hat{s}(x,t) := \hat{s}_i(x,t)$ for $t \in [t_i, t_{i+1}]$.

Similarly to Theorem 3.2, Theorem E.6 yields that the following generalization error bound for $i \ge 1$:

$$\mathbb{E}_{\{x_{0,j}\}_{i=j}^n}\left[\int_{t=t_i}^{t_{i+1}} \mathbb{E}_x\left[\|\hat{s}_i(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right] \le \left(n^{-\frac{2(s+1)}{d+2s}} + \frac{t_i^{-d/2} n^{\frac{\delta d}{(d+2s)}}}{n}\right) \cdot \underbrace{\tilde{\mathcal{O}}(t_i/\sigma_{t_i}^2)}_{=\tilde{\mathcal{O}}(1)}. \tag{78}$$

For $t \lesssim n^{-\frac{2-\delta}{d+2s}}$, we use a network trained via the score matching loss restricted to $[t_i, t_{i+1}]$. Thus, (78) for $i = 0$ is bounded by $\tilde{\mathcal{O}}(n^{-\frac{2s}{d+2s}})$ similarly to Section 3.2.

One may think that the above improvement would be useless because the error caused at $t \le n^{-\frac{2-\delta}{d+2s}}$ has the $n^{-2s/(d+2s)}$ rate and dominates the estimation error. However, another important observation is that the Wasserstain distance is a transportation distance. The score estimation error at time closer to $t = 0$ less contributes to the estimation error, because the distance how much each path evolves is small from that time. As we will see, the idea of improving accuracy for large $t$ indeed yields the minimax optimal rate in $W_1$.

To utilize this observation, let us consider a sequence of stochastic processes. Let $(Y_t)_{[0,\overline{T}]} = (\bar{Y}_t^{(0)})_{[0,\overline{T}]}$, and for $i \ge 1$, let $(\bar{Y}^{(i)})_{[0,\overline{T}]}$ be a stochastic process which uses the true score during $[0, \overline{T} - t_i]$ and the estimated score $\hat{s}$ during $[\overline{T} - t_i, \overline{T} - \underline{T}]$, and $\bar{Y}_0^{(i)} \sim p_{\overline{T}}$. Then, we have that

$$\begin{aligned}\mathbb{E}[W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}})] &\le \mathbb{E}[W_1(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\ &\le \mathbb{E}[W_1(X_0, X_{\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})].\end{aligned} \tag{79}$$

The first term is bounded by $\sqrt{\underline{T}}$ due to Lemma F.7 and the second term is bounded by $\exp(-\overline{T})$ due to Lemma F.8. The last term $\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})]$ is upper bounded by $\sum_{i=1}^{K_*} \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)}, \bar{Y}_{\overline{T}-\underline{T}}^{(i)})]$. Then, we use the following lemma, an informal version of Lemma F.9.

**Lemma F.5.** *For $i = 1, 2, \cdots, K_*$, we have that*

$$W_1(\hat{Y}_{\overline{T}-\underline{T}}^{(i-1)}, \hat{Y}_{\overline{T}-\underline{T}}^{(i)}) \le \tilde{\mathcal{O}}(1) \cdot \sqrt{t_{i-1}\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=t_{i-1}}^{t_i} \mathbb{E}_x\left[\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right]}.$$

RHS is decomposed to the two factors: the score matching loss during $[t_{i-1}, t_i]$ and $\sqrt{t_i}$. The latter corresponds to how much $Y_t$ moves from $t = \overline{T} - t_i$ to $\overline{T} - \underline{T}$. This bound represents that, as $t_i \to 0$, while score matching gets more difficult, its contribution to the $W_1$ error is reduced. The formal proof requires construction of a path-wise transportation map; see the proof for Lemma F.9.

Putting it all together, we finally yields Theorem 3.5, the nearly minimax optimal rate in $W_1$. Specifically, if we ignore logarithmic factors, (79) is bounded by

$$\sqrt{\underline{T}} + \exp(-\overline{T}) + \sqrt{t_0} n^{-\frac{2s}{d+2s}} + \sum_{i=2}^{K_*} \sqrt{t_i} \sqrt{n^{-\frac{2(s+1)}{d+2s}} + \frac{t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}}}{n}} \lesssim n^{-\frac{s+1-\delta}{d+2s}},$$

where we set $\underline{T} = n^{-\frac{2(s+1)}{d+2s}}$ and $\overline{T} = \frac{(s+1)\log n}{\underline{\beta}(d+2s)}$.

**Remark F.6.** Although we used differently optimized multiple networks, it is also possible that such modification is implicitly made in reality. The first evidence is *implicit reguralization*, where sparsify of the solution is induced by learning procedures (Gunasekar et al., 2017; Arora et al., 2019; Soudry et al., 2018). When the sub-networks for differnt time intervals are learned in parallel via the score matching at once (1), these theory suggests the good score network is obtained without explicit regularization like our switching procedure. Another evidence is that in practice the weight function $\lambda(t)$ sometimes increases as $t$ gets large (Song & Ermon, 2019; Song et al., 2020), suggesting that the quality of the score network at larger $t$ is more emphasized.

Now we proceed to the main part of the proof. First, we bound the first term of (79).

**Lemma F.7** (Section 4.3 of De Bortoli (2022))**.** *We can bound $W_1(X_0, X_{\underline{T}})$ as follows.*

$$W_1(X_0, X_{\underline{T}}) \lesssim \sqrt{\underline{T}}$$

*Proof.* Let $X \sim p_0$ and $Z \sim N(0, I_d)$. Then,

$$\begin{aligned} W_1(X_0, X_{\underline{T}}) &\le \mathbb{E}[\|X - m_{T_1} X + \sigma_{T_1} Z\|] \le (1 - m_{\underline{T}})\mathbb{E}[\|X\|] + \sigma_{\underline{T}} \mathbb{E}[\|Z\|] \\ &\le (1 - m_{\underline{T}})\sqrt{d} + \sigma_{\underline{T}}\sqrt{d} \lesssim \sqrt{\underline{T}}, \end{aligned}$$

which concludes the proof. □

Next, we bound the second term of (79).

**Lemma F.8.** *We can bound $\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})]$ as follows.*

$$\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim \mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) \lesssim \exp(-\underline{\beta}\overline{T}).$$

*Proof.* Exponential convergence of the Ornstein–Ulhenbeck process (Bakry et al., 2014) yields that

$$\begin{aligned} &\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) \\ &= \mathrm{TV}(p_{\overline{T}}, \mathcal{N}(0, I_d)) \le \sqrt{2\mathrm{KL}(p_{\overline{T}} \| \mathcal{N}(0, I_d))} \le 2\exp(-\overline{T}\underline{\beta})\sqrt{\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d))} \lesssim \exp(-\underline{\beta}\overline{T}), \end{aligned}$$

because $C_f^{-1} \le p_0 \le C_f$ holds and the density of $\mathcal{N}(0, I_d)$ is lower bounded by $\mathcal{O}(1)$ in $\mathrm{supp}(p_0) = [-1, 1]^d$, which means $\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d)) = \mathcal{O}(1)$. In addition because $\|\hat{Y}^{(k)}_{\overline{T}-\underline{T}}\|_\infty, \|\hat{Y}_{\overline{T}-\underline{T}}\|_\infty \le 2 = \mathcal{O}(1)$, and because the only difference between $\hat{Y}^{(k)}$ and $\hat{Y}$ is the initial distribution, we have $W_1(\hat{Y}^{(k)}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}}) \lesssim \mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) = \mathrm{TV}(p_{\overline{T}}, \mathcal{N}(0, I_d))$. Putting it all together, we obtain that

$$W_1(\hat{Y}^{(k)}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}}) \lesssim \mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) = \mathrm{TV}(p_{\overline{T}}, \mathcal{N}(0, I_d)) \lesssim \exp(-\underline{\beta}\overline{T}),$$

which yields the assertion. □

Finally, we bound the third term of (79). As we sketched in the first part of this subsection,

$$\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \leq \sum_{i=1}^{K_*} \mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})]. \tag{80}$$

We define a sequence of stochastic processes $\{(\bar{Y}^{(i)}_t)_{t=0}^{\overline{T}-\underline{T}}\}_{i=0}^{K_*}$. First, $\bar{Y}^{(0)} = (\bar{Y}^{(0)}_t)_{t\in[0,\overline{T}]} = Y = (Y_t)_{t\in[0,\overline{T}]}$ is defined as a process such that

$$\mathrm{d}Y_t = \beta_{\overline{T}-t}(Y_t + 2\nabla \log p_t(Y_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [0, \overline{T}]), \quad Y^{(0)}_0 \sim p_{\overline{T}}.$$

Then, $Y_{\overline{T}-t} \sim p_t$ holds for all $t \in [0, \overline{T}]$. Next, for $i = 1, 2, \cdots, K_*$, we let $\bar{Y}^{(i)} = (\bar{Y}^{(i)}_t)_{t\in[0,\overline{T}-\underline{T}]}$ to satisfy

$$\begin{aligned}\bar{Y}^{(i)}_0 \sim p_{\overline{T}}, \quad &\mathrm{d}\bar{Y}^{(i)}_t = \beta_{\overline{T}-t}(\bar{Y}^{(i)}_t + 2\nabla \log p_t(\bar{Y}^{(i)}_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [0, \overline{T}-t_i]),\\ &\mathrm{d}\bar{Y}^{(i)}_t = \beta_{\overline{T}-t}(\bar{Y}^{(i)}_t + 2\hat{s}(\bar{Y}^{(i)}_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_i, \overline{T}-\underline{T}]).\end{aligned}$$

Note that $t_0 = \underline{T}$, $t_1 = N^{-\frac{2-\delta}{d}} = n^{-\frac{2-\delta}{d+2s}}$, $1 < \frac{t_{i+1}}{t_i} = \text{const.} \leq 2$, and $t_{K_*} = \overline{T} - \underline{T}$. Then, $\bar{Y}^{(K_*)} = \bar{Y}$ holds. Here $\bar{Y}^{(i)}_{\overline{T}-t} \sim p_t$ holds for all $t \in [0, \overline{T}-t_i]$, but after $t = \overline{T}-t_i$, the true score function is replaced by the estimated one. If $\|\bar{Y}^{(i)}_{\overline{T}-\underline{T}}\|_\infty > 2$ in the original definition, we reset $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$ as $\bar{Y}^{(i)}_{\overline{T}-\underline{T}} := 0$.

Also, we introduce another stochastic process $\bar{Y}^{(i)'}$. We define $d+1$-dimensional set $A \subseteq \mathbb{R}^{d+1}$ as

$$A = \left\{(x, t) \in \mathbb{R}^d \times \mathbb{R} \,\middle|\, \|x\|_\infty \leq m_t + C_{\mathrm{a},1}\sigma_t\sqrt{\log(n)},\ \underline{T} \leq t \leq \overline{T}\right\}.$$

According to Lemma C.1, with probability at least $1 - n^{-\mathcal{O}(1)}$, a path of the backward process $(Y_t)_{t=0}^{\overline{T}}$ satisfies $(Y_t, \overline{T}-t) \in A$ for all $\underline{T} \leq t \leq \overline{T}$. Based on this, for $i = 0, 1, \cdots, K_* - 1$, $\bar{Y}^{(i)'}$ is defined as

$$\begin{aligned}&\bar{Y}^{(i)'}_0 \sim p_{\overline{T}},\\ &\mathrm{d}\bar{Y}^{(i)'}_t = \beta_{\overline{T}-t}(\bar{Y}^{(i)'}_t + 2\nabla \log p_t(\bar{Y}^{(i)'}_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [0, \overline{T}-t_i]),\\ &\mathrm{d}\bar{Y}^{(i)'}_t = \beta_{\overline{T}-t}\Big(\bar{Y}^{(i)'}_t + 2\mathbb{1}[(\bar{Y}^{(i)'}_s, \overline{T}-s) \notin A \text{ for some } s \leq t]\nabla \log p_t(\bar{Y}^{(i)'}_t)\\ &\quad + 2\mathbb{1}[(\bar{Y}^{(i)'}_s, \overline{T}-s) \in A \text{ for all } s \leq t]\hat{s}(\bar{Y}^{(i)'}_t, \overline{T}-t)\Big)\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_{i+1}, \overline{T}-t_i]),\\ &\mathrm{d}\bar{Y}^{(i)'}_t = \beta_{\overline{T}-t}(\bar{Y}^{(i)'}_t + 2\hat{s}(\bar{Y}^{(i)'}_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_i, \overline{T}-\underline{T}]).\end{aligned}$$

**Lemma F.9.** *Suppose that $\|\hat{s}(\cdot, t)\|_\infty \lesssim \frac{\log^{\frac{1}{2}} n}{\sqrt{t}\wedge 1}$ holds. Then, the following holds for all $i = 1, 2, \cdots, K_*$:*

$$W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}}) \lesssim \sqrt{t_i \log n}\sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=t_{i-1}}^{t_i} \mathbb{E}_x\left[\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right]} + n^{-\frac{s+1}{d+2s}}. \tag{81}$$

*Therefore, we have that*

$$\begin{aligned}&\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})]\\ &\lesssim \sqrt{t_i \log n}\sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=t_{i-1}}^{t_i} \mathbb{E}_x\left[\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right]} + n^{-\frac{s+1}{d+2s}}.\end{aligned} \tag{82}$$

*Proof.* We construct the transportation map between $\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ and $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$. Our approach focuses on each path.

Because the Novikov's condition is not satisfied for $\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ and $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$, Proposition F.1 cannot be used to consider the total variation distance between the two paths; Proposition F.1 only gives $\mathrm{KL}(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})$, not $\mathrm{KL}(\bar{Y}^{(i-1)}, \bar{Y}^{(i)}$, and this bound is insufficient for our discussion. Therefore, we first bound $\mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}})]$. According to Lemma C.1, with probability at least $1-n^{-\mathcal{O}(1)}$, a path of the processes $(\bar{Y}^{(i-1)}_t)_{t=0}^{\overline{T}}$ and $(\bar{Y}^{(i-1)'}_t)_{t=0}^{\overline{T}}$ satisfy $(\bar{Y}^{(i-1)}_t, \overline{T}-t), (\bar{Y}^{(i-1)'}_t, \overline{T}-t) \in A$ for all $0 \le t \le \overline{T}-t_{i-1}$. Thus, $\mathbb{E}[\mathrm{TV}(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}})]$ is bounded by $n^{-\mathcal{O}(1)}$ (with a sufficiently large constant in $\mathcal{O}(1)$.). This implies $\mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}})] \lesssim n^{-\mathcal{O}(1)}$, because $\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}} = \mathcal{O}(1)$ (a.s.).

We now discuss $\mathbb{E}[W_1(\bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})]$. Let us write the path measures of $\bar{Y}^{(i-1)'}$ and $\bar{Y}^{(i)}$ be $\mathbb{P}$ and $\mathbb{P}'$, and take some path $p$ that is $y$ at $t=\overline{T}-\underline{T}$ and is $z$ at $t=\overline{T}-t_i$. If $\mathrm{d}\mathbb{P}[p] > \mathrm{d}\mathbb{P}'[p]$, then we move the mass of $\bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}} = y$ that amounts to $\mathrm{d}\mathbb{P}[p]-\mathrm{d}\mathbb{P}'[p]$, to $z$, along the path $p$ by reversing the time until $t=\overline{T}-t_i$. Applying this to all paths $p$, then the total mass of $\bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}}$ that is moved is at most

$$\frac{1}{2}\mathrm{TV}((\bar{Y}^{(i-1)'}), (\bar{Y}^{(i)})) \le \frac{1}{2}\sqrt{\int_{t=t_{i-1}}^{t_i}\int_x p_t(x)\beta_t^{-2}\|\hat{s}(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}x\mathrm{d}t}. \tag{83}$$

according to Proposition F.1. Here we remark that the Novikov's condition certainly holds for this case.

Until now, a part of the mass of $\hat{Y}^{(i-1)'}_{\overline{T}-\underline{T}}$ is moved along each corresponding path, but at this time no coupling measure has been constructed. To realize the coupling measure, we consider the same process for $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$. That is, for each path $p$ with $\bar{Y}^{(i)}_{\overline{T}-\underline{T}} = y$ and $\bar{Y}^{(i)}_{\overline{T}-t_i} = z$, if $\mathrm{d}\mathbb{P}[p] < \mathrm{d}\mathbb{P}'[p]$, then we move the mass of $\bar{Y}^{(i)}_{\overline{T}-\underline{T}} = y$, as much as $\mathrm{d}\mathbb{P}'[p]-\mathrm{d}\mathbb{P}[p]$, to $z$ along the path $p$. The total mass of $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$ affected is bounded by $\frac{1}{2}\mathrm{TV}((\bar{Y}^{(i-1)'}), (\bar{Y}^{(i)'}))$, which is bounded by (83).

Now, we can see that, the same amount of mass is transported from both $\bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}}$ and $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$ to $t=\overline{T}-t_i$. Thus, at each $z$, we can arbitrarily associate the mass from $\bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}}$ to that from $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$. Using this, as much as $\frac{1}{2}\mathrm{TV}((\bar{Y}^{(i-1)'}), (\bar{Y}^{(i)'}))$ of the mass is transported from $\bar{Y}^{(i-1)'}_{\overline{T}-\underline{T}}$ to $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$, by reversing the path to $t=\overline{T}-t_i$.

Now our interest is how far each transport is required to move on average. First we consider when $t_i \lesssim 1$.

First we bound $\|\bar{Y}^{(i)}_{\overline{T}-\underline{T}} - \bar{Y}^{(i)}_{\overline{T}-t_i}\|$. According to Lemma C.1, we have $\|\int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} 2\beta_{\overline{T}-t}\mathrm{d}B_t\| \lesssim \sqrt{t_i\log n}$ for all $t\in[\overline{T}-t_i, \overline{T}-\underline{T}]$, and $\bar{Y}^{(i)}_{\overline{T}-t_i} \lesssim m_{\overline{T}-t_i} + \sigma_{\overline{T}-t_i}\sqrt{\log n} \lesssim \sqrt{\log n}$ with probability $1-n^{-\mathcal{O}(1)}$. We consider the event conditioned on them. Note that $\|s(x,t)\| \lesssim \frac{\sqrt{\log n}}{\sigma_t} \lesssim \frac{\sqrt{\log n}}{\sqrt{t}}$

holds. Then we have that, for all $\overline{T} - t_i \le t \le \overline{T} - \underline{T}$,

$$\begin{aligned}
\|\bar{Y}_t^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\| &= \left\| \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \beta_{\overline{T}-s}(\bar{Y}_s^{(i)} + 2\nabla \log p_t(\bar{Y}_s^{(i)}, \overline{T}-s))\mathrm{d}t + \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \sqrt{2\beta_{\overline{T}-s}}\mathrm{d}B_s \right\| \\
&\lesssim \overline{\beta} \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)}\|\mathrm{d}s + 2\overline{\beta} \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \frac{\sqrt{\log n}}{\sqrt{s}}\mathrm{d}s + \sqrt{t_i \log n}, \\
&\lesssim \overline{\beta} \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)}\|\mathrm{d}s + \sqrt{t_i \log n} + \sqrt{t_i \log n}. \\
&\lesssim \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\|\mathrm{d}s + \sqrt{t_i \log n} + t_i\|\bar{Y}_{\overline{T}-t_i}^{(i)}\| \\
&\lesssim \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\|\mathrm{d}s + \sqrt{t_i \log n} + t_i\sqrt{\log n}
\end{aligned}$$

Now we apply the Gronwall's inequality to obtain

$$\|\bar{Y}_t^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\| \lesssim e^{\overline{\beta} t_i}\sqrt{t_i \log n} \lesssim \sqrt{t_i \log n}.$$

for all $\overline{T} - t_i \le t \le \overline{T} - \underline{T}$. Thus, with probability $1 - n^{-\mathcal{O}(1)}$, $\|\bar{Y}_t^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\|$ is bounded by $\sqrt{t_i \log n}$ up to a constant factor, over all $\overline{T} - t_i \le t \le \overline{T} - \underline{T}$.

Next we bound $\|\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)'} - \bar{Y}_{\overline{T}-t_i}^{(i-1)'}\|$. This is decomposed into

$$\|\bar{Y}_{\overline{T}-t_i}^{(i-1)'} - \bar{Y}_{\overline{T}-t_{i-1}}^{(i-1)'}\| + \|\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)'} - \bar{Y}_{\overline{T}-t_{i-1}}^{(i-1)'}\|.$$

The first term is bounded by $\lesssim \sqrt{t_i \log n}$ with probability at least $1 - n^{-\mathcal{O}(1)}$. This is because $\bar{Y}_t^{(i-1)'} \in A$ holds with probability $1 - n^{-\mathcal{O}(1)}$ due to the first part of Lemma C.1, and for such paths the evolution of $\bar{Y}_t^{(i-1)'}$ is the same as that of $Y_t$, where we apply the second part of Lemma C.1. The second term is bounded by $\sqrt{t_{i-1} \log n}$ with probability $1 - n^{-\mathcal{O}(1)}$, following the discussion on $\|\bar{Y}_t^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\|$. In summary, with probability $1 - n^{-\mathcal{O}(1)}$ we can bound $\|\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)'} - \bar{Y}_{\overline{T}-t_i}^{(i-1)'}\|$ by $\sqrt{t_{i-1} \log n}(\le \sqrt{t_i \log n})$ up to a constant factor.

In summary, when $t_i \lesssim 1$, the transportation map moves at most $\mathcal{O}(\sqrt{t_i \log n})$ with probability $1 - n^{-\mathcal{O}(1)}$. Because the supports of $\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)'}$ and $\bar{Y}_{\overline{T}-\underline{T}}^{(i)}$ are both bounded, for the mass moved more than $\sqrt{t_i \log n}$ affects the Wasserstein distance at most $n^{-\mathcal{O}(1)}$. Therefore, we obtain the desired bound (81) for $t_i \lesssim 1$.

For $t_i \gtrsim 1$, because the supports of $\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)}$ and $\bar{Y}_{\overline{T}-\underline{T}}^{(i)}$ are both bounded,

$$W_1(\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)}, \bar{Y}_{\overline{T}-\underline{T}}^{(i)}) \lesssim \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)}, \bar{Y}_{\overline{T}-\underline{T}}^{(i)}) \lesssim \frac{1}{2}\sqrt{\int_{t=t_{i-1}}^{t_i} \int_x p_t(x)\beta_t^{-2}\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t}$$

holds. Therefore we obtain (81) as well.

From (81), (82) is easily obtained by jensen's inequality.

□

Also, we bound the generalization error of each network $s_i$.

**Lemma F.10.** *For $1 \le i \le K_* - 1$, let $s_i$ be a network that is selected from $\Phi(L, W, S, B)$ with*

$$L = \mathcal{O}(\log^4 n),\ \|W\|_\infty = \mathcal{O}(n^{\frac{d}{d+2s}}),\ S = \mathcal{O}(t_i^{-d/2} n^{\frac{\delta d}{2(2s+d)}}),\ \text{ and } B = \exp(\mathcal{O}(\log^4 n)),$$

*and $\|s_i(\cdot, t)\|_{L^\infty} \lesssim \frac{\log^{\frac{1}{2}} n}{\sigma_t}$. Then, we have that*

$$\mathbb{E}_{\{x_{0,j}\}_{i=j}^n}\left[\int_{t=t_i}^{t_{i+1}} \mathbb{E}_x\left[\|\hat{s}_i(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right] \lesssim n^{-\frac{2(s+1)}{d+2s}} \log n + \frac{t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}} \log^{10} n}{n}.$$

*Moreover, for $i = 0$, let $s_0$ be a network that is selected from $\Phi(L, W, S, B)$ with*

$$L = \mathcal{O}(\log^4 n),\ \|W\|_\infty = \mathcal{O}(n^{\frac{d}{d+2s}} \log^6 n),\ S = \mathcal{O}(n^{\frac{d}{2s+d}} \log^8 n),\ \text{ and } B = \exp(\mathcal{O}(\log^4 n)),$$

*and $\|s_0(\cdot, t)\|_{L^\infty} \lesssim \frac{\log^{\frac{1}{2}} n}{\sigma_t}$. Then, we have that*

$$\mathbb{E}_{\{x_{0,j}\}_{i=j}^n} \left[ \int_{t=t_0}^{t_1} \mathbb{E}_x \left[ \|\hat{s}_0(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t \right] \right] \lesssim n^{-\frac{2s}{d+2s}} \log^{18} n.$$

*Proof.* First we consider the first part. We take $N = n^{\frac{d}{d+2s}}$ and $t_* = t_i/2$ in Lemma D.5. Note that $N$ and $t_* (\geq n^{-\frac{2-\delta}{d+2s}})$ satisfies $t_* \geq N^{-(2-\delta)/d}$(, which is assumed in Lemma D.5). Then, there exists a neural network $\phi \in \Phi(L, W, S, B)$ that satisfies

$$\int_{t=t_i}^{t_{i+1}} \int_x p_t(x) \|\phi(x,t) - s(x,t)\|^2 \mathrm{d}x\mathrm{d}t \lesssim N^{-\frac{2(s+1)}{d}} \log n = N^{-\frac{2(s+1)}{d+2s}} \log n.$$

Specifically, $L = \mathcal{O}(\log^4(n)), \|W\|_\infty = \mathcal{O}(n^{\frac{d}{d+2s}}), S = \mathcal{O}(t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}})$, and $B = \exp(\mathcal{O}(\log^4 n))$. Therefore, we apply Theorem E.6 by replacing $\underline{T}$ and $\overline{T}$ by $t_i$ and $t_{i+1}$, respectively, and with $\varepsilon = n^{-\frac{2(s+1)}{d+2s}}$ to obtain the first assertion as

$$\begin{aligned}
&\mathbb{E}_{\{x_{0,j}\}_{i=j}^n} \left[ \int_{t=t_i}^{t_{i+1}} \mathbb{E}_x \left[ \|\hat{s}_i(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t \right] \right] \\
&\lesssim N^{-\frac{2(s+1)}{d}} \log n + \frac{C_\ell}{n} \log \mathcal{N} + \varepsilon \\
&\lesssim n^{-\frac{2(s+1)}{d+2s}} \log n + \frac{\log^2 n}{n} \left( t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}} \log^8 \right) + n^{-\frac{2(s+1)}{d+2s}} \\
&\lesssim n^{-\frac{2(s+1)}{d+2s}} \log n + \frac{t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}} \log^{10} n}{n}.
\end{aligned}$$

For the second part, we simply follow the discussion that derived the generalization error in Section 3.2, by replacing $\overline{T}$ by $t_1 (< \overline{T})$, which does not increase the generalization error. $\square$

*Proof of Theorem 3.5.* We use the sequence of networks presented in Lemma F.10. Specifically, we consider the following process.

$$\begin{aligned}
\hat{Y}_0^{(i)} \sim \mathcal{N}(0, I), \quad \mathrm{d}\hat{Y}_t^{(i)} &= \beta_{\overline{T}-t} (\hat{Y}_t^{(i)} + 2\hat{s}(\hat{Y}_t^{(i)}, \overline{T} - t)) \mathrm{d}t \\
&\quad + \sqrt{2\beta_{\overline{T}-t}} \mathrm{d}B_t \ (t \in [\overline{T} - t_i, \overline{T} - t_{i+1}], i = 0, 1, \cdots, K_*),
\end{aligned}$$

and we modify $\hat{Y}_{\overline{T}-\underline{T}}^{(i)}$ to 0 if $\|\hat{Y}_{\overline{T}-\underline{T}}^{(i)}\|_\infty > 2$.

Finally, we sum up the errors for the above process. Eq. (80) is further bounded by

$$
\begin{aligned}
&\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\
&\leq \sum_{i=1}^{K_*} \mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})]. \\
&\lesssim \sum_{i=1}^{K_*} \left[ \sqrt{t_{i-1}\log n} \sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \left[ \int_{t=t_i}^{t_i} \mathbb{E}_x [\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t] \right]} + n^{-\frac{s+1}{d+2s}} \right] \text{ (by Lemma F.9)} \\
&\lesssim \sum_{i=2}^{K_*} \left[ \sqrt{t_i \log n} \left( n^{-\frac{(s+1)}{d+2s}} \sqrt{\log n} + \frac{t_i^{-d/4} n^{\frac{\delta d}{4(d+2s)}} \log^5 n}{\sqrt{n}} \right) + n^{-\frac{(s+1)}{d+2s}} \right] \\
&\quad + \sqrt{t_1 \log n} \left[ n^{-\frac{s}{d+2s}} \log^9 n + n^{-\frac{s}{d+2s}} \right] \quad \text{(by Lemma F.10)} \\
&\lesssim \left[ \sqrt{t_1} n^{-\frac{s}{d+2s}} + \sqrt{t_1} \frac{t_1^{-d/4} n^{\frac{\delta d}{4(d+2s)}}}{\sqrt{n}} \right] \cdot \tilde{\mathcal{O}}(1) \\
&\quad \text{(because } K_* = \mathcal{O}(\log n) \text{ and } t_1 \leq \cdots t_{K_*} = \mathcal{O}(\log N) \text{ with } 1 < t_{i+1}/t_i = \text{const.} \leq 2 \ (i \geq 1).) \\
&= \left[ (n^{-\frac{2-\delta}{d+2s}})^{\frac{1}{2}} n^{-\frac{s}{d+2s}} + (n^{-\frac{2-\delta}{d+2s}})^{\frac{1}{2}} \frac{(n^{-\frac{2-\delta}{d+2s}})^{-d/4} n^{\frac{\delta d}{4(d+2s)}}}{\sqrt{n}} \right] \cdot \tilde{\mathcal{O}}(1) \\
&\lesssim n^{-\frac{(s+1-\delta)}{d+2s}}.
\end{aligned} \tag{84}
$$

Therefore, by taking $\underline{T} \lesssim n^{-\frac{2(s+1)}{d+2s}}$ and $\overline{T} = \frac{(s+1)\log n}{\underline{\beta}(d+2s)}$, we obtain that

$$
\begin{aligned}
W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}}) &\leq \mathbb{E}[W_1(X_0, X_{\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\
&\lesssim \sqrt{\underline{T}} + \exp(-\underline{\beta}\overline{T}) + n^{-\frac{(s+1-\delta)}{d+2s}} \quad \text{(by Lemmas F.7 and F.8 and (84))} \\
&\lesssim n^{-\frac{(s+1-\delta)}{d+2s}} + n^{-\frac{(s+1-\delta)}{d+2s}} + n^{-\frac{(s+1-\delta)}{d+2s}} \lesssim n^{-\frac{(s+1-\delta)}{d+2s}},
\end{aligned}
$$

which concludes the proof for Theorem 3.5. □

# G Error analysis with intrinsic dimensionality

This section corresponds to Section 4.

## G.1 Problem settings

We first formalize the problem settings. Let $A \in \mathbb{R}^{d \times d'}$ be a matrix made of orthogonal column vectors with the norm one. We consider the $d'$-dimensional subspace $V \coloneqq \{y \in \mathbb{R}^d \mid \exists x \in \mathbb{R}^{d'} \text{ s.t. } y = Ax\}$ where the true density has its support, i.e., $d'$ represents the intrinsic dimensionality. Together with Assumption 2.3, we assume the followings.

**Assumption G.1.** The true density $p_0$ is a probability measure that is absolutely continuous with respect to the Lebesgue measure on the sub-space $V$. Its probability density function as a function on the canonical coordinate system of the subspace $V$ is denoted by $q$.

**Assumption G.2.** $q$ is upper and lower bounded by $C_f$ and $C_f^{-1}$, respectively. Moreover, $q$ belongs to $U(B^s_{p,q}; [-1,1]^{d'})$.

**Assumption G.3.** $q$ belongs to $U(\mathcal{C}^\infty([-1,1]^{d'} \setminus [-1+a_0, 1-a_0]^{d'}))$ with $a_0 = n^{-\frac{1-\delta}{d'}}$.

## G.2 Proof overview

The generalization error analysis of the score network and how much the score estimation error affects in the final estimation rate in Theorem 4.1 are derived by just replacing $d$ by $d'$ in the previous analysis. Therefore we focus on the approximation error bounds. In order to obtain the counterparts of Theorem 3.1 and Lemma D.5, we aim to decompose the score function into two parts: each of them is determined by the intrinsic structure components (in $V$) and other components (in $V^\perp$). We use $z$ as a $d'$-dimensional vector corresponding to the canonical system of $V$. The first observation to this goal is

$$\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&= \int_V \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} q(z) \exp\left(-\frac{\|A^\top x - m_t z\|^2 + \|(I_d - A^\top)x\|^2}{2\sigma_t^2}\right) \mathrm{d}z \\
&\quad (z \text{ is a } d'\text{-dimensional vector corresponding to the canonical system of } V.) \\
&= \underbrace{\int_V \frac{q(z)}{\sigma_t^{d'} (2\pi)^{\frac{d'}{2}}} \exp\left(-\frac{\|A^\top x - m_t z\|^2}{2\sigma_t^2}\right) \mathrm{d}z}_{p_t^{(1)}(x)} \cdot \underbrace{\frac{1}{\sigma_t^{d-d'} (2\pi)^{\frac{d-d'}{2}}} \exp\left(-\frac{\|(I_d - A^\top)x\|^2}{2\sigma_t^2}\right)}_{p_t^{(2)}(x)}.
\end{aligned}$$

Here $p_t^{(1)}(x)$ and $p_t^{(2)}(x)$ can be seen as the density function with respect to the intrinsic components and remaining space. Note that

$$\nabla \log p_t(x) = \nabla \log(p_t^{(1)}(x) p_t^{(2)}(x)) = \nabla \log p_t^{(1)}(x) + \nabla \log p_t^{(2)}(x).$$

Due to this, we only need to construct the neural networks approximating each term and concatenate them. In addition, $p_t^{(1)}(x)$ can be seen as the density at $A^\top x$, about the diffusion process on the $d'$-dimensional space, where the initial density is defined by $q$. Thus we let

$$q_t(z') = \int_V \frac{q(z)}{\sigma_t^{d'} (2\pi)^{\frac{d'}{2}}} \exp\left(-\frac{\|z' - m_t z\|^2}{2\sigma_t^2}\right) \mathrm{d}z$$

for $z' \in \mathbb{R}^{d'}$. Here $p_t^{(1)}(x) = q_t(A^\top x)$ holds.

### G.3 Proof of Theorem 4.1

We first consider the approximation of $p_t^{(1)}(x)$. We have the following counterpart of Theorem 3.1 and Lemma D.5, where the only difference is that here $d$ is replaced by $d'$.

**Lemma G.4.** *Let $N \gg 1$, $\underline{T} = \mathrm{poly}(N^{-1})$ and $\overline{T} = \mathcal{O}(\log N)$. Then there exists a neural network $\phi_{\mathrm{score},3} \in \Phi(L, W, S, B)$ that satisfies, for all $t \in [\underline{T}, \overline{T}]$,*

$$\int_{x \in \mathbb{R}^d} p_t(x) \|\nabla \log p_t^{(1)}(x) - \phi_{\mathrm{score},3}(A^\top x, t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d'}} \log(N)}{\sigma_t^2}. \tag{85}$$

*Here, $L, W, S$ and $B$ are evaluated as $L = \mathcal{O}(\log^4 N), \|W\|_\infty = \mathcal{O}(N \log^6 N), S = \mathcal{O}(N \log^8 N)$, and $B = \exp(\mathcal{O}(\log^4 N))$. We can take $\phi_{\mathrm{score},3}$ satisfying $\|\phi_{\mathrm{score},3}(\cdot, t)\|_\infty = \mathcal{O}(\sigma_t^{-1} \log^{\frac{1}{2}} N)$.*

*Moreover, let $N' \geq t_*^{-d'/2} N^{\delta/2}$ and $t_* \geq N^{-(2-\delta)/d'}$. Then there exists a neural network $\phi_{\mathrm{score},4} \in \Phi(L, W, S, B)$ that satisfies*

$$\int_{x \in \mathbb{R}^d} p_t(x) \|\nabla \log p_t^{(1)}(x) - A\phi_{\mathrm{score},4}(A^\top x, t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2(s+1)}{d'}}}{\sigma_t^2} \tag{86}$$

*for $t \in [2t_*, \overline{T}]$. Specifically, $L = \mathcal{O}(\log^4(N)), \|W\|_\infty = \mathcal{O}(N), S = \mathcal{O}(N')$, and $B = \exp(\mathcal{O}(\log^4 N))$. We can take $\phi_{\mathrm{score},4}$ satisfying $\|\phi_{\mathrm{score},4}(\cdot, t)\|_\infty = \mathcal{O}(\sigma_t^{-1} \log^{\frac{1}{2}} N)$.*

*Proof.* Let $\phi_{\mathrm{score}} \colon \mathbb{R}^{d'} \times \mathbb{R}_+ \to \mathbb{R}^{d'}$ that approximates $q_t(z)$. Note that

$$\nabla \log p_t^{(1)}(x) = A \nabla \log q_t(A^\top x)$$

and therefore

$$\begin{aligned}
&\int_{x\in\mathbb{R}^d} p_t(x)\|\nabla\log p_t^{(1)}(x) - A\phi_{\rm score}(A^\top x,t)\|^2{\rm d}x\\
&=\int_{x\in\mathbb{R}^d} p_t^{(1)}(x)p_t^{(2)}(x)\|A\nabla\log p_t^{(1)}(A^\top x) - A\phi_{\rm score}(A^\top x,t)\|^2{\rm d}x\\
&=\int_{x\in\mathbb{R}^d} q_t(A^\top x)\|A\nabla\log p_t^{(1)}(A^\top x) - A\phi_{\rm score}(A^\top x,t)\|^2{\rm d}x\\
&=\int_{z\in\mathbb{R}^{d'}} q_t(z)\|\nabla\log q_t(z) - \phi_{\rm score}(z,t)\|^2{\rm d}z,
\end{aligned}$$

where we used the fact that $p_t^{(1)}$ and $p_t^{(2)}$ depend on $A^\top x$ and $(I-A^\top)x$, respectively, and $A^\top x$ and $(I-A^\top)x$ are orthogonal. Moreover, we used $\det(A^\top A)=1$ and orthogonality of the columns of $A$. Thus, we can translate Theorem 3.1 and Lemma D.5. □

We next consider the approximation of $p_t^{(2)}(x)$. As we did in Appendix C, we first show that it suffice to consider the approximation within the bounded region.

**Lemma G.5.** *For $\varepsilon>0$, we define $B_{t,\varepsilon}$ as*

$$B_{t,\varepsilon}=\left\{x\in\mathbb{R}^d \,\middle|\, \|(I_d-A^\top)x\|\le C_{\rm e}\sigma_t\sqrt{\log\varepsilon^{-1}}.\right\}$$

*We sometimes abbreviate this as $B_\varepsilon$. Then, we have that*

$$\int_{x\in\bar B_\varepsilon} p_t(x)\left[1\vee\|\nabla\log(p_t^{(2)}(x))\|^2\right]{\rm d}x\lesssim\varepsilon.$$

*Proof.* The the columns of $A$ are orthogonal. $p_t^{(1)}$ and $p_t^{(2)}$ depend on $A^\top x$ and $(I-A^\top)x$, respectively, and $A^\top x$ and $(I-A^\top)x$ are orthogonal. Thus, we have that

$$\begin{aligned}
&\int_{x\in\bar B_{t,\varepsilon}} p_t(x)\left[1\vee\|\nabla\log(p_t(x))\|^2\right]{\rm d}x \qquad (87)\\
&=\int_{x\in\bar B_{t,\varepsilon}} p_t^{(1)}(x)p_t^{(2)}(x)\left[1\vee\|\nabla\log(p_t(x))\|^2\right]{\rm d}x\\
&=\int_{x\in\bar B_{t,\varepsilon}} p_t^{(2)}(x)\left[1\vee\|\nabla\log(p_t(x))\|^2\right]{\rm d}x\\
&=\int_{w\in\mathbb{R}^{d-d'}:\,\|w\|\ge C_{\rm e}\sigma_t\sqrt{\log\varepsilon^{-1}}} \frac{1\vee\|w\|^2/\sigma_t^2}{\sigma_t^{d-d'}(2\pi)^{\frac{d-d'}{2}}}\exp\left(-\frac{\|w\|^2}{2\sigma_t^2}\right){\rm d}w.
\end{aligned}$$

Applying Corollary J.8, (87) is bounded by $\varepsilon$ with a sufficiently large constant $C_{\rm e}$. □

Now we only need consider the approximation of $\nabla\log p_t^{(2)}(x)$ within $B_{t,\varepsilon}$.

**Lemma G.6.** *Let $N\gg1$, $\underline{T},\varepsilon=\mathrm{poly}(N^{-1})$ and $\overline{T}\simeq\log N$. There exists a neural network $\phi_{\rm score,4}\in\Phi(L,W,S,B)$ such that*

$$\sup_{t\in[\underline{T},\overline{T}]}\int_x p_t(x)\|\nabla\log p_t^{(2)}(x)-\phi_{\rm score,4}(x,t)\|^2{\rm d}x\lesssim\frac{N^{-\frac{2(s+1)}{d'}}}{\sigma_t^2}. \qquad (88)$$

*Specifically, $\phi_{\rm score,4}\in\Phi(L,W,S,B)$ holds, where*

$$L=\mathcal{O}(\log^2N)),\|W\|_\infty=\mathcal{O}(\log^3N),S=\mathcal{O}(\log^4N),\text{ and }B=\exp(\mathcal{O}(\log^2N)). \qquad (89)$$

*Proof.* First note that $\nabla\log p_t^{(2)}(x)=-\frac{1}{\sigma_t^2}(I_d-A)(I_d-A^\top)x$. We approximate this via the following four steps.

1. $\sigma_t$ is approximated by $\phi_\sigma$ from Lemma D.2. Here we set $\varepsilon\leftarrow(\underline{T}^4\wedge\varepsilon^4)\varepsilon^4$.

2. Based on the approximation of $\sigma_t$, $\sigma_t^{-2}$ is approximated by $\phi_{\mathrm{rec}}(\cdot;2)$ from Corollary J.8. Here we set $\varepsilon \leftarrow (\underline{T} \wedge \varepsilon)\varepsilon$.
3. $(I_d - A)(I_d - A^\top)$ is realized by $\mathrm{ReLU}((I_d - A)(I_d - A^\top)\cdot x + 0) - \mathrm{ReLU}(-(I_d - A)(I_d - A^\top)\cdot x + 0)$.
4. According to Lemma J.6 with $\varepsilon \leftarrow \varepsilon$ and $C \leftarrow \underline{T}^{-1} \vee \sqrt{\log \varepsilon^{-1}}$, multiplication of $\sigma_t^{-2}$ and $(I_d - A)(I_d - A^\top)$ is constructed.

By concatenating these networks (using Lemma J.1), the obtained network size is bounded as $L = \mathcal{O}(\log^2 \varepsilon^{-1} + \log^2 \underline{T}^{-1})), \|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1} + \log^3 \underline{T}^{-1}), S = \mathcal{O}(\log^4 \varepsilon^{-1} + \log^4 \underline{T}^{-1})$, and $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1} + \log^2 \underline{T}^{-1}))$.
Then, for $x \in B_{t,\varepsilon}$ with $t \geq \underline{T}$, we have that

$$\|\nabla \log p_t^{(2)}(x) - \phi_{\mathrm{score},4}\| \lesssim \varepsilon.$$

This yields that

$$\int_{B_{t,\varepsilon}} p_t(x)\|\nabla \log p_t^{(2)}(x) - \phi_{\mathrm{score},4}\| \mathrm{d}x \lesssim \varepsilon.$$

Together with Lemma G.5, by taking $\varepsilon = \mathrm{poly}(N^{-1})$, we have the assertion. □

*Proof of Theorem 4.1.* Note that while the error bound (88) in Lemma G.6 is tighter than the bounds (85) and (86) in Lemma G.4, the required network size (89) in Lemma G.6 is smaller than the size bounds in Lemma G.4. Also note that the bounds in Lemma G.4 are the same as those in Theorem 3.1 and Lemma D.5, except for that $d$ is replaced by $d'$. Therefore, by simply aggregating $\phi_{\mathrm{score},3}$ and $\phi_{\mathrm{score},4}$, we obtain the counterpart of the approximation theorems Theorem 3.1 and Lemma D.5, and the rest of the analysis are the same as that of the $d$-dimensional case. Therefore, we obtain the statement. □

# H SAMPLING $t$ AND $x_t$ IN THE EMPIRICAL SCORE MATCHING LOSS

Since our main interest lies in the sample complexity, and for simple presentation, we have considered the situation where $\ell(x)$ can be exactly evaluated. However, in usual implementation (Sohl-Dickstein et al., 2015; Song & Ermon, 2019), two expectations in (1) with respect to $t$ and $x_t$ are also replaced by sampling for computational efficiency. Here we also introduce two ways to replace the expectation by a finite sample of $t$ and $x_t$. As in Section 3, we assume Assumptions 2.2 to 2.4.

**Approximation via polynomial-size sample** Let us sample $(i_j, t_j, x_j)$ from $i_j \sim \mathrm{Unif}(\{1,2,\cdots,n\})$, $t_j \sim \mathrm{Unif}(\underline{T}, \overline{T})$, and $x_j \sim p_{t_j}(x_j|x_{0,i})$. Then we let $\hat{s}$ as

$$\underset{s\in\mathcal{S}}{\mathrm{argmin}} \frac{1}{M}\sum_{j=1}^{M} \|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\|^2$$

and evaluate the difference between

$$\frac{1}{n}\sum_{i=1}^{n} \ell_{\hat{s}}(x_i) - \underset{s\in\mathcal{S}}{\mathrm{argmin}}\ \frac{1}{n}\sum_{i=1}^{n} \ell_s(x_i). \tag{90}$$

The complete proof and formal statement can be found in Theorem H.2 of Appendix H, and here we provide the proof sketch. We first show that $\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\|$ is sub-Gaussian (Lemma H.1). Here, we simply interpret this as $\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\| = \tilde{\mathcal{O}}(t_j^{-\frac{1}{2}}) \lesssim \tilde{\mathcal{O}}(\underline{T}^{-\frac{1}{2}})$ with high probability to proceed. Then, by a similar argument that derived Theorem 3.2, we can bound (90) by $\tilde{O}(\frac{\underline{T}^{-1}\cdot\log\mathcal{N}}{M})$. Here, $\mathcal{N}$ satisfies $\log\mathcal{N} \lesssim \tilde{\mathcal{O}}(n^{\frac{d}{2s+d}})$. In order to make (90) as small as the generalization error $\tilde{\mathcal{O}}(n^{-\frac{2s}{2s+d}})$, we need to take $M \gtrsim n\cdot\underline{T}^{-1}$. Thus, for each $x_{0,i}$, $\tilde{\mathcal{O}}(\underline{T}^{-1}) = \mathrm{poly}(n^{-1})$ sample of $(t_j, x_j|x_{0,i})$ should be considered. We remark that the reason why we need polynomial-size sample is mainly due to the scale of $\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\|^2$.

**Modifying the distribution of** $t$ One may think whether it is possible to consider only one path for each sample $x_{0,i}$. Here, the main problem is that the variance of $\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\|^2$ can grow to infinity as $t_j$ approaches to 0. To address this issue, we sample $t_j$ from $\mu(t) \propto \frac{\mathbb{1}[\underline{T} \le t \le \overline{T}]}{t}$ and modify $\lambda(t)$ as $\lambda(t) = \frac{t \log \overline{T}/\underline{T}}{\overline{T}-\underline{T}}$, while $i_j, x_j$ are sampled as previously. Then, we have that

$$\begin{aligned}&\mathbb{E}_{i_j,t_j,x_j}\left[\lambda(t_j)\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i})\|^2\right] \\ =&\frac{1}{n}\sum_{i=1}^{n} \ell(x_i),\end{aligned}$$

and that $\lambda(t_j)\|s(x_{t_j},t_j) - \nabla \log p_{t_i}(x_{t_j}|x_{0,i})\|^2 = \tilde{\mathcal{O}}(1)$ holds with high probability (because $\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\|^3 = \tilde{\mathcal{O}}(t_j^{-1})$ and that $\lambda(t_j) \lesssim 1/t_j$). In this way of sampling, we let $\hat{s}$ as

$$\operatorname*{argmin}_{s\in\mathcal{S}} \frac{1}{M}\sum_{j=1}^{M} \lambda(t_j)\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\|^2$$

and evaluate the difference (90). Finally, using a similar argument for Theorem 3.2, we again obtain that (90) is bounded by $\tilde{O}(\frac{\log \mathcal{N}}{M}) \lesssim \tilde{O}(\frac{n^{\frac{d}{2s+d}}}{M})$. Taking $M = n$ suffices to make this difference as small as the generalization error $\tilde{\mathcal{O}}(n^{-\frac{2s}{2s+d}})$.

Now we provide justification of two approaches presented here. We first begin with the following lemma. This shows that $\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|$ is sub-Gaussian.

**Lemma H.1.** *Let us sample* $(i_j,t_j,x_j)$ *from* $i_j \sim \mathrm{Unif}(\{1,2,\cdots,\mathrm{n}\})$, $t_j \sim \mathrm{Unif}(\underline{\mathrm{T}},\overline{\mathrm{T}})$, *and* $x_j \sim p_{t_j}(x_j|x_{0,i_j})$. *Then, we have that, for all* $t > 0$,

$$\mathbb{P}\left[\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| \ge \sup_{(x,t)}\|s(x,t)\| + \frac{\sqrt{d}t}{\sigma_{\underline{T}}}\right] \le 2\exp\left(-t^2/2\right).$$

*Proof.* First note that

$$\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| \le \|s(x_j,t_j)\| + \|\nabla p_{t_j}(x_j|x_{0,i_j})\| \le \sup_{x,t}\|s(x,t)\| + \|\nabla p_{t_j}(x_j|x_{0,i_j})\|.$$

Because $\nabla p_{t_j}(x_j|x_{0,i_j}) = \frac{x_j - m_t x_{0,i_j}}{\sigma_t^2}$ and $x_j \sim p_{t_j}(x_j|x_{0,i_j}) = \mathcal{N}\left(m_t x_{0,i_j}, \sigma_t^2\right)$, we have that $[\nabla p_{t_j}(x_j|x_{0,i_j})]_i$ is sub-Gaussian with $\sigma_t^{-1}$. Thus, $\|\nabla p_{t_j}(x_j|x_{0,i_j})\|$ is sub-Gaussian with $\sqrt{d}\sigma_t^{-1}$. Now, applying $\sigma_t \ge \sigma_{\underline{T}}$, we have the assertion. □

Now, we give the following theorem for the first approach.

**Theorem H.2.** *Let us sample* $(i_j,t_j,x_j)$ *from* $i_j \sim \mathrm{Unif}(\{1,2,\cdots,\mathrm{n}\})$, $t_j \sim \mathrm{Unif}(\underline{\mathrm{T}},\overline{\mathrm{T}})$, *and* $x_j \sim p_{t_j}(x_j|x_{0,i})$. *Let* $s_1$ *be the minimizer of*

$$\frac{1}{M}\sum_{j=1}^{M}\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i})\|^2$$

*and* $s_2$ *be the minimizer of*

$$\frac{1}{n}\sum_{i=1}^{n}\ell(x_i) = \frac{1}{n}\sum_{i=1}^{n}\int_{t=\underline{T}}^{\overline{T}} \|s(x_t,t) - \nabla p_t(x_t|x_{0,i})\|^2 p_t(x_t|x_{0,i})\mathrm{d}x_t\mathrm{d}t,$$

*over* $\mathcal{S} \subseteq \Phi(L,W,S,B)$, *where* $s \in \mathcal{S}$ *satisfies* $\|\|s(\cdot,t)\|_2\|_{L^\infty} = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} n) \lesssim \mathcal{O}(\sigma_{\underline{T}}^{-1}\log^{\frac{1}{2}} n) =: C_s$. *Then, we have that*

$$\mathbb{E}_{\{(i_j,t_j,x_j)\}_{i=1}^n}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_1(x_i) - \frac{1}{n}\sum_{i=1}^{n}\ell_2(x_i)\right| \lesssim \frac{C_s^2 + \sigma_{\underline{T}}^{-2}}{M} 2SL\log(\varepsilon^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \varepsilon.$$

*Proof.* We denote $(i_j, t_j, x_j) = y_j$ for simplicity and $Y = \{(i_j, t_j, x_j)\}_{j=1}^M = \{y_j\}_{j=1}^M$. Let $Y' = \{(i'_j, t'_j, x'_j)\}_{j=1}^M = \{y'_j\}_{j=1}^M$ be a copy of $Y$, which is independent of $Y$. We write $\kappa(y_j) = \|s(x_j, t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|^2$. Then, we have that

$$\mathbb{E}_Y \left| \frac{1}{M}\sum_{j=1}^M \kappa_1(y_j) - \frac{1}{M}\sum_{j=1}^M \kappa_2(y_j) - \frac{1}{n}\sum_{i=1}^n \ell_1(x_i) - \frac{1}{n}\sum_{i=1}^n \ell_2(x_i) \right| \tag{91}$$

$$= \mathbb{E}_Y \left| \frac{1}{M}\sum_{j=1}^M (\kappa_1(y_j) - \kappa_2(y_j)) - \mathbb{E}_{Y'}\left[ \frac{1}{M}\sum_{j=1}^M (\kappa_1(y'_j) - \kappa_2(y'_j)) \right] \right|$$

$$\leq \mathbb{E}_{Y,Y'} \left| \frac{1}{M}\sum_{j=1}^M ((\kappa_1(y_j) - \kappa_2(y_j)) - (\kappa_1(y'_j) - \kappa_2(y'_j))) \right|. \tag{92}$$

Next, we let $C_s$ be the minimum integer that satisfies $C_s \geq \sup_{s\in\mathcal{C}} \sup_{x,t} \|s(x,t)\|$, and for $i = 1, 2, \cdots$, we define $\mathcal{E}_i$ as an event where $C_s + \frac{\sqrt{d}(i-1)}{\sigma_{\underline{T}}} \leq \sup_{s\in\mathcal{C}} \max_j \max\{\|s(x_j, t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|, \|s(x'_j, t'_j) - \nabla p_{t'_j}(x'_j|x_{0,i'_j})\|\} < C_s + \frac{\sqrt{d}i}{\sigma_{\underline{T}}}$ holds. For $i = 0$, we define $\mathcal{E}_0$ as an event where $\sup_{s\in\mathcal{S}} \max_j \max\{\|s(x_j, t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|, \|s(x'_j, t'_j) - \nabla p_{t'_j}(x'_j|x_{0,i'_j})\|\} < C_s$ holds. We let $a_i = \mathbb{P}[\mathcal{E}_i]$ and $\mathbb{E}_i$ be the expectation conditioned by the event $\mathcal{E}_i$. Then, (92) is bounded by

$$\mathbb{E}_0 \left| \frac{1}{M}\sum_{j=1}^M ((\kappa_1(y_j) - \kappa_2(y_j)) - (\kappa_1(y'_j) - \kappa_2(y'_j))) \right|$$

$$+ \sum_{i=1}^\infty a_i \mathbb{E}_i \left| \frac{1}{M}\sum_{j=1}^M ((\kappa_1(y_j) - \kappa_2(y_j)) - (\kappa_1(y'_j) - \kappa_2(y'_j))) \right|. \tag{93}$$

We remark that $\frac{1}{M}\sum_{j=1}^M ((\kappa_1(y_j) - \kappa_2(y_j)) - (\kappa_1(y'_j) - \kappa_2(y'_j)))$ is bounded by $8C_s^2 + \frac{8di^2}{\sigma_t^2}$ for each $\mathbb{E}_i$. Here, $\kappa_1$ is the minimizer of $\frac{1}{M}\sum_{j=1}^M \kappa(y_j)$ and $\kappa_2$ is the minimizer of $\mathbb{E}[\kappa(y)]$. Moreover, because $\|(x_j - x_{0,i_j})/\sigma_t\| = \|\nabla p_{t_j}(x_j|x_{0,i_j})\| \leq \|s(x_j, t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| + \|s(x_j, t_j)\|$, we have that $\|s(x_j, t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| \leq C_s + \frac{\sqrt{d}i}{\sigma_{\underline{T}}}$ implies $\|x_j\| \leq 2C_s + \sqrt{d}i$. We apply the same argument as that in Theorem E.6 to obtain that

$$\mathbb{E}_i \left| \frac{1}{M}\sum_{j=1}^M \kappa_1(y_j) - \frac{1}{M}\sum_{j=1}^M \kappa_2(y_j) - \frac{1}{n}\sum_{i=1}^n \ell_1(x_i) - \frac{1}{n}\sum_{i=1}^n \ell_2(x_i) \right|$$
$$\lesssim \frac{C_s^2 + \sigma_{\underline{T}}^{-2} i^2}{M} \log \mathcal{N}(\mathcal{S}, L^\infty([-(2C_s + \sqrt{d}i), 2C_s + \sqrt{d}i]^{d+1}), \varepsilon/(C_s + i\sigma_{\underline{T}}^{-1})) + \varepsilon.$$
$$\lesssim \frac{C_s^2 + \sigma_{\underline{T}}^{-2} i^2}{M} 2SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s + i)) + \varepsilon.$$

We remark that, $y_j$ and $y'_j$ are not independent, when conditioned by $\mathcal{E}_i$. However, the similar argument still holds in (67), where we used the independentness of $x_i$ and $x'_i$ in the original proof, because the symmetry of $y_j$ and $y'_j$ is not collapsed by taking the conditional expectation. Based on this, and $a_i \leq 2\exp(-(i-1)^2/2)$ $(i \geq 1)$ due to Lemma H.1, we evaluate (93) as

$$\begin{aligned}
(93) &\lesssim \frac{C_s^2 + \sigma_{\underline{T}}^{-2}}{M} SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s)) + \varepsilon \\
&\quad + \sum_{i=1}^{\infty} a_i \left[ \frac{C_s^2 + \sigma_{\underline{T}}^{-2} i^2}{M} SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s + i)) + \varepsilon \right] \\
&\lesssim \frac{C_s^2 + \sigma_{\underline{T}}^{-2}}{M} SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s)) + \varepsilon \\
&\quad + \sum_{i=1}^{\infty} \exp\left(-\frac{(i-1)^2}{2}\right) \left[ \frac{C_s^2 + \sigma_{\underline{T}}^{-2} i^2}{M} 2SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s + i)) + \varepsilon \right] \\
&\lesssim \frac{C_s^2 + \sigma_{\underline{T}}^{-2}}{M} SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s)) + \varepsilon.
\end{aligned}$$

This bounds (91). Thus, we finally obtain that

$$\begin{aligned}
&\mathbb{E}_{\{y_i\}_{i=1}^n} \left[ \frac{1}{n} \sum_{i=1}^n \ell_1(x_i) - \frac{1}{n} \sum_{i=1}^n \ell_2(x_i) \right] \\
&\leq \mathbb{E}_{\{y_i\}_{j=1}^M} \left[ \frac{1}{M} \sum_{j=1}^M \kappa_1(y_j) - \sum_{j=1}^M \kappa_2(y_j) \right] + \frac{C_s^2 + \sigma_{\underline{T}}^{-2}}{M} SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s)) + \varepsilon \\
&\leq \frac{C_s^2 + \sigma_{\underline{T}}^{-2}}{M} SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s)) + \varepsilon,
\end{aligned}$$

because $\kappa_1$ is the minimizer of $\frac{1}{M} \sum_{j=1}^M \kappa(y_j)$. Now, we obtain the assertion. □

**Remark H.3.** When $\|s(x,t)\| = \sqrt{\log N}/\sigma_t$ holds, $\underline{T} = \mathrm{poly}(N^{-1}), \overline{T} = \mathcal{O}(\log N)$, we have $\sup_{(x,t)} \|s(x,t)\| = C_s \lesssim \sqrt{\underline{T}^{-1} \log N}$. we set $N = n^{\frac{d}{2s+d}}$, $\varepsilon = n^{-\frac{2s}{d+2s}}$ and use the network class in Theorem 3.1 to obtain that

$$\begin{aligned}
&\mathbb{E}_{(i_j, t_j, x_j)} \left[ \frac{1}{n} \sum_{i=1}^n \ell_1(x_i) \right] - \int_{\ell_s:\, s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^n \ell_s(x_i) \\
&\lesssim \frac{C_s^2 + \sigma_{\underline{T}}^{-2}}{M} 2SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s)) + \varepsilon \\
&\lesssim \frac{\underline{T}^{-1} \log n + \underline{T}^{-1}}{M} n^{-\frac{d}{2s+d}} \log^{16} n \lesssim \frac{n^{-\frac{d}{2s+d}} \log^{17} n}{\underline{T} M}.
\end{aligned}$$

Next, we show the proof for the second approach.

**Theorem H.4.** *We sample $t_j$ from $\mu(t) \propto \frac{\mathbb{1}[\underline{T} \leq t \leq \overline{T}]}{t}$ and modify $\lambda(t)$ as $\lambda(t) = \frac{t \log \overline{T}/\underline{T}}{\overline{T} - \underline{T}}$, while $i_j, x_j$ are sampled as $i_j \sim \mathrm{Unif}(\{1, 2, \cdots, \mathrm{n}\})$ and $x_j \sim p_{t_j}(x_j | x_{0,i})$. Then, the minimizer $s_1$ over $\mathcal{S} \subseteq \Phi(L, W, S, B)$ of*

$$\frac{1}{M} \sum_{j=1}^M \lambda(t_j) \|s(x_j, t_j) - \nabla p_{t_j}(x_j | x_{0,i})\|^2$$

*satisfies*

$$\mathbb{E}_{(i_j, t_j, x_j)} \left[ \frac{1}{n} \sum_{i=1}^n \ell_1(x_i) \right] - \int_{\ell_s:\, s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^n \ell_s(x_i) \lesssim \frac{C_s^2 + \overline{T}}{M} SL \log(\varepsilon^{-1} L \|W\|_\infty (B \vee 1)(C_s)) + \varepsilon,$$

*Here, $C_s = \sup_{t,x} \sqrt{\lambda(t)} \|s(x,t)\|$.*

*Proof.* We just replace $\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i})\|$ by $\sqrt{\lambda(t_j)}\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i})\|$ in the previous lemma. Similarly to Lemma H.1, we have that, for all $t > 0$,

$$\mathbb{P}\left[\lambda^{\frac{1}{2}}(t_j)\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| \geq \sup_{(x,t)} \lambda^{\frac{1}{2}}(t)\|s(x,t)\| + \frac{\sqrt{d}\lambda^{\frac{1}{2}}(t_j)t}{\sigma_{t_j}}\right] \leq 2\exp\left(-t^2/2\right).$$

Then, we replace $\sup_{(x,t)}\|s(x,t)\|$ by $\sup_{(x,t)}\lambda^{\frac{1}{2}}(t)\|s(x,t)\|$, and $\frac{\sqrt{d}}{\sigma_{\underline{T}}}$ by $\sup_t \frac{\sqrt{d}\lambda^{\frac{1}{2}}(t)}{\sigma_t}$, respectively, to obtain that

$$\begin{aligned}
&\mathbb{E}_{i_j,t_j,x_j}\mathbb{E}_{i_j',t_j',x_j'}\left[\lambda(t_j)\|s_1(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|^2\right] \\
&\quad - \inf_{s\in\mathcal{S}} \mathbb{E}_{i_j,t_j,x_j}\left[\lambda(t_j)\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|^2\right] \\
&\lesssim \frac{C_s^2 + \overline{T}}{M} SL\log(\varepsilon^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \varepsilon,
\end{aligned} \tag{94}$$

where $(i_j', t_j', x_j')$ are the independent copy of $(i_j, t_j, x_j)$. Note that

$$\mathbb{E}_{i_j,t_j,x_j}\left[\lambda(t_j)\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|^2\right] = \frac{1}{n}\sum_{i=1}^n \ell(x_i) \tag{95}$$

for all (fixed) $s$. (94) and (95) yield that

$$\begin{aligned}
&\mathbb{E}_{(i_j,t_j,x_j)}\left[\frac{1}{n}\sum_{i=1}^n \ell_1(x_i)\right] - \int_{\ell_s:\, s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^n \ell_s(x_i) \\
&\leq \frac{C_s^2 + \overline{T}}{M} SL\log(\varepsilon^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \varepsilon,
\end{aligned}$$

which concludes the proof. □

**Remark H.5.** When $\|s(x,t)\| = \sqrt{\log N}/\sigma_t$ holds, $\underline{T} = \mathrm{poly}(N^{-1}), \overline{T} = \mathcal{O}(\log N)$, we have $\sup_{(x,t)}\sqrt{\lambda(t)}\|s(x,t)\| = C_s \lesssim \sqrt{\log N}$. we set $N = n^{\frac{d}{2s+d}}$, $\varepsilon = n^{-\frac{2s}{d+2s}}$ and use the network class in Theorem 3.1 to obtain that

$$\mathbb{E}_{(i_j,t_j,x_j)}\left[\frac{1}{n}\sum_{i=1}^n \ell_1(x_i)\right] - \int_{\ell_s:\, s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^n \ell_s(x_i) \lesssim n^{-\frac{2s}{d+2s}}\log^{17} n.$$

# I Discussion on the discretization error

Although the continuous time SDE is mainly focused on for simple presentation, we can also take the discretization error into consideration. As in Section 3, we assume Assumptions 2.2 to 2.4. Let $t_0 = \underline{T} < t_1 < \cdots < t_{K_*} = \overline{T}$ be the time steps with $\eta \equiv t_{k+1} - t_k$. We train the score network as the minimizer of

$$\sum_{i=1}^n \sum_{k=0}^{K-1} \eta\mathbb{E}[\|s(x_{t_k}, t_k) - \nabla\log p_{\overline{T}-t_k}(x_{t_k}|x_{0,i})\|^2].$$

Here the expectation is taken with respect to $x_{\overline{T}-t_k} \sim p_{\overline{T}-t_k}(x_{\overline{T}-t_k}|x_{0,i})$. Then consider the following process $(Y_t^{\mathrm{d}})_{t=0}^{\eta K}$ with $Y_0^{\mathrm{d}} \sim \mathcal{N}(0, I_d)$: for $t \in [\overline{T} - t_i, \overline{T} - t_{i+1}]$,

$$\mathrm{d}Y_t^{\mathrm{d}} = \beta_t(Y_t^{\mathrm{d}} + 2\hat{s}(Y_{\overline{T}-t_i}^{\mathrm{d}}, \overline{T} - t_i))\mathrm{d}t + \beta_{\overline{T}-t}\mathrm{d}B_t$$

This is just replacement of the drift term at $t$ by that at the last discretized step, and we can obtain $\bar{Y}_{\eta(k+1)}$ from $\bar{Y}_{\eta k}$ as easy as the classical Euler-Maruyama discretization because $\bar{Y}_{\eta(k+1)}$ conditioned on $\bar{Y}_{\eta k}$ is a Gaussian. This is also adopted in De Bortoli (2022); Chen et al. (2022). However, De Bortoli (2022) requires $\eta_i \leq \exp(-n^{\mathcal{O}(1)})$ and Chen et al. (2022) assumes Lipschitzness of the score, which does not necessarily hold in our setting.

The following discretization error bound holds:

**Theorem I.1.** *Let* $\underline{T} = n^{-\mathcal{O}(1)}$, $\overline{T} = \frac{s \log n}{2s+d}$, *and* $\eta = \mathrm{poly}(n^{-1})$. *Then,*

$$\mathbb{E}[\mathrm{TV}(X_0, \bar{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-\frac{2s}{d+2s}} \log^{18} n + \eta^2 \underline{T}^{-3} \log^3 n + \eta \underline{T}^{-1} \log^3 n + \eta \log^4 n.$$

Thus, taking $\eta = \underline{T}^{-1.5} n^{-s/(2s+d)} = \mathrm{poly}(n^{-1})$ suffices to ignore the discretization error.

*Proof of Appendix I.* We first show that the minimizer $\hat{s}$ over $\Phi'$ (given in Section 3.2) of

$$\hat{s} \in \mathrm{argmin}\, \frac{1}{n} \sum_{i=1}^{n} \sum_{k=}^{K} \eta \mathbb{E}[\|s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k}|x_{0,i})\|^2].$$

satisfies

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \left[ \sum_{k=1}^{K} \eta \mathbb{E}_{x_{t_k} \sim p_{t_k}} [\|\hat{s}(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k})\|^2] \right] \lesssim n^{-2s/(2s+d)} \log^{18} n. \tag{96}$$

We take $N = n^{\frac{d}{d+2s}}$ According to Theorem 3.1, for $N \gg 1$, there exists a neural network $\phi_{\mathrm{score}}$ with $L = \mathcal{O}(\log^4 N)$, $\|W\|_\infty = \mathcal{O}(N \log^6 N)$, $S = \mathcal{O}(N \log^8 N)$, and $B = \exp(\mathcal{O}(\log^4 N))$ that satisfies

$$\int_x p_t(x) \|\phi_{\mathrm{score}}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d}} \log(N)}{\sigma_t^2}. \tag{97}$$

for all $t \in [\underline{T}, \overline{T}]$. By summing up this for all $t = t_k$, we have that

$$\sum_{k=1}^{K} \eta \mathbb{E}_{x_{t_k} \sim p_{t_k}} [\|\phi_{\mathrm{score}}(x_{t_k}, t_k) - \nabla \log p_{\eta k}(X_{t_k})\|^2] \lesssim \sum_{k=1}^{K} \eta \frac{N^{-\frac{2s}{d}} \log(N)}{1 \wedge t_k} \tag{98}$$

$$\leq N^{-\frac{2s}{d}} \log(N) \left( \eta K + \eta \sum_{k=1}^{K} \frac{1}{t_k} \right) \lesssim N^{-\frac{2s}{d}} \log(N)(\overline{T} + \log(\overline{T}/\underline{T})) \lesssim N^{-\frac{2s}{d}} \log^2(N).$$

In order to convert this into the generalization bound, we need to evaluate the following two things. First, $\hat{s}$ can be taken so that

$$\sup_x \|\phi_{\mathrm{score}}(x,t)\| \mathrm{d}x \lesssim \frac{\log^{\frac{1}{2}}(N)}{\sigma_t},$$

and therefore we clip $s$ as in Section 3.2. Because such $s$ satisfies

$$\int_x p_t(x) \|\phi_{\mathrm{score}}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{\log(N)}{\sigma_t^2},$$

we have that

$$\sum_{k=1}^{K} \eta \mathbb{E}_{x_{t_k} \sim p_{t_k}} [\|\phi_{\mathrm{score}}(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k})\|^2] \leq C_\ell = \mathcal{O}(\log^2(n))$$

(follow the argument for Lemma E.3 and how we derived (98) from (97)). Second, the covering number of the network class of $\ell(x) = \sum_{k=1}^{K} \eta \mathbb{E}[\|s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k}|x)\|^2]$ over all $s$ with $\delta = n^{-\frac{2s}{d+2s}}$ is bounded by $n^{\frac{d}{d+2s}} \log^{16} n$, by following Appendix E.3. Thus, Theorem E.6 can be modified to this setting and we obtain that

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \left[ \sum_{k=1}^{K} \eta \mathbb{E}_{x_{t_k} \sim p_{t_k}} [\|s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k})\|^2] \right] \lesssim n^{-s/(2s+d)} \log^2 n.$$

holds. Therefore, following the discussion in Section 3.2, we have that

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \left[ \sum_{k=1}^{K} \eta_k \mathbb{E}_{x_{t_k} \sim p_{t_k}} [\|s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k})\|^2] \right]$$

$$\lesssim \sum_{k=1}^{K} \eta \mathbb{E}_{x_{t_k} \sim p_{t_k}} [\|\phi_{\mathrm{score}}(x_{t_k}, t_k) - \nabla \log p_{\eta k}(X_{t_k})\|^2] + \frac{C_\ell}{n} \log \mathcal{N} + \delta$$

$$\lesssim n^{\frac{d}{d+2s}} \log^2 n + \frac{\log^2 n}{n} \cdot n^{\frac{d}{d+2s}} \log^{18} n + n^{-\frac{2s}{d+2s}} \lesssim n^{-\frac{2s}{d+2s}} \log^{18} n,$$

which proves (96).

From now, we bound $\mathrm{TV}(Y_0, Y^{\mathrm{d}}_{\overline{T}-\underline{T}})$. We introduce the following processes. $\bar{Y}^{\mathrm{d}} = (\bar{Y}^{\mathrm{d}}_t)_{t=0}^{\overline{T}-\underline{T}}$ is defined in the same way as $Y^{\mathrm{d}}$, except for the initial distribution of $\bar{Y}^{\mathrm{d}}_0 \sim p_{\overline{T}}$. At $t = \overline{T} - \underline{T}$, if the $\int$-norm is more than 2, then we reset it to 0. $\bar{Y} = (\bar{Y}_t)_{t=0}^{\overline{T}-\underline{T}}$ is defined as $\bar{Y}_0 \sim p_{\overline{T}}$, and

$$\begin{aligned}
&\bar{Y}_0 \sim p_{\overline{T}},\\
&\mathrm{d}\bar{Y}_t = \beta_{\overline{T}-t}\big(Y_t + 2\mathbb{1}[(\bar{Y}_s, \overline{T}-s) \notin A \text{ for some } s \le t]\nabla \log p_t(\bar{Y}_t)\\
&\quad + 2\mathbb{1}[(Y_s, \overline{T}-s) \in A \text{ for all } s \le t]\hat{s}(\bar{Y}_{\overline{T}-t_k}, \overline{T}-t_k)\big)\,\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_i, \overline{T}-t_{i-1}]).
\end{aligned}$$

At $t = \overline{T} - \underline{T}$, if the $\infty$-norm is more than 2, then we reset it to 0. Here, $A \subseteq \mathbb{R}^{d+1}$ is defined as

$$A = \left\{(x,t) \in \mathbb{R}^d \times \mathbb{R} \,\middle|\, \|x\|_\infty \le m_t + C_{\mathrm{a}}\sigma_t\sqrt{\log(n)},\ \underline{T} \le t \le \overline{T}\right\}.$$

Then, we have that

$$\begin{aligned}
\mathrm{TV}(Y_{\overline{T}}, Y^{\mathrm{d}}_{\overline{T}-\underline{T}}) &\le \mathrm{TV}(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}}) + \mathrm{TV}(Y_0, \bar{Y}_{\overline{T}-\underline{T}}) + \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}}) + \mathrm{TV}(\bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}})\\
&\le \mathrm{TV}(X_0, X_{\underline{T}}) + \mathrm{TV}(Y_0, \bar{Y}_{\overline{T}-\underline{T}}) + \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}}) + \mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d)).
\end{aligned}$$

The first term is bounded by $n^{-\frac{2s}{d+2s}}$, by setting $\underline{T} = n^{-\mathcal{O}(1)}$ in Theorem F.2. The second term is bounded by $n^{-\frac{2s}{d+2s}}$, by taking $C_{\mathrm{a}}$ sufficient large, according to Lemma C.1. The forth term is bounded by $\exp(-\underline{\beta}\overline{T})$ by Lemma F.3, and thus setting $\underline{T} = \mathcal{O}(\log n)$ yields $\exp(-\underline{\beta}\overline{T}) \lesssim n^{-\frac{2s}{d+2s}}$.

Now, we bound the third term. Proposition F.1 yields that

$$\begin{aligned}
&\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}})\\
&\lesssim \sum_{k=1}^K \int_{t=\overline{T}-t_k}^{\overline{T}-t_{k-1}} \mathbb{E}_{\bar{Y}}[\mathbb{1}[(\bar{Y}_s, \overline{T}-s) \in A \text{ for all } s \le t]\|\hat{s}(\bar{Y}_{\overline{T}-t_k}, \overline{T}-t_k) - \nabla \log p_t(\bar{Y}_t)\|^2]\mathrm{d}t\\
&\le \sum_{k=1}^K \int_{t=\overline{T}-t_k}^{\overline{T}-t_{k-1}} \mathbb{E}_{\bar{Y}}[\mathbb{1}[(\bar{Y}_t, \overline{T}-t) \in A, (\bar{Y}_{\overline{T}-t_k}, t_k) \in A]\|\hat{s}(\bar{Y}_{\overline{T}-t_k}, \overline{T}-t_k) - \nabla \log p_t(\bar{Y}_t)\|^2]\mathrm{d}t\\
&\le \sum_{k=1}^K \int_{t=t_{k-1}}^{t_k} \mathbb{E}_X[\mathbb{1}[(X_t, t) \in A, (X_{t_k}, t_k) \in A]\|\hat{s}(X_{t_k}, t_k) - \nabla \log p_t(X_t)\|^2]\mathrm{d}t\\
&\lesssim \sum_{k=1}^K \int_{t=t_{k-1}}^{t_k} \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|\hat{s}(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k})\|^2]\mathrm{d}t \qquad (99)\\
&\quad + \sum_{k=1}^K \int_{t=t_{k-1}}^{t_k} \mathbb{E}_X[\mathbb{1}[(X_t, t) \in A, (X_{t_k}, t_k) \in A]\|\nabla \log p_t(X_t) - \nabla \log p_{t_k}(X_t)\|^2]\mathrm{d}t \qquad (100)\\
&\quad + \sum_{k=1}^K \int_{t=t_{k-1}}^{t_k} \mathbb{E}_X[\mathbb{1}[(X_t, t) \in A, (X_{t_k}, t_k) \in A]\|\nabla \log p_{t_k}(X_t) - \nabla \log p_{t_k}(X_{t_k})\|^2]\mathrm{d}t \qquad (101)
\end{aligned}$$

First, we consider (100). Because $(X_t, t) \in A$, $(\|X_t\|_\infty - m_t)_+ \lesssim \sigma_t\sqrt{\log(n)}$. Over all $t \le s \le t_k$, $|\partial_s \sigma_s| \lesssim \frac{1}{\sqrt{t}}$, $|\partial_s m_s| \lesssim 1$, and

$$\begin{aligned}
\|\partial_s \nabla \log p_s(x)\| &\lesssim \frac{|\partial_s\sigma_s| + |\partial_s m_s|}{\sigma_s^3}\left(\frac{(\|x\|_\infty - m_s)_+^2}{\sigma_s^2} \vee 1\right)^{\frac{3}{2}}\\
&\lesssim \frac{|\partial_t\sigma_{t_k}| + |\partial_t m_{t_k}|}{\sigma_{t_k}^3}\left(\frac{(\|x\|_\infty - m_{t_k})_+^2}{\sigma_{t_k}^2} \vee 1\right)^{\frac{3}{2}},
\end{aligned}$$

according to Lemma C.3. Therefore, (100) is bounded by $\sum_{k=1}^K \eta(\eta(t_k^{-2} \vee 1)\log^{\frac{3}{2}} n)^2 = \eta^2(t_k^{-4} \vee 1)\log^3 n$.

Next, for (101), we first note that $\|X_t\|_\infty - m_{t_k}, \|X_{t_k}\|_\infty - m_{t_k} \lesssim \sigma_{t_k}\sqrt{\log(n)} = \tilde{\mathcal{O}}(1)$. Therefore, according to Lemma C.3, $\|\partial_{x_i}\nabla \log p_{t_k}(x)\|$ is bounded by $\frac{1}{\sigma_{t_k}^2}\left(\frac{(\|X_{t_k}\|_\infty - m_{t_k})_+^2}{\sigma_{t_k}^2} \vee 1\right) \lesssim t_k^{-1}\log n$. With probability at least $1-n^{-\mathcal{O}(1)}$, $\|X_t - X_{t_k}\|_\infty \lesssim \sqrt{\eta \log n}$, according to Lemma J.15. Therefore,

$$(101) \lesssim \sum_{k=1}^{K} \eta(\sqrt{\eta \log n}\cdot(t_k^{-1}\vee 1)\log n)^2 + n^{-\mathcal{O}(1)}\cdot \tilde{\mathcal{O}}(1) \lesssim \sum_{k=1}^{K}\eta^2(t_k^{-2}\vee 1)\log^3 n.$$

Finally, for (101), we apply (96). Now, all three terms of (99), (100), and (101) are bounded and we obtain that

$$\begin{aligned}\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}})\right] &\lesssim n^{-\frac{2s}{d+2s}}\log^{18} n + \sum_{k=1}^{K}(\eta^3(t_k^{-4}\vee 1)\log^3 n + \eta^2(t_k^{-2}\vee 1)\log^3 n)\\ &\lesssim n^{-\frac{2s}{d+2s}}\log^{18} n + \eta^2\underline{T}^{-3}\log^3 n + \eta\underline{T}^{-1}\log^3 n + \eta\overline{T}\log^3 n\\ &\lesssim n^{-\frac{2s}{d+2s}}\log^{18} n + \eta^2\underline{T}^{-3}\log^3 n + \eta\underline{T}^{-1}\log^3 n + \eta\log^4 n.\end{aligned}$$

Therefore, by setting $\eta = \underline{T}^{-1.5}n^{-\frac{s}{d+2s}}$ yields the assersion.

□

# J Auxiliary lemmas

This final section summarizes existing results and prepares basic tools for the main parts of the proofs. A large part of this section (Appendix J.1 to J.4) is devoted to introduction of basic tools for the function approximation with neural networks, and thus those familiar with such topics (Yarotsky, 2017; Petersen & Voigtlaender, 2018; Schmidt-Hieber, 2019) can skip these subsections (although they contain some refinement and extension). Lemma J.14 is for elementary bounds on the Gaussian distribution and hitting time of the Brownian motion.

In the following we will define constants $C_{\mathrm{f},1}$ and $C_{\mathrm{f},2}$. Other than in this section, they are denoted by $C_{\mathrm{f}}$, and sometimes other constants that comes from this section can be also denoted by $C_{\mathrm{f}}$.

## J.1 Construction of a larger neural network

Through construction of the desired neural network, we often need to combine sub-networks that approximates simpler functions to realize more complicated functions. We prepare the following lemmas, whose direct source is Nakada & Imaizumi (2020) but similar ideas date back to earlier literature such as Yarotsky (2017); Petersen & Voigtlaender (2018).

First we consider construction of composite functions. Although the bound on the sparsity $S$ was not given in the original version, we can verify it by carefully checking their proof.

**Lemma J.1** (Concatenation of neural networks (Remark 13 of Nakada & Imaizumi (2020)))**.** *For any neural networks $\phi^1\colon \mathbb{R}^{d_1}\to\mathbb{R}^{d_2}, \phi^2\colon\mathbb{R}^{d_2}\to\mathbb{R}^{d_3},\cdots,\phi^k\colon\mathbb{R}^{d_k}\to\mathbb{R}^{d_{k+1}}$ with $\phi^i \in \Psi(L^i, W^i, S^i, B^i)$ $(i=1,2,\cdots,d)$, there exists a neural network $\phi\in\Phi(L,W,S,B)$ satisfying $\phi(x) = \phi^k\circ\phi^{k-1}\cdots\circ\phi^1(x)$ for all $x\in\mathbb{R}^{d_1}$, with*

$$L = \sum_{i=1}^{k} L^i, \quad W \le 2\sum_{i=1}^{k} W^i, \quad S \le \sum_{i=1}^{k} S^i + \sum_{i=1}^{k-1}(\|A^i_{L^i}\|_0 + \|b^i_{L^i}\|_0 + \|A^{i+1}_1\|_0) \le 2\sum_{i=1}^{k} S^i,$$

*and $B \le \max_{1\le i\le k} B^i$.*

*Here $A^i_j$ is the parameter matrix and $b^i_j$ is the bias vector at the $j$th layer of the $i$th neural network $\phi^i$.*

Next we introduce the identity function.

**Lemma J.2** (Identity function (p.19 of Nakada & Imaizumi (2020)))**.** *For $L \geq 2$ and $d \in \mathbb{N}$, there exists a neural network $\phi_{\mathrm{Id}}^{d,L} \in \Phi(L, W, S, B)$ with parameters $(A_1, b_1) = ((I_d, -I_d)^\top, 0), (A_i, b_i) = (I_{2d}, 0)(i = 1, 2, \cdots, L-2), (A_L) = ((I_d, -I_d), 0)$, that realize d-dimensional identity map. Here,*

$$\|W\|_\infty = 2d, \quad S = 2dL, \quad B = 1.$$

*For $L = 1$, a neural network $\phi_{\mathrm{Id}}^{d,1} \in \Phi(1, (d), d, 1)$ with parameters $(A_1, b_1) = (I_d, 0)$ realizes $d$-dimensional identity map.*

We then consider parallelization of neural networks. The following lemmas are Remarks 14 and 15 of Nakada & Imaizumi (2020) with a modification to allow sub-networks to have different depths.

**Lemma J.3** (Parallelization of neural networks)**.** *For any neural networks $\phi^1, \phi^2, \cdots, \phi^k$ with $\phi^i \colon \mathbb{R}^{d_i} \to \mathbb{R}^{d_i'}$ and $\phi^i \in \Psi(L^i, W^i, S^i, B^i)$ $(i = 1, 2, \cdots, d)$, there exists a neural network $\phi \in \Phi(L, W, S, B)$ satisfying $\phi(x) = [\phi^1(x^1)^\top\ \phi^2(x^2)^\top\ \cdots\ \phi^k(x^k)^\top]^\top \colon \mathbb{R}^{d_1+d_2+\cdots+d_k} \to \mathbb{R}^{d_1'+d_2'+\cdots+d_k'}$ for all $x = (x_1^\top\ x_2^\top\ \cdots\ x_k^\top)^\top \in \mathbb{R}^{d_1+d_2+\cdots+d_k}$ (here $x_i$ can be shared), with*

$$L = L, \|W\|_\infty \leq \sum_{i=1}^k \|W^i\|_\infty, S \leq \sum_{i=1}^k S^i, \text{and } B \leq \max_{1 \leq i \leq k} B^i \text{ (when } L = L_i \text{ holds for all } i),$$

$$L = \max_{1 \leq i \leq k} L^i, \|W\|_\infty \leq 2 \sum_{i=1}^k \|W^i\|_\infty, S \leq 2 \sum_{i=1}^k (S^i + LW_L^i), \text{and } B \leq \max\{\max_{1 \leq i \leq k} B^i, 1\} \text{ (otherwise)}.$$

*Moreover, there exists a network $\phi_{\mathrm{sum}}(x) \in \Phi(L, W, S, B)$ that realizes $= \sum_{i=1}^k \phi^i(x)$, with*

$$L = \max_{1 \leq i \leq k} L^i + 1, \quad \|W\|_\infty \leq 4 \sum_{i=1}^k \|W^i\|_\infty, \quad S \leq 4 \sum_{i=1}^k (S^i + LW_L) + 2W_L,$$

$$\text{and } B \leq \max\{\max_{1 \leq i \leq k} B^i, 1\}.$$

*Proof of Lemma J.3.* Let us consider the first part. For the case when $L = L_i$ holds for all $i$, the assertions are exactly the same as Remarks 14 and 15 Nakada & Imaizumi (2020). Otherwise, we first prepare a network $\phi'^i$ realizing $\phi_{\mathrm{Id}}^{d, L-L_i} \circ \phi^i$ for all $i$, so that every network have the same depth without changing outputs of the networks. From Lemmas J.1 and J.2, $\phi'^i \in \Phi(L, W'^i, S'^i, B'^i)$ holds, with $L = \max_{1 \leq i \leq k} L^i, \|W'^i\|_\infty = \max\{\|W^i\|_\infty, 2W_L\} \leq 2\|W^i\|_\infty, S'^i \leq 2S^i + 2(L - L_i)W_L^i \leq 2(S^i + LW_L^i)$, and $B'^i = \max\{B^i, 1\}$. We then apply the results for the case of $L = L_i$ $(i = 1, 2, \cdots, k)$.

For the second part, since summation of the outputs of $k$ neural networks can be realized by a 1 layer neural network with the width of $k$, Lemma J.3 together with Lemma J.1 gives the bound to realize $\sum_{i=1}^k \phi^i(x)$. □

In the analysis of the score-based diffusion model, we often face unbounded functions. To resolve difficulty coming from the unboundedness, the clippling operation is often be adopted.

**Lemma J.4** (Clipping function)**.** *For any $a, b \in \mathbb{R}^d$ with $a_i \leq b_i$ $(i = 1, 2, \cdots, d)$, there exists a clipping function $\phi_{\mathrm{clip}}(x; a, b) \in \Phi(2, (d, 2d, d)^\top, 7d, \max_{1 \leq i \leq d} \max\{|a_i|, b_i\})$ such that*

$$\phi_{\mathrm{clip}}(x; a, b)_i = \min\{b_i, \max\{x_i, a_i\}\} \quad (i = 1, 2, \cdots, d)$$

*holds. When $a_i = c$ and $b_i = C$ for all $i$, we sometimes denote $\phi_{\mathrm{clip}}(x; a, b)$ as $\phi_{\mathrm{clip}}(x; c, C)$ using scaler values $c$ and $C$.*

*Proof.* Because, for each coordinate $i$, $\min\{b_i, \max\{x_i, a_i\}\}$ is realized as

$$\min\{b_i, \max\{x_i, a_i\}\} = \mathrm{ReLU}(x_i - a_i) - \mathrm{ReLU}(x_i - b_i) + a_i \in \Phi(2, (1, 2, 1), 7, \max\{|a_i|, b_i\}),$$

parallelizing this for all $i$ with Lemma J.3 yields the assertion. □

With the above clipping function, we prepare switching functions, which gives the way to construct approximation in the combined region when there are two different approximations valid for different regions.

**Lemma J.5** (Switching function)**.** *Let $\underline{t}_1 < \underline{t}_2 < \bar{t}_1 < \bar{t}_2$, and $f(x,t)$ be some scaler-valued function (for a vector-valued function, we just apply this coordinate-wise). Assume that $\phi^1(x,t)$ and $\phi^2(x,t)$ approximate $f(x,t)$ up to an additive error of $\epsilon$ but approximation with $\phi^1(x,t)$ and $\phi^2(x,t)$ are valid for $[\underline{t}_1, \bar{t}_1]$ and $[\underline{t}_2, \bar{t}_2]$, respectively. Then, there exist neural networks $\phi^1_{\rm swit}(t;\underline{t}_2,\bar{t}_1), \phi^2_{\rm swit}(t;\underline{t}_2,\bar{t}_1) \in \Phi(3,(1,2,1,1)^\top, 8, \max\{\bar{t}_1, (\bar{t}_1-\underline{t}_2)^{-1}\})$, and $\phi^1_{\rm swit}(t;\underline{t}_2,\bar{t}_1)\phi^1(x,t) + \phi^2_{\rm swit}(t;\underline{t}_2,\bar{t}_1)\phi^2(x,t)$ approximates $f(x,t)$ up to an additive error of $\epsilon$ in $[\underline{t}_1, \bar{t}_2]$.*

*Proof.* We define

$$\phi^1_{\rm swit}(t;\underline{t}_2,\bar{t}_1) = \frac{1}{\bar{t}_1-\underline{t}_2}\mathrm{ReLU}(\phi_{\rm clip}(t;\underline{t}_2,\bar{t}_1)-\underline{t}_2),$$
$$\text{and } \phi^2_{\rm swit}(t;\underline{t}_2,\bar{t}_1) = \frac{1}{\bar{t}_1-\underline{t}_2}\mathrm{ReLU}(\bar{t}_1-\phi_{\rm clip}(t;\underline{t}_2,\bar{t}_1)).$$

Here $\phi^1_{\rm swit}(t;\underline{t}_2,\bar{t}_1), \phi^2_{\rm swit}(t;\underline{t}_2,\bar{t}_1) \in [0,1]$, $\phi^1_{\rm swit}(t;\underline{t}_2,\bar{t}_1) + \phi^2_{\rm swit}(t;\underline{t}_2,\bar{t}_1) = 1$ for all $t$, $\phi^1_{\rm swit}(t;\underline{t}_2,\bar{t}_1) = 0$ for all $t \geq \bar{t}_1$, and $\phi^2_{\rm swit}(t;\underline{t}_2,\bar{t}_1)$ for $t \leq \underline{t}_2$. From this construction, the assertion follows. □

### J.2 Basic neural network structure that approximates rational functions

When approximating a function in the Besov space with a neural network, the most basic structure of the network is that of approximating polynomials (Suzuki, 2018). In our construction of the diffused B-spline basis, we need to approximate rational functions.

We begin with monomials. Although the traditional fact that we can approximate monomials with neural networks with an arbitrary additive error of $\epsilon$ using only $\mathcal{O}(\log\varepsilon^{-1})$ non-zero parameters has been very famous (Yarotsky, 2017; Petersen & Voigtlaender, 2018; Schmidt-Hieber, 2020), we could not find the result that explicitly states the dependency on parameters including the degree and the range of the input. Therefore, just to be sure, we revisit Lemma A.3 of Schmidt-Hieber (2020) and here gives the extended version of that lemma.

**Lemma J.6** (Approximation of monomials)**.** *Let $d \geq 2$, $C \geq 1$, $0 < \varepsilon_{\rm error} \leq 1$. For any $\varepsilon > 0$, there exists a neural network $\phi_{\rm mult}(x_1,x_2,\cdots,x_d) \in \Psi(L,W,S,B)$ with $L = \mathcal{O}(\log d(\log\varepsilon^{-1} + d\log C))$, $\|W\|_\infty = 48d$, $S = \mathcal{O}(d\log\varepsilon^{-1} + d\log C))$, $B = C^d$ such that*

$$\left|\phi_{\rm mult}(x'_1,x'_2,\cdots,x'_d) - \prod_{d'=1}^d x_{d'}\right| \leq \varepsilon + dC^{d-1}\varepsilon_{\rm error},$$
$$\text{for all } x \in [-C,C]^d \text{ and } x' \in \mathbb{R} \text{ with } \|x-x'\|_\infty \leq \varepsilon_{\rm error},$$

*$|\phi_{\rm mult}(x)| \leq C^d$ for all $x \in \mathbb{R}^d$, and $\phi_{\rm mult}(x'_1,x'_2,\cdots,x'_d) = 0$ if at least one of $x'_i$ is 0.*

*We note that some of $x_i, x_j$ $(i \neq j)$ can be shared. For $\prod_{i=1}^I x_i^{\alpha_i}$ with $\alpha_i \in \mathbb{Z}_+$ $(i = 1,2,\cdots,I)$ and $\sum_{i=1}^I \alpha_i = d$, there exists a neural network satisfying the same bounds as above, and the network is denoted by $\phi_{\rm mult}(x;\alpha)$.*

*Proof.* First of all, it is known from Schmidt-Hieber (2020) that there exists a neural network $\bar{\phi}'_{\rm mult}(x,y) \in \Psi(L,W,S,B)$ with $L = i+5$, $\|W\|_\infty = 6$, $B = 1$ such that

$$|\bar{\phi}'_{\rm mult}(x,y) - xy| \leq 2^{-i}, \quad \text{for all } (x,y) \in [0,1]^2,$$

and $|\bar{\phi}'_{\rm mult}(x,y)| \leq 1$ for all $(x,y) \in \mathbb{R}^2$, and $\bar{\phi}'_{\rm mult}(x,y) = 0$ if either $x$ or $y$ is 0. With this network, we can see that $|\mathrm{sign}(xy)\bar{\phi}'_{\rm mult}(|x|,|y|) - xy| \leq 2^{-i}$ holds for all $(x,y) \in [-1,1]^2$,

$|\bar{\phi}'_{\mathrm{mult}}(x,y)| \le 1$ for all $(x,y) \in \mathbb{R}^2$, and $\bar{\phi}_{\mathrm{mult}}(x,y) = 0$ if either $x$ or $y$ is 0. Because

$$\begin{aligned}\mathrm{sign}(xy)\bar{\phi}'_{\mathrm{mult}}(|x|,|y|) &= \mathrm{ReLU}(\bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(x),\mathrm{ReLU}(y)) + \bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(-x),\mathrm{ReLU}(-y)) \\ &\quad - \bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(-x),\mathrm{ReLU}(y)) - \bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(x),\mathrm{ReLU}(-y))) \\ &\quad - \mathrm{ReLU}(-\bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(x),\mathrm{ReLU}(y)) - \bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(-x),\mathrm{ReLU}(-y)) \\ &\quad + \bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(-x),\mathrm{ReLU}(y)) + \bar{\phi}'_{\mathrm{mult}}(\mathrm{ReLU}(x),\mathrm{ReLU}(-y))) \\ &=: \bar{\phi}_{\mathrm{mult}}(x,y)\end{aligned}$$

holds, we can realize the function $xy$ for $[-1,1]^d$, by a neural network $\bar{\phi}_{\mathrm{mult}}(x,y) \in \Psi(L,W,S,B)$ with $L = i+7, \|W\|_\infty = 48, S \le L\|W\|_\infty(\|W\|_\infty+1) = 48(i+7), B = 1$ with an approximation error up to $2^{-i}$.

Then, following Schmidt-Hieber (2020), we recursively construct $\bar{\phi}_{\mathrm{mult}}(x_1,x_2,\cdots,x_{2^{j+1}})$ using

$$\bar{\phi}_{\mathrm{mult}}(x_1,x_2,\cdots,x_{2^{j+1}}) = \bar{\phi}_{\mathrm{mult}}(\bar{\phi}_{\mathrm{mult}}(x_1,x_2,\cdots,x_{2^j}),\bar{\phi}_{\mathrm{mult}}(x_{2^j+1},x_{2^j+2},\cdots,x_{2^{j+1}})).$$

By filling extra dimensions of $(x_1,x_2,\cdots,x_{2^j})$ with 1, we obtain the neural network $\phi_{\mathrm{mult}}(x_1,x_2,\cdots,x_d) \in \Psi(L,W,S,B)$ for all $d \ge 2$ and $L = \mathcal{O}(\log d(\log\varepsilon^{-1}+\log d)), \|W\|_\infty = 48d, S = \mathcal{O}(d(\log\varepsilon^{-1}+\log d)), B = 1$ such that

$$\left|\bar{\phi}_{\mathrm{mult}}(x_1,x_2,\cdots,x_d) - \prod_{d'=1}^{d} x_{d'}\right| \le \varepsilon, \quad \text{for all } x \in [-1,1]^d.$$

We then construct $\phi_{\mathrm{mult}}$ as follows:

$$\phi_{\mathrm{mult}}(x) = C^d\bar{\phi}_{\mathrm{mult}}(\phi_{\mathrm{clip}}(x;-C,C)/C).$$

Here the approximation error over $[-C,C]^d$ is bounded by $C^{-d}\varepsilon$. We reset $\varepsilon \leftarrow C^{-d}\varepsilon$ so that the approximation error is smaller than $\varepsilon$, and then we have $\phi_{\mathrm{mult}} \in \Phi(L,W,S,B)$ with $L = \mathcal{O}(\log d(\log d + \log\varepsilon^{-1} + d\log C)), \|W\|_\infty = 48d, S = \mathcal{O}(d(\log d + \log\varepsilon^{-1} + d\log C)), B = 1$. Therefore, the bounds on $L, \|W\|_\infty, B, S$ in the assertion follows from Lemmas J.1 and J.4.

When the input fluctuates, we have

$$\begin{aligned}&\left|C^d\bar{\phi}_{\mathrm{mult}}(\phi_{\mathrm{clip}}(x';-C,C)/C) - \prod_{i=1}^{d} x_i\right| \\ &\le \left|C^d\bar{\phi}_{\mathrm{mult}}(\phi_{\mathrm{clip}}(x';-C,C)/C) - \prod_{i=1}^{d}\min\{C,\max\{x'_i,-C\}\}\right| \\ &\quad + \left|\prod_{i=1}^{d}\min\{C,\max\{x'_i,-C\}\} - \prod_{i=1}^{d} x_i\right| \\ &\le C^d \cdot C^{-d}\varepsilon + C^{d-1}\sum_{i=1}^{d}|x_i - \min\{C,\max\{x'_i,-C\}\}| = \varepsilon + dC^{d-1}\varepsilon_{\mathrm{error}},\end{aligned}$$

which yields the first part of the assertion.

Finally, we note that some of $x_i, x_j$ $(i \ne j)$ can be shared because all we need is to identify columns in the first layer of $\bar{\phi}_{\mathrm{mult}}(x_1,\cdots,x_d)$ that correspond to the same coordinate. □

We next provide how to approximate the reciprocal function $y = \frac{1}{x}$. Approximation of rational functions has already investigated in (Telgarsky, 2017; Boullé et al., 2020). However, we found that their bounds (in Lemma 3.5 of Telgarsky (2017)) of $L = \mathcal{O}(\log^7\varepsilon^{-1})$ and $\mathcal{O}(\log^4\varepsilon^{-1})$ nodes can be improved with careful use of local Taylor expansion up to the order of $\mathcal{O}(\log\varepsilon^{-1})$, so we provide our own proof.

**Lemma J.7** (Approximating the reciprocal function)**.** *For any* $0 < \varepsilon < 1$*, there exists* $\phi_{\mathrm{rec}} \in \Psi(L,W,S,B)$ *with* $L \le \mathcal{O}(\log^2\varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log^3\varepsilon^{-1}), S = \mathcal{O}(\log^4\varepsilon^{-1})$*, and* $B = \mathcal{O}(\varepsilon^{-2})$ *such that*

$$\left|\phi_{\mathrm{rec}}(x') - \frac{1}{x}\right| \le \varepsilon + \frac{|x'-x|}{\varepsilon^2}, \quad \textit{for all } x \in [\varepsilon,\varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

*Proof.* We approximate the inverse function $y = \frac{1}{x}$ with a piece-wise polynomial function. We take $x_i = 1.5^i \cdot \varepsilon$ $(i = 0, 1, \cdots, i^* := \lceil 2\log_{1.5}\varepsilon^{-1}\rceil)$ so that $x_{i^*} \geq \varepsilon^{-1}$ and approximate $y = \frac{1}{x}$ in the following way:

$$\frac{1}{x} \fallingdotseq \sum_{i=1}^{i^*} f_i(\phi_{\text{clip}}(x; x_{i-1}, x_i)) + \frac{1}{\varepsilon},$$

where $f_i(x)$ is a function that satisfies $f_i(x) = 0$ for $x \leq x_{i-1}$, $f_i(x) = -\frac{1}{x_{i-1}} + \frac{1}{x_i}$ for $x_i \leq x$, and

$$\max_{x_{i-1}\leq x\leq x_i} |f_i(x) - 1/x + 1/x_{i-1}| \leq \frac{\varepsilon}{2}.$$

Now we show construction of such functions. First, by $\frac{1}{x} = \frac{1}{x_{i-1}}\frac{x_{i-1}}{x} = \frac{1}{x_{i-1}}\sum_{l'=1}^{\infty}(-\frac{x}{x_{i-1}} + 1)^{l'}$ $(1 \leq \frac{x}{x_{i-1}} \leq 1.5)$, let

$$\tilde{f}_i(x) = \frac{1}{x_{i-1}}\sum_{l'=1}^{l}(-x/x_{i-1} + 1)^{l'} - \frac{1}{x_{i-1}}.$$

The difference between $\tilde{f}_i(x)$ and $\frac{1}{x} - \frac{1}{x_{i-1}}$ is $((x_{i-1} - x)/x_{i-1})^{l+1}/x$, which is bounded by $2^{-l-1}/x$. Moreover, by adding $\frac{(\frac{1}{x_i} - \tilde{f}_i(x_i))(x - x_{i-1})}{x_i - x_{i-1}} = \frac{((x_{i-1}-x_i)/x_{i-1})^{l+1}(x-x_{i-1})}{x_i(x_i - x_{i-1})}$ to $\tilde{f}_i(x)$, we have $f_i(x)$, with $f_i(x_{i-1}) = 0$, $f_i(x_i) = -\frac{1}{x_{i-1}} + \frac{1}{x_i}$, and

$$\max_{x_{i-1}\leq x\leq x_i} |f_i(x) - 1/x + 1/x_{i-1}| \leq 2^{-l}/x \leq 2^{-l}\varepsilon^{-1}.$$

Thus, we take $l = \lceil \log_2 2\varepsilon^{-1}\rceil$ so that RHS is smaller than $\frac{\varepsilon}{2}$. Therefore, we finally have the explicit approximation of $y = \frac{1}{x}$:

$$\begin{aligned} f(x) =& \underbrace{\sum_{i=1}^{i^*}\frac{1}{x_{i-1}}\sum_{l'=1}^{l}(-\phi_{\text{clip}}(x; x_{i-1}, x_i))/x_{i-1} + 1)^{l'}}_{\text{(a)}} - \sum_{i=1}^{i^*}\frac{1}{x_{i-1}} \\ &+ \underbrace{\sum_{i=1}^{i^*}\frac{((x_{i-1} - x_i)/x_{i-1})^{l+1}(\phi_{\text{clip}}(x; x_{i-1}, x_i)) - x_{i-1})}{x_i(x_i - x_{i-1})}}_{\text{(b)}} + \frac{1}{\varepsilon}. \end{aligned} \tag{102}$$

From Lemma J.6, $(-\phi_{\text{clip}}(x; x_{i-1}, x_i))/x_{i-1} + 1)^{l'}$ is realized by $L = \mathcal{O}((\log\log\varepsilon^{-1} + \log\varepsilon^{-1})\log\log\varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log\varepsilon^{-1}), S = \mathcal{O}(\log\varepsilon^{-1}(\log\log\varepsilon^{-1} + \log\varepsilon^{-1})), B = 1.5^{\lceil\log_2 2\varepsilon^{-1}\rceil} = \mathcal{O}(\varepsilon^{-1})$ so that approximation error for each is bounded by $\mathcal{O}(\varepsilon^2/li^*)$. Because there are $\mathcal{O}(li^*)$ terms in (a) of (102), from Lemmas J.1 and J.3, the final approximation error of $f(x)$ using a neural network $\phi_{\text{rec}}$ is $\frac{\varepsilon}{2}$, where $\phi_{\text{rec}} \in \Phi(L, W, S, B)$ with $L \leq \mathcal{O}((\log\log\varepsilon^{-1} + \log\varepsilon^{-1})\log\log\varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log^3\varepsilon^{-1}), S = \mathcal{O}(\log^3\varepsilon^{-1}(\log\log\varepsilon^{-1} + \log\varepsilon^{-1}))$, and $B = \mathcal{O}(\varepsilon^{-2})$. (Here $B = \mathcal{O}(\varepsilon^{-2})$ is calculated because in (b) we need to bound the coefficient $\frac{((x_{i-1}-x_i)/x_{i-1})^{l+1}}{x_i(x_i-x_{i-1})}$ by $\varepsilon^{-2}$.)

The sensitivity analysis follows from $|\phi_{\text{rec}}(x') - \frac{1}{x}| \leq |\phi_{\text{rec}}(x') - \frac{1}{\max\{x', \varepsilon\}}| + |\frac{1}{\max\{x',\varepsilon\}} - \frac{1}{x}|$. $\square$

Combining Lemmas J.6 and J.7, we have the following corollary.

**Corollary J.8.** *For any* $0 < \varepsilon < 1$, *there exists* $\phi_{\text{rec}} \in \Psi(L, W, S, B)$ *with* $L \leq \mathcal{O}(\log^2 l + \log^2\varepsilon)), \|W\|_\infty = \mathcal{O}(l + \log^3\varepsilon^{-1}), S = \mathcal{O}(l\log l + l\log\varepsilon^{-1} + \log^4\varepsilon^{-1})$, *and* $B = \mathcal{O}(\varepsilon^{-(2\vee l)})$ *such that*

$$\left|\phi_{\text{rec}}(x'; l) - \frac{1}{x^l}\right| \leq \varepsilon + l\frac{|x' - x|}{\varepsilon^{l+1}}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

*Proof.* Consider $\phi_{\mathrm{mult}}(\cdot; l) \circ \phi_{\mathrm{rec}}$. The result directly follows from Lemma J.6 and Lemma J.7. $\square$

In the same way, by using Taylor expansion of $\sqrt{1+x}$ at each interval defined in the above proof, we can obtain a similar result for $y = \sqrt{x}$.

**Lemma J.9** (Approximating the root function)**.** *For any* $0 < \varepsilon < 1$*, there exists* $\phi_{\mathrm{root}} \in \Psi(L, W, S, B)$ *with* $L \leq \mathcal{O}(\log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1}), S = \mathcal{O}(\log^4 \varepsilon^{-1})$*, and* $B = \mathcal{O}(\varepsilon^{-1})$ *such that*

$$\left|\phi_{\mathrm{root}}(x') - \sqrt{x}\right| \leq \varepsilon + \frac{|x'-x|}{\sqrt{\varepsilon}}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

### J.3 How to deal with exponential functions

We sometimes need to approximate certain types of integrals where the integrand contains a density function of some Gaussian distribution and the integral interval is $\mathbb{R}^d$. for example, the diffused B-spline basis is a typical example of them. To deal with them, we adopt the following two-step argument: first we clip the integral interval, and next we approximate the integrand with rational functions. We need rational functions because the density function depends on the inverse of (the squared-root of) the variance, which depends on $t$ and should be approximated. The first lemma corresponds to the first step, and the second and third correspond to the second step, respectively.

**Lemma J.10** (Clipping of integrals)**.** *Let* $x \in \mathbb{R}^d$*,* $0 < m_t \leq 1$*,* $\alpha \in \mathbb{Z}_+^d$ *with* $\sum_{i=1}^d \alpha_i \leq k$*, and* $f$ *be an any function on* $\mathbb{R}^d$ *whose absolute value is bounded by* $C_f$*. For any* $0 < \varepsilon < \frac{1}{2}$*, there exists a constant* $C_{\mathrm{f},1}$ *that only depends on* $k$ *and* d*, such that*

$$\Bigg|\int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t}\right)^{\alpha_i} f(y) \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2}\right) dy \\ - \int_{A^x} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t}\right)^{\alpha_i} f(y) \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2}\right) dy\Bigg| \lesssim \varepsilon,$$

*where* $A^x = \prod_{i=1}^d a_i^x$ *with* $a_i^x = [\frac{x_i}{m_t} - \frac{\sigma_t C_{\mathrm{f},1}}{m_t}\sqrt{\log \varepsilon^{-1}}, \frac{x_i}{m_t} + \frac{\sigma_t C_{\mathrm{f},1}}{m_t}\sqrt{\log \varepsilon^{-1}}]$.

*Proof.*

$$\begin{aligned}
&\frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}}\Bigg|\int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t}\right)^{\alpha_i} f(y) \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2}\right) dy \\
&\qquad - \int_{A^x} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t}\right)^{\alpha_i} f(y) \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2}\right) dy\Bigg| \\
&\leq \frac{C_f}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d \backslash A^x} \prod_{i=1}^d \left(\frac{|m_t y_i - x_i|}{\sigma_t}\right)^{\alpha_i} \mathbb{1}[\|y\|_\infty \leq 1] \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2}\right) dy \quad (\text{by } |f(y)| \leq C_f) \\
&\leq \frac{C_f}{\sigma^d (2\pi)^{\frac{d}{2}}} \sum_{i=1}^d \int_{\underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{i-1 \text{ times}} \times (\mathbb{R}\backslash a_i^x) \times \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{d-i \text{ times}}} \mathrm{d}y \\
&\qquad \prod_{j=1}^d \left(\frac{|m_t y_j - x_j|}{\sigma_t}\right)^{\alpha_j} \mathbb{1}[|y_j| \leq 1] \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2}\right) \\
&= C_f \sum_{i=1}^d \prod_{j=1}^d \Bigg(\mathbb{1}[i \neq j] \int_{\mathbb{R}} \left(\frac{|m_t y_j - x_j|}{\sigma_t}\right)^{\alpha_j} \frac{\mathbb{1}[|y_j| \leq 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2}\right) dy_j \\
&\qquad + \mathbb{1}[i = j] \int_{\mathbb{R}\backslash a_i^x} \left(\frac{|m_t y_j - x_j|}{\sigma_t}\right)^{\alpha_j} \frac{\mathbb{1}[|y_j| \leq 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2}\right) dy_j\Bigg). \quad (103)
\end{aligned}$$

We now bound each term. First,

$$
\begin{aligned}
&\int_{\mathbb{R}} \left(\frac{|m_t y_j - x_j|}{\sigma_t}\right)^{\alpha_j} \frac{\mathbb{1}[|y_j| \le 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2}\right) dy_j \\
&\le \begin{cases} \frac{1}{m_t}\int_{\mathbb{R}} |y_j'|^{\alpha_j} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{y_j'^2}{2}\right) dy_j' & \left(\frac{m_t y_j - x_j}{\sigma_t} = y_j'\right) \\ \frac{2^{d+\alpha_j}}{\sigma_t^{\alpha_j+1}(2\pi)^{\frac{1}{2}}} & \text{(because of the term of } \mathbb{1}[|y_j| \le 1].) \end{cases}
\end{aligned}
$$

Thus, LHS can be bounded by $\lesssim \max\left\{\frac{1}{m_t}, \frac{1}{\sigma_t^{\alpha_j+1}}\right\} \lesssim 1$.

Next,

$$
\begin{aligned}
&\int_{\mathbb{R}\setminus a_i^x} \left(\frac{|m_t y_j - x_j|}{\sigma_t}\right)^{\alpha_j} \frac{\mathbb{1}[|y_j| \le 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2}\right) dy_j \qquad (104) \\
&\le \frac{2}{m_t}\int_{C_{\mathrm{f},1}\sqrt{\log\varepsilon^{-1}}}^{\infty} |y_j|^{\alpha_j} \exp\left(-\frac{y_j^2}{2}\right) \mathrm{d}y_i \quad \left(\text{by letting } \frac{m_t y_j - x_j}{\sigma_t} \mapsto y_j\right) \\
&\le \begin{cases} \frac{2}{m_t}\sum_{l=0}^{\frac{\alpha_j-1}{2}} \frac{(\alpha_j-1)!!}{(2l)!!}(C_{\mathrm{f},1}^2\log\varepsilon^{-1})^l \varepsilon^{\frac{C_{\mathrm{f},1}}{2}} & (\text{ if } \alpha_j \text{ is odd}) \\ \frac{2}{m_t}\sum_{l=1}^{\frac{\alpha_j}{2}} \frac{(\alpha_j-1)!!}{(2l-1)!!}(C_{\mathrm{f},1}^2\log\varepsilon^{-1})^l \varepsilon^{\frac{C_{\mathrm{f},1}}{2}} + \frac{2}{m_t}\int_{C_{\mathrm{f},1}\sqrt{\log\varepsilon^{-1}}}^{\infty} \exp\left(-\frac{y_j^2}{2}\right) \mathrm{d}y_j & (\text{ if } \alpha_j \text{ is even}). \end{cases}
\end{aligned}
$$

Therefore, by setting $C_{\mathrm{f},1}$ sufficiently large, in a way that $C_{\mathrm{f},1}$ depends on $\alpha_j(\le k)$ and $d$, this can be bounded by $\frac{\varepsilon}{m_t}$. Moreover, if $m_t \gtrsim 1$, then the integral interval does not overlap with $-1 \le y_j \le 1$, and in this case (104) is alternatively bounded by $0$.

Therefore, (103) can further be bounded by

$$
(103) \lesssim \sum_{i=1}^{d}\prod_{j=1}^{d} 1^{d-1}\cdot\varepsilon \lesssim \varepsilon,
$$

which gives the assertion. □

Next we give the ways of Taylor expansion of exponential functions with polynomials (Lemma J.11) and with neural networks (Lemma J.12), respectively.

**Lemma J.11** (Approximating an exponential function with polynomials)**.** *Let* $A > 0$ *and* $0 \le m_t \le 1$*. For* $t \ge \max\{4eA^2, \lceil\log_2\varepsilon^{-1}\rceil\}$*, we have that*

$$
\left|\exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right) - \sum_{s=0}^{t-1}\frac{(-1)^s}{s!}\frac{(x-m_t y)^{2s}}{2^s\sigma_t^{2s}}\right| \le \varepsilon
$$

*for all* $y \in [\frac{-\sigma_t A + x}{m_t}, \frac{\sigma_t A + x}{m_t}]$.

*Proof.* By standard Taylor expansion of $e^z$ up to degree $t-1$, we have

$$
\exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right) = \sum_{s=0}^{t-1}\frac{(-1)^s}{s!}\frac{(x-m_t y)^{2s}}{2^s\sigma_t^{2s}} + \frac{(-1)^t}{t!}\frac{(\theta(x-m_t y))^{2t}}{2^t\sigma_t^{2t}}
$$

with some $\theta \in (0,1)$. We bound the second term of the residual. When $y \in [\frac{-\sigma_t A + x}{m_t}, \frac{\sigma_t A + x}{m_t}]$ and $t$ is the minimum integer satisfying $t \ge \max\{4eA^2, \lceil\log_2\varepsilon^{-1}\rceil\}$, we have

$$
\frac{1}{t!}\frac{(\theta(x-m_t y)+(1-\theta)x)^{2t}}{2^t\sigma_t^{2t}} \le \frac{(2\sigma_t A)^{2t}}{t!2^t\sigma_t^{2t}} \le \frac{(2\sigma_t A)^{2t}}{(t/e)^t\cdot 2^t\sigma_t^{2t}} \le \frac{2^t A^{2t}}{(4A^2)^t} \le \frac{1}{2^t} \le \varepsilon,
$$

where we used the fact $t! \ge (t/e)^t$. □

**Lemma J.12** (Approximating an exponential function with a neural network)**.** *Take* $\varepsilon > 0$ *arbitrarily. There exists a neural network* $\phi_{\exp} \in \Phi(L, W, S, B)$ *such that*

$$\sup_{x,x' \geq 0} \left| e^{-x'} - \phi_{\exp}(x) \right| \leq \varepsilon + |x - x'|$$

*holds, where* $L = \mathcal{O}(\log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(\log \varepsilon^{-1}), S = \mathcal{O}(\log^2 \varepsilon^{-1}), B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$*. Moreover,* $|\phi_{\exp}(x)| \leq \varepsilon$ *for all* $x \geq \log 3\varepsilon^{-1}$*.*

*Proof.* Let us take $A = \log 3\varepsilon^{-1}$. From Taylor expansion, for all $x$ in $0 \leq x \leq A$, we have

$$\left| e^{-x} - \sum_{i=0}^{k-1} \frac{(-1)^i}{i!} x^i \right| \leq \frac{A^k}{k!}.$$

Moreover, we can evaluate RHS as $\frac{A^k}{k!} \leq \left(\frac{eA}{k}\right)^k$, so by taking $k = \max\{2eA, \lceil \log_2 3\varepsilon^{-1} \rceil\}$, we can bound the RHS by $\frac{\varepsilon}{3}$. Now we approximate each $x^i$ using Lemma J.6 with $d = \mathcal{O}(A + \log \varepsilon^{-1}), C = \mathcal{O}(A), \varepsilon = \frac{\varepsilon}{3k}$ and aggregate them using Lemma J.3. This gives the neural network with $L = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(A + \log \varepsilon^{-1}), S = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}), B = \exp(\log A \cdot \mathcal{O}(A + \log \varepsilon^{-1}))$. Finally, we add two layers $\phi_{\mathrm{clip}}(x; 0, A)$ before this neural network to limit the input within $x > 0$. Then, we obtain a neural network $\phi_{\exp}$ that approximates $e^{-x}$ with an additive error up to $\frac{2\varepsilon}{3}$ in $[0, A]$. Moreover, for $x > A$, we have $|\phi_{\exp}(x) - e^{-x}| \leq |e^{-x} - e^{-A}| + |\phi_{\exp}(A) - e^{-A}| \leq \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon$.

The sensitivity analysis follows from $|\phi_{\exp}(x') - e^{-x}| \leq |\phi_{\exp}(\max\{x', 0\}) - e^{-x}| \leq |\phi_{\exp}(\max\{x', 0\}) - e^{-\max\{x', 0\}}| + |e^{-\max\{x', 0\}} - e^{-x}| \leq \varepsilon + |\max\{x', 0\} - x| \leq \varepsilon + |x' - x|$. □

### J.4 Existing results for approximation

Our diffused B-spline basis decomposition (Section 3 and Appendix D) is built on the B-spline basis decomposition of the Besov space (DeVore & Popov, 1988; Suzuki, 2018). The following fact can be found in Lemma 2 of Suzuki (2018) (although the original version adopts $\Omega = [0, 1]^d$, we can easily adjust the difference by dividing the domain into cubes with each side length 1). The magnitude of $|\alpha_{k,j}|$ is evaluated in p.17 of Suzuki (2018).

**Lemma J.13** (Approximability of the Besov space (Suzuki (2018)))**.** *Let* $C > 0$*. Under* $s > d(1/p - 1/r)_+$ *and* $0 < s < \min\{l, l - 1 + 1/p\}$ *where* $l \in \mathbb{N}$ *is the order of the cardinal B-spline bases, for any* $f \in B^s_{p,q}([-C, C]^d)$*, there exists* $f_N$ *that satisfies*

$$\|f - f_N\|_{L^r([-C,C]^d)} \lesssim C^s N^{-s/d} \|f\|_{B^s_{p,q}([-C,C]^d)}$$

*for* $N \gg 1$*, and has the following form:*

$$f_N(x) = \sum_{k=0}^{K} \sum_{j \in J(k)} \alpha_{k,j} M^d_{k,j}(x) + \sum_{k=K+1}^{K^*} \sum_{i=1}^{n_k} \alpha_{k,j_i} M^d_{k,j_i}(x) \quad \text{with} \sum_{k=0}^{K} |J(k)| + \sum_{k=K+1}^{K^*} n_k = N,$$

*where* $J(k) = \{-C2^k - l, -C2^k - l + 1, \cdots C2^k - 1, C2^k\}$*,* $(j_i)_{i=1}^{n_k} \subseteq J(k)$*,* $K = \mathcal{O}(d^{-1} \log(N/C^d))$*,* $K^* = (\mathcal{O}(1) + \log(N/C^d))\nu^{-1} + K, n_k = \mathcal{O}((N/C^d) 2^{-\nu(k-K)})$ $(k = K + 1, \cdots, K^*)$ *for* $\delta = d(1/p - 1/r)_+$ *and* $\nu = (s - \delta)/(2\delta)$*. Moreover,* $|\alpha_{k,j}| \lesssim N^{(\nu^{-1} + d^{-1})(d/p - s)_+}$*.*

### J.5 Elementary bounds for the Gaussian and hitting time

**Lemma J.14.** *Let* $0 < \varepsilon \ll 1$*,* $l \in \mathbb{Z}^d_+$*, and* $p(x)$ *be the density funciton of* $\mathcal{N}(0, \sigma_t^2 I_d)$*, i.e.,* $p(x) = \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x\|^2}{\sigma_t^2}\right)$*. Then, the following bound holds:*

$$\int_{\|x\|_\infty \geq \sigma_t \sqrt{4 \log dl\varepsilon^{-1}}} \frac{\prod_{i=1}^d x_i^{l_i}}{\sigma^{\sum_{i=1}^d l_i}} p(x) \mathrm{d}x \lesssim \varepsilon.$$

*We sometimes write* $\sqrt{4 \log dl\varepsilon^{-1}} = C_{\mathrm{f},2} \sqrt{\log \varepsilon^{-1}}$*.*

*Proof.* Let us denote $x^l = \prod_{i=1}^d x_i^{l_i}$ and $|l| = \sum_{i=1}^d l_i$ for simple presentation. Let $r = \|x\|_\infty$, and we get

$$
\begin{aligned}
&\int_{\|x\|_\infty \geq \sigma_t \sqrt{4\log \varepsilon^{-1}}} \frac{x^l}{\sigma_t^{|l|}} p(x)\mathrm{d}x \\
&\int_{\|x\|_1 \geq \sigma_t \sqrt{4\log \varepsilon^{-1}}} \frac{x^l}{\sigma_t^{|l|}} p(x)\mathrm{d}x \\
&\leq \int_{r=\sigma_t\sqrt{4\log\varepsilon^{-1}}}^\infty \frac{r^{|l|}}{\sigma_t^{|l|}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{r^2}{2\sigma^2}\right)(d-1)r^{d-1}\mathrm{d}r \\
&= \int_{s=\sqrt{4\log\varepsilon^{-1}}}^\infty s^{|l|+d-1} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s \quad \text{(by letting } s = r/\sigma_t\text{)} \\
&= \frac{(4\log\varepsilon^{-1})^{(|l|+d-1)/2}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{4\log\varepsilon^{-1}}{2}\right)(d-1) \\
&\quad + \int_{s=\sqrt{4\log\varepsilon^{-1}}}^\infty \frac{(|l|+d-1)s^{|l|+d-2}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s \\
&= \cdots = \sum_{0\leq i\leq \lfloor \frac{|l|+d-1}{2}\rfloor} \frac{\frac{(|l|+d-1)!!}{(|l|+d-1-2i)!!}(4\log\varepsilon^{-1})^{(|l|+d-1-2i)/2}(d-1)}{(2\pi)^{\frac{d}{2}}}\varepsilon^2 \\
&\qquad\qquad + \begin{cases} \int_{s=\sqrt{4\log\varepsilon^{-1}}}^\infty \frac{(|l|+d-1)!!}{(2\pi)^{\frac{d}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s & (|l|+d\text{: even}) \\ 0 & (|l|+d\text{: odd}) \end{cases} \\
&\qquad\qquad \text{(by iterating integration by parts)} \\
&\lesssim \varepsilon^2 \log^{\frac{d+|l|-1}{2}} \varepsilon^{-1}.
\end{aligned} \tag{105}
$$

Replacing $\varepsilon$ by $\varepsilon/dl$, RHS of (105) is bounded by

$$
\frac{\varepsilon^2}{d^2 l^2} \log^{\frac{dn+|l|-1}{2}} (\varepsilon/dl)^{-1} \lesssim \varepsilon,
$$

which yields the conclusion. □

**Lemma J.15.** *Let $(B_s)_{[0,t]}$ be the 1-dimensional Brownian motion and $X_t = \int_0^t \beta_s \mathrm{d}B_s$, with $\beta_s \leq \overline{\beta}$. Then, we have that*

$$
\mathbb{P}\left[\sup_{s\in[0,t]} |X_t| \geq 2\sqrt{\overline{\beta}t\log(2\varepsilon^{-1})}\right] \leq \varepsilon.
$$

*Proof.* We bound the case $\beta_s \equiv \overline{\beta}$ because it maximize the hitting probability. According to Karatzas et al. (1991), for $x > 0$,

$$
\mathbb{P}\left[\sup_{s\in[0,t]} |X_t| \geq x\right] = \frac{4}{\sqrt{2\pi}} \int_{\frac{x}{\sqrt{2\overline{\beta}t}}}^\infty e^{-y^2/2}\mathrm{d}y = \frac{4}{\sqrt{2\pi}} \int_{\frac{x}{\sqrt{4\overline{\beta}t}}}^\infty e^{-z^2}\sqrt{2}\mathrm{d}z \leq 2e^{-x^2/4\overline{\beta}t}.
$$

For the second equality, we simply replaced $y/\sqrt{2}$ with $z$. For the last inequality, we used $\frac{4}{\sqrt{2\pi}} \cdot \sqrt{2} \leq 2$ and $\int_x^\infty e^{-y^2}\mathrm{d}y \leq e^{-x^2}$. Therefore, setting $x = 2\sqrt{\overline{\beta}t\log(2\varepsilon^{-1})}$ yields the assertion. □