DIFFUSION MODELS ARE MINIMAX OPTIMAL DISTRIBUTION ESTIMATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We provide the first rigorous analysis on estimation error bounds of diffusion modeling 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. We expect these results advance theoretical understandings of diffusion modeling and its ability to generate verisimilar outputs.

1 INTRODUCTION

Diffusion modeling, or 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 number of sample from that, and therefore the score of the diffusion process from the empirical distribution is used instead (e.g., Sohl-Dickstein et al. (2015)). Moreover, for computational efficiency, the empirical score is further replaced by a neural network (score network) trained by the score matching (Hyvärinen & Dayan, 2005; Vincent, 2011). These modifications cause 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). Concurrently, with the L^∞-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 is 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 $\delta > 0$ under the 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 $\Phi(L, W, S, B) := \{(A^{(L)}\mathrm{ReLU}(\cdot) + b^{(L)}) \circ \cdots \circ (A^{(1)}x + b^{(1)}) |\ A^{(i)} \in \mathbb{R}^{W_i \times W_{i+1}}, b^{(i)} \in \mathbb{R}^{W_{i+1}}, \sum_{i=1}^{l}(\|A^{(i)}\|_0 + \|b^{(i)}\|_0) \le S, \max_i \|A^{(i)}\|_\infty \vee \|b^{(i)}\|_\infty \le B\}$.

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 ∞ for any neural network. We can use finite sample of t and x_t, instead of taking expectation, which is explained in Appendix D.1.

2.1 ASSUMPTIONS

We evaluate $TV(X_0, \hat{Y}_{\overline{T}-\underline{T}})$ 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 $U(B^s_{p,q}([-1, 1]^d); C)$ 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 β_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 assume $n \gg 1$ and take $a_0 = n^{-\frac{1-\delta}{d+2s}}$ for an arbitrary fixed $\delta > 0$ in Assumption 2.4. We 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.

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 $t \in [\underline{T}, \overline{T}]$,*

$$\int_x p_t(x) \|\phi_{\mathrm{score}}(x, t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d}} \log^{\frac{d+16s}{d}}(N)}{\sigma_t^2}.$$

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

The formal proof can be found in Appendix C.

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 be valid for different t because otherwise we would need to switch polynomially many neural networks only valid at each time t, which would increase the required network size. 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 the score approximation.

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 $\nabla p_t(x) = \frac{\nabla p_t(x)}{p_t(x)}$.

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 Φ given in Theorem 3.1 into Φ', which consists of a network ϕ satisfying $\|\phi(\cdot,t)\|_\infty = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} n)$. Let $\mathcal{N}_\varepsilon = \mathcal{N}(\Phi, \|\cdot\|_{L^\infty([-C_1,C_1]^d)}, \varepsilon)$ be the ε-covering number with the L^∞ norm of the function space realized by elements of Φ, where $C_1 = \tilde{O}(1)$ is given in Lemma D.3.

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

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{x_t\sim p_t}[\|\hat{s}(x_t,t) - \nabla \log p_t(x_t)\|^2]\mathrm{d}t\right] \tag{2}$$
$$\lesssim \inf_{s\in\Phi'}\int_{\underline{T}}^{\overline{T}} \mathbb{E}_{x_t\sim p_t}[\|s(x_t,t) - \nabla \log p_t(x_t)\|_2^2]\mathrm{d}t + \frac{\log^2(n)\log(\mathcal{N}_{\varepsilon N^{-1}})}{n} + \varepsilon.$$

The proof is inspired by Schmidt-Hieber (2020); Hayakawa & Suzuki (2020). Evaluating $\mathcal{N}$, Theorem 3.1, and adjusting N yields (2) $\lesssim n^{-\frac{2s}{d+2s}}\log^{\frac{10d+16s}{d}}(n)$. For more details, see Appendix D.

3.3 Estimation error analysis

This section finally provides the estimation error bounds. As a small modification, if $\|\hat{Y}_{\overline{T}-\underline{T}}\|_\infty \geq 2$, then we reset it to $\hat{Y}_{\overline{T}-\underline{T}} = 0$. Formal proofs for this section are found in Appendix E.

Theorem 3.3. *The estimation error in the total variation distance is bounded as*

$$\mathbb{E}[\mathrm{TV}(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-s/(2s+d)}\log^{\frac{5d+8s}{d}} n. \tag{3}$$

Moreover, we can train the score network based on n sample such that the estimation error in the Wasserstein distance of order one is bounded as

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

Theorem E.1 shows that $\inf_{\hat{\mu}}\sup_{p\in B^s_{p,q}} \mathbb{E}[\mathrm{TV}(\hat{\mu},p)] \gtrsim n^{-s/(2s+d)}$, for $0 < p,q \leq \infty$, $s > 0$, and $s > \{d(\frac{1}{p} - \frac{1}{2}), d(\frac{1}{p} - 1), 0\}$, where $\hat{\mu}$ runs over all estimators based on n observations. Therefore, (3) is minimax optimal up to logatithmic factors. Moreover, 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. For (4), 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 E.2.

Moreover, 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. The formal statement and its proof are found in Appendix F.

In summary, we showed that diffusion modeling can achieve nearly minimax estimation rates both in TV and W_1, providing theoretical supports for the real-world success of diffusion modeling.

REFERENCES

H Amann, JP Bourguignon, K Grove, PL Lions, H Araki, F Brezzi, KC Chang, N Hitchin, H Hofer, H Knörrer, et al. Monographs in mathematics vol. 99. 1983.

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.

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. 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.

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. 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

1 Introduction **1**
1.1 Our contributions . . . 2

2 Problem settings **2**
2.1 Assumptions . . . 3

3 Main results **3**
3.1 Approximation of the true score . . . 3
3.2 Generalization of the score network . . . 4
3.3 Estimation error analysis . . . 4

A Additional related works **9**
A.1 Distribution and function estimation . . . 9
A.2 Analysis under the manifold hypothesis . . . 9

B The Besov space **9**

C Approximation of the score function **10**
C.1 Detailed proof sketch . . . 10
C.2 Approximation of m_t and σ_t . . . 12
C.3 Several upper and lower bounds regarding $p_t(x)$ and its derivatives . . . 13
C.4 Diffused B-spline basis . . . 18
C.5 Approximation error bound . . . 23
C.6 Discussion on the discretization error . . . 30

D Generalization bound **31**
D.1 Sampling t and x_t instead of taking expectation . . . 35

E Estimation error analysis **36**
E.1 Estimation rate in the TV distance . . . 36
E.2 Estimation rate in the W_1 distance . . . 39

F Error analysis with intrinsic dimensionality **43**
F.1 Brief proof overview . . . 43
F.2 Proof of Theorem F.1 . . . 44

G Auxiliary lemmas **47**
G.1 Construction of a larger neural network . . . 47
G.2 Basic neural network structure that approximates rational functions . . . 49
G.3 How to deal with exponential functions . . . 52

G.4 Existing results for approximation 54

G.5 Elementary bounds for the Gaussian and hitting time 54

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 F, 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 Ω 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

$$w_{r,p}(f,t) = \sup_{\|h\|_2 \le t} \|\Delta_h^r(f)\|_p, \quad \text{where } \Delta_h^r(f)(x)$$
$$= \begin{cases} \sum_{j=0}^r \binom{r}{j}(-1)^{r-j} f(x+jh) & (\text{if } x+jh \in \Omega \text{ for all } j) \\ 0 & (\text{otherwise}). \end{cases}$$

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 $\partial^\alpha = \frac{\partial^{|\alpha|} f}{\partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d}}(x)$. 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 Amann et al. (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 Approximation of the score function

This section corresponds to Section 3.1, and discusses approximation error in the score matching loss minimization. We construct a neural network that approximates $\nabla \log p_t(x)$ and bound the approximation error at each time t.

C.1 Detailed proof sketch

Here we provide detailed proof sketch of Theorem 3.1.

Approximation via the diffused B-spline Basis 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 C.1 (Informal version of Lemma G.15;Suzuki (2018))**.** *For any $p_0 \in U(B^s_{p,q})$, there exists a super-position of N tensor-product B-spline bases that satisfies*

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

With m_t and σ_t, we can explicitly 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, we come up with approximation of p_t as

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

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

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

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. First, we construct networks approximating m_t and σ_t.

Lemma C.2 (See also Lemma C.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 σ_t up to ε 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, σ_t). Then the following is obtained as an informal version of Lemma C.11.

Lemma C.3. *For $\varepsilon > 0$, there exists a neural network $\phi_{\mathrm{TDB}} \in \Phi(L, W, S, B)$, with $L = \mathcal{O}(\log^4(\varepsilon^{-1})), \|W\|_\infty = \mathcal{O}(\log^6(\varepsilon^{-1})), S = \mathcal{O}(\log^8(\varepsilon^{-1})), B = \mathcal{O}(\exp(\mathcal{O}(\log^2(\varepsilon^{-1}))))$ such that $\|\phi_{\mathrm{TDB}}(x, m_t, \sigma_t) - E_{k,j}(x,t)\|_\infty \leq \varepsilon$.*

Here ϕ_{TDB} approximates $E_{k,j}(x,t)$ given (x, m_t, σ_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 of $\tilde{\mathcal{O}}(N^{-2s/d}/\sigma_t^2)$ as in Theorem 3.1. In this construction, the error is amplified in the area where $p_t(x) \ll 1$, we need the smoothness in the area with distance less than $\tilde{\mathcal{O}}(\sqrt{t})$ from the edge of the support, as Assumption 2.4. Because this approach is used during $t \in [\underline{T}, 2N^{-\frac{2-\delta}{d}}]$, a_0 is set to $a_0 = N^{-\frac{1-\delta}{d}}$. Now the proof sketch for Theorem 3.1 has completed for $t \leq 2N^{-\frac{2-\delta}{d}}$.

We remark that the above network sizes are slightly larger than that for the B-spline basis (Suzuki (2018)) because approximating integrals and exponential functions (appendix G.3) and rational functions (appendix G.2) is more difficult than realizing the B-spline basis via polynomials.

Utilizing the smoothness induced by the noise The previous method 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. Thus, there should be the "swithing point" T_*, after which the smoothness induced by the Gaussian noise gives a tighter bound than approximation based on the initial structure of the data distribution.

The first step is to bound the derivatives of p_t w.r.t. t.

Lemma C.4. *For $k \in \mathbb{Z}_+$, there exists a constant $C_{\mathrm{b},2}$ depending only on k, d, and C_f such that*

$$|\partial_{x_{i_1}} \partial_{x_{i_2}} \cdots \partial_{x_{i_k}} p_t(x)| \leq \frac{C_{\mathrm{b},2}}{\sigma_t^k}.$$

This gives $\|p_{t_*}\|_{W_p^k} = \mathcal{O}(1/t_*^{\frac{k}{2}})$ for $t_* > 0$. For $t > t_*$, p_t can also be seen as the diffusion from p_{t_*}. Together with $W_p^k \hookrightarrow B_{p,\infty}^k$, this approach gives approximation of p_t with an error $N'^{-k/d}/t_*^{\frac{k}{2}}$, where N' is the size of the network. The error gets smaller than $N^{-\frac{2(s+1)}{d}}\sigma_t^{-1}$ when k is sufficiently large. This argument is formalized as follows.

Lemma C.5. *Let $N \gg 1$ and $N' \geq t_*^{-d/2}N^{\delta/2}$. Suppose $1 \geq t_* > T_* = N^{-(2-\delta)/d}$ and $t > t_*$. Then there exists a neural network $\phi'_{\text{score}} \in \Phi(L, W, S, B)$ that satisfies*

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

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 $t_* = N^{-\frac{2-\delta}{d}} =: T_*$ and $N' = N$ in this lemma, we obtain the bound in Theorem 3.1 after $t \geq 2T_*$, without Assumption 2.4. Moreover, this later plays an important role in bounding the W_1 distance.

C.2 Approximation of m_t and σ_t

Now we proceed to the main part of the proof. In addition to the true data distribution $p_0(x)$, the score $\nabla \log p_t(x)$ also depends on m_t and σ_t. Indeed, in our construction, each tensor-product diffused B-spline basis is approximated as a rational function of x, m_t and σ_t. Here, m_t and σ_t are as important as x, because we use exponentiation of m_t and σ_t, as well as that of x, while exact values of m_t and σ_t are unavailable. In other words, because approximation errors of m_t and σ_t are amplified via such exponentiation, approximating m_t and σ_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 σ_t. The following is the formal version of Lemma C.2.

Lemma C.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 ε, 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 σ_t for all $t \geq \varepsilon$, within the additive error of ε, 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 $\beta \geq \underline{\beta}$, $\int_0^t \beta_s \mathrm{d}s \geq \log 3\varepsilon^{-1}$ for all $t \geq A := \log 3\varepsilon^{-1}/\underline{\beta}$. We limit ourselves within $[0, A]$. Then, from Assumption 2.3, we can expand β_s as $\beta_s = \sum_{i=0}^{k-1} \frac{\beta^{(i)}}{i!}s^i + \frac{\beta^{(k)}}{k!}(\theta s)^i$ 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 $k = \max\{2eA, \lceil \log_2 3\varepsilon^{-1}\rceil\} - 1$ so that we have $\frac{A^{k+1}}{(k+1)!} \leq \left(\frac{eA}{k+1}\right)^{k+1} \leq \frac{\varepsilon}{3}$. $\int_0^t \sum_{i=1}^{k-1} \frac{\beta^{(i)}}{i!}s^i = \sum_{i=1}^{k-1} \frac{\beta^{(i)}}{(i+1)!}t^{i+1}$ can be realized with an additive error up to $\frac{\varepsilon}{3}$ 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 Lemma G.8 and remark G.5. From the definition of A, we can easily check that $e^{-A} \leq \frac{\varepsilon}{3}$ holds. We clip the input with $[0, A]$ to obtain the neural network ϕ_1, which approximates $\int_0^t \beta_s \mathrm{d}s$ with an additive error of $\frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \frac{2\varepsilon}{3}$ for $x \in [0, A]$, and satisfies $|\phi_1(x)| = |\phi_1(A)|$ for all $x \geq A$.

Then we apply Lemma G.14 with $\varepsilon = \frac{\varepsilon}{3}$. Then we obtain the neural network ϕ_m of the desired size, which approximates $m_t = \exp(-\int_0^t \beta_s \mathrm{d}s)$ with an additive error of $\frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$ for $x \in [0, A]$ and $|\phi_m(x) - e^{-x}| \leq |\phi_m(x) - \phi_m(A)| + |\phi_m(A) - e^{-A}| + |e^{-A} - e^{-x}| \leq 0 + \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$ 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 G.11 with $\varepsilon = c\varepsilon$ and finally obtain a neural network $\phi_\sigma(t)$ that approximates σ_t with an additive error of $c\varepsilon + \frac{\varepsilon^{1.5}}{\sqrt{c\varepsilon}} = \mathcal{O}(\varepsilon)$, 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 $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$. Adjusting hidden constants can make the approximation error smaller than ε, and concludes the proof. □

C.3 Several upper and lower bounds regarding $p_t(x)$ and its derivatives

When dealing with $p_t(x)$, its unboundedness sometimes causes troubles. This subsection aims to provide several treatments for such issues.

We first give upper and lower bounds on $p_t(x)$. They are rather loose, but they suffice to derive bounds on the score function $s^\circ(x,t)$ and its derivatives allowing logarithmic terms later. We do not pursue tightening these bounds here.

Lemma C.7 (Upper and lower bounds on the density $p_t(x)$)**.** *The following upper and lower bounds on $p_t(x)$ holds for a constant $C_{\mathrm{b},1}$ depending on C_f and d.*

$$C_{\mathrm{b},1}^{-1} \exp\left(-\frac{d(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\right) \leq p_t(x) \leq \frac{C_{\mathrm{b},1}}{m_t^d} \exp\left(-\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2}\right).$$

Proof. We first consider the case when $x \in [-m_t, m_t]^d$. The upper bound is trivial, since $f(y) \leq C_f$ means

$$p_t(x) = \int \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y \leq C_f \int \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y = \frac{C_f}{m_t^d},$$

and therefore letting $C_f \leq C_{\mathrm{b},1}$ yields the bound.

The lower bound can be understood as follows. We have

$$\begin{aligned} p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\ &\geq \frac{1}{(2\pi)^{\frac{d}{2}}} \int f(x/m_t - \sigma_t y) \exp\left(-\frac{\|y\|^2}{2}\right) \mathrm{d}y \quad \text{(by letting } (x - m_t y)/\sigma_t \mapsto m_t y). \end{aligned} \tag{5}$$

Since $x \in [-m_t, m_t]^d$, we have $x/m_t \in [-1,1]^d$. Thus, $|\{y \in [-1,1]^d | x/m_t - \sigma y \in [-1,1]\}| \geq 1$. Moreover, $\exp\left(-\frac{\|y\|^2}{2}\right) \geq \exp(-d^2/2)$ in $y \in [-1,1]^d$. Therefore, the integral (5) 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

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

where we used $\int_z^\infty e^{-a^2}\mathrm{d}a \le e^{-z^2}$ (Chang et al., 2011) for the last inequality.

On the other hand,

$$
\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&\ge C_f^{-1} \prod_{i=1}^d \underbrace{\int_{y_i \in [-1,1]} \frac{1}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y}_{\text{(a)}} \\
&= C_f^{-1} \left(\int_{y_{i^*} \in [-1,1]} \frac{1}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_{i^*} - m_t y_{i^*})^2}{2\sigma_t^2}\right) \mathrm{d}y \right)^d \quad \text{(because (a) is minimized when } i = i_*) \\
&\ge \frac{C_f^{-1}}{m_t^d} \left(\int_{a=r/\sqrt{2}}^{r/\sqrt{2}+\sqrt{2}m_t/\sigma_t} \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}y \right)^d \quad \text{(by } (x_{i^*} - m_t y_{i^*})\sigma_t^{-1}/\sqrt{2}) \\
&\ge \frac{C_f^{-1}}{m_t^d} \left(\int_{a=r/\sqrt{2}}^{r/\sqrt{2}+\sqrt{2}m_t} \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}y \right)^d \\
&\ge \frac{C_f^{-1}}{m_t^d} \left(\frac{\sqrt{2}m_t}{\sqrt{\pi}} \exp\left(-(r/\sqrt{2} + \sqrt{2}m_t)^2\right) \right)^d \\
&\text{(by lower bounding } \exp(-a^2) \text{ in the integral interval and just multiplying the width of the interval)} \\
&\ge \frac{C_f^{-1}}{m_t^d} \left(\frac{\sqrt{2}m_t}{\sqrt{\pi}} \exp\left(-r^2 - 4\right) \mathrm{d}a \right)^d \\
&\ge \frac{C_f^{-1} 2^{d/2}}{e^{4d}\pi^{d/2}} \exp\left(-dr^2\right),
\end{aligned}
$$

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

Lemma C.7 allows us to evaluate the density $p_t(x)$, the score function $\nabla \log p_t(x)$, and their derivatives. $\partial^k p_t(x) = \mathcal{O}(1/\sigma_t^k) = \mathcal{O}(t^{-k/2})$ can be shown by rather straightforward argument. On the other hand, $\sup_{x\in\mathbb{R}^d} \|\nabla \log p_t(x)\| = \infty$ holds in general (see the proof of the next lemma below). Nevertheless, we can show that $\|\nabla \log p_t(x)\| = \tilde{\mathcal{O}}(1/\sigma_t) = \tilde{\mathcal{O}}(1 \vee 1/\sqrt{t})$ when $p_t(x)$ is not too small. Therefore, we exclude the region with small $p_t(x)$ from evaluation of differentials in the following part. Excluding such points does not cause a major problem after all, since in such a case $p_t(x)$ is so small that the contribution of the approximation error at x is negligible when considering the total approximation error in $L^2(p_t)$.

Lemma C.8 (Boundedness of derivatives)**.** *For $k \in \mathbb{Z}_+$, there exists a constant $C_{\mathrm{b},2}$ depending only on k, d, and C_f such that*

$$
|\partial_{x_{i_1}} \partial_{x_{i_2}} \cdots \partial_{x_{i_k}} p_t(x)| \le \frac{C_{\mathrm{b},2}}{\sigma_t^k}. \tag{6}
$$

Moreover, we have that

$$
\|\nabla \log p_t(x)\| \le \frac{C_{\mathrm{b},2}}{\sigma_t} \sqrt{\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2} + \log m_t^{-1} + 1}, \tag{7}
$$

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

$$
\|\partial_{x_i} \nabla \log p_t(x)\| \le \frac{C_{\mathrm{b},2}}{\sigma_t^2} \left(\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2} + \log m_t^{-1} + 1 \right). \tag{8}
$$

and that

$$
\|\partial_t \nabla \log p_t(x)\| \le \frac{C_{\mathrm{b},2}}{\sigma_t^3} \left[|\partial_t \sigma_t| + |\partial_t m_t|\right] \left(\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2} + \log m_t^{-1} + 1 \right)^{\frac{3}{2}}. \tag{9}
$$

Proof. First, we consider (6). Let $g_1(x) = p_t(x) = \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y$. 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 $B_s^1 = \{s' \in \mathbb{Z}_+^d | s'_i \le s_i \ (i = 1, \cdots, d)\}$ and a constant c_s^1 such that $\partial_{x_1}^{s_1} \partial_{x_2}^{s_2} \cdots \partial_{x_d}^{s_d} e^{-\|x\|^2} = \sum_{s' \in B_s^1} c_{s'}^1 x_1^{s'_1} x_2^{s'_2} \cdots x_d^{s'_d} e^{-\|x\|^2}$ holds. Then, because of $\partial_{x_i} = \frac{1}{\sigma} \partial_{\frac{x_i}{\sigma}}$, we can write down

$$g_1^{(s)}(x) = \frac{\sum_{s' \in B_s^1} c_{s'}^1}{\sigma^{\sum_{i=1}^d s_i}} \underbrace{\int \prod_{i=1}^d \left(\frac{x_i - m y_i}{\sigma_t}\right)^{s'_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}_{\text{(a)}}. \tag{10}$$

Note that $\max_{s:\ \sum s_i \le k} \{\sum_{s' \in B_s^1} c_{s'}^1\}$ is bounded by a constant that only depends on k. Moreover, when $t \le 1$, (a) in (10) can be bounded by $\mathcal{O}(1/m_t^d) \simeq \mathcal{O}(1)$ where we hide dependency on $\sum_{i=1}^d s'_i \le k$ and C_f, $|g_1^{(s)}(x)|$ is bounded by $\lesssim \sigma_t^{-k}$, which gives the first bound (6). On the other hand, when $t \ge 1$, $\sigma_t \gtrsim 1$ holds, and in this case we can also bound (a) by $\mathcal{O}(1)$. Now, the first statement is 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

$$[\nabla \log p_t(x)]_1 = \frac{1}{\sigma_t} \cdot \frac{g_2(x)}{g_1(x)}, \quad [\partial_{x_i} \nabla \log p_t(x)]_1 = \frac{1}{\sigma_t} \cdot \frac{\partial_{x_i} g_2(x)}{g_1(x)} - \frac{1}{\sigma_t} \cdot \frac{g_2(x)(\partial_{x_i} g_1(x))}{g_1^2(x)}. \tag{11}$$

Moreover,

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

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

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

In order to bound them, we consider the following quantity. We assume $p_t(x) \ge \frac{\varepsilon}{m_t^d}$ and let $s \in \mathbb{Z}_+$ with $k = \sum_{i=1}^d s_i$.

$$\frac{\int \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int_{A^x} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{\sigma_t^2}\right) \mathrm{d}y} \tag{12}$$

According to Lemma G.12, we have that

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

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

Therefore, when $p_t(x) = g_1(x) \geq \frac{\varepsilon}{m_t^d}$,

$$
\begin{aligned}
(12) &\leq \frac{2\int_{A^x} \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int_{A^x} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{\sigma_t^2}\right) \mathrm{d}y} + \frac{m_t^d \frac{\varepsilon}{2}}{m_t^d \varepsilon} \\
&\leq 2 \max_{y \in A_x} \left[\prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i'}\right] + \frac{1}{2} \\
&\leq 2\left(C_{\mathrm{clip},1}^2 \log \varepsilon^{-1}\right)^{(\sum_{i=1}^d s_i)/2} + \frac{1}{2} \leq 3\left(C_{\mathrm{clip},1}^2 \log \varepsilon^{-1}\right)^{k/2} \quad \text{(when } C_{\mathrm{clip},1} \text{ sufficiently large).}
\end{aligned} \tag{13}
$$

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

$$
3\left(C_{\mathrm{clip},1}^2 \log \varepsilon^{-1}\right)^{1/2}, 3\left(C_{\mathrm{clip},1}^2 \log \varepsilon^{-1}\right)^{1/2} \sigma_t^{-1}, \text{ and } 3\left(C_{\mathrm{clip},1}^2 \log \varepsilon^{-1}\right) \sigma_t^{-1},
$$

respectively. Finally, we apply this to (11) and obtain that

$$
\|\nabla \log p_t(x)\| \leq \frac{3C_{\mathrm{clip},1} \log^{1/2} \varepsilon^{-1}}{\sigma_t} \text{and, } \|\partial_{x_i} \nabla \log p_t(x)\| \lesssim \frac{C_{\mathrm{clip},1}^2 \log \varepsilon^{-1}}{\sigma_t^2}.
$$

Now we apply Lemma C.7 to this. Because of $C_{\mathrm{b},1}^{-1} \exp\left(-\frac{d(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\right) \leq p_t(x)$, we have that

$$
\begin{aligned}
\|\nabla \log p_t(x)\| &\leq \frac{C_{\mathrm{b},2}}{\sigma_t}\sqrt{\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2} + \log m_t^{-1} + 1} \quad \text{and,} \\
\|\nabla \log p_t(x)\| &\leq \frac{C_{\mathrm{b},2}}{\sigma_t^2}\left(\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2} + \log m_t^{-1} + 1\right),
\end{aligned}
$$

with $C_{\mathrm{b},2}$ depending on k and C_f. Thus, we obtain (8) and (C.8) of the assertion.

Finally, we have that

$$
\begin{aligned}
&\partial_t \frac{g_2(x)}{g_1(x)} \\
&= \frac{\int \frac{(\partial_t m_t) y_1 + (x_1 - m_t y_1)((d+1)(\partial_t \sigma_t)\sigma_t^{-1} + \|x - m_t y\|^2 (\nabla_t \sigma_t)\sigma_t^{-3} + (\partial_t m_t) y^\top (m_t y - x)\sigma_t^{-2})}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y} \\
&\quad + \frac{\int \frac{-d(\partial_t \sigma_t)\sigma_t^{-1} + \|x - m_t y\|^2 (\partial_t \sigma_t)\sigma_t^{-3} + (\partial_t m_t) y^\top (m_t y - x)\sigma_t^{-2}}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\left(\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y\right)^2} \\
&\qquad \cdot \int \frac{x_1 - m_t y_1}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y.
\end{aligned}
$$

By carefully decomposing this into the sum of (12), and applying (13) and Lemma C.7, we have the final bound (9). □

Now 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.9 (Error bounds due to clipping operations)**.** *Let $0 < \varepsilon < \frac{1}{2}$ and $t > 0$. There exists a constant $C_{\mathrm{b},3}$ depending on d and C_f, we have*

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

Moreover, for x such that $\|x\|_\infty \leq m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}$, we have

$$\|\nabla \log p_t(x)\| \leq \frac{C_{\mathrm{b},3}}{\sigma_t}\sqrt{\log\varepsilon^{-1}\log m_t^{-1}}.$$

Therefore,

$$\begin{aligned}
&\int_{\|x\|_\infty \leq m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}} p_t(x)\mathbb{1}[p_t(x)\leq\varepsilon]\|\nabla \log p_t(x)\|^2 \mathrm{d}x \\
&\leq \varepsilon \cdot \frac{C_{\mathrm{b},3}^2}{\sigma_t^2}(\log\varepsilon^{-1} + \log m_t^{-1} + 1)(2m_t + 2\sigma_t\sqrt{2\log\varepsilon^{-1}})^d, \\
&\int_{\|x\|_\infty \leq m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}} p_t(x)\mathbb{1}[p_t(x)\leq\varepsilon]\mathrm{d}x \leq \varepsilon \cdot C_{\mathrm{b},3}(m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}})^d.
\end{aligned}$$

Proof. According to Lemma C.7 and Lemma C.8,

$$\begin{aligned}
p_t(x)\|\nabla \log p_t(x)\|^2 &\leq \frac{C_{\mathrm{b},1}}{m_t^d}\exp\left(-\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2}\right)\cdot\frac{C_{\mathrm{b},2}^2}{\sigma_t^2}\left(\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2} + \log m_t^{-1} + 1\right) \\
&\leq \frac{C_{\mathrm{b},1}C_{\mathrm{b},2}^2}{m_t^d\sigma_t^2}\exp\left(-\frac{r^2}{2}\right)(r + \sqrt{\log m_t^{-1}} + 1),
\end{aligned}$$

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

$$\begin{aligned}
&\int_{\|x\|_\infty \geq m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}} p_t(x)\|\nabla \log p_t(x)\|^2 \mathrm{d}x \\
&\leq \int_{\sqrt{2\log\varepsilon^{-1}}}^{\infty} \frac{C_{\mathrm{b},1}C_{\mathrm{b},2}^2}{m_t^d\sigma_t}\exp\left(-\frac{r^2}{2}\right)(r + \log m_t^{-1} + 1)(d-1)(\sigma_t r + m_t)^{d-1}\mathrm{d}r \\
&\leq \frac{C_1}{m_t^d\sigma_t}\left(\log^{(d-1)/2}\varepsilon^{-1} + (1+\log m_t^{-1})\log^{d/2-1}\varepsilon^{-1}\right)\varepsilon,
\end{aligned}$$

where we let $C_1 = C_1(d, C_f)$ and used $\int_a^\infty e^{-x^2}\mathrm{d}x \leq e^{-x^2}$ (Chang et al., 2011).

In the same way,

$$\begin{aligned}
\int_{\|x\|_\infty \geq m + \sigma\sqrt{2\log\varepsilon^{-1}}} p_t(x)\mathrm{d}x &\leq \int_{\sqrt{2\log\varepsilon^{-1}}}^{\infty} \frac{C_{\mathrm{b},1}\sigma_t}{m_t^d}\exp\left(-\frac{r^2}{2}\right)(d-1)(\sigma_t r + m)^{d-1}\mathrm{d}r \\
&\leq \frac{C_1\sigma_t}{m_t^d}\varepsilon\log^{(d-2)/2}\varepsilon^{-1}.
\end{aligned}$$

We then consider the second part of the lemma. Due to Lemma C.8, for x with $\|x\|_\infty \leq m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}$

$$\|\nabla \log p_t(x)\| \leq \frac{C_{\mathrm{b},2}}{\sigma_t}\sqrt{\log\varepsilon^{-1} + \log m_t^{-1} + 1}$$

Using this, we have

$$\int_{\|x\|_\infty \le m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}} p_t(x)\mathbb{1}[p_t(x)\le\varepsilon]\|\nabla\log p_t(x)\|^2\mathrm{d}x$$
$$\le \varepsilon\cdot\frac{C_{\mathrm{b},2}^2}{\sigma_t^2}(\log\varepsilon^{-1}+\log m_t^{-1}+1)(2m_t+2\sigma_t\sqrt{2\log\varepsilon^{-1}})^d$$

and

$$\int_{\|x\|_\infty \le m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}} p_t(x)\mathbb{1}[p_t(x)\le\varepsilon]\mathrm{d}x \le \varepsilon\cdot(2m_t+2\sigma_t\sqrt{2\log\varepsilon^{-1}})^d.$$

Therefore, letting $C_{\mathrm{b},3} = C_{\mathrm{b},3}(C_1, d, C_{\mathrm{b},2})$ concludes the proof. □

C.4 Diffused B-spline basis

This subsection introduces the *diffused B-spline basis*, which enables us to approximate $s^\circ(x,t)$ in the space of $\mathbb{R}^d\times[\underline{T},\overline{T}]$ with the same size of the bases as that of the B-spline bases required for approximation of $p_0(x)$. Before beginning the formal proof, we provide extended proof outline about the approximation via the diffusion B-spline basis, which is more detailed than that in Section 3 and the first part of this section.

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

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

(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

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

are decomposed into the sum of

$$E_{k,j}^{(1)}(x) := \int\frac{1}{\sigma^d(2\pi)^{\frac{d}{2}}}\mathbb{1}[\|y\|_\infty\le 2-\varepsilon_{\mathrm{a}}]M_{k,j}^d(y)\exp\left(-\frac{\|x-my\|^2}{2\sigma^2}\right)\mathrm{d}y \tag{14}$$

and

$$E_{k,j}^{(2)}(x) := \int\frac{x-my}{\sigma^{d+1}(2\pi)^{\frac{d}{2}}}\mathbb{1}[\|y\|_\infty\le 2-\varepsilon_{\mathrm{a}}]M_{k,j}^d(y)\exp\left(-\frac{\|x-my\|^2}{2\sigma^2}\right)\mathrm{d}y, \tag{15}$$

respectively. Note that here in the following $E_{k,j}^{(1)}(x)$ is the same as $E_{k,j}(x)$ used in Section 3. Also note that ε_{a} $(0\le\varepsilon_{\mathrm{a}}\le 2)$ be a scaler value determined later. We then approximate each of the denominator and numerator of $\nabla\log p_t(x)$ combining sub-networks that approximates each $E_{k,j}^{(1)}(x)$ or $E_{k,j}^{(2)}(x)$.

Here we briefly remark why $\mathbb{1}[\|y\|_\infty\le 2-\varepsilon_{\mathrm{a}}]$ appears. Let us assume $\varepsilon_{\mathrm{a}}=1$ and approximate $p_t(x)$ based on basis decomposition of $f^\circ(x)=p_0(x)$, although later we need to consider other situations. If we use basis decomposition as $f^\circ(x)\fallingdotseq f_N(x)=\sum M_{k,j}^d(x)$, existing results such as Lemma G.15 only assure that the approximation is valid within $[-1,1]^d$ and do not guarantee

anything about outside the region. This might harm the approximation accuracy when we integrate $f_N(x)$ over all $\mathbb{R}^d$. Therefore, we need to force $f_N(x) = 0$ if $\|x\|_\infty > 1$ by this indicator function.

From now, we realize these basic components with neural networks. We take $E^{(1)}_{k,j}$ as an example, and the procedures for $E^{(2)}_{k,j}$ is essentially the same. Remind that in Section 3 and the first part of this section we decomposed $E^{(1)}_{k,j}$ into the produce of the diffused B-spline basis,

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

But here, more formally, we first truncate the integral intervals. We clip the integral interval as

$$\begin{aligned} E^{(1)}_{k,j}(x) &\coloneqq \int_{y\in A^x} \frac{1}{\sigma^d(2\pi)^{\frac{d}{2}}} \mathbb{1}[\|y\|_\infty \le 2-\varepsilon_{\mathrm{a}}] M^d_{k,j}(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma^2}\right)\mathrm{d}y \\ &= \prod_{i=1}^d \left(\sum_{l'=0}^{l+1} \frac{(-1)^{l'}\,{}_{l+1}\mathrm{C}_{l'}}{l!} \int_{y_i\in a_i^x} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \mathbb{1}[|y_i| \le 2-\varepsilon_{\mathrm{a}}] \mathbb{1}[0 \le 2^{k_i}y_i - j_i \le l+1] \right. \\ &\qquad \left. \times (2^{k_i}y_i - l' - j_i)_+^l \exp\left(-\frac{(x_i - my_i)^2}{2\sigma^2}\right)\mathrm{d}y_i \right), \end{aligned} \tag{16}$$

$A^x = \prod_{i=1}^d a_i^x$ with $a_i^x = [\frac{x_i}{m} - \frac{\sigma C_{\mathrm{clip},1}}{m}\sqrt{\log\varepsilon^{-1}}, \frac{x_i}{m} + \frac{\sigma C_{\mathrm{clip},1}}{m\sigma}\sqrt{\log\varepsilon^{-1}}]$, $C_{\mathrm{clip},1} = \mathcal{O}_d(1)$, and $0 < \varepsilon < 1$. This clipping causes the error at most $\frac{\varepsilon\left((l+1)^{l+1}2^{l+1}\right)^d}{m^d}$ according to Lemma G.12 and the observation $\mathbb{1}[\|y\|_\infty \le 2-\varepsilon_{\mathrm{a}}]M^d_{k,j}(y) \le \left((l+1)^{l+1}2^{l+1}\right)^d$. Owing to the fact that $M^d_{k,j}(x)$ is a product of univariate functions of x_i $(i = 1, 2, \cdots, d)$, the integral over $\mathbb{R}^d$ is now decomposed into that 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 $\int_{y_i\in a_i^x} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\mathbb{1}[|y_i| \le 2-\varepsilon_{\mathrm{a}}]\mathbb{1}[0 \le 2^{k_i}y_i - j_i \le l+1](2^{k_i}y_i - l' - j_i)_+^l \exp\left(-\frac{(x_i-my_i)^2}{2\sigma^2}\right)\mathrm{d}y_i$. Note that $\mathbb{1}[|y_i| \le 2-\varepsilon_{\mathrm{a}}]\mathbb{1}[0 \le 2^{k_i}y_i - j_i \le l+1] \equiv 0$ or $= \mathbb{1}[a \le 2^{k_i}y_i \le b]$ holds with a, b satisfying

$$-C2^k - l \le \min_i j_i \le j_i \le a, b \le j_i + l + 1 \le \max j_i + l + 1 \le C2^k + l + 1,$$

if we assume $\mathrm{supp}(p_0) = [-C, C]^d$ (see Lemma G.15).

We sometimes omit the subscript i for the coordinates, for simple presentation.

Lemma C.10. *Let $j, \underline{j}, \overline{j}, k, l$ be integers satisfying $-C2^k - l \le j \le C2^k + l + 1$, $j \le \underline{j} \le \overline{j} \le j + l + 1$, and $k, l \ge 0$. Assume that $\underline{\sigma} \le \sigma, \sigma' \le 1$, $\underline{m} \le m, m' \le 1$, $|\sigma' - \sigma|, |m' - m| \le \varepsilon_{\mathrm{error}}$, and take ε from $0 < \varepsilon < \frac{1}{2}$ and $C > 0$ arbitrarily. Then, there exists a neural network $\phi^{j,\overline{j},\underline{j},k}_{\mathrm{dif},1} \in \Phi(L, W, S, B)$ with*

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

such that

$$\left| \phi^{j,\overline{j},\underline{j},k}_{\mathrm{dif},1}(x,\sigma',m') - \int_{-\frac{\sigma C_{\mathrm{clip},1}}{m}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m}}^{\frac{\sigma C_{\mathrm{clip},1}}{m}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m}} \frac{1}{\sqrt{2\pi}\sigma}\mathbb{1}[\underline{j} \le 2^k y \le \overline{j}](2^k y - j)^l \exp\left(-\frac{(x-my)^2}{2\sigma^2}\right)\mathrm{d}y \right|$$

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

holds for all x in $-C \le x \le C$.

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

$$\left|\phi_{\text{dif},\underline{2}}^{j,\bar{j},\underline{j},k}(x,\sigma',m') - \int_{-\frac{\sigma C_{\text{clip},1}}{m}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m}}^{\frac{\sigma C_{\text{clip},1}}{m}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m}} \frac{[x-my]_i}{\sqrt{2\pi}\sigma^2}\mathbb{1}[\underline{j}\le 2^k y\le \bar{j}](2^k y-j)^l \exp\left(-\frac{(x-my)^2}{2\sigma^2}\right)\mathrm{d}y\right|$$

$$\le \mathcal{O}(\varepsilon\log^{-1}\varepsilon) + \varepsilon_{\text{error}} 2^{(k+1)l}(2^{(k+1)l(l+2)}\underline{\sigma}^{-2(l+2)} + \underline{m}^{-4l} + C^{4l}2^{4kl})\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}).$$

holds for all x *in* $-C\le x\le C$.

Furthermore, we have $\phi_{\text{dif},\underline{1}}^{j,\bar{j},\underline{j},k}(x,\sigma',m') = \phi_{\text{dif},\underline{2}}^{j,\bar{j},\underline{j},k}(x,\sigma',m') = 0$ *if* $C_{\text{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1} \le \frac{x-m2^{-k}\bar{j}}{\sigma}$ *or* $-C_{\text{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1} \ge \frac{x-m2^{-k}\underline{j}}{\sigma}$ *holds and* $\|\phi_{\text{dif},\underline{1}}^{j,\bar{j},\underline{j},k}\|_\infty, \|\phi_{\text{dif},\underline{2}}^{j,\bar{j},\underline{j},k}\|_\infty = \mathcal{O}(1)$.

Proof. Here we only prove the first part, because the second part follows by simply changing $l'+2s+1$ to $l'+2s+2$ in (17) and following the subsequent argument.

First, we approximate the exponential function within the closed interval, using polynomials of degree at most $\mathcal{O}(\log\varepsilon^{-1})$. Note that $\mathbb{1}[\underline{j}\le 2^k y\le\bar{j}](2^k y-j)^l$ is bounded by $(l+1)^l$, from the assumption of $j\le\underline{j}\le\bar{j}\le j+l+1$. Therefore, according to Lemma G.13, there exists a constant $A = C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}$, we have that

$$\left|\exp\left(-\frac{(x-my)^2}{2\sigma^2}\right) - \sum_{s=0}^{t-1}\frac{(-1)^s}{s!}\frac{(x-my)^{2s}}{2^s\sigma^{2s}}\right| \le (l+1)^l\varepsilon$$

for all $y\in[-m\sigma C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}+x, m\sigma C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}+x]$, where $t = \max\{4eC_{\text{clip},1}^2\log\varepsilon^{-1}, \lceil\log_2\varepsilon^{-1}\rceil\}$. Replacing the exponential function by polynomials within $[-\frac{\sigma C_{\text{clip},1}}{m}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m}, \frac{\sigma C_{\text{clip},1}}{m}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m}]$ causes the error at most $\frac{C_{\text{clip},1}^2(l+1)^l}{m\sqrt{2\pi}}\varepsilon\log\varepsilon^{-1}$.

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

$$\begin{aligned}
&\sum_{s=0}^{t-1}\int_{-m\sigma C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}+x}^{m\sigma C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}+x}\frac{1}{\sqrt{2\pi}\sigma}\mathbb{1}[\underline{j}\le 2^k y\le\bar{j}](2^k y-j)^l\cdot\frac{(-1)^s}{s!}\frac{(x-my)^{2s}}{2^s\sigma^{2s}}\mathrm{d}y\\
&=\sum_{s=0}^{t-1}\sum_{l'=0}^{l}\frac{-(-1)^{s+l}}{\sqrt{2\pi}m^{l+1}s!2^s}\left[{}_l\mathrm{C}_{l'}(2^k\sigma)^{l'}(jm-2^kx)^{l-l'}\int_{-C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}}^{C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}}\mathbb{1}\left[\frac{x-m2^{-k}\bar{j}}{\sigma}\le y\le\frac{x-m2^{-k}\underline{j}}{\sigma}\right]y^{l'+2s}\mathrm{d}y\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left(\text{by resetting }\frac{x-my}{\sigma}\text{ as }y\right)\\
&=\sum_{s=0}^{t-1}\sum_{l'=0}^{l}\frac{-(-1)^{s+l}{}_l\mathrm{C}_{l'}2^{kl'}\sigma^{l'}(jm-2^kx)^{l-l'}}{\sqrt{2\pi}m^{l+1}s!2^s(l'+2s+1)}\left[\left(\min\left\{C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}},\max\left\{\frac{x-m2^{-k}\underline{j}}{\sigma},-C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}\right\}\right\}\right)^{l'+2s+1}\right.\\
&\qquad\left.-\left(\min\left\{C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}},\max\left\{\frac{x-m2^{-k}\bar{j}}{\sigma},-C_{\text{clip},1}\sqrt{\log\varepsilon^{-1}}\right\}\right\}\right)^{l'+2s+1}\right].
\end{aligned}\tag{17}$$

We decompose (17) into the following sub-modules for convenience. We let

$$\begin{aligned}
f_1^{l',s}(x,\sigma,m) &= (\min\{C_{\text{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1},\max\{\frac{x-m2^{-k}\underline{j}}{\sigma},-C_{\text{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1}\}\})^{l'+2s+1},\\
f_2^{l',s}(x,\sigma,m) &= (\min\{C_{\text{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1},\max\{\frac{x-m2^{-k}\bar{j}}{\sigma},-C_{\text{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1}\}\})^{l'+2s+1},\\
f_3^{l',s}(x,\sigma,m) &= f_1^{l',s}(x,\sigma,m) - f_2^{l',s}(x,\sigma,m)\\
f_4^{l'}(x,m) &= (jm-2^kx)^{l-l'},\\
f_5^{l'}(\sigma) &= \sigma^{l'},\\
f_6(m) &= m^{-(l+1)},\\
f_7^{l',s}(x,\sigma,m) &= f_3^{l',s}(x,\sigma,m)f_4^{l'}(x,m)f_5^{l'}(\sigma)f_6(m).
\end{aligned}$$

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

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

by setting $\varepsilon = \min\{\underline{\sigma}, \varepsilon_1\}$ in corollary G.10 for ϕ_{inv}, $\varepsilon = \varepsilon_1, C = \max\{2C + 3 + l, \underline{\sigma}^{-1}\} \geq \max\{|x| + m2^{-k}\underline{j}, \underline{\sigma}^{-1}\}$ in Lemma G.8 for the first ϕ_{mult}, $a = -C_{\mathrm{clip},1} \log^{\frac{1}{2}} \varepsilon^{-1}, b = C_{\mathrm{clip},1} \log^{\frac{1}{2}} \varepsilon^{-1}$ in Lemma G.6 for ϕ_{clip}, and $\varepsilon = \varepsilon_1, C = C_{\mathrm{clip},1} \log^{\frac{1}{2}} \varepsilon^{-1}$ in Lemma G.8 for the second ϕ_{mult}. Then, using Lemmas G.1, G.6, G.8 and G.9 the size of the network is at most

$$\begin{aligned} L &= \mathcal{O}(\log^2 \varepsilon_1^{-1} + \log^2 \underline{\sigma}^{-1} + \log^2 \varepsilon^{-1} + \log C), \\ \|W\|_\infty &= \mathcal{O}(\log^3 \underline{\sigma}^{-1} + \log^3 \varepsilon_1^{-1} + \log \varepsilon^{-1}), \\ S &= \mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^4 \underline{\sigma}^{-1} + \log^2 \varepsilon^{-1} + \log C), \\ B &= \mathcal{O}(\varepsilon_1^{-3} + \underline{\sigma}^{-3} + C^2) + \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}. \end{aligned} \tag{18}$$

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

$$\begin{aligned} &\varepsilon_1 + \max\{4eC_{\mathrm{clip},1}^2 \log \varepsilon^{-1}, \lceil \log_2 \varepsilon^{-1} \rceil\}(C_{\mathrm{clip},1} \log^{\frac{1}{2}} \varepsilon^{-1})^{l+2\max\{4eC_{\mathrm{clip},1}^2 \log \varepsilon^{-1}, \lceil \log_2 \varepsilon^{-1} \rceil\}-2} \\ &\quad \times (\varepsilon_1 + 2\max\{C + l + 3, \underline{\sigma}^{-1}\}^2(\varepsilon_1 + \varepsilon_{\mathrm{error}}\varepsilon_1^{-2})) \\ &= (\varepsilon_1 \underline{\sigma}^{-2} + \varepsilon_{\mathrm{error}}(\varepsilon_1^{-2} + C^2)) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}. \end{aligned}$$

$f_2^{l',s}(x, \sigma, m)$ is also approximated in the same way, and therefore $f_3^{l',s}(x, \sigma, m)$ is approximated by $\phi_3^{l',s}(x, \sigma', m')$ with the error up to an additive error of $(\varepsilon_1 \underline{\sigma}^{-2} + \varepsilon_{\mathrm{error}}(\varepsilon_1^{-2} + C^2)) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}$ using a neural network with the same size as that of (18).

Next, we consider $f_4^{l'}(x, m)$. Since $2^k x = \mathcal{O}(C2^k)$ and $|jm - jm'| \leq \mathcal{O}(C2^k \varepsilon_{\mathrm{error}})$, we approximate $f_4^{l'}(x, m)$ 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 G.1 and G.8 (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 $f_4^{l'}(x, m)$ and $\phi_4^{l'}(x, m')$ is bounded as $\varepsilon_1 + \mathcal{O}(C^l 2^{kl})\varepsilon_{\mathrm{error}}$, using Lemma G.8.

The arguments for $f_5^{l'}(\sigma)$ and $f_6(m)$ are just setting appropriate parameters in Lemma G.8 and corollary G.10, respectively. For $f_5^{l'}(\sigma)$, 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 $d = l', \varepsilon = \varepsilon_1$ in Lemma G.8. For $f_6(m)$, there exists a neural network $\phi^{l+1}(\sigma)$ with $L = \mathcal{O}(\log^2 \varepsilon_1^{-1} + \log^2 \underline{m}^{-1}), \|W\|_\infty = \mathcal{O}(\log^3 \varepsilon_1^{-1} + \log^3 \underline{m}^{-1}), S = \mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^4 \underline{m}^{-1}), B = \mathcal{O}(\varepsilon_1^{-l-1} + \underline{m}^{-l-1})$ and the approximation error between $f_6(m)$ and $\phi_6(m')$ is bounded by $\varepsilon_1 + (l+1)\varepsilon_1^{-l-2}\varepsilon_{\mathrm{error}} + (l+1)\underline{m}^{-l-2}\varepsilon_{\mathrm{error}}$, by setting $d = l + 1, \varepsilon = \min\{\varepsilon_1, \underline{m}\}$ in corollary G.10.

Therefore, Lemma G.8 with $\varepsilon = \varepsilon_1$ says that there exists a neural network $\phi_7^{l',s}(x, m, \sigma)$ such that

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

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

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

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 (17) and using remark G.5. 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 (17) is bounded as

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

In order to bound the first term by ε, we take $\varepsilon_1 = \mathcal{O}(2^{-(k+1)l} \underline{\sigma}^2 \log^{-\mathcal{O}_l(\log \varepsilon^{-1})} \varepsilon^{-1})$. Then, the total approximation error is bounded by $\mathcal{O}(\varepsilon) + \varepsilon_{\text{error}} 2^{(k+1)l} (2^{(k+1)l(l+2)} \underline{\sigma}^{-2(l+2)} + \underline{m}^{-4l} + C^{4l} 2^{4kl}) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1})$ and this is achieved by a neural network with

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

Note that $\phi_3^{l',s}(x, \sigma, m) = 0$ holds if $C_{\text{clip},1} \log^{\frac{1}{2}} \varepsilon^{-1} \leq \frac{x - m2^{-k}\bar{j}}{\sigma}$ or $-C_{\text{clip},1} \log^{\frac{1}{2}} \varepsilon^{-1} \geq \frac{x - m2^{-k}\underline{j}}{\sigma}$ holds. Therefore, according to Lemma G.8, $\phi_7^{l',s}(x, m', \sigma')$ also satisfies this property, and so $\phi_{\text{dif},1}$ does. Also, because $\int_{-\frac{\sigma C_{\text{clip},1}}{m}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m}}^{\frac{\sigma C_{\text{clip},1}}{m}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m}} \frac{1}{\sqrt{2\pi}\sigma} \mathbb{1}[\underline{j} \leq 2^k y \leq \bar{j}](2^k y - j)^l \exp\left(-\frac{(x-my)^2}{2\sigma^2}\right) \mathrm{d}y$ can bounded by $\mathcal{O}(C_f)$, we can clip $\phi_{\text{dif},1}$ so that it is bounded by $\mathcal{O}(1)$. □

We now approximate the (modified) tensor product diffused B-spline bases. The following is the formal version of Lemma C.3. Without $\left(\prod_{i=1}^d \mathbb{1}[-2 + \varepsilon_{\text{a}} \leq y_i \leq 2 - \varepsilon_{\text{a}}]\right)$, the statement matches that of Lemma C.3. This network $\phi_{\text{dif},3}$ corresponds to ϕ_{TDB} in Lemma C.3.

Lemma C.11 (Approximation of the diffused B-spline bases)**.** *Let $k \in \mathbb{Z}_+^d, j \in \mathbb{Z}^d, l \in \mathbb{Z}_+$ with $k_i \leq \bar{k}$ $(i = 1, 2, \cdots, d)$ and $-C2^{k_i} - l \leq j_i \leq C2^{k_i} + l + 1$ $(i = 1, 2, \cdots, d)$, and take ε $(0 < \varepsilon < \frac{1}{2})$ and $C > 0$ arbitrarily. Assume that $1 \geq \sigma, \sigma' \geq \underline{\sigma}$, $1 \geq m, m' \geq \underline{m}$, $|\sigma' - \sigma|, |m' - m| \leq \varepsilon_{\text{error}}$. There exists a neural network $\phi_{\text{dif},3}(x) \in \Phi(L, W, S, B)$ with*

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

such that

$$\left| \phi_{\text{dif},3}^{k,j}(x, \sigma', m') - \int_{\mathbb{R}^d} \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \left(\prod_{i=1}^d \mathbb{1}[-2 + \varepsilon_{\text{a}} \leq y_i \leq 2 - \varepsilon_{\text{a}}] \right) M_{k,j}^d(y) \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y \right|$$
$$\leq \mathcal{O}(\varepsilon \log^{-1} \varepsilon) + \varepsilon_{\text{error}} 2^{(k+1)l} (2^{(k+1)l(l+2)} \underline{\sigma}^{-2(l+2)} + \underline{m}^{-4l} + C^{4l} 2^{4kl}) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}$$

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

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

$$\left\| \phi_{\text{dif},4}^{k,j}(x, \sigma', m') - \int_{\mathbb{R}^d} \frac{x - my}{\sigma^{d+1} (2\pi)^{\frac{d}{2}}} \left(\prod_{i=1}^d \mathbb{1}[-2 + \varepsilon_{\text{a}} \leq y_i \leq 2 - \varepsilon_{\text{a}}] \right) M_{k,j}^d(y) \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y \right\|$$
$$\leq \mathcal{O}(\varepsilon \log^{-1} \varepsilon) + \varepsilon_{\text{error}} 2^{(k+1)l} (2^{(k+1)l(l+2)} \underline{\sigma}^{-2(l+2)} + \underline{m}^{-4l} + C^{4l} 2^{4kl}) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}.$$

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

Furthermore, $\phi_{\text{dif},3}^{k,j}(x, \sigma', m') = 0$ and $\phi_{\text{dif},4}^{k,j}(x, \sigma', m') = 0$ hold if $\|x\|_\infty \geq m'(2 - \varepsilon_{\text{a}}) + \sigma' C_{\text{clip},1} \log^{\frac{1}{2}} \varepsilon^{-1}$.

Proof. Here we only prove the first part, because the second part follows in the same way.

From the discussion (16), we approximate

$$\prod_{i=1}^{d}\left(\sum_{l'=0}^{l+1}\frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!}\int_{y_i\in a_i^x}\frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\mathbb{1}[|y_i|\le 2-\varepsilon_{\mathrm{a}}]\mathbb{1}[0\le 2^{k_i}y_i-j_i\le l+1]\right.$$
$$\left.\times(2^{k_i}y_i-l'-j_i)_+^l\exp\left(-\frac{(x_i-my_i)^2}{2\sigma^2}\right)\mathrm{d}y_i\right) \tag{19}$$

is equal to $D_{k,j}^d(x)$ within an additive error of $\mathcal{O}(\varepsilon)$, so we approximate (19). Here $a_i^x=[\frac{x_i}{m}-\frac{\sigma C_{\mathrm{clip},1}}{m}\sqrt{\log\varepsilon^{-1}},\frac{x_i}{m}+\frac{\sigma C_{\mathrm{clip},1}}{m\sigma}\sqrt{\log\varepsilon^{-1}}]$.

We let $f_i(y_i;j_i,k_i,l'):=\mathbb{1}[|y_i|\le 2-\varepsilon_{\mathrm{a}}]\mathbb{1}[0\le 2^{k_i}y_i-j_i\le l+1](2^{k_i}y_i-l'-j_i)_+^l\exp\left(-\frac{(x_i-my_i)^2}{2\sigma^2}\right)\mathrm{d}y_i$. First, $\sum_{l'=0}^{l+1}\frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!}f_i(y_i;j_i,k_i,l')$ is approximated by $\sum_{l'=0}^{l+1}\frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!}\phi_{\mathrm{dif},1}^{j_i,\overline{j}_{l'},\underline{j}_{l'},k_i}(y_i,\sigma',m')$ (see remark G.5 for aggregation of the networks). Here, $\overline{j}_{l'}$ and $\underline{j}_{l'}$ are defined so that $\mathbb{1}[\underline{j}_{l'}\le 2^k y\le\overline{j}_{l'}]=\mathbb{1}[|y_i|\le 2-\varepsilon_{\mathrm{a}}]\mathbb{1}[0\le 2^{k_i}y_i-j_i\le l+1]$ holds.

Now we multiply $\sum_{l'=0}^{l+1}\frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!}\phi_{\mathrm{dif},1}^{j_i,\overline{j}_{l'},\underline{j}_{l'},k_i}(y_i,\sigma',m')$ over $i=1,2,\cdots,d$ using ϕ_{mult} to obtain the desired network $\phi_{\mathrm{dif},3}^{k,j}$. According to Lemma C.10 with $\varepsilon=\varepsilon$ and Lemma G.8 with $\varepsilon=\varepsilon$ and $C=\mathcal{O}(1)$, there exists a neural network $\phi_{\mathrm{dif},3}(x)\in\Phi(L,W,S,B)$ with

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

and we can bound the approximation error for (19) with

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

Finally, because $\mathbb{1}[\underline{j}_{l'}\le 2^k y\le\overline{j}_{l'}]=\mathbb{1}[|y_i|\le 2-\varepsilon_{\mathrm{a}}]\mathbb{1}[0\le 2^{k_i}y_i-j_i\le l+1]$ holds, we have $2^k\overline{j}_{l'}\le 2-\varepsilon_{\mathrm{a}}$ and $2^k\underline{j}_{l'}\ge -2+\varepsilon_{\mathrm{a}}$. Therefore, according to Lemma C.10, $\phi_{\mathrm{dif},1}^{j_i,\overline{j}_{l'},\underline{j}_{l'},k_i}(y_i,\sigma',m')=0$ whenever $|y_i-m'(2-\varepsilon_{\mathrm{a}})|\ge\sigma' C_{\mathrm{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1}$. Together with Lemma G.8, the resulting network $\phi_{\mathrm{dif},3}^{k,j}(x,\sigma',m')$ is equal to 0 if $\|y\|_\infty\ge m'(2-\varepsilon_{\mathrm{a}})+\sigma' C_{\mathrm{clip},1}\log^{\frac{1}{2}}\varepsilon^{-1}$.

□

C.5 Approximation error bound

Now we put it all together and derive the approximation theorems. Throughout this subsection, we take $N\gg 1$, $T_1=\underline{T}=N^{-2(s+1)/d}$ and $T_5=\overline{T}=\mathcal{O}(\log N)$. However, as long as $\underline{T}=\mathrm{poly}(N)$ and $\overline{T}=\mathcal{O}(\log N)$, the following still holds by just adjusting constants hidden in the $\mathcal{O}$ notation. Moreover, we take $T_2=T_*=N^{-(2-\delta)/d}$, $T_3=2T_2$, $T_4=3T_2$.

We first limit the discussion into the bounded region. Let's take $\varepsilon_{\mathrm{b}}=N^{-2(2s+4)/d}m_t^{-d}$. According to Lemma C.9, we have that

$$\begin{aligned}&\int_{\|x\|_\infty\ge m_t+\sigma_t\sqrt{2\log\varepsilon_{\mathrm{b}}^{-1}}}p_t(x)\|\hat{s}(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}x\\ &\le\frac{2C_{\mathrm{b},3}}{m_t^d\sigma_t}\left(\log^{(d-1)/2}\varepsilon_{\mathrm{b}}^{-1}+(1+\log m_t^{-1})\log^{d/2-1}\varepsilon_{\mathrm{b}}^{-1}\right)\varepsilon_{\mathrm{b}}\\ &\quad+2\left(\frac{C_{\mathrm{b},3}\sigma_t}{m_t^d}\log^{(d-2)/2}\varepsilon_{\mathrm{b}}^{-1}\right)\varepsilon_{\mathrm{b}}\|s^\circ(\cdot,t)\|_\infty^2.\end{aligned} \tag{20}$$

Because $\sigma_t^{-1} = \mathcal{O}(\sqrt{T_1}) = \mathcal{O}(N^{-(s+1)/d})$ and $m_t^{-1} = \exp(\mathcal{O}(\log N))$ for $t \in [T_1, T_5]$, (20) is bounded by $\lesssim N^{-(3s+7)/d} \log^{(d+1)/2} N + N^{-2(2s+4)/d} \|\hat{s}(\cdot, t)\|_\infty^2 \log^{d/2} N \lesssim N^{-3(s+2)/d} + N^{-2(2s+3)/d} \|\hat{s}(\cdot, t)\|_\infty^2$.

Let us then consider $s^\circ(x, t)$ in $x \in [-m_t - \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}, m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}]^d$. Then, according to (7) in Lemma C.8, $s^\circ(x, t)$ can be taken so that it satisfies

$$\|s^\circ(\cdot, t)\|_\infty \leq \frac{C_{\mathrm{b},2}}{\sigma_t} \sqrt{\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2} + \log m_t^{-1} + 1} \lesssim \frac{\sqrt{\log N}}{\sigma_t}$$

Therefore, we take $\|s^\circ(\cdot, t)\|_\infty \lesssim \frac{\sqrt{\log N}}{\sigma_t} \lesssim N^{(s+2)/d}$ (just using clipping (Lemma G.6)), so that (20) is bounded by $\lesssim N^{-2(s+1)/d}$.

In the following, we consider approximation of $s^\circ(x, t)$ within $[-m_t - \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}, m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}]^d$, where $s^\circ(x, t)$ is bounded with $\lesssim \sigma_t^{-1} \sqrt{\log N}$. Therefore we clip $\hat{s}(x, t)$ so that $\|\hat{s}(x, t)\| \lesssim \sigma_t^{-1} \log N$ holds.

Also, we ignore the place where $p_t(x) \leq N^{-2(s+2)/d}$. We can bound the error caused by this using Lemma C.9, as

$$\begin{aligned} &\int_{\|x\|_\infty \leq m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}} \mathbb{1}[p_t(x) \leq N^{-2(s+2)/d}] p_t(x) \|\hat{s}(x, t) - s^\circ(x, t)\|^2 \mathrm{d}x \\ &\lesssim N^{-2(s+2)/d} \cdot \frac{\log N}{\sigma_t^2} (2m_t + 2\sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}})^d \lesssim \frac{N^{-2(s+1)/d}}{\sigma_t^2}. \end{aligned} \tag{21}$$

From now, we bound the approximation error in the area that satisfies $\|x\|_\infty \leq m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}$ and $p_t(x) \geq N^{-2(s+2)/d}$.

We first consider the approximation for $[T_1, T_4]$. We start with the following lemma.

Lemma C.12 (Basis decomposition)**.** *Under Assumptions 2.2, 2.3, 2.4 with $a_0 = N^{-(1-\delta)/d}$, and $N \gg 1$, there exists f_N such that*

$$\begin{aligned} &\|f - f_N\|_{L^2([-1,1]^d)} \lesssim N^{-s/d}, \\ &\|f - f_N\|_{L^2([-1,1]^d \setminus [-1+N^{-(1-\delta)/d}, 1-N^{-(1-\delta)/d}]^d)} \lesssim N^{-(3s+7)/d}, \end{aligned}$$

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

$$f_N(x) = \sum_{i=1}^{N} \alpha_i \mathbb{1}[\|x\|_\infty \leq 1] M_{k_i, j_i}^d(x) + \sum_{i=N+1}^{3N} \alpha_i \mathbb{1}[\|x\|_\infty \leq 1 - N^{-(1-\delta)/d}] M_{k_i, j_i}^d(x), \tag{22}$$

where $-2^{(k_i)_m} - l \leq (j_i)_m \leq 2^{(k_i)_m}$ $(i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$, $\|k\|_\infty \leq K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ 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 $f \in B_{p,q}^{2s}([-1, 1]^d \setminus [-1+N^{-(1-\delta)/d}, 1-N^{-(1-\delta)/d}]^d)$, according to Lemma G.15, we have f_1 such that

$$\|f - f_1\|_{L^2([-1,1]^d \setminus [-1+N^{-(1-\delta)/d}, 1-N^{-(1-\delta)/d}]^d)} \lesssim N^{-(3s+7)/d}.$$

and has the following form:

$$f_1(x) = \sum_{i=1}^{N} \alpha_i M_{k_i, j_i}^d(x),$$

where $-2^{(k_i)_m} - l \leq (j_i)_m \leq 2^{(k_i)_m}$ $(i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$, $\|k\|_\infty \leq K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ 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 $\|f\|_{B^s_{p,q}} \leq 1$, we have f_2 such that

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

and has the following form:

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

where $-2^{(k_i)_j} - l \leq (j_i)_j \leq 2^{(k_i)_j}$ $(i = 1, 2, \cdots, N,\ j = 1, 2, \cdots, d)$, $\|k\|_\infty \leq K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1}\log N)$ 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,

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

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 C.13 (Approximation of score function for $T_1 \leq t \leq T_4$)**.** *There exists a neural network $\phi_{\mathrm{score},1} \in \Phi(L, W, S, B)$ that satisfies*

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

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

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

Proof. The construction is straightforward. Based on (22) of Lemma C.12, we let

$$\begin{aligned}
\int \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y &\eqqcolon \int \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y \\
&= \sum_{i=1}^{N} \alpha_i E^{(1)}_{k_i,j_i}(x) =: f_1((x,\sigma,m))
\end{aligned}$$

and

$$\begin{aligned}
\int \frac{x - my}{\sigma^{d+1} (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y &\eqqcolon \int \frac{x - my}{\sigma^{d+1} (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x - my\|^2}{2\sigma^2}\right) \mathrm{d}y \\
&= \sum_{i=1}^{N} \alpha_i E^{(2)}_{k_i,j_i}(x) =: f_2(x,\sigma,m),
\end{aligned}$$

so that α_i, $E^{(1)}_{k_i,j_i}(x)$ and $E^{(2)}_{k_i,j_i}(x)$ correspond to the basis decomposition in Lemma C.12. Thus, $|\alpha_i| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)}$ and $\|k\|_\infty = \mathcal{O}(\log N)$. We remark that ε_{a} is set to be 1 or $1 + N^{-(1-\delta)/d}$ in (14) and (15). We approximate $E^{(1)}_{k_i,j_i}$ and $E^{(2)}_{k_i,j_i}$ by Lemma C.11 with $\varepsilon = \varepsilon_1$ and $C = m_t + \sigma_t\sqrt{2\log \varepsilon_{\mathrm{b}}^{-1}} = \mathcal{O}(1)$, where ε_1 is a scaler value smaller than $N^{-2(s+2)/d}$ and will be adjusted later. This gives realization of $f_1(x,\sigma,m)$ and $f_2(x,\sigma,m)$ by neural networks $\phi_{\mathrm{dif},5}(x,\sigma,m)$ and $\phi_{\mathrm{dif},6}(x,\sigma,m)$. Because $\int \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma^2}\right) \mathrm{d}y$

and $\int \frac{x-my}{\sigma^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma^2}\right) \mathrm{d}y$ are bounded with $\mathcal{O}(1)$, we can assume $\phi_{\mathrm{dif},5}$ and $\phi_{\mathrm{dif},6}$ are clipped with $\mathcal{O}(1)$. Moreover, we lower bound $\phi_{\mathrm{dif},5}$ with $N^{-2(s+2)/d}$ because of (21).

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

$$[\phi_{\mathrm{dif},7}(x,\sigma,m)]_i := \phi_{\mathrm{clip}}(\phi_{\mathrm{mult}}(\phi_{\mathrm{inv}}(\phi_{\mathrm{dif},5}(x,\sigma,m)), [\phi_{\mathrm{dif},6}(x,\sigma,m)]_i); -C_1, C_1).$$

Because $s^\circ(x,t)$ is bounded with $\lesssim \sigma_t^{-1}\log N$, we take $C_1 = \mathcal{O}(\log^{\frac{1}{2}} N)$ above. For ϕ_{inv} we let $\varepsilon = \varepsilon_1$ in Lemma G.9 and for ϕ_{mult} we let $\varepsilon = \varepsilon_1$ and $C = \mathcal{O}(1)$.

Now the structure of $\phi_{\mathrm{dif},7}$ is evaluated as

$$\begin{aligned} L &= \mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^2 N), \\ \|W\|_\infty &= \mathcal{O}(N(\log^6 \varepsilon_1^{-1} + \log^3 N)), \\ S &= \mathcal{O}(N(\log^8 \varepsilon_1^{-1} + \log^4 N)), \\ B &= \exp(\mathcal{O}(\log^2 \varepsilon_1^{-1} + \log N)). \end{aligned}$$

The sensitivity analysis by Lemma G.9 is too loose to consider the approximation bound here. We refine the bound using the boundedness of $\phi_{\mathrm{dif},5}$, $\phi_{\mathrm{dif},6}$, and $\phi_{\mathrm{dif},7}$. We divide into the two cases. First, we consider when $x \in [-m_t, m_t]^d$. Since, $p_t(x)$ is lower bounded by $C_{\mathrm{b},1}^{-1}$ according to Lemma C.7, as long as $|\phi_{\mathrm{dif},5}(x,\sigma',m') - p_t(x)| \le C_{\mathrm{b},1}^{-1}/2$ and ε_1 is sufficiently small, we can say that the approximation error is bounded by $\lesssim |\phi_{\mathrm{dif},6}(x,\sigma',m') - p_t(x)|$. On the other hand, if $|\phi_{\mathrm{dif},6}(x,\sigma',m') - p_t(x)| \ge C_{\mathrm{b},1}^{-1}/2$, we no longer have such bound, but this time we can use the fact that $\phi_{\mathrm{dif},7}$ and $\frac{\sigma_t \nabla p_t(x)}{p_t(x)}$ is bounded by $\log^{\frac{1}{2}} N$. Therefore, when $x \in [-m_t, m_t]^d$, we can bound the approximation error as

$$\begin{aligned} &\left\| \phi_{\mathrm{dif},7}(x,\sigma,m) - \frac{\sigma_t \nabla p_t(x)}{p_t(x)} \right\| \\ &\lesssim \log^{\frac{1}{2}} N(|\phi_{\mathrm{dif},5}(x,\sigma',m') - p_t(x)| + \|\phi_{\mathrm{dif},6}(x,\sigma',m') - \sigma_t \nabla p_t(x)\|). \end{aligned}$$

Next, we consider the case when $x \in [-m_t - \sigma_t\sqrt{2\log \varepsilon_{\mathrm{b}}^{-1}}, m_t + \sigma_t \sqrt{2\log \varepsilon_{\mathrm{b}}^{-1}}]^d$ holds and at the same time x is outside of $[-m_t, m_t]^d$. In this case, let $C_1 = \mathcal{O}(\log^{\frac{1}{2}} N)$ be the upper bound of $\sigma_t s^\circ(x,t)$. Then, when ε_1 is sufficiently small, we have that

$$\begin{aligned} &\left\| \phi_{\mathrm{dif},7}(x,\sigma',m') - \frac{\sigma_t \nabla p_t(x)}{p_t(x)} \right\| \\ &\lesssim \frac{\|\phi_{\mathrm{dif},6}(x,\sigma',m') - \sigma_t \nabla p_t(x)\|}{p_t(x)} + \|\sigma_t \nabla p_t(x)\| \left| \frac{1}{\phi_{\mathrm{dif},5}(x,\sigma',m')} - \frac{1}{p_t(x)} \right|. \end{aligned}$$

Then, because it should hold that $\|\sigma_t \nabla p_t(x)\| \le C_1 p_t(x)$, $\|\sigma_t \nabla p_t(x)\| \left| \frac{1}{\phi_{\mathrm{dif},5}(x,\sigma',m')} - \frac{1}{p_t(x)} \right|$ is further bounded by $C_1 |\phi_{\mathrm{dif},5}(x,\sigma',m') - p_t(x)|/p_t(x)$. Because we are considering x with $p_t(x) \ge N^{-2(s+2)/d}$, the approximation error is bounded by

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

Now we let

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

By setting $\varepsilon = \varepsilon_1$ and $C = \sigma_t^{-1}$ in Lemma G.8 for ϕ_{mult} and $\varepsilon = \varepsilon_{\mathrm{error}}$ in Lemma C.6 for ϕ_σ and ϕ_m. To summarize, the approximation error is evaluated as

$$\begin{aligned} &\|\phi_{\mathrm{score},1}(x,t) - s^\circ(x,t)\| \\ &\lesssim N^{(2s+5)/d}[\mathcal{O}(\varepsilon_1 \log^{-1} \varepsilon_1) + \varepsilon_{\mathrm{error}} 2^{(k+1)l}(2^{(k+1)l(l+2)} \underline{\sigma}^{-2(l+2)} + \underline{m}^{-4l} + C^{4l} 2^{4kl}) \log^{\mathcal{O}(\log \varepsilon_1^{-1})} \varepsilon_1^{-1}]/\sigma_t \\ &\quad + \begin{cases} \log^{\frac{1}{2}} N(|f_1(x) - p_t(x)| + \|f_2(x) - \sigma_t \nabla p_t(x)\|)/\sigma_t & (\|x\|_\infty \le m_t) \\ N^{(2s+5)/d}(|f_1(x) - p_t(x)| + \|f_2(x) - \sigma_t \nabla p_t(x)\|)/\sigma_t & (\|x\|_\infty \le m_t + \sigma_t\sqrt{2\log \varepsilon_{\mathrm{b}}^{-1}}). \end{cases} \end{aligned}$$

We take $\varepsilon_1 = N^{-3(s+3)/d} < N^{(2s+5)/d}$, and $\varepsilon_{\text{error}}^{-1} = N^{-3(s+3)/d} 2^{(k+1)l}(2^{(k+1)l(l+2)}\underline{\sigma}^{-2(l+2)} + \underline{m}^{-4l} + C^{4l}2^{4kl}) \log^{\mathcal{O}(\log \varepsilon_1^{-1})} \varepsilon_1^{-1}] = \exp(\log^2 N)$ so that $[\mathcal{O}(\varepsilon_1 \log^{-1} \varepsilon_1) + \varepsilon_{\text{error}} 2^{(k+1)l}(2^{(k+1)l(l+2)}\underline{\sigma}^{-2(l+2)} + \underline{m}^{-4l} + C^{4l}2^{4kl}) \log^{\mathcal{O}(\log \varepsilon_1^{-1})} \varepsilon_1^{-1}]$ is bounded by $N^{-(s+3)/d}$. The network structure is then evaluated as

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

Finally, we evaluate $\int p_t(x) \left\|\phi_{\text{score},1}(x,t) - \nabla \log p_t(x)\right\|^2 \mathrm{d}x$.

$$\begin{aligned}
&\int_{m_t \le \|x\|_\infty \le \sigma_t \sqrt{2\log \varepsilon_{\text{b}}^{-1}}} \mathbb{1}[p_t(x) \ge N^{-2(s+2)/d}] p_t(x) \left\|\phi_{\text{score},1}(x,t) - \nabla \log p_t(x)\right\|^2 \mathrm{d}x \\
&\le \int_{\|x\|_\infty \le m_t} \mathbb{1}[p_t(x) \ge N^{-2(s+2)/d}] \left\|\phi_{\text{score},1}(x,t) - \nabla \log p_t(x)\right\|^2 \mathrm{d}x \\
&\quad + \int_{m_t \le \|x\|_\infty \le \sigma_t \sqrt{2\log \varepsilon_{\text{b}}^{-1}}} \mathbb{1}[p_t(x) \ge N^{-2(s+2)/d}] \left\|\phi_{\text{score},1}(x,t) - \nabla \log p_t(x)\right\|^2 \mathrm{d}x \\
&\lesssim N^{-2(s+3)/d} \\
&+ \log N \int_{\|x\|_\infty \le m_t} p_t(x)(|f_1(x,\sigma,m) - p_t(x)|^2 + \|f_2(x,\sigma,m) - \sigma_t \nabla p_t(x)\|^2)/\sigma_t^2 \mathrm{d}x \qquad (23) \\
&\quad + N^{(4s+10)/d} \int_{m_t \le \|x\|_\infty \le \sigma_t \sqrt{2\log \varepsilon_{\text{b}}^{-1}}} p_t(x)(|f_1(x,\sigma,m) - p_t(x)|^2 + \|f_2(x,\sigma,m) - \sigma_t \nabla p_t(x)\|^2)/\sigma_t^2 \mathrm{d}x. \qquad (24)
\end{aligned}$$

We first bound (23).

$$\begin{aligned}
&\log N \int_{\|x\|_\infty \le m_t} p_t(x)(|f_1(x,\sigma_t,m_t) - p_t(x)|^2 + \|f_2(x,\sigma_t,m_t)(x,\sigma,m) - \sigma_t \nabla p_t(x)\|^2)/\sigma_t^2 \mathrm{d}x \\
&\lesssim \log N \sigma_t^{-2} \int_{\|x\|_\infty \le m_t} p_t(x) \int_y \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} |f_N(y) - f(y)|^2 \left(1 + \frac{\|x - my\|^2}{\sigma_t^2}\right) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \mathrm{d}x \\
&\lesssim \log N \sigma_t^{-2} \int_{\|x\|_\infty \le m_t} p_t(x) \int_{y:\ \|y - x/m\|_\infty \le \frac{\sigma_t}{m_t}\sqrt{2\log \varepsilon_{\text{b}}^{-1}}} \\
&\qquad\qquad \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} |f_N(y) - f(y)|^2 \left(1 + \frac{\|x - my\|^2}{\sigma_t^2}\right) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \mathrm{d}x + N^{-2s/d} \\
&\lesssim \log N \sigma_t^{-2} \int_{\|y\|_\infty \le 2} \int_x p_t(x) \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} |f_N(x) - f(x)|^2 \left(1 + \frac{\|x - my\|^2}{\sigma_t^2}\right) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}x \mathrm{d}y \\
&\quad + N^{-2s/d} \\
&\lesssim N^{-2s/d} \log N / \sigma_t^2.
\end{aligned}$$

For the first inequality, we used Jensen-type inequality. More specifically, because we have $\int_y \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y = \mathcal{O}(1)$, we obtain that $|f_1(y,\sigma_t,m_t) - p_t(y)|^2 = |\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} (f_N(x) - p_0(x)) \exp(-\|x - my\|^2/2\sigma_t^2) \mathrm{d}y|^2 \lesssim \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} |f_1(x,\sigma_t,m_t) - p_t(x)|^2 \exp(-\|x - my\|^2/2\sigma_t^2) \mathrm{d}y$ and that $\|f_2(y,\sigma_t,m_t) - \sigma_t \nabla p_t(y)\|^2 = \|\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} (f_2(y,\sigma_t,m_t) - \sigma_t \nabla p_t(y)) \exp(-\|x - my\|^2/2\sigma_t^2) \mathrm{d}y\|^2 \lesssim \int \frac{\|x - my\|^2}{\sigma_t^{d+2} (2\pi)^{\frac{d}{2}}} \|f_2(x,\sigma_t,m_t) - \sigma_t \nabla p_t(x)\|^2 \exp(-\|x - my\|^2/2\sigma_t^2) \mathrm{d}y \lesssim \int \frac{\|x - my\|^2}{\sigma_t^{d+2} (2\pi)^{\frac{d}{2}}} |f_1(x,\sigma_t,m_t) - p_t(x)|^2 \exp(-\|x - my\|^2/2\sigma_t^2) \mathrm{d}y$. Moreover, we used Lemma G.12 with $\varepsilon \lesssim m_t C_f^{-1} N^{-2s/d}$ for the second inequality. This allows us to clip the integral interval by $\|y - x/m\|_\infty \le \frac{\sigma_t C_{\text{clip},1}}{m_t} \sqrt{\log \varepsilon^{-1}}$ and to roughly bound the range of $\|y\|_\infty$ after clipping by 2. The final inequality is due to Lemma C.12.

Next, we consider (24).

$$
\begin{aligned}
&N^{(4s+10)/d}\int_{m_t\le\|x\|_\infty\le\sigma_t\sqrt{2\log\varepsilon_{\mathrm{b}}^{-1}}} p_t(x)(|f_1(x,\sigma_t,m_t)-p_t(x)|^2+\|f_2(x,\sigma_t,m_t)(x,\sigma,m)-\sigma_t\nabla p_t(x)\|^2)/\sigma_t^2\mathrm{d}x\\
&\lesssim N^{(4s+10)/d}\sigma_t^{-2}\int_{m_t\le\|x\|_\infty\le\sigma_t\sqrt{2\log\varepsilon_{\mathrm{b}}^{-1}}} p_t(x)\int_{y:\ \|y-x/m\|_\infty\le\frac{\sigma_t}{m_t}\sqrt{2\log\varepsilon_{\mathrm{b}}^{-1}}}\\
&\qquad\qquad \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}|f_N(y)-f(y)|^2\left(1+\frac{\|x-my\|^2}{\sigma_t^2}\right)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y\mathrm{d}x\\
&\lesssim N^{(4s+10)/d}\sigma_t^{-2}\int_{1-\sigma_t\sqrt{2\log\varepsilon_{\mathrm{b}}^{-1}}\le\|y\|_\infty\le 2}\int_x\frac{p_t(x)}{\sigma_t^d(2\pi)^{\frac{d}{2}}}|f_N(x)-f(x)|^2\left(1+\frac{\|x-my\|^2}{\sigma_t^2}\right)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}x\mathrm{d}y\\
&\lesssim N^{(4s+10)/d}\cdot N^{-2(3s+7)/d}/\sigma_t^2=N^{-2(s+2)/d}/\sigma_t^2\le N^{-2s/d}/\sigma_t^2.
\end{aligned}
$$

The final inequality is due to Lemma C.12 and $\sigma_t\sqrt{2\log\varepsilon_{\mathrm{b}}^{-1}}\lesssim N^{-(2-\delta)/2}\log N\lesssim N^{-(1-\delta)/d}$ when $t\le T_4=3N^{-(2-\delta)/d}$.

Together with (20) and (21), we finally have that

$$
\int_x p_t(x)\left\|\phi_{\mathrm{score},1}(x,t)-s^\circ(x,t)\right\|^2\mathrm{d}x\lesssim N^{-2s/d}\log N/\sigma_t^2.
$$

□

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

Lemma C.14 (Basis decomposition of p_t at $t=t_*$)**.** *If $N,N'\gg 1$ and $N'\ge t_*^{-\frac{d}{2}}N^{\frac{\delta}{2}}$, there exists $f_{N'}$ such that*

$$
\|p_{t_*}-f_{N'}\|_{L^2(\mathbb{R}^d)}\lesssim N^{-(2s+8)/d},
$$

$f_{N'}(x)=0$ for x with $\|x\|_\infty\ge C_{\mathrm{b},8}:=\sqrt{2\log N^{(2s+8)/d}}+1$, and has the following form:

$$
f_N(x)=\sum_{i=1}^N\mathbb{1}[\|x\|_\infty\le C_{\mathrm{b},8}]M^d_{k_i,j_i}(x),
$$

where $-C_{\mathrm{b},8}2^{(k_i)_m}-l\le(j_i)_l\le C_{\mathrm{b},8}2^{(k_i)_l}$ $(i=1,2,\cdots,N,\ m=1,2,\cdots,d)$, $\|k_i\|_\infty\le K=\mathcal{O}(d^{-1}\log N)$ and $|\alpha_i|\lesssim N^{\frac{4(s*4)(2-\delta)}{2\delta}}$.

Proof. Let $\alpha=\frac{2(2s+8)}{\delta}+1$. According to Lemma C.8, for any $x\in[-C_{\mathrm{b},8},C_{\mathrm{b},8}]^d$, we have

$$
\|\partial_{x_{i_1}}\partial_{x_{i_2}}\cdots\partial_{x_{i_k}}p_{T_2}(x)\|\le\frac{C_{\mathrm{b},2}}{\sigma_{t_*}^k}.
$$

Because all derivatives up to order α is bounded by $\sigma_{t_*}^{-\alpha}\lesssim t_*^{-\frac{\alpha}{2}}$, $\frac{p_{t_*}(x)}{t_*^{-\frac{\alpha}{2}}}$ belongs to W^α_∞ and its norm in W^α_∞ is bounded by a constant depending on α, and hence to $B^\alpha_{\infty,\infty}$. Therefore, according to Lemma G.15, there exists a basis decomposition with the order of the B-spline basis $l=\alpha+2$:

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

such that

$$
\|p_{t_*}-f_{N'}\|_{L^2([-2,2]^d)}\lesssim C^d_{\mathrm{b},2}N'^{-\alpha/d}t_*^{-\frac{\alpha}{2}}=C^d_{\mathrm{b},2}tN^{\alpha\delta/2d}=C^d_{\mathrm{b},2}N^{-(2s+8)/d}\lesssim N^{-(2s+7)/d},
$$

where $-C_{\mathrm{b},8}2^{(k_i)_m}-l\le(j_i)_l\le C_{\mathrm{b},8}2^{(k_i)_l}$ $(i=1,2,\cdots,N,\ m=1,2,\cdots,d)$, $\|k_i\|_\infty\le K=\mathcal{O}(d^{-1}\log N)$, and $|\alpha_{k,j}|\lesssim 1$. Also, Lemma C.9 with $\varepsilon=N^{\mathcal{O}(1)}$ $m_{t_*}+\sigma_{t_*}\sqrt{2\log N^{(2s+8)/d}}\le C_{\mathrm{b},2}$ guarantees that $\|p_{T_2}-f_N\|_{L^2(\mathbb{R}^d\subseteq[-C_{\mathrm{b},2},C_{\mathrm{b},2}]^d)}\lesssim N^{-(2s+7)/d}$. Therefore, by resetting $N^{\alpha(2-\delta)/2}\alpha_i$ as α_i, the assertion holds. □

Lemma C.14 gives a concrete construction of the neural network for $T_3 \le t \le T_5$.

Lemma C.15 (Approximation of score function for $T_3 \le t \le T_5$)**.** *Let $N, N' \gg 1$, $N' \ge t_*^{-\frac{d}{2}} N^{\frac{\delta}{2}}$, and $T_3 \vee 2t_* \le t \le T_5$. There exists a neural network $\phi_{\mathrm{score},1} \in \Phi(L, W, S, B)$ that satisfies*

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

for $t_* \le t \le T_5$. Here, $L, \|W\|_\infty, S, B$ are evaluated as

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

Proof. The proof is essentially the same as that of Lemma C.13. Let us reset the time $t - t_*$ as t in the following proof. Observe that in the new definition $t \ge T_2 \ge T_1$ holds, which is necessary to apply the argument of Lemma C.13. We decompose $f_1(x) = p_t(x)$ and $f_2(x) = \int \frac{x - my}{\sigma^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma^2}\right) \mathrm{d}y$ into the diffused B-spline bases and approximate f_1 and f_2 with $\phi_{\mathrm{dif},8}$ and $\phi_{\mathrm{dif},9}$ by using Lemma C.11, where we let $\varepsilon = \varepsilon_1$ and $C = m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}} = \mathcal{O}(\log^{\frac{1}{2}} N)$. Then we define $\phi_{\mathrm{dif},10}$ as

$$[\phi_{\mathrm{dif},10}(x, \sigma, m)]_i := \phi_{\mathrm{clip}}(\phi_{\mathrm{mult}}(\phi_{\mathrm{inv}}(\phi_{\mathrm{dif},8}(x, \sigma, m)), [\phi_{\mathrm{dif},9}(x, \sigma, m)]_i); -C_1, C_1)$$

with $C_1 = \mathcal{O}(\log^{\frac{1}{2}} N)$, and $\phi_{\mathrm{score},2} \in \Phi(L, W, S, B)$ as

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

where

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

The remaining part is exactly the same as that of Lemma C.11, except for slight difference in evaluation of $\int p_t(x) \left\|\phi_{\mathrm{score},1}(x,t) - s^\circ(x,t)\right\|^2 \mathrm{d}x$. We do not need to care about the boundary area. We proceed as

$$\begin{aligned}
&\int_{\|x\|_\infty \le m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}} \mathbb{1}[p_t(x) \ge N^{-2(s+2)/d}] p_t(x) \left\|\phi_{\mathrm{score},1}(x,t) - s^\circ(x,t)\right\|^2 \mathrm{d}x \\
&\lesssim N^{(4s+10)/d} \int_{\|x\|_\infty \le m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}} p_t(x) (|f_1(x, \sigma, m) - p_t(x)|^2 + \|f_2(x, \sigma, m) - \sigma_t \nabla p_t(x)\|^2)/\sigma_t^2 \mathrm{d}x \\
&\lesssim N^{(4s+10)/d} \int_{\|x\|_\infty \le m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}} \int_y \frac{p_t(x)}{\sigma_t^d (2\pi)^{\frac{d}{2}}} |f_N(y) - f(y)|^2 \left(1 + \frac{\|x - my\|^2}{\sigma_t^2}\right) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \mathrm{d}x / \sigma_t^2 \\
&\lesssim N^{(4s+10)/d} \int_{\|x\|_\infty \le m_t + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}} \int_{y \colon \|y - x/m\|_\infty \le \frac{\sigma_t}{m_t} \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}} \\
&\qquad \frac{p_t(x)}{\sigma_t^d (2\pi)^{\frac{d}{2}}} |f_N(y) - f(y)|^2 \left(1 + \frac{\|x - my\|^2}{\sigma_t^2}\right) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \mathrm{d}x / \sigma_t^2 \\
&\lesssim N^{(4s+10)/d} \int_{\|y\|_\infty \ge 1 + \sigma_t \sqrt{2 \log \varepsilon_{\mathrm{b}}^{-1}}} \int_x \frac{p_t(x)}{\sigma_t^d (2\pi)^{\frac{d}{2}}} |f_N(x) - f(x)|^2 \left(1 + \frac{\|x - my\|^2}{\sigma_t^2}\right) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}x \mathrm{d}y / \sigma_t^2 \\
&\lesssim \frac{N^{(4s+10)/d} \cdot N^{-3(2s+7)/d} \log^{\frac{d}{2}} N^{-1}}{\sigma_t^2} \lesssim \frac{N^{-2(s+2)/s} \log^{\frac{d}{2}} N^{-1}}{\sigma_t^2}.
\end{aligned}$$

Now we reset the time $t + T_2$ as t. Therefore, we have

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

for all $2t_* \le t \le T_5$.

□

Combining Lemmas C.13 and C.15, where we use Lemma C.13 for $T_1 \le t \le T_4$ and Lemma C.15 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 C.15. According to Lemmas C.13 and C.15, we have two neural networks $\phi_{\text{score},1}(x,t)$ and $\phi_{\text{score},2}(x,t)$, 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 G.7, $\phi^1_{\text{swit}}(t; \underline{t}_2, \bar{t}_1)\phi_{\text{score},1}(x,t) + \phi^2_{\text{swit}}(t; \underline{t}_2, \bar{t}_1)\phi_{\text{score},2}(x,t)$ approximates the approximation error in $L^2(p_t)$ with an additive error of $\frac{N^{-2s/d}\log N}{\sigma_t^2}$. Here, multiplication is realized by Lemma G.8 with $C = \mathcal{O}(\sigma_t^{-1}) \lesssim T_1^{-1/2}$ and $\varepsilon = N^{-2s/d}$. Note that $\phi^1_{\text{swit}}, \phi^2_{\text{swit}} \in [0,1]$, this multiplication does not amplify the approximation error of $\phi_{\text{score},1}$ and $\phi_{\text{score},2}$. Finally, the size of the network is evaluated by Lemmas C.13, C.15 and G.8 and remark G.5.

Finally, we slightly modify the definition of N. At this time we have that

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

with $L = \mathcal{O}(\log^4 N), \|W\|_\infty = \mathcal{O}(N\log^6 N), S = \mathcal{O}(N\log^8 N)$, and $B = \exp(\mathcal{O}(\log^4 N))$. We set $N\log^8 N$ to be N in the new definition. Then, we have that

$$\int_x p_t(x)\|\phi_{\text{score},1}(x,t) - s(x,t)\|^2 \mathrm{d}x\mathrm{d}t \lesssim \frac{N^{-2s/d}\log^{\frac{d+16s}{d}} N}{\sigma_t^2},$$

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

In order to convert the approximation error to the estimation error, we integrate the bound of Theorem 3.1 over $t \in [\underline{T}, \overline{T}] = [T_1, T_5]$ to obtain the following theorem.

Theorem C.16 (Approximation theorem)**.** *Suppose Assumptions 2.2, 2.3, 2.4 with $a_0 = N^{-(1-\delta)/d}$, and $N \gg 1$. Then there exists a neural network $\phi_{\text{score}} \in \Phi(L, W, S, B)$ that satisfies*

$$\int_{t=T_1}^{T_5}\int_x p_t(x)\|\phi_{\text{score}}(x,t) - s(x,t)\|^2 \mathrm{d}x\mathrm{d}t \lesssim N^{-2s/d}\log^{\frac{2d+16s}{d}} N.$$

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

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

C.6 Discussion on the discretization error

Although the continuous time SDE is mainly considered for simple presentation, we can also take the discretization error into consideration.

We first provide the summary. Let η be the step size and take $k = 1, 2, \cdots, K = \lfloor \frac{\overline{T}-\underline{T}}{\eta} \rfloor$, and consider the following process $(\bar{Y}_t)_{t=0}^{\eta K}$ with $\bar{Y}_0 \sim \mathcal{N}(0, I_d)$:

$$\mathrm{d}\bar{Y}_t = \beta_t(\bar{Y}_{\eta k} + 2\bar{s}(\bar{Y}_{\eta k}, t))\eta + \beta_t \mathrm{d}B_t \ (t \in [\eta k, \eta(k+1)]).$$

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. Then, we can say that $\nabla \log p_t(x)$ is $\tilde{\mathcal{O}}(\underline{T}^{-2})$-Lipschitz about x and t with high probability. Then, for all $\underline{T} \le t \le \overline{T}$, the discretization error is bounded as

$$\mathbb{E}_{t\sim p_t}[\|\nabla \log p_t(x_{\eta k}) - \nabla \log p_t(x_t)\|^2] \lesssim \varepsilon \cdot \tilde{\mathcal{O}}(\underline{T}^{-2}).$$

Setting $\eta = \tilde{\mathcal{O}}(\varepsilon_0 \underline{T}^2)$ achieves the error smaller than ε_0.

Now we formalize the argument. Let t satisfies $t \in [\eta k, \eta(k+1)]$. We need to bound the following error for $t \in [\eta k, \eta(k+1)]$.

$$\mathbb{E}_{t\sim p_t}[\|\log p_t(x_{\overline{T}-\eta k}) - \nabla \log p_t(x_{\overline{T}-t})\|^2].$$

With probability $1-\varepsilon$ with $\varepsilon > 0$, according to Lemma C.8, $\|X_{\overline{T}-\eta(k+1)}\|_\infty$ is bounded by $\tilde{\mathcal{O}}(m_{\overline{T}-\eta(k+1)} + \sigma_{\overline{T}-\eta(k+1)})$. Then, we bound $\|X_{\overline{T}-t}\|_\infty$ for $t \in [\eta k, \eta(k+1)]$. According to

Lemma G.17, $\int_{\overline{T}-\eta(k+1)}^{\overline{T}-\eta k} \sqrt{2\beta_t}\mathrm{d}B_t$ is bounded by $\tilde{\mathcal{O}}(\sqrt{\eta})$ with probability $1-\varepsilon$. Then, with probability $1-2\varepsilon$, the effect from the drift term $\int_{\overline{T}-\eta(k+1)}^{\overline{T}-\eta k} -\beta_t X_t \mathrm{d}t$ is also bounded by $\tilde{\mathcal{O}}(\eta)$. Therefore, $\|X_{\overline{T}-t}\|_\infty$ is uniformly bounded by $\tilde{\mathcal{O}}(m_{\overline{T}-\eta(k+1)} + \sigma_{\overline{T}-\eta(k+1)}\sqrt{\eta}) \lesssim \tilde{\mathcal{O}}(m_{\overline{T}-t} + \sigma_{\overline{T}-t})$ with probability $1-2\varepsilon$. Thus, at least with probability $1-2\varepsilon$, $\|Y_t\|_\infty = \tilde{\mathcal{O}}(m_{\overline{T}-t} + \sigma_{\overline{T}-t})$ holds for all $t \in [\eta k, \eta(k+1)]$.

For such paths, according to Lemma C.8, $\nabla \log p_t(x_t)$ is $\tilde{\mathcal{O}}(\sigma_t^{-3}(|\partial_t m_t| + |\partial_t \sigma_t|))$-Lipschitz. This Lipschitz constant is bounded by $\tilde{\mathcal{O}}(\sigma_t^{-3}(|\partial_t m_t| + |\partial_t \sigma_t|)) \lesssim \mathcal{O}(t^{-2})$. On the other hand, for the remaining paths of the mass $\tilde{\mathcal{O}}(\varepsilon)$, Lemma C.8 yields that the approximation error of the score at each t is bounded at most $\tilde{\mathcal{O}}(\varepsilon/\sigma_t^2)$ by separately considering the region with $\|x_t\|_\infty \le m_t + \sigma_t\sqrt{2\log\varepsilon^{-1}}$ and the other. Therefore, the discretization error is bounded by

$$\mathbb{E}_{t\sim p_t}[\|\log p_t(x_{\overline{T}-\eta k}) - \nabla \log p_t(x_{\overline{T}-t})\|^2] \lesssim \underbrace{\eta \cdot \mathcal{O}(t^{-2})}_{\text{from the Lipschitz continuous region}} + \underbrace{\varepsilon/\sigma_t^2}_{\text{from the other}}.$$

Now we set $\varepsilon = \mathcal{O}(\varepsilon_0 \underline{T})$ and $\eta = \mathcal{O}(\varepsilon_0 \underline{T}^2)$, so that the discretization error is smaller than ε_0 for all $\underline{T} \le t \le \overline{T}$.

D Generalization bound

This section considers the generalization error, via theory of function estimation.

Lemma D.1 (Equivalence of explicit score matching and denoising score matching (Vincent (2011))). *The following equality holds for all $s(x,t)$ and $t > 0$:*

$$\int_x \|s(x,t) - \nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}x = \int_x \int_y \|s(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y)p_0(x)\mathrm{d}y\mathrm{d}x + C,$$

where $C = \int_x \|\nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}x - \int_x \int_y \|\nabla \log p_t(x|y)\|^2 p_t(x|y)p_0(y)\mathrm{d}x\mathrm{d}y$.

Proof. The proof follows Vincent (2011).

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

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

□

According to this, the score matching loss at each t

$$\int_x \|s(x,t) - \nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}x$$

is equal to

$$\begin{aligned}&\mathbb{E}_{x_0\sim p_0}[\mathbb{E}_{x_t\sim p_t(x_t|x_0)}[\|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2]]\\ &- \mathbb{E}_{x_0\sim p_0}[\mathbb{E}_{x_t\sim p_t(x_t|x_0)}[\|\nabla \log p_t(x_t) - \nabla \log p_t(x_t|x_0)\|^2]].\end{aligned}$$

Based on this, we evaluate the generalization error based on the covering number evaluation. In the following, $x_{0,i}$ is sometimes written as x_i for simple presentation.

Theorem D.2 (Generalization error bound based on the covering number)**.** *Let $\hat{s}$ be the minimizer of*

$$\frac{1}{n}\sum_{i=1}^n \int_{t=\underline{T}}^{\overline{T}} \int_x \|s(x,t) - \nabla \log p_t(x|x_i)\|_2^2 p_t(x|x_i)\mathrm{d}x\mathrm{d}t, \tag{25}$$

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 ℓ corresponding to each $s \in \mathcal{S}$. Suppose every element $\ell \in \mathcal{L}$ satisfies $\|\ell\|_{L^\infty([-1,1]^d)} \le C_\mathrm{s}$ for some fixed $C_\mathrm{s} > 0$. For an arbitrary $\delta > 0$, if $\mathcal{N} := \mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty([-1,1]^d)}, \delta) \ge 3$, then we have that

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

Proof. (25) 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

$$\begin{aligned}R(\hat{\ell}, \ell^\circ) &:= \int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}t\mathrm{d}x\\ &= \int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}t\mathrm{d}x\\ &\quad - \underbrace{\int_x \int_{t=\underline{T}}^{\overline{T}} \|s^\circ(x,t) - \nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}t\mathrm{d}x}_{=0}\\ &= \int_y \int_{t=\underline{T}}^{\overline{T}} \int_x \|s(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y)p_0(x)\mathrm{d}y\mathrm{d}t\mathrm{d}x + C(\overline{T} - \underline{T})\\ &\qquad - \int_y \int_{t=\underline{T}}^{\overline{T}} \int_x \|s^\circ(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y)p_0(x)\mathrm{d}y\mathrm{d}t\mathrm{d}x - C(\overline{T} - \underline{T})\\ &= \mathbb{E}_{\{x_i'\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\ell}(x_i') - \ell^\circ(x_i'))\right]\end{aligned} \tag{26}$$

with $\{x_i'\}_{i=1}^n$, i.i.d. sample from f° and independent of $\{x_i\}_{i=1}^n$. For the second equality, we used Lemma D.1.

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 (26), 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 $\mathcal{L}_d = \{\ell_1, \ell_2, \cdots, \ell_{\mathcal{N}}\}$ be a δ-covering of $\mathcal{L}$ with the minimum cardinality in the $L^\infty([-1,1]^d)$ metric. From the assumption of $\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta) \geq 3$, we have $\log \mathcal{N} \geq 1$. 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 $\{1, 2, \cdots, \mathcal{N}\}$ such that $\|\hat{\ell} - f_J\|_\infty \leq \delta$, so that we have

$$\begin{aligned} D \leq & \frac{1}{n} \mathbb{E}_{x_i, x_i'} \left[\left| \sum_{i=1}^n g_J(x_i, x_i') \right| \right] + \|(\hat{\ell}_j(x) - \ell_J(x)) - (\hat{\ell}_j(x') - \ell_J(x')\|_\infty \\ &\leq \frac{1}{n} \mathbb{E}_{x_i, x_i'} \left[\left| \sum_{i=1}^n g_J(x_i, x_i') \right| \right] + \delta. \end{aligned} \tag{27}$$

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 eq. (27) as

$$D \leq \frac{1}{n} \mathbb{E}_{x_i, x_i'}[r_J G] + \delta \leq \frac{1}{n} \sqrt{\mathbb{E}_{x_i, x_i'}[r_J^2] \mathbb{E}_{x_i, x_i'}[G^2]} + \delta \leq \frac{1}{2} \mathbb{E}_{x_i, x_i'}[r_J^2] + \frac{1}{2n^2} \mathbb{E}_{x_i, x_i'}[G^2] + \delta, \tag{28}$$

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

$$\mathbb{E}_{x_i, x_i'}[r_J^2] \leq A^2 + \mathbb{E}_{x'}[\ell_J(x') - \ell^\circ(x')] \leq A^2 + \mathbb{E}_{x'}[\hat{\ell}(x') - \ell^\circ(x')] + \delta = R(\hat{\ell}, \ell^\circ) + A^2 + \delta. \tag{29}$$

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

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

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_{s}. Also, $\frac{g_j(x_i, x_i')}{r_j}$ is bounded with C_{s}/A. Then, using Bernstein's inequality, we have that

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

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

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

These two integrals are computed as

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

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

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

$$\begin{aligned}\mathbb{E}_{x_i, x_i'}[G^2] &\leq t_0 + 2\mathcal{N}\left(8C_{\mathrm{s}}n + 16C_{\mathrm{s}}n + \frac{128C_{\mathrm{s}}n^2}{t_0}\right) \exp\left(-\frac{t_0}{8C_{\mathrm{s}}n}\right) \\ &\leq t_0 + 16\mathcal{N} \exp\left(-\frac{3A\sqrt{t_0}}{4C_{\mathrm{s}}}\right) n(3 + 16n/t_0) \exp\left(-\frac{t_0}{8C_{\mathrm{s}}n}\right)\end{aligned}$$

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

$$\mathbb{E}_{x_i, x_i'}[G^2] \leq 18C_{\mathrm{s}}n \left(\log \mathcal{N} + 6 + \frac{2}{C_{\mathrm{s}} \log \mathcal{N}}\right). \tag{30}$$

Now, we combine (28), (29), (30), and $A^2 = \frac{2C_{\mathrm{s}} \log \mathcal{N}}{9n}$ to obtain

$$\begin{aligned}D &\leq \left(\frac{1}{2}R(\hat{\ell}, \ell^\circ) + \frac{1}{2}A^2 + \frac{1}{2}\delta\right) + \frac{4C_{\mathrm{s}}}{n}\left(\log N + 6 + \frac{2}{C_{\mathrm{s}} \log \mathcal{N}}\right) + \delta \\ &\leq \frac{1}{2}R(\hat{\ell}, \ell^\circ) + \frac{C_{\mathrm{s}}}{n}\left(\frac{37}{9}\log \mathcal{N} + 32\right) + \frac{3}{2}\delta,\end{aligned}$$

where we have used that $\log \mathcal{N} \geq 1$. Therefore, we obtain

$$R(\hat{\ell}, \ell^\circ) \leq 2\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\ell}(x_i) - \ell^\circ(x_i))\right] + \frac{2C_{\mathrm{s}}}{n}\left(\frac{37}{9}\log \mathcal{N} + 32\right) + 3\delta. \tag{31}$$

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

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

RHS is minimized as $\inf_{\ell \in \mathcal{L}} \mathbb{E}_x[\ell(x) - \ell^\circ(x)]$. Finally, combining this with eq. (31), we obtain

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

According to Lemma D.1, we have

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

□

We evaluate the covering number of $\mathcal{L}$ as follows.

Lemma D.3 (Covering number of $\mathcal{L}$)**.** *We assume $\mathcal{L}$ is made from the network class in Theorem C.16. Assume that $s(x,t)$ is uniformly bounded by $\lesssim \sigma_t^{-1} \log^{\frac{1}{2}} N$ for all $s \in \mathcal{S}$. The covering number of $\mathcal{L}$ is evaluated by*

$$\log \mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty([-1,1]^d)}, \delta) \lesssim \log \mathcal{N}(\mathcal{S}, \|\cdot\|_{L^\infty([-C_1,C_1]^d)}, \delta N^{-1})$$

with $C_1 = \tilde{O}(1)$.

Proof. We take $C_1 = \mathcal{O}(\sigma_t\sqrt{\log \sigma_t^{-1}\delta^{-1}})$ that satisfies $\int_{y:\ \|y-m_t x\|_\infty \geq C_1} p(y|x)\mathrm{d}y \lesssim \sigma_t^2\delta$, $\int_{y:\ \|y-m_t x\|_\infty \geq C_1} \|\nabla \log p_t(y|x)\|^2 p(y|x)\mathrm{d}y \lesssim \tilde{\mathcal{O}}(\delta)$, and if $\|y\|_\infty \leq C_1$ then $\nabla \log p_t(y|x) = \mathcal{O}(\sqrt{\log \delta^{-1}}/\sigma_t)$ holds. Also note that $s(x,t)$ is bounded by $\lesssim \sigma_t^{-1}\log^{\frac{1}{2}} N$.

Let s_1 and s_2 satisfies $\|\max_i |[s_1 - s_2]_i|\|_{L^\infty([-C,C]^d)} \leq \delta$. Then we have that

$$
\begin{aligned}
&|\ell_1(x) - \ell_2(x)| \\
&= \int_{t=T_1}^{T_5}\int_y \max_i |[s_1(y,t) - s_2(y,t)]_i| \max_i |[s_1(y,t) + s_2(y,t) + \nabla \log p_t(y|x)]_i| p_t(y|x)\mathrm{d}y\mathrm{d}t \\
&= \int_{t=T_1}^{T_5}\int_{y:\ \|y-m_t\|_\infty \leq C} \\
&\quad \max_i |[s_1(y,t) - s_2(y,t)]_i| \max_i |[s_1(y,t) + s_2(y,t) + \nabla \log p_t(y|x)]_i| p_t(y|x)\mathrm{d}y\mathrm{d}t \qquad (32) \\
&\quad + \int_{t=T_1}^{T_5}\int_{y:\ \|y-m_t\|_\infty \geq C} \\
&\quad \max_i |[s_1(y,t) - s_2(y,t)]_i| \max_i |[s_1(y,t) + s_2(y,t) + \nabla \log p_t(y|x)]_i| p_t(y|x)\mathrm{d}y \qquad (33)
\end{aligned}
$$

(32) is bounded by $\lesssim C^d\delta(T_5 - T_1)(\log^{\frac{1}{2}} N + \log^{\frac{1}{2}} \delta^{-1})\log T_1$, because of $|s(x,t)| \lesssim \sigma_t^{-1}\log N = \mathcal{O}(t^{-\frac{1}{2}}\log^{\frac{1}{2}} N)$ and $\nabla \log p_t(y|x) = \mathcal{O}(\sqrt{\log \delta^{-1}}/\sigma_t)$.

For (33), $s(x,t) \lesssim \sigma_t^{-1}\log N$, $\int_{y:\ \|y-m_t x\|_\infty \geq C} p(y|x)\mathrm{d}y \lesssim \sigma_t^2\delta$, and $\int_{y:\ \|y-m_t x\|_\infty \geq C} \|\nabla \log p_t(y|x)\|^2 p(y|x)\mathrm{d}y \lesssim \tilde{\mathcal{O}}(\delta)$ imply the bound of $\tilde{\mathcal{O}}(\delta)$.

Therefore, using a $\delta/\tilde{\mathcal{O}}(1)$-net for $\mathcal{S}$ with respect to $L^\infty([-C_1, C_1]^d)$, we obtain a δ-net for $\mathcal{L}$. Note that because logarithmic terms hidden in $\tilde{\mathcal{O}}(1)$ is bounded with $N\delta^{-1}$, we need to take take δ/N-net of $\mathcal{S}$ to obtain the δ-net in $\mathcal{L}$. □

Combining eq. (25) and Lemma D.3, we obtain Theorem 3.3. Moreover, $\log \mathcal{N}(\mathcal{S}, \|\cdot\|_{L^\infty([-C_1,C_1]^d)}, \delta N^{-1})$ is further bounded as follows.

Lemma D.4. *For $\mathcal{S} \subset \Phi(L, W, S, B)$, we can bound the covering number as*

$$
\log \mathcal{N}(\mathcal{S}, \|\cdot\|_{L^\infty([-C_1,C_1]^d)}, \delta N^{-1}) \lesssim SL\log(\delta^{-1}L\|W\|_\infty(B \vee 1)C_1).
$$

Proof. Lemma 3 of Suzuki (2018), by changing the domain $[0,1]^d$ to $[-C_1, C_1]^d$, yields the result. □

Therefore, together with the bound of Theorem 3.3, we have that

$$
\begin{aligned}
&\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{x_t \sim p_t}[\|\hat{s}(x_t,t) - \nabla \log p_t(x_t)\|^2]\mathrm{d}t\right] \\
&\lesssim N^{-\frac{2s}{d}}\log^{\frac{2d+16s}{d}}(N) + \frac{\log^2(n)N(\log^8(n) + \log^4(n)\log(\varepsilon^{-1}))}{n} + \varepsilon.
\end{aligned}
$$

By letting $N = \delta = n^{-d/(2s+d)}$ yields that

$$
\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{x_t \sim p_t}[\|\hat{s}(x_t,t) - \nabla \log p_t(x_t)\|^2]\mathrm{d}t\right] \lesssim n^{-\frac{2s}{d+2s}}\log^{\frac{10d+16s}{d}}(n). \qquad (34)
$$

D.1 Sampling t and x_t instead of taking expectation

Because our main interest lies in the sample complexity, and for simple presentation, we basically do not care about the expectation in $\ell(x) = \int_{t=\underline{T}}^{\overline{T}}\int_x \|s(x,t) - \nabla \log p_t(y|x)\|_2^2 p_t(y|x)\mathrm{d}y\mathrm{d}t$. However, here we also introduce two ways to replace the expectation by a sample of t and x_t.

Approximation via polynomial-size sample 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 $x_j \sim p_{t_j}(x_j|x_{0,i})$. Then we consider the difference between

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

Using that $s(x_j, t_j)$ and $\nabla p_{t_j}(x_j|x_{0,i})$ are bounded by $\mathcal{O}(\frac{\log n}{\sigma_t})$ with high probability, we can use the standard generalization bounds for the squared-loss such as in Schmidt-Hieber (2020); Hayakawa & Suzuki (2020) to bound the difference by $\tilde{O}(\frac{n^{-\frac{2s}{d+2s}}}{M\underline{T}})$.

Modifying the distribution of t In this case, it suffices to consider only one path for each sample $x_{0,i}$. We sample t from $\mu(t) \propto \frac{\mathbb{1}[\underline{T} \le t \le \overline{T}]}{t}$ and modify $\lambda(t)$ as $\lambda(t) = t$. This score matching loss is equivalent to that of $t \sim \mathrm{Unif}(\underline{T}, \overline{T})$ and $\lambda(t) \equiv 1$, up to a logarithmic factor. Moreover, when we sample $(x_{0,i}, t_i, x_{t_j,i})$ from $x_{0,i} \sim p_0$, $t_i \sim \mu(t)$, and $x_{t_i,i} \sim p_{t_i}(x_{t_i,i}|x_{0,i})$, with high probability

$$\lambda(t_i)\|s(x_{t_i,i}, t_i) - \nabla p_{t_i}(x_{t_i,i}|x_{0,i})\|^2 = \tilde{\mathcal{O}}(1).$$

Therefore, following the standard generalization bounds (Schmidt-Hieber, 2020; Hayakawa & Suzuki, 2020), we can bound the generalization error by $\tilde{\mathcal{O}}(n^{-\frac{2s}{d+2s}})$.

E ESTIMATION ERROR ANALYSIS

This section evaluates the estimation rates, or the goodness of diffusion modeling as a density estimator. As a small modification, if $\|\hat{Y}_{\overline{T}-\underline{T}}\|_\infty \ge 2$, then we reset it to $\hat{Y}_{\overline{T}-\underline{T}} = 0$.

E.1 ESTIMATION RATE IN THE TV DISTANCE

First, we consider the bound in the total variation distance in the same manner as Song & Ermon (2019); Chen et al. (2022). The estimation error in TV is decomposed as

$$\begin{aligned}
\mathbb{E}[\mathrm{TV}(\mathrm{X}_0, \hat{\mathrm{Y}}_{\overline{\mathrm{T}}-\underline{\mathrm{T}}})] &\lesssim \mathbb{E}[\mathrm{TV}(\mathrm{Y}_{\overline{\mathrm{T}}}, \mathrm{Y}_{\overline{\mathrm{T}}-\underline{\mathrm{T}}})] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] \\
&\lesssim \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{\mathrm{T}}}) + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0)] \\
&= \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{\mathrm{T}}}) + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d))] \qquad (35)
\end{aligned}$$

Here $\bar{Y}$ is $\hat{Y}$ with replacing $\hat{Y}_0 \sim \mathcal{N}(0, I_d)$ by $\bar{Y}_0 \sim p_t$. The first term is bounded by $\sqrt{\underline{T}}n^{\mathcal{O}(1)}$ according to Theorem E.3, the second term is bounded by $\exp(-\overline{T})$ due to Lemma E.4, and the final term is bounded by

$$\begin{aligned}
&\mathbb{E}_{x_i}\sqrt{\left[\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_x\left[\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right]} \\
&\lesssim n^{-\frac{s}{d+2s}} \log^{\frac{5d+8s}{d}}(n)
\end{aligned}$$

from (34) in the previous section and Proposition E.2. Therefore, by appropriately choosing $\underline{T}$ and $\overline{T}$, we obtain the first bound (3) of Theorem 3.3. On the other hand, estimation of the Besov space has the following lower bound.

Theorem E.1. *Assume that* $0 < p, q \le \infty$, $s > 0$, *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

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

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

Therefore, we have proved that diffusion modeling achieves the minimax estimation rate in in TV up to the logarithmic factor for the Besov space $B^s_{p,q}$.

Now we move to the main part of the proof. The following Girsanov theorem is useful when converting the error of the score matching to the estimation error.

Proposition E.2 (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*

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

We denote the path measures of Z and Z' as $\mathbb{P}$ and $\mathbb{P}'$.

Suppose the following Novikov's condition:

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

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

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

We then bound the three terms of (35) in a row.

Theorem E.3. *We have that*

$$\mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{T}}) \lesssim \sqrt{\underline{T}}n^{\mathcal{O}(1)} \lesssim n^{-s/(d+2s)}$$

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

Proof. We need to evaluate the difference $\mathbb{E}[\mathrm{TV}(X_0, X_{\underline{T}})]$. Remember that p_0 is decomposed as

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

in Lemma G.15, 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 $\|p_0 - f_N\| \lesssim N^{-s/d} \simeq n^{-s/(2s+d)}$ 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.

First note that at each x,

$$\begin{aligned} p_{\underline{T}}(x) &= \int \frac{p_0(y)}{\sigma^d_{\underline{T}}(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x - m_{\underline{T}}y\|^2}{2\sigma^2_{\underline{T}}}\right)\mathrm{d}y \qquad (36) \\ &\simeq \int_{A_x} \frac{p_0(y)}{\sigma^d_{\underline{T}}(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x - m_{\underline{T}}y\|^2}{2\sigma^2_{\underline{T}}}\right)\mathrm{d}y \pm n^{-s/(d+2s)}, \end{aligned}$$

where $A^x = \prod_{i=1}^d a^x_i$ with $a^x_i = [\frac{x_i}{m_{\underline{T}}} - \frac{\sigma_{\underline{T}}\mathcal{O}(1)}{m_{\underline{T}}}\sqrt{\log n}, \frac{x_i}{m_{\underline{T}}} + \frac{\sigma_{\underline{T}}\mathcal{O}(1)}{m_{\underline{T}}}\sqrt{\log n}]$. 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 $x \in [-m_{\underline{T}} - \mathcal{O}(\sqrt{\underline{T}\log n}), m_{\underline{T}} + \mathcal{O}(\sqrt{\underline{T}\log n})]^d \supset [-1-\mathcal{O}(\sqrt{\underline{T}\log n}), 1+\mathcal{O}(\sqrt{\underline{T}\log n})]^d$,

$$\begin{aligned}
&\int_{A_x} \frac{p_0(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \\
&\simeq \int_{A_x} \frac{p_0(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \pm \mathcal{O}(\sqrt{\underline{T}\log n}) \cdot n^{\mathcal{O}(1)}
\end{aligned} \tag{37}$$

By taking $\underline{T}$ polynomially small w.r.t. n, we have that $\mathcal{O}(\sqrt{\underline{T}\log n}) \cdot n^{\mathcal{O}(1)} \lesssim n^{-s/(d+2s)}$. Moreover,

$$\begin{aligned}
&\int_{A_x} \frac{p_0(x)}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \simeq \int \frac{p_0(x)}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&\simeq p_0(x)\left(1 - \frac{1}{m_t^d}\right) = p_0(x) - \mathcal{O}(\underline{T}),
\end{aligned} \tag{38}$$

because of $m_{\underline{T}}^d = (1 - \mathcal{O}(\underline{T}))^d = 1 - \mathcal{O}(\underline{T})$.

On the other hand, $\|p_0 - f_N\|_1 \lesssim N^{-s/d} \simeq n^{-s/(d+2s)}$ holds. Therefore, combining this with (36), (37), and (38),

$$\|p_0 - p_{\underline{T}}\|_1 \lesssim \sqrt{\underline{T}} n^{\mathcal{O}(1)} \lesssim n^{-s/(d+2s)}.$$

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

The second term $\mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})]$ in (35) is bounded by Girsanov's theorem Proposition E.2 and (34) from appendix D:

$$\begin{aligned}
\mathbb{E}_{x_i} \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}}) &\lesssim \mathbb{E}_{x_i} \sqrt{\int_{t=\underline{T}}^{\overline{T}} p_t(x) \beta_t^2 \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t} \\
&\lesssim \sqrt{\mathbb{E}_{x_i} \int_{t=\underline{T}}^{\overline{T}} p_t(x) \beta_t^2 \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t} \\
&\lesssim \sqrt{n^{-\frac{2s}{d+2s}} \log^{\frac{10d+16s}{d}} n} \\
&\lesssim n^{-\frac{s}{d+2s}} \log^{\frac{10d+16s}{2d}} n.
\end{aligned}$$

Lemma E.4. *We can bound* $\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d))$ *as follows.*

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

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

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

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

Therefore, all three terms in (35) are bounded as above and Theorem 3.3 follows.

We conclude this subsection with the proof of Theorem E.1.

Proof of Theorem E.1. Theorem 10 of Triebel (2011) showed that, for a bounded domain $\Omega \subset \mathbb{R}^d$,

$$\log N(U(B_{p,q}^s(\Omega)), \|\cdot\|_r, \varepsilon) \simeq \varepsilon^{-d/s}, \tag{39}$$

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 (39) is transferred to the entropy. The condition 2 of the theorem is checked directly from eq. (39). 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 ε_n is chosen as $\log N(U(B^s_{p,q}(\Omega)), \|\cdot\|_r, \varepsilon_n) = n\varepsilon_n^2$ holds. Together with (39), we obtain the assertion. □

E.2 Estimation rate in the W_1 distance

We also consider the estimation rate in W_1. Because both X_0 and $\hat{Y}_{\overline{T}-\underline{T}}$ have bounded supports, Theorem 3.3 directly yields the convergence rate of $n^{-s/(2s+d)}$. However, it is known from Niles-Weed & Berthet (2022) that the minimax estimation rate in W_1 is faster than this.

Proposition E.5 (Niles-Weed & Berthet (2022))**.** *Let $p, q \geq 1$, $s > 0$ and $d \geq 2$.*

$$\int_{\hat{\mu}} \sup_{p \in B^s_{p,q}} \mathbb{E}[W_1(\hat{\mu}, p)] \gtrsim n^{-(s+1)/(2s+d)},$$

where $\hat{\mu}$ runs over all estimators based on n observations. Moreover, if $1 \leq p < \infty$, $1 \leq q \leq \infty$, $s > 0$, and $d \geq 3$, there exists an estimator $\hat{\mu}_*$ that achieves this minimax rate.

Then are diffusion models sub-optimal in W_1? In the following, we show the surprising fact that diffusion modeling also achieves this nearly minimax optimal rate, if some modification applied.

Theorem E.6. *For any fixed $\delta > 0$, we can train the score network with $n(\gg 1)$ sample so that*

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

This corresponds to the second part (4) of Theorem 3.3.

Switching score networks To prove Theorem E.6, let us carefully consider where we lose the estimation rate. First, remember the approximation error analysis. Although we used Theorem 3.1 for all $\underline{T} \leq t \leq \overline{T}$, Lemma C.5 tells us that if $t \geq 2T_* = 2N^{-\frac{2-\delta}{d}}$ ($N = n^{-\frac{d}{2s+d}}$ is applied), we can make the approximation error smaller than $N^{-\frac{2(s+1)}{d}}\sigma_t^{-2} = n^{-\frac{2(s+1)}{d+2s}}\sigma_t^{-2}$. Moreover, Lemma C.5 also tells us that we can use sparser networks for $t \geq 2T_*$, which makes the generalization error much smaller.

Based on these discussion, we divide the time $[T_*, \overline{T}]$ into $T_* = t_0 < t_1 < \cdots < t_{K_*+1} > \overline{T}$ with $t_{i+1}/t_i = 2$. And we modify $t_{K_*+1} = \overline{T}$. The number of intervals amounts to $K_* = \mathcal{O}(\log n)$. Lemma C.5 yields that for $i \geq 1$ these exists a network $s_i \in \Phi(L_i, W_i, S_i, W_i)$ such that

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

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

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

Similarly to (34) of appendix D, Theorem D.2 and Lemmas D.3 and D.4 yields that the following generalization error bound for $i \geq 1$:

$$\mathbb{E}_{x_i}\left[\int_{t=t_i}^{t_{i+1}} \mathbb{E}_x\left[\|\hat{s}_i(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right] \lesssim n^{-\frac{2(s+1)}{d+2s}} + \frac{t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}} \log^{10} n}{n}. \tag{40}$$

Because $t_i \geq 2T_* = 2n^{-(2-\delta)/d}$, this gives the tighter bound on the error for $t \leq 2T_*$.

One may think that this improvement would be useless because the error caused at $t \leq 2T_*$ has still the $n^{-2s/(d+2s)}$ rate and dominates the error. However, another important observation is that the Wasserstain distance is a transportation distance, and therefore the estimation error caused at $t \approx 0$ less contributes to the error and that at $t \approx 1$ more because the time determines how much the mass goes further from the correct path.

To formalize this observation, let us consider a sequence of stochastic processes. First, $\hat{Y}^{(0)} = Y = (Y_t)_{t\in[0,\overline{T}]}$ is defined as a process such that

$$\mathrm{d}Y_t = \beta_{\overline{T}-t}(Y_t + 2\nabla \log p_t(Y_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [0,\overline{T}]), \quad Y_0^{(1)} \sim p_{\overline{T}}.$$

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

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

Here $\hat{Y}_{\overline{T}-t}^{(i)} \sim p_t$ 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. And we modify $\hat{Y}_{\overline{T}-\underline{T}}^{(i)} := 0$ if $\|\hat{Y}_{\overline{T}-\underline{T}}^{(i)}\|_\infty \geq 2$ in the original definition.

Finally, $\hat{Y} = (\hat{Y}_t)_{t\in[0,\overline{T}-\underline{T}]}$ is defined, which corresponds to the actual diffusion process.

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

And we modify $\hat{Y}_{\overline{T}-\underline{T}} := 0$ if $\|\mathrm{d}\hat{Y}_{\overline{T}-\underline{T}}\|_\infty \geq 2$ in the original definition.

We bound $W_1(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}}) = W_1(X_0, X_{\underline{T}}), W_1(\hat{Y}_{\overline{T}-\underline{T}}^{(i-1)}, \hat{Y}_{\overline{T}-\underline{T}}^{(i)})$ $(i = 1, 2, \cdots, k)$, and $W_1(\hat{Y}_{\overline{T}-\underline{T}}^{(K_*)}, \hat{Y}_{\overline{T}-\underline{T}}) \lesssim \mathrm{TV}(X_{\overline{T}}, \hat{Y}_0)$ in order.

Lemma E.7. *We can bound $W_1(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}})$ as follows.*

$$W_1(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}}) \lesssim \sqrt{\overline{\beta}\underline{T}d}$$

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

$$W_1(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}}) \leq \mathbb{E}[\|X - m_{T_1}X + \sigma_{T_1}Z\|] \leq (1-m_{\underline{T}})\mathbb{E}[\|X\|] + \sigma_{\underline{T}}\mathbb{E}[\|Z\|]$$
$$\leq (1-m_{\underline{T}})\sqrt{d} + \sigma_{\underline{T}}\sqrt{d} \lesssim \sqrt{\overline{\beta}\underline{T}}\sqrt{d},$$

which concludes the proof. □

Lemma E.8. *Suppose that $\hat{s}(x,t) \leq \frac{C_s}{\sqrt{t}\wedge 1}$ holds for all x and t. Then, the following holds for all $i = 2, 3, \cdots, K_*$:*

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

Also,

$$W_1(\hat{Y}^{(0)}_{\overline{T}-\underline{T}}, \hat{Y}^{(1)}_{\overline{T}-\underline{T}}) \lesssim C_s\sqrt{\underline{T}}\log \underline{T}^{-1}\sqrt{\int_{t=\underline{T}}^{t_1}\int_x p_t(x)\beta_t^2\|(\hat{s}-s^\circ)(x,t)\|^2 \mathrm{d}x\mathrm{d}t} + n^{-\frac{(s+1)}{d+2s}}\log^2 n + \underline{T}\log^{(d-2)/2}\underline{T}^{-1}.$$

Proof. We only consider the first assertion, because the latter can be derived in the same way. Until the final part of the proof, we ignore the modification, that is, $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ is modified to 0 due to $\|\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}\|_\infty \geq 2$ in the original definition, and we consider the original path. Also, if $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$ is modified to 0, we consider the original path of it as well.

We construct the transportation between $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ and $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$. Our approach focuses on each path. Specifically, let us write the path measures of $\hat{Y}^{(i-1)}$ and $\hat{Y}^{(i)}$ be $\mathbb{P}$ and $\mathbb{P}'$, and take some path p with $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}} = y$ and $\hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}} = z$. If $\mathrm{d}\mathbb{P}[p] > \mathrm{d}\mathbb{P}'[p]$, then we move the mass of $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ corresponding to p, that amounts to $\mathrm{d}\mathbb{P}[p] - \mathrm{d}\mathbb{P}'[p]$, to z. Applying this to all paths p, then the total mass of $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ that is moved by is at most

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

according to Proposition E.2 and Pinsker's inequality.

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 $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$. That is, for each path p with $\hat{Y}^{(i)}_{\overline{T}-\underline{T}} = y$ and $\hat{Y}^{(i)}_{\overline{T}-t_i} = z$, if $\mathrm{d}\mathbb{P}[p] < \mathrm{d}\mathbb{P}'[p]$, then we move the mass of $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$ corresponding to p, as much as $\mathrm{d}\mathbb{P}'[p] - \mathrm{d}\mathbb{P}[p]$, to z. This affects to the mass of $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$, amounting to $\frac{1}{2}\mathrm{TV}(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \hat{Y}^{(i)}_{\overline{T}-\underline{T}})$.

Now, we can see that, to each value of $\hat{Y}_{\overline{T}-t_i} = z$, the same amount of mass is transported from both $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ and $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$. Thus, at each z, we arbitrarily associate the mass from $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ to that from $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$. Using this, as much as $\frac{1}{2}\mathrm{TV}(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \hat{Y}^{(i)}_{\overline{T}-\underline{T}})$ of the mass is transported from $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}$ to $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$, along each corresponding path.

Now our interest is how far each transport is required to move on average. First we bound $\hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}}$ and $\hat{Y}^{(i)}_{\overline{T}-t_i}$ According to Lemma C.9, we can bound them with probability at least $1-\mathcal{O}(\underline{T_1})$ by $\mathcal{O}(\log^{\frac{1}{2}}\underline{T_1}^{-1})$.

Then we consider $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \hat{Y}^{(i-1)}_{\overline{T}-t_i} = (\hat{Y}^{(i-1)}_{\overline{T}-t_i} - \hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}}) + (\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}})$. Roughly speaking, $\|(\hat{Y}^{(i-1)}_{\overline{T}-t_i} - \hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}})\|$ is bounded by $\tilde{\mathcal{O}}(\sqrt{t_i})$ with high probability. Note that $\hat{Y}^{(i-1)}_{\overline{T}-t} = m_t\hat{Y}^{(i-1)}_{\overline{T}} + \sigma_t Z$ with $Z \sim \mathcal{N}(0, I_d)$, so $\hat{Y}^{(i-1)}_{\overline{T}-t_i} - \hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}}$ is decomposed into the drift term and the diffusion term. The drift term is $(m_{\overline{T}-t_i} - m_{\overline{T}-t_{i-1}})\hat{Y}^{(i-1)}_{\overline{T}}$, which is bounded by $\mathcal{O}(t_i) \lesssim \mathcal{O}(\sqrt{t_i})$. We then consider the diffusion term. According to Lemma G.16, $\sigma_{t_i}Z$ and $\sigma_{t_{i-1}}Z'$ with $Z, Z' \sim \mathcal{N}(0, I_d)$ are bounded by $\mathcal{O}(\sqrt{t_i}\log^{\frac{1}{2}}\underline{T}^{-1})$, with probability at least $1-\underline{T}$. Therefore, the probability where $(\hat{Y}^{(i-1)}_{\overline{T}-t_i} - \hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}})$ moves more than $\mathcal{O}(\sqrt{t_i\log^{\frac{1}{2}}\underline{T}^{-1}})$ is bounded by $\underline{T}$. Next we consider $(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}})$. The effect of the drift term $\sup_{t\in[\overline{T}-t_{i-1},\overline{T}-\underline{T}]}\int_{\overline{T}-t_{i-1}}^{\overline{T}-\underline{T}}\sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t$ is bounded with $\mathcal{O}(\sqrt{t_i}\log^{\frac{1}{2}}\underline{T}^{-1})$ with probability at least $1-\mathcal{O}(\underline{T})$ according to Lemma G.17. Conditioning by this event, the drift term is bounded with $\mathcal{O}(\frac{C_s}{\sigma_{\overline{T}-t}})$ at each t. Therefore, the drift term

causes the move at most

$$\int_{t=\overline{T}-t_{i-1}}^{\overline{T}-t} \frac{C_s}{\sigma_{\overline{T}-t}} \mathrm{d}t \lesssim C_s \sqrt{t_{i-1}}.$$

In summary, $\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \hat{Y}^{(i-1)}_{\overline{T}-t_i}$ is bounded by $\mathcal{O}(C_s\sqrt{t_i}\log \underline{T}^{-1})$ with probability at least $1-\mathcal{O}(\underline{T})$. Also, $\hat{Y}^{(i)}_{\overline{T}-\underline{T}} - \hat{Y}^{(i)}_{\overline{T}-t_i}$ is bounded in the same way we used for $(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \hat{Y}^{(i-1)}_{\overline{T}-t_{i-1}})$. Therefore, with probability at least $1-\mathcal{O}(\underline{T})$ the mass moves at most $\mathcal{O}(C_s\sqrt{t_i}\log \underline{T}^{-1})$.

Finally we consider the effect of the modification moving $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$ with its $\inf$-norm more than 2 to 0. At $t = \overline{T} - 2T_*$, according to Proposition E.2, we can bound that

$$\mathrm{TV}(Y_{\overline{T}-2T_*}, \hat{Y}^{(i)}_{\overline{T}-2T_*}) \lesssim \log \underline{T}^{-1} \sqrt{\int_{t=2T_*}^{t_i} \int_x p_t(x)\beta_t^2 \|(\hat{s}-s^\circ)(x,t)\|^2 \mathrm{d}x\mathrm{d}t} \lesssim n^{-\frac{(2s+2)}{d+2s}} \log^2 n.$$

Moreover, according to Lemma C.9, $\int_{\|x\|_\infty \geq m_{2T_*}+\sigma_{2T_*}\sqrt{2\log \underline{T}^{-1}}} p_t(x)\mathrm{d}x \lesssim \underline{T}\log^{(d-2)/2}\underline{T}^{-1}$. Therefore, with probability at least $1 - n^{-\frac{(s+1)}{d+2s}}\log^2 n - \underline{T}\log^{(d-2)/2}\underline{T}^{-1}$, $\|\hat{Y}^{(i)}_{\overline{T}-2T_*}\| \leq m_{2T_*} + \sigma_{2T_*}\sqrt{2\log \underline{T}^{-1}} \leq 1.5$ (when T_* is sufficiently small). As we did previously, $\hat{Y}^{(i)}_{\overline{T}-\underline{T}} - \hat{Y}^{(i)}_{\overline{T}-2T_*}$ is bounded by $\mathcal{O}(C_s\sqrt{2T_*}\log \underline{T}^{-1})$ with probability at least $1-\mathcal{O}(\underline{T})$, and when T_* is sufficiently small, $\mathcal{O}(C_s\sqrt{2T_*}\log \underline{T}^{-1}) = o(1)$. Therefore, with probability at least $1 - n^{-\frac{(s+1)}{d+2s}}\log^2 n - \underline{T}\log^{(d-2)/2}\underline{T}^{-1}$, $\|\hat{Y}^{(i)}_{\overline{T}-\underline{T}}\|_\infty \leq 2$ holds. In summary, after modification, the mass moved more than $\mathcal{O}(C_s\sqrt{t_i}\log \underline{T}^{-1})$ is bounded by $1 - n^{-\frac{(s+1)}{d+2s}}\log^2 n - \underline{T}\log^{(d-2)/2}\underline{T}^{-1}$, which yields the conclusion.

□

Lemma E.9. *We can bound $W_1(\hat{Y}^{(K_*)}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})$ as follows.*

$$W_1(\hat{Y}^{(K_*)}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}}) \lesssim \mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) \lesssim \exp(-\overline{T}).$$

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

$$\mathrm{TV}(p_{\overline{T}}, \mathcal{N}(0, I_d)) \leq \sqrt{2\mathrm{KL}(p_{\overline{T}}\|\mathcal{N}(0,I_d))} \leq 2\exp(-\overline{T}\sqrt{\mathrm{KL}(p_0\|\mathcal{N}(0,I_d))} \lesssim \exp(-\overline{T}),$$

because $C_f^{-1} \leq p_0 \leq C_f$ holds and the density of $\mathcal{N}(0, I_d)$ is lower bounded by $\mathcal{O}(1)$ in $\mathrm{supp}(p_0) = [-1,1]^d$, which means $\mathrm{KL}(p_0\|\mathcal{N}(0,I_d)) = \mathcal{O}(1)$. Therefore, because $\|\hat{Y}^{(k)}_{\overline{T}-\underline{T}}\|_\infty, \|\hat{Y}_{\overline{T}-\underline{T}}\|_\infty \leq 2 = \mathcal{O}(1)$, and $\hat{Y}^{(k)}$ and $\hat{Y}$ evolves in the same way over $[0, \overline{T}-\underline{T}]$, we have

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

which yields the assertion. □

Finally, we sum up the errors.

$$\begin{aligned} W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}}) &\lesssim W_1(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}}) + \sum_{i=1}^{K_*} W_1(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \hat{Y}^{(i)}_{\overline{T}-\underline{T}}) + W_1(\hat{Y}^{(K_*)}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}}) \\ &\lesssim \sqrt{\underline{T}} + \sum_{i=1}^{K_*} W_1(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \hat{Y}^{(i)}_{\overline{T}-\underline{T}}) + \exp(-\overline{T}). \end{aligned}$$

The sum is bounded by Lemma E.8 ($C_s = \log^{\frac{1}{2}} n$ from Theorem 3.1) together with (34) and eq. (40) as

$$
\begin{aligned}
&\sum_{i=1}^{K_*} W_1(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \hat{Y}^{(i)}_{\overline{T}-\underline{T}}) \\
&\lesssim \log^{\frac{3}{2}} n \sum_{i=2}^{K_*} \sqrt{t_i \left(n^{-\frac{2(s+1)}{d+2s}} + \frac{t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}} \log^{10} n}{n} \right)} + \log^{\frac{3}{2}}(n) \sqrt{T_* n^{-\frac{2s}{d+2s}} \log^{\frac{10d+16s}{d}}(n)} \\
&\quad + K_* (n^{-\frac{(s+1)}{d+2s}} \log^2 n + \underline{T} \log^{(d-2)/2} \underline{T}^{-1}) \\
&\lesssim \left(\sum_{i=2}^{k} t_i^{\frac{1}{2}} n^{-\frac{(s+1)}{d+2s}} + \frac{t_i^{-d/4+1/2} n^{\frac{\delta d}{2(d+2s)}} \log^5 n}{\sqrt{n}} \right) + n^{-\frac{(s+1)}{d+2s}} \log^3 n \\
&\quad + n^{-\frac{s+1-\delta/2}{d+2s}} \log^{\frac{13d+16s}{2d}}(n) + \underline{T} \log n \log^{(d-2)/2} \underline{T}^{-1} \\
&\lesssim n^{-\frac{(s+1)}{d+2s}} \log^{\frac{3}{2}} n + \frac{T_*^{-d/4+1/2} n^{\frac{\delta d}{2(d+2s)}} \log^5 n}{\sqrt{n}} + n^{-\frac{s+1-\delta/2}{d+2s}} \log^{\frac{13d+16s}{2d}}(n) + \underline{T} \log n \log^{(d-2)/2} \underline{T}^{-1} \\
&\lesssim \frac{(n^{-\frac{2-\delta}{d+2s}})^{-d/4+1/2} n^{\frac{\delta d}{2(d+2s)}} \log^5 n}{\sqrt{n}} + n^{-\frac{s+1-\delta/2}{d+2s}} \log^{\frac{13d+16s}{2d}}(n) + \underline{T} \log n \log^{(d-2)/2} \underline{T}^{-1} \\
&\lesssim n^{-\frac{s+1-\delta/2}{d+2s}} \log^5 n + n^{-\frac{s+1-\delta/2}{d+2s}} \log^{\frac{13d+16s}{2d}}(n) + \underline{T} \log n \log^{(d-2)/2} \underline{T}^{-1} \\
&\lesssim n^{-\frac{(s+1-\delta)}{d+2s}} \quad \text{(by taking } \underline{T} \lesssim n^{-\frac{s+1}{d+2s}}\text{)}.
\end{aligned}
$$

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

$$
W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}}) \lesssim n^{-\frac{(s+1-\delta)}{d+2s}},
$$

which concludes the proof for Theorem E.6.

F Error analysis with intrinsic dimensionality

This section considers the case when the data distribution lies in the lower-dimensional plane. Let p_0 is a probability measure that is absolutely continuous with respect to the Lebesgue measure restricted on the sub-space V with a density upper and lower bounded by C_f and C_f^{-1}, respectively, on the support. Moreover we assume that the density belongs to $U(B^s_{p,q})$ as a function on the canonical coordinate system of the subspace V, and especially in the boundary region $([-1,1]^d \setminus [1-a_0, 1-a_0]^d) \cap V$ with $a_0 = n^{-\frac{1-\delta}{d'}}$, the density belongs to $U(\mathcal{C}^\infty)$ in the same sense. Then, we state our result as follows:

Theorem F.1. *For any fixed $\delta > 0$, we can train the score network with $n(\gg 1)$ sample so that*

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

Contrary to Theorem 3.3, the upper bound here depends on d' (not on d). Thus, we can conclude that the diffusion models also have the adaptivity: if the smaller intrinsic dimensionality is, the smaller estimation error will be.

F.1 Brief proof overview

Because the generalization error analysis of the score network and the final estimation rate (Theorem F.1) are derived by just replacing d by d' in the previous analysis, here we focus on the approximation error bound. In order to obtain Theorem F.1, 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 $V^\perp$). The first observation to this goal is

$$\begin{aligned} p_t(x) &= \int_{\bar{V}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) dy \\ &= \int_{\bar{V}} \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x_V - m_t y\|^2 + \|x_{V^\perp}\|^2}{2\sigma_t^2}\right) dy \\ &= \underbrace{\int_{\bar{V}} \frac{f(y)}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x_V - m_t y\|^2}{2\sigma_t^2}\right) dy}_{p_t^{(1)}(x)} \cdot \underbrace{\exp\left(-\frac{\|x_{V^\perp}\|^2}{2\sigma_t^2}\right)}_{p_t^{(2)}(x)}, \end{aligned}$$

where $\bar{V} := V \cap [-1,1]^d$ and $x =: x_V + x_{V^\perp}$ is the orthogonal decomposition of x with respect to the subspace V. $p_t^{(1)}(x)$ and $p_t^{(2)}(x)$ can be seen as the density function with respect to the intrinsic components and remaining space. Since

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

we only need to construct the neural networks approximating each term and concatenate them [1]. For approximating $\frac{\nabla p_t^{(1)}(x)}{p_t^{(1)}(x)}$, we can show that the neural networks only suffers from d', the intrinsic dimensionality. The other term can be easily approximated since $\frac{\nabla p_t^{(2)}(x)}{p_t^{(2)}(x)} = -\frac{1}{\sigma_t^2} x_{V^\perp}$, which is the simple linear function.

F.2 Proof of Theorem F.1

Now we move to the main part of the proof. Let $d' \leq d$ be an integer and $A \in \mathbb{R}^{d \times d'}$ be a column full-rank matrix. Then, we consider the d'-dimensional subspace $V := \{y \in \mathbb{R}^d \mid \exists x \in \mathbb{R}^{d'} \text{ s.t. } y = Ax\}$. Although we assume the subspace passing through the origin in $\mathbb{R}^{d'}$, we can expand the following results to the general subspace with small modification.

Assumption F.2. The true density is supported on $[-1,1]^d \cap V$, where it is upper and lower bounded by C_f and C_f^{-1}, respectively. Moreover, it belongs to $U(B^s_{p,q}([-1,1]^d \cap V)$.

Here $U(\cdot)$ means the unit ball with respect to the Besov norm. We note that the Besov space $B^s_{p,q}([-1,1]^d \cap U)$ can be defined by the same way as definition B.1.

Lemma F.3. *Let $\bar{V} := [-1,1]^d \cap V$. There exists a matrix $P_V \in \mathbb{R}^{d \times d}$ such that*

$$P_V \bar{V} := \{y \in \mathbb{R}^d \mid \exists x \in \bar{V} \text{ s.t. } y = P_V x\} = \{y \in \mathbb{R}^d \mid \exists z \in [-1,1]^{d'} \text{ s.t. } y = (z, 0, \ldots, 0)\}$$

holds. Moreover, we can take P_V to be the injective map from $\bar{V}$ to $\mathbb{R}^d$.

Proof. Since A is column full-rank, we can take the submatrix $\bar{A} \in \mathbb{R}^{d' \times d'}$ of A with $\text{rank}(A) = d'$. Without loss of generality, we assume A' coincides with the first d' rows of A. Then, we show that $P_V = \begin{pmatrix} I_{d'} & O \\ O & O \end{pmatrix}$ is the desired matrix.

($\supset$) follows from $\text{rank}(A) = d'$: for any $z \in [-1,1]^d$, there exists $x \in \mathbb{R}^{d'}$ such that $Ax = (z, 0, \ldots, 0)^T$.

($\subset$) is direct consequence from the definition of $\bar{V}$ and P_V.

Moreover, this P_V is injective as follows: for Ax_1, $Ax_2 \in \bar{V}$, we have $P_V(Ax_1 - Ax_2) = \begin{pmatrix} A' & O \\ O & O \end{pmatrix}(x_1 - x_2) \neq \mathbf{0}$ as long as $x_1 \neq x_2$. Thus, we get the assertion. □

[1] In the completed proof, we further decompose $p_t^{(1)}(x)$ into two parts. Therefore, this is different from what we take as $p_t^{(1)}(x)$ in the proof but we employ it for the simple explanation.

Then, we have that

$$p_t(x) = \int_{\bar{V}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) dy$$
$$= \int_{\bar{V}} \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x_V - m_t y\|^2 + \|x_{V^\perp}\|^2}{2\sigma_t^2}\right) dy,$$

where $x =: x_V + x_{V^\perp}$ is the orthogonal decomposition of x w.r.t. the subspace V. Moreover, this can be calculated as

$$\int_{\bar{V}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x_V - m_t y\|^2 + \|x_{V^\perp}\|^2}{2\sigma_t^2}\right) dy$$
$$= \int_{\bar{V}} \frac{1}{\sigma_t^{d'} (2\pi)^{\frac{d'}{2}}} f(y) \exp\left(-\frac{\|x_V - m_t y\|^2}{2\sigma_t^2}\right) dy \cdot \frac{1}{\sigma_t^{d-d'} (2\pi)^{\frac{d-d'}{2}}} \exp\left(-\frac{\|x_{V^\perp}\|^2}{2\sigma_t^2}\right)$$
$$= \int_{[-1,1]^{d'}} \frac{1}{\sigma_t^{d'} (2\pi)^{\frac{d'}{2}}} \bar{f}(z) \exp\left(-\frac{\|P_V x_V - m_t z\|^2}{2\sigma_t^2}\right) dz$$
$$\cdot \exp\left(-\frac{\|(I - P_V)x_V\|^2}{2\sigma_t^2}\right) \cdot \frac{1}{\sigma_t^{d-d'} (2\pi)^{\frac{d-d'}{2}}} \exp\left(-\frac{\|x_{V^\perp}\|^2}{2\sigma_t^2}\right),$$

where $\bar{f}(z) \coloneqq f((z, 0, \ldots, 0))$. The last equality follows from the transformation $z \mapsto y$. Thus, we can decompose p_t into the product of two ingredients: the principal component

$$p_t^{(1)}(x) \coloneqq \int_{[-1,1]^{d'}} \frac{1}{\sigma_t^{d'} (2\pi)^{\frac{d'}{2}}} \bar{f}(z) \exp\left(-\frac{\|P_V x_V - m_t z\|^2}{2\sigma_t^2}\right) dz$$

and the Gaussian component

$$p_t^{(2)}(x) \coloneqq \frac{1}{\sigma_t^{d-d'} (2\pi)^{\frac{d-d'}{2}}} \exp\left(-\frac{\|(I - P_V)x_V\|^2 + \|x_{V^\perp}\|^2}{2\sigma_t^2}\right).$$

Since it holds that

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

it is suffices to construct the neural networks such that each of them approximates $\frac{\nabla p_t^{(1)}(x)}{p_t^{(1)}(x)}$ and $\frac{\nabla p_t^{(2)}(x)}{p_t^{(2)}(x)}$ for approximating the score function.

Moreover, the straightforward calculation shows

$$\frac{\nabla p_t^{(2)}(x)}{p_t^{(2)}(x)} = -\frac{1}{\sigma_t^2}[(I - P_V)x_V + x_{V^\perp}].$$

According to the above observation, we construct $\phi_{score} \in \Phi(L, W, S, B)$ as follows:

1. At the first and second layer, the input x is decomposed into three ingredients: $P_V x_V$, $(I - P_V)x_V$ and $x_{V^\perp}$.
2. As for the $p^{(1)}(x)$, we can use the same technique as before, where the dependence on d is replaced with d'.
3. As for the Gaussian component $p^{(2)}(x)$, we can easily represent it.
4. By combining 1–3, we get the desired result.

Then, we move to state the approximation theorem. Here, we take $T_1 = \underline{T} = N^{-2(s+1)/d'}$,$T_2 = T_* = N^{-(2-\delta)/d'}$,$T_3 = 2T_2$,$T_4 = 3T_2$, and $1 < T_5 = \overline{T} = \mathcal{O}(\log N)$.

Theorem F.4 (Approximation theorem with lower dimension)**.** *There exists a neural network* $\phi^{d'}_{\text{score},1} \in \Phi(L, W, S, B)$ *that satisfies*

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

for $\underline{T} \le t \le T_*$ *with*

$$L = \mathcal{O}(\log^4 N), \quad \|W\|_\infty = \mathcal{O}(N\log^6 N), \quad S = \mathcal{O}(d^2 + N\log^8 N), \quad \textit{and } B = \exp(\mathcal{O}(\log^4 N)).$$

Moreover, for $N' \ge t_*^{-\frac{d'-\delta}{2}}$*, there exists a neural network* $\phi^{d'}_{\text{score},2} \in \Phi(L, W, S, B)$ *that satisfies*

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

for $T_3 \vee 2t_* \le 2 \le T_5$*, with*

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

Proof. According to the decomposition $\nabla \log p_t(x) = \frac{\nabla p_t^{(1)}(x)}{p_t^{(1)}(x)} + \frac{\nabla p_t^{(2)}(x)}{p_t^{(2)}(x)} =: s^{(1)}(x,t) + s^{(2)}(x,t)$, we aim to construct $\phi^{(1)}_{\text{score}}$ and $\phi^{(2)}_{\text{score}}$ to approximate each term. We consider the construction of $\phi^{d'}_{\text{score},1}$, because that of $\phi^{d'}_{\text{score},2}$ can be done in the same way, where the only difference is to use Lemma C.15 instead of Lemma C.13. Let U_1, U_2 be the projection matrix to V and $V^\perp$, respectively.

As for the first term, we consider the following neural network $\phi^{(1)}_{\text{score}} \colon \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$: for each fixed t,

$$\phi^{(1)}_{\text{score}}(\cdot, t) = \bar{\phi}_{\text{score}}(\cdot, t) \circ P_V \circ U_1,$$

where $\bar{\phi}_{\text{score}} \in \Phi(L, W, S, B) : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^D$ is what taken by Lemma C.13 with

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

satisfying

$$\int_x p_t(x)\|\bar{\phi}_{\text{score}}(P_V x_V, t) - s^{(1)}(x,t)\|^2 \mathrm{d}x\mathrm{d}t \lesssim \frac{N^{-2s/d'}\log N}{\sigma_t^2}$$

We note that $\bar{f}(\cdot)$ is also included in $B^s_{p,q}([-1,1]^{d'})$. Since the linear operators P_V and U_1 is described by

$$P_V x = \mathrm{ReLU}(P_V x) - \mathrm{ReLU}(-P_V x), \quad U_1 x = \mathrm{ReLU}(U_1 x) - \mathrm{ReLU}(-U_1 x), \tag{41}$$

we can conclude that $\phi^{(1)}_{\text{score}} \in \Phi(L_1', W_1', S_1', B_1')$ with

$$L_1' = \mathcal{O}(\log^4 N), \quad \|W_1'\|_\infty = \mathcal{O}(N\log^6 N), \quad S_1' = \mathcal{O}(N\log^8 N), \quad \text{and } B_1' = \exp(\mathcal{O}(\log^4 N))$$

(note that each component of the $U_1 \in \mathbb{R}^{d\times d}$ is included in $[-1,1]$). As for the second term, we consider the decomposition

$$s^{(2)}(t,x) = -\frac{1}{\sigma_t^2} I \circ (I - P_V) \circ U_1(x) - \frac{1}{\sigma_t^2} I \circ U_2(x)$$

The terms $(I - P_V) \circ U_1(x)$ and $U_2(x)$ are described by an element of $\Phi(2, (2d, 2d), \mathcal{O}(d^2), 1)$ in the same way as eq. (41). Moreover, the inverse function $-\frac{1}{x}$ can be approximate up to ϵ by an element of $\Phi(L, W, S, B)$ with $L = \mathcal{O}(\log^2 \epsilon^{-1})$, $\|W\|_\infty = \mathcal{O}(\log^3 \epsilon^{-1})$, $S = \mathcal{O}(\log^4 \epsilon^{-1})$, $B = \mathcal{O}(\epsilon^{-3})$ for any $x \in [\epsilon, \epsilon^{-1}]$ Telgarsky (2017); Boullé et al. (2020). By combining this fact with Lemma C.6 (approximation of σ_t) we can approximate the operator $-\frac{1}{\sigma_t^2} I$ up to $\frac{N^{-2(s+1)/d'}}{\sigma_t^2}$ by an element of $\Phi(L, W, S, B)$ with

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

Then, we can conclude that there exists $\phi_{\text{score}}^{(2)} \in \Phi(L_2', W_2', S_2', B_2')$ with

$L_2' = \mathcal{O}(\log^3 N), \|W_2'\|_\infty = \mathcal{O}(\max\{d, \log^3 N\}), S_2' = \mathcal{O}(d^2 + \log^4 N),$ and $B_2' = \exp(\mathcal{O}(\log^2 N))$

satisfying

$$\int_x p_t(x)\|\phi_{\text{score}}^{(2)}(x,t) - s^{(2)}(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-2s/d'}\log N}{\sigma_t^2}.$$

Finally, by concatenating $\phi_{\text{score}}^{(1)}$ and $\phi_{\text{score}}^{(2)}$ as $\phi_{\text{score},1}^{d'} \coloneqq \phi_{\text{score}}^{(1)} + \phi_{\text{score}}^{(2)}$, we obtain the desired result. □

Then, we obtained the counterparts of Lemma C.13 and Lemma C.15, where d' is replaced by d. The following argument to derive the estimation rate in W_1 is just replacing d with d' whenever d appears. Therefore, we finally obtained Theorem F.1.

G AUXILIARY LEMMAS

This final section summarizes existing results and prepares basic tools for the main parts of the proofs.

G.1 CONSTRUCTION OF A LARGER NEURAL NETWORK

Through construction of the desired neural network, we often need to combine sub-networks 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 G.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*

$$L = \sum_{i=1}^k L^i, \quad W \le 2\sum_{i=1}^k W^i, \quad S \le \sum_{i=1}^k S^i + \sum_{i=1}^{k-1}(\|A_{L^i}^i\|_0 + \|b_{L^i}^i\|_0 + \|A_1^{i+1}\|_0) \le 2\sum_{i=1}^k S^i,$$
$$\text{and } B \le \max_{1\le i\le k} B^i.$$

Here A_j^i is the parameter matrix and b_j^i is the bias vector at the jth layer of the ith neural network ϕ^i.

Next we introduce the identity function.

Lemma G.2 (Identity function (p.19 of Nakada & Imaizumi (2020))). *For $L \ge 2$ and $d \in \mathbb{N}$, there exists a neural network $\phi_{\text{Id}}^{d,L} \in \Phi(L, W, S, B)$ 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 $\phi_{\text{Id}}^{d,1} \in \Phi(1, (d), d, 1)$ with parameters $(A_1, b_1) = (I_d, 0)$ realize 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 G.3 (Parallelization of neural networks with the same input)**.** *For any neural networks $\phi^1, \phi^2, \cdots, \phi^k$ with $\phi^i \colon \mathbb{R}^d \to \mathbb{R}^{d'_i}$ 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 $\phi(x) = [\phi^1(x)^\top \ \phi^2(x)^\top \ \cdots \ \phi^k(x)^\top]^\top \cdot \mathbb{R}^{d_1} \to \mathbb{R}^{d'_1 + d'_2 + \cdots + d'_k}$ for all $x \in \mathbb{R}^d$, with*

$$L = L, \|W\|_\infty \leq \sum_{i=1}^k \|W^i\|_\infty, S \leq \sum_{i=1}^k S^i, \textit{and } B \leq \max_{1 \leq i \leq k} B^i \ \textit{(when } L = L_i \textit{ holds for all } i\textit{)},$$

$$L = \max_{1 \leq i \leq k} L^i, \|W\|_\infty \leq 2 \sum_{i=1}^k \|W^i\|_\infty, S \leq 2 \sum_{i=1}^k (S^i + LW^i_L), \textit{and } B \leq \max\{\max_{1 \leq i \leq k} B^i, 1\}$$

(otherwise).

Lemma G.4 (Parallelization of neural networks with different inputs)**.** *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 $x = (x_1^\top \ x_2^\top \ \cdots \ x_k^\top)^\top \in \mathbb{R}^{d_1 + d_2 + \cdots + d_k}$, with*

$$L = L, \|W\|_\infty \leq \sum_{i=1}^k \|W^i\|_\infty, S \leq \sum_{i=1}^k S^i, \textit{and } B \leq \max_{1 \leq i \leq k} B^i \ \textit{(when } L = L_i \textit{ holds for all } i\textit{)},$$

$$L = \max_{1 \leq i \leq k} L^i, \|W\|_\infty \leq 2 \sum_{i=1}^k \|W^i\|_\infty, S \leq 2 \sum_{i=1}^k (S^i + LW^i_L), \textit{and } B \leq \max\{\max_{1 \leq i \leq k} B^i, 1\}$$

(otherwise).

Proof of Lemmas G.3 and G.4. 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 ϕ'^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 G.1 and G.2, $\phi'^i \in \Phi(L, W'^i, S'^i, B'^i)$ holds, with $\max_{1 \leq i \leq k} L^i, \|W'^i\|_\infty = \max\{\|W^i\|_\infty, 2W_L\} \leq 2\|W^i\|_\infty, S'^i \leq 2S^i + 2(L - L_i)W^i_L \leq 2(S^i + LW^i_L)$, and $B'^i = \max\{B^i, 1\}$. We then apply the results for the case of $L = L_i$ $(i = 1, 2, \cdots, k)$. □

Remark G.5 (Aggregation of the outputs of different neural networks)**.** Since summation of the outputs of k neural networks can be realized by a 1 layer neural network with the width of k, Lemmas G.3 and G.4 together with Lemma G.1 give the bound to realize $\sum_{i=1}^k \phi^i(x)$ (when the input is the same and $d'_i = W^i_L =: W_L$ for all i) and $\sum_{i=1}^k \phi^i(x^i)$ (when the inputs are different and and $d'_i = W^i_L =: W_L$ for all i). Concretely, the bounds on L, $\|W\|_\infty$, S, and B are

$$L = \max_{1 \leq i \leq k} L^i + 1, \quad \|W\|_\infty \leq 4 \sum_{i=1}^k \|W^i\|_\infty, \quad S \leq 4 \sum_{i=1}^k (S^i + LW_L) + 2W_L, \quad \text{and}$$
$$B \leq \max\{\max_{1 \leq i \leq k} B^i, 1\}.$$

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 G.6 (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 G.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 G.7 (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 ϵ 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 $\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1), \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1) \in \Phi(3, (1,2,1,1)^\top, 9, \max\{\bar{t}_1, (\bar{t}_1 - \underline{t}_2)^{-1}\})$, and $\phi^1_{\mathrm{swit}}(\underline{t}; \underline{t}_2, \bar{t}_1)\phi^1(x,t) + \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1)\phi^2(x,t)$ approximates $f(x,t)$ up to an additive error of ϵ in $[\underline{t}_1, \bar{t}_2]$.*

Proof. We define

$$\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1) = \frac{1}{\bar{t}_1 - \underline{t}_2}\mathrm{ReLU}(\phi_{\mathrm{clip}}(t; \underline{t}_2, \bar{t}_1) - \underline{t}_2),$$
$$\text{and } \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1) = \frac{1}{\bar{t}_1 - \underline{t}_2}\mathrm{ReLU}(\bar{t}_1 - \phi_{\mathrm{clip}}(t; \underline{t}_2, \bar{t}_1)).$$

Here $\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1), \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1) \in [0,1]$, $\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1) + \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1) = 1$ for all t, $\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1) = 0$ for all $t \geq \bar{t}_1$, and $\phi^2_{\mathrm{swit}}(t; \underline{t}_2, \bar{t}_1)$ for $t \leq \underline{t}_2$. From this construction, the assertion follows. □

G.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. In our construction of the diffused B-spline basis, we need to approximate rational functions.

We begin with monomial functions. Although the traditional fact that we can approximate monomial functions with neural networks with an arbitrary additive error of ϵ 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 G.8 (Multiplication function)**.** *Let $d \geq 2$, $C \geq 1$, $0 < \varepsilon_{\mathrm{error}} < 1$. For any $\varepsilon > 0$, there exists a neural network $\phi_{\mathrm{mult}}(x_1, x_2, \cdots, x_d) \in \Psi(L, W, S, B)$ 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*

$$\left|\phi_{\mathrm{mult}}(x'_1, x'_2, \cdots, x'_d) - \prod_{d'=1}^{d} x_{d'}\right| \leq \varepsilon + dC^{d-1}\varepsilon_{\mathrm{error}},$$
$$\text{for all } x \in [-C, C]^d \text{ and } x' \in \mathbb{R} \text{ with } \|x - x'\|_\infty \leq \varepsilon_{\mathrm{error}},$$

$|\phi_{\mathrm{mult}}(x)| \leq C^d$ for all $x \in \mathbb{R}^d$, and $\phi_{\mathrm{mult}}(x'_1, x'_2, \cdots, x'_d) = 0$ if at least one of x'_i is 0.

Moreover, 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 in a neural network $\phi_{\mathrm{mult}}(x; \alpha) \in \Psi(L, W, S, B)$ and the same bounds on the error and $L, \|W\|_\infty, S, B$ as above hold.

Proof. First of all, it is known from Schmidt-Hieber (2020) that there exists a neural network $\bar{\phi}'_{\mathrm{mult}}(x,y) \in \Psi(L, W, S, B)$ with $L = i + 5$, $\|W\|_\infty = 6$, $B = 1$ such that

$$|\bar{\phi}'_{\mathrm{mult}}(x,y) - xy| \leq 2^{-i}, \quad \text{for all } (x,y) \in [0,1]^2,$$

and $|\bar{\phi}'_{\mathrm{mult}}(x,y)| \leq 1$ for all $(x,y) \in \mathbb{R}^2$, and $\bar{\phi}'_{\mathrm{mult}}(x,y) = 0$ if either x or y is 0. With this network, we can see that $|\mathrm{sign}(xy)\bar{\phi}'_{\mathrm{mult}}(|x|, |y|) - xy| \leq 2^{-i}$ holds for all $(x,y) \in [-1,1]^2$,

$|\bar{\phi}'_{\text{mult}}(x,y)| \le 1$ for all $(x,y) \in \mathbb{R}^2$, and $\bar{\phi}_{\text{mult}}(x,y) = 0$ if either x or y is 0. Because

$$\begin{aligned}\text{sign}(xy)\bar{\phi}'_{\text{mult}}(|x|,|y|) &= \text{ReLU}(\bar{\phi}'_{\text{mult}}(\text{ReLU}(x),\text{ReLU}(y)) + \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x),\text{ReLU}(-y)) \\ &\quad - \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x),\text{ReLU}(y)) - \bar{\phi}'_{\text{mult}}(\text{ReLU}(x),\text{ReLU}(-y))) \\ &\quad - \text{ReLU}(-\bar{\phi}'_{\text{mult}}(\text{ReLU}(x),\text{ReLU}(y)) - \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x),\text{ReLU}(-y)) \\ &\quad + \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x),\text{ReLU}(y)) + \bar{\phi}'_{\text{mult}}(\text{ReLU}(x),\text{ReLU}(-y))) \\ &=: \bar{\phi}_{\text{mult}}(x,y)\end{aligned}$$

holds, we can realize the function xy for $[-1,1]^d$, by a neural network $\bar{\phi}_{\text{mult}}(x,y) \in \Psi(L,W,S,B)$ with $L = i+7, \|W\|_\infty = 48, S \le L\|W\|_\infty(\|W\|_\infty + 1), B = 1$ with an approximation error up to 2^{-i}.

Then, following Schmidt-Hieber (2020), we recursively construct $\bar{\phi}_{\text{mult}}(x_1, x_2, \cdots, x_{2^{j+1}})$ using

$$\bar{\phi}_{\text{mult}}(x_1, x_2, \cdots, x_{2^{j+1}}) = \bar{\phi}_{\text{mult}}(\bar{\phi}_{\text{mult}}(x_1, x_2, \cdots, x_{2^j}), \bar{\phi}_{\text{mult}}(x_{2^j+1}, x_{2^j+2}, \cdots, x_{2^{j+1}})),$$

and we obtain the neural network $\phi_{\text{mult}}(x_1, x_2, \cdots, x_d) \in \Psi(L,W,S,B)$ with $L = \mathcal{O}(\log d(\log \varepsilon^{-1} + \log d)), \|W\|_\infty = 48d, S = \mathcal{O}(d(\log \varepsilon^{-1} + \log d)), B = 1$ such that

$$\left|\bar{\phi}_{\text{mult}}(x_1, x_2, \cdots, x_d) - \prod_{d'=1}^{d} x_{d'}\right| \le \varepsilon, \quad \text{for all } x \in [-1,1]^d.$$

We then construct ϕ_{mult} as follows:

$$\phi_{\text{mult}}(x) = C^d \bar{\phi}_{\text{mult}}(\phi_{\text{clip}}(x; -C, C)/C).$$

Here $\bar{\phi}_{\text{mult}}$ is taken so that the error guarantee in the above part is smaller than $C^{-d}\varepsilon$, thus $L = \mathcal{O}(\log d(\log d + \log \varepsilon^{-1} + d \log C)), \|W\|_\infty = 48d, S = \mathcal{O}(d(\log d + \log \varepsilon^{-1} + d \log C)), B = 1$. Therefore, the bounds on $L, \|W\|_\infty, B, S$ in the assertion follows from Lemmas G.1 and G.6. For the approximation error,

$$\begin{aligned}&\left|C^d \bar{\phi}_{\text{mult}}(\phi_{\text{clip}}(x'; -C, C)/C) - \prod_{i=1}^{d} x_i\right| \\ &\le \left|C^d \bar{\phi}_{\text{mult}}(\phi_{\text{clip}}(x'; -C, C)/C) - \prod_{i=1}^{d} \min\{C, \max\{x'_i, -C\}\}\right| \\ &\quad + \left|\prod_{i=1}^{d} \min\{C, \max\{x'_i, -C\}\} - \prod_{i=1}^{d} x_i\right| \\ &\le C^d \cdot C^{-d}\varepsilon + C^{d-1} \sum_{i=1}^{d} |x_i - \min\{C, \max\{x'_i, -C\}\}| = \varepsilon + dC^{d-1}\varepsilon_{\text{error}},\end{aligned}$$

which yields the first part of the assertion.

The second assertion immediately follows from the above construction. All we need is to identify $\alpha_{d'}$ columns in the first layer of $\bar{\phi}_{\text{mult}}(x_1, \cdots, x_d)$ that correspond to each coordinate d'. □

We next provide how to approximate the inverse 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 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 G.9 (Approximating the inverse function)**.** *For any $0 < \varepsilon < 1$, there exists $\phi_{\text{inv}} \in \Psi(L,W,S,B)$ 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 $B = \mathcal{O}(\varepsilon^{-3})$ such that*

$$\left|\phi_{\text{inv}}(x') - \frac{1}{x}\right| \le \varepsilon + \frac{|x' - x|}{\varepsilon^2}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

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 $y = \frac{1}{x}$:

$$\begin{aligned} f(x) = &\sum_{i=1}^{i^*}\frac{1}{x_{i-1}}\sum_{l'=1}^{l}(-\phi_{\text{clip}}(x; x_{i-1}, x_i))/x_{i-1} + 1)^{l'} - \sum_{i=1}^{i^*}\frac{1}{x_{i-1}} \\ &+ \sum_{i=1}^{i^*}\frac{((x_{i-1} - x_i)/x_{i-1})^{l+1}(\phi_{\text{clip}}(x; x_{i-1}, x_i)) - x_{i-1})}{x_i(x_i - x_{i-1})} + \frac{1}{\varepsilon}. \end{aligned} \tag{42}$$

From Lemma G.8, $(-\phi_{\text{clip}}(x; x_{i-1}, x_i))/x_{i-1} + 1)^{l'}$ 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 (42), from Lemmas G.1 and G.3, the final approximation error of $f(x)$ using a neural network ϕ_{inv} is $\frac{\varepsilon}{2}$, where $\phi_{\text{inv}} \in \Phi(L, W, S, B)$ 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 $B = \mathcal{O}(\varepsilon^{-3})$.

The sensitivity analysis follows from $|\phi_{\text{inv}}(x') - \frac{1}{x}| \leq |\phi_{\text{inv}}(x') - \frac{1}{\max\{x', \varepsilon\}}| + |\frac{1}{\max\{x', \varepsilon\}} - \frac{1}{x}|$. □

Combining Lemmas G.8 and G.9, we have the following corollary.

Corollary G.10. *For any $0 < \varepsilon < 1$, there exists $\phi_{\text{inv}} \in \Psi(L, W, S, B)$ 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 $B = \mathcal{O}(\varepsilon^{-(3\vee l)})$ such that*

$$\left|\phi_{\text{inv}}(x'; l) - \frac{1}{x^l}\right| \leq \varepsilon + l\frac{|x' - x|}{\varepsilon^{l+1}}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

Proof. Consider $\phi_{\text{mult}}(\cdot; l) \circ \phi_{\text{inv}}$. □

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 G.11 (Approximating the root function)**.** *For any $0 < \varepsilon < 1$, there exists $\phi_{\mathrm{inv}} \in \Psi(L, W, S, B)$ 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*

$$\left|\phi_{\mathrm{root}}(x') - \sqrt{x}\right| \leq \varepsilon + \frac{|x' - x|}{\sqrt{\varepsilon}}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

G.3 How to deal with exponential functions

We sometimes need to approximate certain types of integrals, for example, the integral in the diffused B-spline basis. 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. The first and second lemmas corresponds to the first and second steps, respectively.

Lemma G.12 (Clipping of integrals)**.** *Let $x \in \mathbb{R}^d$, $0 < m \leq 1$, $\alpha \in \mathbb{Z}_+^d$ 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 $C_{\mathrm{clip},1}$ that only depends on k and d, such that*

$$\left|\int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{my_i - x_i}{\sigma}\right)^{\alpha_i} f(y) \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|my - x\|^2}{2\sigma^2}\right) dy \right.$$
$$\left. - \int_{A^x} \prod_{i=1}^d \left(\frac{my_i - x_i}{\sigma}\right)^{\alpha_i} f(y) \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|my - x\|^2}{2\sigma^2}\right) dy \right| \leq \frac{\varepsilon C_f}{m^d},$$

where $A^x = \prod_{i=1}^d a_i^x$ with $a_i^x = [\frac{x_i}{m} - \frac{\sigma C_{\mathrm{clip},1}}{m}\sqrt{\log \varepsilon^{-1}}, \frac{x_i}{m} + \frac{\sigma C_{\mathrm{clip},1}}{m}\sqrt{\log \varepsilon^{-1}}]$.

Proof.

$$\frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \left|\int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{my_i - x_i}{\sigma}\right)^{\alpha_i} f(y) \exp\left(-\frac{\|my - x\|^2}{2\sigma^2}\right) dy \right.$$
$$\left. - \int_{A^x} \prod_{i=1}^d \left(\frac{my_i - x_i}{\sigma}\right)^{\alpha_i} f(y) \exp\left(-\frac{\|my - x\|^2}{2\sigma^2}\right) dy \right|$$
$$\leq \frac{C_f}{\sigma^d (2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d \setminus A^x} \prod_{i=1}^d \left(\frac{|my_i - x_i|}{\sigma}\right)^{\alpha_i} \exp\left(-\frac{\|my - x\|^2}{2\sigma^2}\right) dy \quad (\text{by } |f(y)| \leq C_f)$$
$$\leq \frac{C_f}{\sigma^d (2\pi)^{\frac{d}{2}}} \sum_{i=1}^d \int_{\underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{i-1 \text{ times}} \times \mathbb{R} \setminus a_i^x \times \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{d-i \text{ times}}} \prod_{j=1}^d \left(\frac{|my_j - x_j|}{\sigma}\right)^{\alpha_j} \exp\left(-\frac{\|my - x\|^2}{2\sigma^2}\right) dy$$
$$= C_f \sum_{i=1}^d \prod_{j=1}^d \left(\mathbb{1}[i \neq j] \int_{\mathbb{R}} \left(\frac{|my_j - x_j|}{\sigma}\right)^{\alpha_j} \frac{1}{\sigma (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(my_j - x_j)^2}{2\sigma^2}\right) dy_j \right.$$
$$\left. + \mathbb{1}[i = j] \int_{\mathbb{R} \setminus a_i^x} \left(\frac{|my_j - x_j|}{\sigma}\right)^{\alpha_j} \frac{1}{\sigma (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(my_j - x_j)^2}{2\sigma^2}\right) dy_j \right). \tag{43}$$

We now bound each term. First,

$$\int_{\mathbb{R}} \left(\frac{|my_j - x_j|}{\sigma}\right)^{\alpha_j} \frac{1}{\sigma (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(my_j - x_j)^2}{2\sigma^2}\right) dy_j = \frac{1}{m} \int_{\mathbb{R}} |y_j'|^{\alpha_j} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{y_j'^2}{2}\right) dy_j'$$

This can be bounded by $\frac{C_1}{m}$, where C_1 only depends on $\alpha_j (\leq k)$.

Next,

$$\int_{\mathbb{R}\setminus a_i^x} \left(\frac{|my_j - x_j|}{\sigma}\right)^{\alpha_j} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(my_j - x_j)^2}{2\sigma^2}\right) dy_j$$

$$\leq \frac{2}{m}\int_{C_{\mathrm{clip},1}\sqrt{\log\varepsilon^{-1}}}^{\infty} |y_j|^{\alpha_j} \exp\left(-\frac{y_j^2}{2}\right) \mathrm{d}y_i \quad \text{(by letting } \frac{my_j - x_j}{\sigma} \mapsto y_j\text{)}$$

$$\leq \begin{cases} \frac{2}{m}\sum_{l=0}^{\frac{\alpha_j-1}{2}} \frac{(\alpha_j-1)!!}{(2l)!!}(C_{\mathrm{clip},1}^2 \log\varepsilon^{-1})^l \varepsilon^{\frac{C_{\mathrm{clip},1}}{2}} & \text{(if } \alpha_j \text{ is odd)} \\ \frac{2}{m}\sum_{l=1}^{\frac{\alpha_j}{2}} \frac{(\alpha_j-1)!!}{(2l-1)!!}(C_{\mathrm{clip},1}^2 \log\varepsilon^{-1})^l \varepsilon^{\frac{C_{\mathrm{clip},1}}{2}} + \frac{2}{m}\int_{C_{\mathrm{clip},1}\sqrt{\log\varepsilon^{-1}}}^{\infty} \exp\left(-\frac{y_j^2}{2}\right) \mathrm{d}y_j & \text{(if } \alpha_j \text{ is even).} \end{cases}$$

Therefore, by setting $C_{\mathrm{clip},1}$ that only depends on $\alpha_j (\leq k)$ and d, this can be bounded by $\frac{\varepsilon}{dmC_1^{d-1}}$.

Therefore, (43) can further be bounded by

$$(43) \leq C_f \sum_{i=1}^{d}\prod_{j=1}^{d}\left(\frac{C_1}{m}\right)^{d-1} \cdot \frac{\varepsilon}{dmC_1^{d-1}} \leq \frac{\varepsilon C_f}{m^d},$$

which gives the assertion. □

Here we give the ways of Taylor expansion accuracy for different two cases.

Lemma G.13 (Approximating an exponential function with polynomials)**.** *Let $A > 0$ and $0 \leq m \leq 1$. For $t \geq \max\{4eA^2, \lceil \log_2 \varepsilon^{-1}\rceil\}$, we have that*

$$\left|\exp\left(-\frac{(x-my)^2}{2\sigma^2}\right) - \sum_{s=0}^{t-1}\frac{(-1)^s}{s!}\frac{(x-my)^{2s}}{2^s\sigma^{2s}}\right| \leq \varepsilon$$

for all $y \in [\frac{-\sigma A + x}{m}, \frac{\sigma A + x}{m}]$.

Proof. By standard Taylor expansion about x around $x = y$ up to degree $t-1$, we have

$$\exp\left(-\frac{(x-my)^2}{2\sigma^2}\right) = \sum_{s=0}^{t-1}\frac{(-1)^s}{s!}\frac{(x-my)^{2s}}{2^s\sigma^{2s}} + \frac{(-1)^t}{t!}\frac{(\theta(x-my)+(1-\theta)x)^{2t}}{2^t\sigma^{2t}}$$

with some $\theta \in (0,1)$. We bound the second term of the residual. When $x \in [-\sigma A + y, \sigma A + y]$ and t is the minimum integer satisfying $t \geq \max\{4eA^2, \lceil \log_2 \varepsilon^{-1}\rceil\}$, we have

$$\frac{1}{t!}\frac{(\theta(x-my)+(1-\theta)x)^{2t}}{2^t\sigma^{2t}} \leq \frac{(2\sigma A)^{2t}}{t!2^t\sigma^{2t}} \leq \frac{(2\sigma A)^{2t}}{(t/e)^t \cdot 2^t\sigma^{2t}} \leq \frac{2^tA^{2t}}{(4A^2)^t} \leq \frac{1}{2^t} \leq \varepsilon,$$

where we used the fact $t! \geq (t/e)^t$. □

Lemma G.14 (Approximating an exponential function with polynomials)**.** *Take $\varepsilon > 0$ arbitrarily. There exists a neural network $\phi_{\exp} \in \Phi(L, W, S, B)$ such that*

$$\sup_{x\geq 0, x'\in\mathbb{R}^d}\left|e^{-x'} - \phi_{\exp}(x)\right| \leq \varepsilon + |x - x'|$$

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 G.8 with $d =$

$\mathcal{O}(A + \log \varepsilon^{-1}), C = \mathcal{O}(A), \varepsilon = \frac{\varepsilon}{3k}$ and aggregate them using remark G.5. This gives the neural network with $L = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}), \|W\|_\infty = \mathcal{O}(A + \log \varepsilon^{-1}), S = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}), B = \exp(\log A \mathcal{O}(A + \log \varepsilon^{-1}))$. Finally, we add two layer $\phi_{\text{clip}}(x; 0, A)$ 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 $|\phi_{\exp}(x') - e^{-x}| \leq |\phi_{\exp}(\max\{x', 0\}) - e^{-x}| \leq |\phi_{\exp}(\max\{x', 0\}) - e^{-\max\{x',0\}}| + |e^{-\max\{x',0\}} - e^{-x}| \leq \varepsilon + |\max\{x', 0\} - x| \leq \varepsilon + |x' - x|$. □

G.4 Existing results for approximation

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 G.15 (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*

$$\|f - f_N\|_{L^2([-C,C]^d)} \lesssim C^s N^{-s/d} \|f\|_{B^s_{p,q}([-C,C]^d)}$$

for $N \gg 1$, and has the following form:

$$f_N(x) = \sum_{k=0}^{K} \sum_{j \in J(k)} \alpha_{k,j} M^d_{k,j}(x) + \sum_{k=K+1}^{K^*} \sum_{i=1}^{n_k} \alpha_{k,j_i} M^d_{k,j_i}(x),$$

where $(j_i)_{i=1}^{n_k} \in J(k) = \{-C2^k - l, -C2^k - l + 1, \cdots C2^k - 1, C2^k\}$, $K = \mathcal{O}(d^{-1} \log N)$, $K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + K, n_k = \mathcal{O}(N 2^{-\nu(k-K)})$ $(k = K+1, \cdots, K^*)$ for $\delta = d(1/p - 1/r)_+$ and $\nu = (s - \delta)/(2\delta)$. Moreover, $|\alpha_{k,j}| \lesssim N^{(\nu^{-1} + d^{-1})(d/p - s)_+}$.

G.5 Elementary bounds for the Gaussian and hitting time

Lemma G.16. *Let $0 < \varepsilon \ll 1$, $l \in \mathbb{Z}^d_+$, and $p(x)$ be the density funciton of $\mathcal{N}(0, \sigma^2 I_d)$ i.e., $p(x) = \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x\|^2}{\sigma^2}\right)$. Then, the following bound holds:*

$$\int_{\|x\|_\infty \geq \sigma \sqrt{2 \log d l \varepsilon^{-1}}} \frac{x^l}{\sigma^{|l|}} p(x) \mathrm{d}x \leq \varepsilon.$$

Proof. For the first part, let $r = \|x\|_\infty$.

$$
\begin{aligned}
&\int_{\|x\|_\infty \geq \sigma\sqrt{2\log\varepsilon^{-1}}} \frac{x^l}{\sigma^{|l|}} p(x)\mathrm{d}x \\
&\leq \int_{r=\sigma\sqrt{2\log\varepsilon^{-1}}}^{\infty} \frac{r^{|l|}}{\sigma^{|l|}} \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{r^2}{2\sigma^2}\right)(d-1)r^{d-1}\mathrm{d}r \\
&= \int_{s=\sqrt{2\log\varepsilon^{-1}}}^{\infty} s^{|l|+d-1} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s \\
&= \frac{(2\log\varepsilon^{-1})^{(|l|+d-1)/2}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{2\log\varepsilon^{-1}}{2}\right)(d-1) \\
&\quad + \int_{s=\sqrt{2\log\varepsilon^{-1}}}^{\infty} \frac{(|l|+d-1)s^{|l|+d-2}}{(2\pi)^{\frac{d}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s \\
&= \cdots = \sum_{0\leq i\leq \lfloor\frac{|l|+d-1}{2}\rfloor} \frac{\frac{(|l|+d-1)!!}{(|l|+d-1-2i)!!}(2\log\varepsilon^{-1})^{(|l|+d-1-2i)/2}(d-1)}{(2\pi)^{\frac{d}{2}}}\varepsilon \\
&\quad + \begin{cases} \int_{s=\sqrt{2\log\varepsilon^{-1}}}^{\infty} \frac{(|l|+d-1)!!}{(2\pi)^{\frac{d}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s & (|l|+d\text{: even}) \\ \int_{s=\sqrt{2\log\varepsilon^{-1}}}^{\infty} \frac{(|l|+d-1)!!s}{(2\pi)^{\frac{d}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s & (|l|+d\text{: odd}) \end{cases} \\
&\lesssim \varepsilon \log^{\frac{d+|l|-1}{2}} \varepsilon^{-1}. \qquad (44)
\end{aligned}
$$

Replacing ε by ε/dl, RHS of eq. (44) is bounded by

$$\varepsilon/dl \log^{\frac{dn+|l|-1}{2}} (\varepsilon/dl)^{-1} \lesssim \varepsilon.$$

□

Lemma G.17. *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 $x > 0$,

$$mathbbP[\sup_{s\in[0,t]} |X_t| \geq x] = \frac{4}{\sqrt{2\pi}} \int_{\frac{x}{\sqrt{2\overline{\beta}t}}}^{\infty} e^{-y^2/2}\mathrm{d}y \leq 2e^{-x^2/4\overline{\beta}t}.$$

Therefore, setting $x = 2\sqrt{\overline{\beta}t\log(2\varepsilon^{-1})}$ yields the assertion. □