Wasserstein Convergence of Score-based Generative Models under Semiconvexity and Discontinuous Gradients

Anonymous authors
Paper under double-blind review

Abstract

Score-based Generative Models (SGMs) approximate a data distribution by perturbing it with Gaussian noise and subsequently denoising it via a learned reverse diffusion process. These models excel at modeling complex data distributions and generating diverse samples, achieving state-of-the-art performance across domains such as computer vision, audio generation, reinforcement learning, and computational biology. Despite their empirical success, existing Wasserstein-2 convergence analysis typically assume strong regularity conditions–such as smoothness or strict log-concavity of the data distribution–that are rarely satisfied in practice. In this work, we establish the first non-asymptotic Wasserstein-2 convergence guarantees for SGMs targeting semiconvex distributions with potentially discontinuous gradients. Our upper bounds are explicit and sharp in key parameters, achieving optimal dependence of $O(\sqrt{d})$ on the data dimension d and convergence rate of order one. The framework accommodates a wide class of practically relevant distributions, including symmetric modified half-normal distributions, Gaussian mixtures, double-well potentials, and elastic net potentials. By leveraging semiconvexity without requiring smoothness assumptions on the potential such as differentiability, our results substantially broaden the theoretical foundations of SGMs, bridging the gap between empirical success and rigorous guarantees in non-smooth, complex data regimes.

1 Introduction

Score-based Generative Models (SGMs), also known as diffusion-based generative models (Song & Ermon, 2019; Song et al., 2021; Sohl-Dickstein et al., 2015; Ho et al., 2020), have rapidly emerged over the past few years as a popular approach in modern generative modelling due to their remarkable capabilities in generating complex data, surpassing previous state-of-the-art models, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational AutoEncoders (VAEs) (Kingma & Welling, 2014). These models are now widely adopted in computer vision and audio generation tasks (Kong et al., 2020; Chen et al., 2021; Mittal et al., 2021; Avrahami et al., 2021; Kim et al., 2021; Bansal et al., 2023; Saharia et al., 2022; Po et al., 2023; Zhang et al., 2023), text generation (Li et al., 2022; Yu et al., 2022; Lovelace et al., 2023), sequential data modeling (Alcaraz & Strodthoff, 2023; Tashiro et al., 2021; Tevet et al., 2023), reinforcement learning and control (Pearce et al., 2023; Chi et al., 2023; Hansen-Estruch et al., 2023; Reuss et al., 2023; Zhu et al., 2023; Ding & Jin, 2024), as well as life-science (Chung & Ye, 2021; Jing et al., 2022; Watson et al., 2023; Song et al., 2022; Weiss et al., 2023). We refer the reader to the survey papers Yang et al. (2023); Chen et al. (2024) for a more comprehensive exposition of their applications.

The primary goal of SGMs is to generate synthetic data that closely match a target data distribution π_{D}, given a sample set. In particular, these models generate approximate data samples from high-dimensional data distributions by combining two diffusion processes, a forward and a backward process in time. The forward process is used to iteratively and smoothly transform samples from the unknown data distribution into (Gaussian) noise, while the associated backward process reverses the noising procedure to generate new samples from the starting unknown data distribution. A key role in these models is played by the score function, i.e. the gradient of the log-density of the solution of the forward process, which appears in the drift of the stochastic differential equation (SDE) associated with the backward process. Since this quantity depends on the unknown data distribution, an estimator of the score must be constructed during the noising step using score-matching techniques (Hyvärinen, 2005; Vincent, 2011).

The widespread applicability and success of SGMs have been accompanied by a growing interest in the theoretical understandings of these models, particularly in the convergence analysis under different metrics such as Total Variation (TV) distance, Kullback Leibler (KL) divergence, Wasserstein distance, e.g., Block et al. (2020); De Bortoli et al. (2021); Bortoli (2022); Lee et al. (2022); Yang & Wibisono (2022); Kwon et al. (2022); Liu et al. (2022); Oko et al. (2023); Lee et al. (2023); Chen et al. (2023a;b); Li et al. (2024); Pedrotti et al. (2024); Conforti et al. (2025); Benton et al. (2024); Strasman et al. (2025); Bruno et al. (2025); Tang & Zhao (2024); Mimikos-Stamatopoulos et al. (2024); Gentiloni-Silveri & Ocello (2025); Yu & Yu (2025). In this work, we provide a non-asymptotic convergence analysis in Wasserstein distance of order two, as this metric is often considered more practical and informative for estimation tasks (see e.g., equation 5), and is closely connected to the popular Fréchet Inception Distance (FID) used to assess the quality of images in generative modeling (see, e.g., Section 4). A significant limitation of prior analysis in Wasserstein-2, e.g., Strasman et al. (2025); Gao et al. (2025); Bruno et al. (2025); Tang & Zhao (2024); Yu & Yu (2025), is their reliance on strong regularity conditions–such as smoothness or strict log-concavity – of the data distribution and its potential. These assumptions facilitate mathematical tractability but limit the applicability of theoretical results to more general settings, especially when the data distribution is only semiconvex and the potential's gradient may be discontinuous. The only exception outside the strict log-concavity regime is the recent contribution in Gentiloni-Silveri & Ocello (2025), where the authors assumes that the data distribution is weakly convex. However, their analysis still requires the potential to be twice continuously differentiable (see, e.g., Gentiloni-Silveri & Ocello (2025, Proofs of Propositions B.1 and B.2)), and the stepsize of their generative algorithm must be bounded by a quantity inversely proportional to the one-sided Lipschitz constant of the potential (see Gentiloni-Silveri & Ocello (2025, equation (30))). Still, such conditions on π_{D} in existing Wasserstein-2 convergence analysis do not fully reflect the complexity of real-world data, which often exhibit non-smooth or non-log-concave distributions. Therefore, the aim of this work is to address the following fundamental question:

Can Score-based Generative Models be guaranteed to converge in Wasserstein-2 distance when the data distribution is only semiconvex and the potential admits discontinuous gradients?

We provide a positive answer to this question by combining recent findings in non-smooth, non-log-concave sampling, with standard stochastic analysis tools, thereby presenting the first contributions in the Score-based generative modeling literature for non-smooth potentials. We establish explicit, non-asymptotic Wasserstein-2 convergence bounds for SGMs under semiconvexity assumptions on the data distribution, accommodating potentials with discontinuous gradients. This framework covers a variety of practically relevant distributions arising in Bayesian statistical methods, including symmetric modified half-normal distributions, Gaussian mixtures, double-well potentials, and elastic net potentials, all of which satisfy our relaxed assumptions.

In addition, our estimates are explicit and exhibit the best known optimal dependencies in terms of data dimension, i.e., $O(\sqrt{d})$ in Theorem 13, and rate of convergence, i.e., $O(\gamma)$ in Theorem 15. In contrast to prior works under the same metric Gentiloni-Silveri & Ocello (2025); Gao et al. (2025); Strasman et al. (2025); Tang & Zhao (2024), our estimates in Theorems 13 and Theorem 15 are derived without imposing any restrictions on the stepsize of the generative algorithm, making them more suitable to practical implementation. By circumventing the need for strict regularity conditions on the score function and allowing discontinuities in the gradients of the potentials, our work significantly expands the theoretical foundation of SGMs. This advancement not only bridges the gap between empirical success and theoretical guarantees but also opens new avenues for the application of diffusion models to data distributions with non-smooth potentials.

One source of error in the construction of the generative algorithm arises from replacing the initial condition of the backward process with the invariant measure of the forward process. To ensure this error remains small, the drift terms of both SDEs must satisfy, for instance, a monotonicity property with a time-dependent bound that meets an appropriate integrability condition (see, e.g., equation 19 and equation 23 below). To address this, we identify a time horizon for the generative algorithm that ensures the paths of the two backward processes become contractive. Notably, the integrability condition on the monotonicity bound depends only on the known constants in Assumption 2, making it significantly easier to verify in practice compared to the analogous condition in Gentiloni-Silveri & Ocello (2025, Appendix C), which relies on weak convexity constants that are often difficult to estimate.

In conclusion, we present the first explicit, dimension- and parameter-dependent W_2-convergence guarantees for Score-based Generative models operating on data distributions having potentials with discontinuous gradients. Our results mark an important step forward in the rigorous analysis of SGMs, providing both theoretical insights and practical tools for advancing generative modeling in challenging, non-smooth regimes.

Notation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a fixed probability space. We denote by $\mathbb{E}[X]$ the expectation of a random variable X. For $1 \leq p < \infty$, L^p is used to denote the usual space of p-integrable real-valued random variables. The L^p-integrability of a random variable X is defined as $\mathbb{E}[|X|^p] < \infty$. Fix an integer $d \geq 1$. For an $\mathbb{R}^d$-valued random variable X, its law on $\mathcal{B}(\mathbb{R}^d)$, i.e. the Borel sigma-algebra of $\mathbb{R}^d$ is denoted by $\mathcal{L}(X)$. Let $T > 0$ denote a time horizon. For a positive real number b, we denote its integer part by $\lfloor b \rfloor$. The Euclidean scalar product is denoted by $\langle \cdot, \cdot \rangle$, with $|\cdot|$ standing for the corresponding norm (where the dimension of the space may vary depending on the context). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. The gradient of f is denote by ∇f. For any integer $q \geq 1$, let $\mathcal{P}(\mathbb{R}^q)$ be the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$. For μ, $\nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ and ν. For any μ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, the Wasserstein distance of order 2 is defined as

$$W_2(\mu, \nu) = \left(\inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x - y|^2 \, \mathrm{d}\zeta(x, y) \right)^{\frac{1}{2}}.$$

For any $x \in A \subseteq \mathbb{R}^d$ and any function $U : A \to \mathbb{R}$, the subdifferential $\partial U(x)$ of U at x is defined as

$$\partial U(x) = \left\{ p \in \mathbb{R}^d : \liminf_{y \to x} \frac{U(y) - U(x) - \langle p, y - x \rangle}{|y - x|} \geq 0 \right\}. \tag{1}$$

At the points where U is differentiable, it holds that $\partial U(x) = \{\nabla U(x)\}$. For the sake of presentation, an element of $\partial U(x)$ is denoted by $h(x)$ for any $x \in \mathbb{R}^d$.

2 Technical Background for OU-based SGMs

In this section, we briefly summarize the construction of score-based generative models (SGMs) via diffusion processes, as introduced by Song et al. (2021). The core idea behind SGMs is to employ an ergodic (forward) diffusion process that gradually transforms the unknown data distribution $\pi_{\mathsf{D}} \in \mathcal{P}(\mathbb{R}^d)$ into a known prior distribution. A backward (in time) process is then learned to transform the prior back to the target distribution π_{D} by estimating the score function of the forward process. In our analysis, we consider the forward process $(X_t)_{t \in [0,T]}$ to be an Ornstein-Uhlenbeck (OU) process, consistent with the choice in the original paper Song et al. (2021)

$$\mathrm{d}X_t = -X_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t, \quad X_0 \sim \pi_{\mathsf{D}}, \tag{2}$$

where $(B_t)_{t \in [0,T]}$ is an d-dimensional Brownian motion and we assume that $\mathbb{E}[|X_0|^2] < \infty$. Under mild assumptions on the target data distribution π_{D} (Haussmann & Pardoux, 1986; Cattiaux et al., 2023), the backward process $(Y_t)_{t \in [0,T]} = (X_{T-t})_{t \in [0,T]}$ is given by

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log p_{T-t}(Y_t)) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\bar{B}_t, \quad Y_0 \sim \mathcal{L}(X_T), \tag{3}$$

where $\{p_t\}_{t \in [0,T]}$ is the family of densities of $\{\mathcal{L}(X_t)\}_{t \in (0,T]}$ with respect to the Lebesgue measure, $\bar{B}_t$ is an another Brownian motion independent of B_t in 2 defined on $(\Omega, \mathcal{F}, \mathbb{P})$. In practice, however, the initial distribution is taken to be the invariant measure of the forward process, which corresponds to the standard Gaussian distribution. As a result, the backward process in 3 becomes

$$\mathrm{d}\widetilde{Y}_t = (\widetilde{Y}_t + 2 \, \nabla \log p_{T-t}(\widetilde{Y}_t)) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\bar{B}_t, \quad \widetilde{Y}_0 \sim \pi_\infty = \mathcal{N}(0, I_d). \tag{4}$$

Since the target distribution π_{D} is unknown, the score function $\nabla \log p_t$ in 3 cannot be computed exactly. To overcome this limitation, an estimator $s(\cdot, \theta^*, \cdot)$ is *learned* based on a family of functions $s : [0, T] \times \mathbb{R}^M \times \mathbb{R}^d \to$

$\mathbb{R}^d$ parametrized in θ, aiming at approximating the score of the ergodic forward process 5 over a fixed time window $[0, T]$. In practice, s are neural networks and in particular cases, e.g., the motivating example in Bruno et al. (2025, Section 3.1), the functions s can be carefully designed. The optimal value θ^* of the parameter θ is determined by optimizing the following score-matching objective

$$\mathbb{R}^d \ni \theta \mapsto \mathbb{E}\left[\int_0^T |\nabla \log p_t(X_t) - s(t, \theta, X_t)|^2 \, \mathrm{d}t\right]. \tag{5}$$

An explicit expression of the stochastic gradient of 5 derived via denoising score matching (Vincent, 2011) is provided in Bruno et al. (2025, equation (8), Section 2). Following Bruno et al. (2025, Section 2), we define an auxiliary process $(Y_t^{\mathrm{aux}})_{t\in[0,T]}$ that incorporates the approximating function s, which depends on the (random) estimator of θ^* denoted by $\hat{\theta}$. For $t \in [0, T]$, this process is given by

$$\mathrm{d}Y_t^{\mathrm{aux}} = (Y_t^{\mathrm{aux}} + 2\ s(T-t, \hat{\theta}, Y_t^{\mathrm{aux}}))\ \mathrm{d}t + \sqrt{2}\ \mathrm{d}\bar{B}_t, \quad Y_0^{\mathrm{aux}} \sim \pi_\infty = \mathcal{N}(0, I_d). \tag{6}$$

The auxiliary process 6 serves as a bridge between the backward process 4 and the numerical scheme 8, and it facilitates the analysis of the convergence of the diffusion model (see the upper bounds involving Y_t^{aux} in the proof of Theorem 13 in Appendix C for further details). We now introduce the numerical scheme. Let the step size $\gamma_k = \gamma \in (0,1)$ for each $k = 0, \ldots, K$, where $K \in \mathbb{N}$. The discrete process $(Y_k^{\mathrm{EM}})_{k\in\{0,\ldots,K+1\}}$ of the Euler–Maruyama approximation of 6 is given, for any $k \in \{0, \ldots, K\}$, as follows

$$Y_{k+1}^{\mathrm{EM}} = Y_k^{\mathrm{EM}} + \gamma(Y_k^{\mathrm{EM}} + 2\ s(T - t_k, \hat{\theta}, Y_k^{\mathrm{EM}})) + \sqrt{2\gamma}\ \bar{Z}_{k+1}, \quad Y_0^{\mathrm{EM}} \sim \pi_\infty = \mathcal{N}(0, I_d), \tag{7}$$

where $\{\bar{Z}_k\}_{k\in\{0,\ldots,K+1\}}$ is a sequence of independent d-dimensional Gaussian random variables with zero mean and identity covariance matrix. The continuous-time interpolation of 7, for $t \in [0, T]$, is given by

$$\mathrm{d}\widehat{Y}_t^{\mathrm{EM}} = (\widehat{Y}_{\lfloor t/\gamma\rfloor\gamma}^{\mathrm{EM}} + 2\ s(T - \lfloor t/\gamma\rfloor\gamma, \hat{\theta}, \widehat{Y}_{\lfloor t/\gamma\rfloor\gamma}^{\mathrm{EM}}))\ \mathrm{d}t + \sqrt{2}\ \mathrm{d}\bar{B}_t, \qquad \widehat{Y}_0^{\mathrm{EM}} \sim \pi_\infty = \mathcal{N}(0, I_d), \tag{8}$$

where $\mathcal{L}(\widehat{Y}_k^{\mathrm{EM}}) = \mathcal{L}(Y_k^{\mathrm{EM}})$ at grid points for each $k \in \{0, \ldots, K+1\}$.

3 Wasserstein Convergence Analysis for SGMs

In this section, we provide the full non-asymptotic estimates in Wasserstein distance of order two between the target data distribution π_{D} and the generative distribution of the diffusion model under the assumptions stated below. As discussed in Bruno et al. (2025, Section 2 and Appendix A), it may be necessary to restrict $t \in [\epsilon, T]$ for $\epsilon \in (0,1)$ in 5 to account for numerical instabilities that can arise during training and sampling near $t = 0$ as also observed in practice in Song et al. (2021, Appendix C), and for the possibility that the integral of the score function in 5 may diverge when $t = 0$. Therefore, we truncate the integration in the backward diffusion at $T - \epsilon$ and consider the process $(Y_t)_{t\in[0,T-\epsilon]}$.

3.1 Assumptions

We begin by stating the main assumptions of our setting. The optimization problem in 5 can be solved using algorithms such as stochastic gradient descent (Jentzen et al., 2021), ADAM (Kingma & Ba, 2015), Stochastic Gradient Langevin Dynamics (Bruno et al., 2025, Section 3.1), and TheoPouLa (Lim & Sabanis, 2024), provided they satisfy the following assumption.

Assumption 1. *Let θ^* be a minimiser*[1] *of 5 and let $\hat{\theta}$ be the (random) estimator of θ^* obtained through some approximation procedure such that $\mathbb{E}[|\hat{\theta}|^2] < \infty$. There exists $\widetilde{\varepsilon}_{AL} > 0$ such that*

$$\mathbb{E}[|\hat{\theta} - \theta^*|^2] < \widetilde{\varepsilon}_{AL}.$$

Remark 1. *As a consequence of Assumption 1, one obtains $\mathbb{E}[|\hat{\theta}|^2] < 2\widetilde{\varepsilon}_{AL} + 2|\theta^*|^2$.*

[1]The score-matching optimization problem 5 is not necessarily (strongly) convex.

Recall that for any $x \in \mathbb{R}^d$, $h(x)$ denotes an element of the subdifferential $\partial U(x)$ defined in 1.

Assumption 2. *The data distribution π_{D} has a finite second moment and it is absolutely continuous with respect to the Lebesgue measure with $\pi_{\mathsf{D}}(\mathrm{d}x) = \exp(-U(x))\ \mathrm{d}x$ for some $U : \mathbb{R}^d \to \mathbb{R}$. Moreover,*

(i) The potential U is continuous and its gradient exists almost everywhere.

(ii) The potential U is K-semiconvex. That is, there exists $K, R \geq 0$, such that for all $x, \bar{x} \in \mathbb{R}^d$,

$$\langle h(x) - h(\bar{x}), x - \bar{x} \rangle \geq -K|x - \bar{x}|^2, \qquad \textit{when} \quad |x - \bar{x}| < R,$$

or equivalently $U + \frac{K}{2}|\cdot|^2$ is convex.

(iii) The potential U is μ-strongly convex at infinity. That is, there exists $\mu > 0$ and $R \geq 0$, such that for all $x, \bar{x} \in \mathbb{R}^d$,

$$\langle h(x) - h(\bar{x}), x - \bar{x} \rangle \geq \mu|x - \bar{x}|^2, \qquad \textit{when} \quad |x - \bar{x}| \geq R. \tag{9}$$

Remark 2. *As a consequence of Proposition 17, due to Conforti et al. (2025, Proposition 3.1), and Assumption 2-(i), $\nabla \log p_t(x)$ is continuous for $t \in (0, T]$ and $x \in \mathbb{R}^d$. Moreover, Assumption 2 implies that the processes in 3 and 4 have a unique strong solution.*

Next, we consider the following assumption on the approximating function s.

Assumption 3.a. The function $s : [0,T] \times \mathbb{R}^M \times \mathbb{R}^d \to \mathbb{R}^d$ is continuously differentiable in $x \in \mathbb{R}^d$. Let $D_1 : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}_+$, $D_2 : [0,T] \times [0,T] \to \mathbb{R}_+$ and $D_3 : [0,T] \times [0,T] \to \mathbb{R}_+$ be such that $\int_\epsilon^T \int_\epsilon^T D_2(t, \bar{t})\ \mathrm{d}t\ \mathrm{d}\bar{t} < \infty$ and $\int_\epsilon^T \int_\epsilon^T D_3(t, \bar{t})\ \mathrm{d}t\ \mathrm{d}\bar{t} < \infty$. For $\alpha \in \left[\frac{1}{2}, 1\right]$ and for all $t, \bar{t} \in [0,T]$, $x, \bar{x} \in \mathbb{R}^d$, and $\theta, \bar{\theta} \in \mathbb{R}^M$, we have that

$$|s(t, \theta, x) - s(\bar{t}, \bar{\theta}, \bar{x})| \leq D_1(\theta, \bar{\theta})|t - \bar{t}|^\alpha + D_2(t, \bar{t})|\theta - \bar{\theta}| + D_3(t, \bar{t})|x - \bar{x}|,$$

where D_1, D_2 and D_3 have the following growth in each variable: i.e., there exist K_1, K_2, and $\mathsf{K}_3 > 0$ such that for each $t, \bar{t} \in [0,T]$ and $\theta, \bar{\theta} \in \mathbb{R}^M$,

$$|D_1(\theta, \bar{\theta})| \leq \mathsf{K}_1(1 + |\theta| + |\bar{\theta}|), \qquad |D_2(t, \bar{t})| \leq \mathsf{K}_2(1 + |t|^\alpha + |\bar{t}|^\alpha),$$
$$|D_3(t, \bar{t})| \leq \mathsf{K}_3(1 + |t|^\alpha + |\bar{t}|^\alpha).$$

Remark 3. *Assumption 3.a implies that the process in 6, 7, and 8 have a unique strong solution. For a discussion on the practical justification of this assumption in the context of neural network-based approximations, we refer the reader to Bruno et al. (2025, Remark 6).*

Remark 4. *Let $\mathsf{K}_{Total} := \mathsf{K}_1 + \mathsf{K}_2 + \mathsf{K}_3 + |s(0,0,0)| > 0$. Using Assumption 3.a, one obtains*

$$|s(t, \theta, x)| \leq \mathsf{K}_{Total}(1 + |t|^\alpha)(1 + |\theta| + |x|).$$

The proof of Remark 4 can be found, e.g., in Bruno et al. (2025, Appendix D.3). By imposing an additional condition on the gradient of s in Assumption 3.a, we obtain the optimal convergence rate established in Theorem 15 below.

Assumption 3.b. Let s be as in Assumption 3.a and there exists $\mathsf{K}_4 > 0$ such that, for all $x, \bar{x} \in \mathbb{R}^d$ and for any $k = 1, \ldots d$,

$$|\nabla_x s^{(k)}(t, \theta, x) - \nabla_{\bar{x}} s^{(k)}(t, \theta, \bar{x})| \leq \mathsf{K}_4(1 + 2|t|^\alpha)|x - \bar{x}|.$$

For the following assumption on the score approximation, we let $\hat{\theta}$ be as in Assumption 1 and we let $(Y_t^{\mathrm{aux}})_{t \in [0,T]}$ be the auxiliary process defined in 6.

Assumption 4. *There exists $\varepsilon_{SN} > 0$ such that*

$$\mathbb{E} \int_0^{T-\epsilon} |\nabla \log p_{T-r}(Y_r^{aux}) - s(T - r, \hat{\theta}, Y_r^{aux})|^2\ dr < \varepsilon_{SN}. \tag{10}$$

Remark 5. *Assumption 4 is now a standard assumption considered in the literature, see, e.g., Gao et al. (2025); Bruno et al. (2025); Strasman et al. (2025); Gentiloni-Silveri & Ocello (2025), and its theoretical and practical soundedness is discussed, e.g., in Bruno et al. (2025, Remark 7, 8, and 9).*

3.2 Assumption 2 and Weak Convexity of the Data Distribution

We show that Assumption 2-(ii) and Assumption 2-(iii) are related to the notion of weak convexity in the sense made precise in Proposition 8 below. We start by introducing the definition of weak convexity for subgradients.

Definition 6. The potential $U : \mathbb{R}^d \to \mathbb{R}$ is weakly convex if its weak convexity profile $\kappa_U : [0, \infty) \to \mathbb{R}$ defined as

$$\kappa_U(r) = \inf_{x,\bar{x}\in\mathbb{R}^M:\ |x-\bar{x}|=r} \left\{ \frac{\langle \partial U(x) - \partial U(\bar{x}), x - \bar{x}\rangle}{|x-\bar{x}|^2} \right\} \tag{11}$$

satisfies

$$\kappa_U(r) \geq \beta - r^{-1} f_L(r), \quad \text{for all} \ \ r > 0, \tag{12}$$

for some constants $\beta, L > 0$, where the function $f_L : [0, \infty] \to [0, \infty]$ is defined as

$$f_L(r) = 2L^{1/2} \tanh((rL^{1/2})/2). \tag{13}$$

We modify Conforti et al. (2023, Lemma 5.9) to our setting, namely when $\beta > 0$[2] to have an explicit expression of the weak convexity constant at each $t \in [0, T]$.

Lemma 7. *(Conforti et al., 2023, Modification of Lemma 5.9) Assume that U is weakly convex as in Definition 6 and fix $t \in [0, T]$. Then, the function $x \mapsto -\log p_t(x)$ is weakly convex with weak convexity profile $\kappa_{-\log p_t(x)}$ satisfying*

$$\kappa_{-\log p_t}(r) \geq \frac{\beta}{\beta + (1-\beta)e^{-2t}} - \frac{e^{-t}}{\beta + (1-\beta)e^{-2t}} \frac{1}{r} f_L \left(\frac{e^{-t}}{\beta + (1-\beta)e^{-2t}} r \right).$$

In particular, the score function satisfies

$$\langle \nabla \log p_t(x) - \nabla \log p_t(\bar{x}), x - \bar{x} \rangle \leq -\widehat{C}_t |x - \bar{x}|^2, \quad \textit{for } x, \bar{x} \in \mathbb{R}^d, \tag{14}$$

with

$$\widehat{C}_t = \frac{\beta}{\beta + (1-\beta)e^{-2t}} - \frac{e^{-2t}}{(\beta + (1-\beta)e^{-2t})^2} L. \tag{15}$$

An overview of the proof of Proposition 8 below can be found in Appendix B.

Proposition 8. *Let the data distribution π_{D} be in Assumption 2 and let f_L be as in 13. Then*

$$\kappa_U(r) \geq \mu - r^{-1} f_L(r), \quad \textit{for all } r > 0, \tag{16}$$

where $\mu > 0$ is the strong convexity at infinity constant from Assumption 2-(iii). Conversely, if U is weakly convex as in Definition 6 with lower bound 16 for some known constants μ and $L > 0$, then

1. *The potential U is $\widetilde{\mu}$-strongly convex at infinity with $\widetilde{\mu} := \mu - R^{-1} f_L(R) > 0$, such that for all $x, \bar{x} \in \mathbb{R}^d$, we have*

   $$\langle h(x) - h(\bar{x}), x - \bar{x} \rangle \geq \widetilde{\mu} |x - \bar{x}|^2, \qquad \textit{when} \quad |x - \bar{x}| \geq R, \tag{17}$$

   which holds for all $R > 0$ when $\mu > L$ and for $R \geq R_0 = \frac{2z_0}{L^{1/2}}$ with z_0 being the solution of 43 when $\mu \leq L$.

2. *The potential U is K-semiconvex, such that there exists $K, R \geq 0$ for all $x, \bar{x} \in \mathbb{R}^d$,*

   $$\langle h(x) - h(\bar{x}), x - \bar{x} \rangle \geq -K |x - \bar{x}|^2, \qquad \textit{when} \quad |x - \bar{x}| \leq R. \tag{18}$$

As a consequence of Proposition 8 and Lemma 7, one obtains the explicit form of $\widehat{C}_t$ in 14 in our setting, which is given in the following corollary.

[2]See Gentiloni-Silveri & Ocello (2025, Lemma B.4) for a similar statement.

Corollary 9. *Let U be K-semiconvex as in Assumption 2-(ii) and be μ-strongly convex at infinity as in Assumption 2-(iii) and fix $t \in [0,T]$. Then*

$$\langle \nabla \log p_t(x) - \nabla \log p_t(\bar{x}), x - \bar{x} \rangle \le -\beta_t^{OS} |x - \bar{x}|^2, \qquad \text{for} \quad x, \bar{x} \in \mathbb{R}^d, \tag{19}$$

where

$$\beta_t^{OS} = \frac{\mu}{\mu + (1-\mu)e^{-2t}} - \frac{e^{-2t}}{(\mu + (1-\mu)e^{-2t})^2} L, \tag{20}$$

for some $L > 0$ satisfying 16.

Remark 10. *By Corollary 9 and the proof of Proposition 8, we have*

$$\lim_{t \to 0} \beta_t^{OS} = \mu - L < -K. \tag{21}$$

We emphasize that the gap between the limit on the left-hand side of 21 and the semiconvexity constant K is due to the particular choice of f_L in 13 in Proposition 8. This gap may vanish for different functions $f \in \widetilde{\mathcal{F}}$, where

$$\widetilde{\mathcal{F}} := \Bigg\{ f \in C^2((0,\infty), \mathbb{R}_+) : \ r \mapsto r^{1/2} f(r^{1/2}), \ \text{non-decreasing, concave, bounded such that} \\ \lim_{r \downarrow 0} r f(r) = 0, \ f^{'} \ge 0, \quad 2f^{''} + f f^{'} \le 0 \Bigg\}.$$

For this reason, we use the constant $K + \mu$ as a proxy of the constant L and replace 20 with the following monotonicity bound

$$\beta_t^{OS,K,\mu} = \frac{\mu}{\mu + (1-\mu)e^{-2t}} - \frac{e^{-2t}}{(\mu + (1-\mu)e^{-2t})^2} (K + \mu). \tag{22}$$

Moreover, it holds that

$$\lim_{t \to 0} \beta_t^{OS,K,\mu} = -K,$$

and

$$\lim_{t \to \infty} \beta_t^{OS} = \lim_{t \to \infty} \beta_t^{OS,K,\mu} = 1,$$

which is consistent with $\pi_\infty \sim \mathcal{N}(0, I_d)$, the invariant distribution of the OU process.

Using the explicit expression of 22, we are able to find a time for which the integral of the monotonicity bound[3] $\beta_t^{\mathrm{OS},K,\mu}$ is positive. The proof of the following result is postponed to Appendix B.

Proposition 11. *Let $\mu > 0$ and $K \ge 0$. The time integral of $\beta_t^{OS,K,\mu}$ from Remark 10 is*

$$\begin{aligned} B(t, 0, \mu, K) &= \int_0^t \left(\frac{\mu}{\mu + (1-\mu)e^{-2s}} - \frac{e^{-2s}}{(\mu + (1-\mu)e^{-2s})^2} (K + \mu) \right) \, \mathrm{d}s \\ &= \frac{1}{2} \left[\log\left(\mu(e^{2t} - 1) + 1 \right) + \left(\frac{K}{\mu} + 1 \right) \left(\frac{1}{\mu(e^{2t} - 1) + 1} - 1 \right) \right] > 0, \end{aligned} \tag{23}$$

when $t > t^\star > \ln\left(\sqrt{1 + \frac{K}{\mu^2}} \right)$ with $t^\star := \inf\{t > 0 : B(t, 0, \mu, K) > 0\}$.

Remark 12. *If we consider the case when $K = 0$ in Assumption 2-(ii), then 23 is satisfied for all $t > 0$.*

3.3 Main Results - Optimal Data Dimensional Dependence and Rate of Convergence

The main results are stated as follows. An overview of their proofs can be found in Appendix C.

[3] Note that $\beta_t^{\mathrm{OS},K,\mu}$ is a function of time.

Theorem 13. *Let Assumptions 1, 2, 3.a and 4 hold. Then, there exist constants C_1, C_2, C_3 and $C_4 > 0$ such that for any $T > 0$ and $\gamma, \epsilon \in (0,1)$,*

$$W_2(\mathcal{L}(Y_K^{EM}), \pi_{\mathsf{D}}) \le C_1\sqrt{\epsilon} + C_2 e^{-2\int_\epsilon^T \beta_t^{OS,K,\mu}\ \mathrm{d}t - \epsilon} + C_3(T,\epsilon)\sqrt{\varepsilon_{SN}} + C_4(T,\epsilon)\gamma^{1/2}, \tag{24}$$

where C_1, C_2, C_3 and C_4 are given explicitly in Table 2 (Appendix E), $\beta_t^{OS,K,\mu}$ is defined in 22, and its integral is computed in Proposition 11. In addition, the result in 24 implies that for any $\delta > 0$, if we choose $0 < \epsilon < \epsilon_\delta$, $T > T_\delta$, $0 < \varepsilon_{SN} < \varepsilon_{SN,\delta}$ and $0 < \gamma < \gamma_\delta$ with ϵ_δ, T_δ, $\varepsilon_{SN,\delta}$, and γ_δ given in Table 2, then

$$W_2(\mathcal{L}(Y_K^{EM}), \pi_{\mathsf{D}}) < \ \delta.$$

Remark 14. *The constant $C_4(T,\epsilon)$ in the error bound in 24 contains the optimal dependence of the data dimension, i.e. $O(\sqrt{d})$, which has been found under the more strict assumption of strong-log concavity of π_{D} in Bruno et al. (2025, Theorem 1 and Remark 12). However, the optimal dependence of the dimension is achieved at the expenses of a worst rate of covergence of order* $1/2$.

The optimal rate of convergence of order $\alpha \in [\frac{1}{2}, 1]$ for the Euler or Milstein scheme of SDEs with constant diffusion coefficients can be attained in Theorem 13 provided that $\mathbb{E}[|\hat{\theta}|^4] < \infty$ and that Assumption 3.a is replaced by Assumption 3.b, as stated in Theorem 15 below.

Theorem 15. *Let Assumptions 1, 2, 3.b and 4 hold, and assume that $\mathbb{E}[|\hat{\theta}|^4] < \infty$ Then, there exist constants C_1, C_2, C_3 and $\widetilde{C}_4 > 0$ such that for any $T > 0$ and $\gamma, \epsilon \in (0,1)$,*

$$W_2(\mathcal{L}(Y_K^{EM}), \pi_{\mathsf{D}}) \le C_1\sqrt{\epsilon} + C_2 e^{-2\int_\epsilon^T \beta_t^{OS,K,\mu}\ \mathrm{d}t - \epsilon} + C_3(T,\epsilon)\sqrt{\varepsilon_{SN}} + \widetilde{C}_4(T,\epsilon)\gamma^{\alpha}, \tag{25}$$

where C_1, C_2, C_3 and $\widetilde{C}_4$ are given explicitly in Table 2 (Appendix E), $\beta_t^{OS,K,\mu}$ is defined in 22, and its integral is computed in Proposition 11. In addition, the result in 25 implies that for any $\delta > 0$, if we choose $0 < \epsilon < \epsilon_\delta$, $T > T_\delta$, $0 < \varepsilon_{SN} < \varepsilon_{SN,\delta}$ and $0 < \gamma < \widetilde{\gamma}_\delta$ with ϵ_δ, T_δ, $\varepsilon_{SN,\delta}$, and $\widetilde{\gamma}_\delta$ given in Table 2, then

$$W_2(\mathcal{L}(Y_K^{EM}), \pi_{\mathsf{D}}) < \ \delta. \tag{26}$$

Remark 16. *The explicit expression of $\widetilde{C}_4$ in Table 2 (Appendix E) exhibits an $O(d)$ dependence of the data dimension, resulting from numerical techniques introduced in Kumar & Sabanis (2019) and employed in the proof of Theorem 15 to achieve the optimal convergence rate of order $\alpha \in [\frac{1}{2}, 1]$.*

3.4 Examples of potentials satisfying by Assumption 2

We present several examples to demonstrate the wide applicability of our Assumption 2 to a broad class of data distributions, some of which are not covered by previous results in Wasserstein distance of order two (Gentiloni-Silveri & Ocello, 2025; Strasman et al., 2025; Gao et al., 2025; Bruno et al., 2025; Tang & Zhao, 2024; Yu & Yu, 2025).

3.4.1 Symmetric modified half-normal distribution

We consider the case of a one-dimensional symmetric modified half-normal distribution

$$\pi_{\mathsf{D}}(\mathrm{d}x) = \frac{\sqrt{\xi}\exp\left(-\xi x^2 - |x|\right)}{\Psi\left(\frac{1}{2}, \frac{-1}{\sqrt{\xi}}\right)}\ \mathrm{d}x, \quad x \in \mathbb{R}, \tag{27}$$

for some unknown $\xi > 0$ and normalizing constant

$$\Psi\left(\frac{1}{2}, \frac{-1}{\sqrt{\xi}}\right) := \sum_{n=0}^{\infty} \frac{\Gamma\left(\frac{1}{2} + \frac{n}{2}\right)}{\Gamma(n)} \frac{(-1)^n \xi^{-n/2}}{n!},$$

where $\Gamma(n)$ is the Gamma function. We refer the reader to Appendix D for additional details about the derivation of 27. As highlighted in Sun et al. (2023, Section 2), the modified half-normal distribution appears in several Bayesian statistical methods as a posterior distribution to sample from in Bayesian Binary regression, analysis of directional data, and Bayesian graphical models.

Assumption 2-(i) is satisfied for $U(x) = \xi x^2 + |x|$. In addition, we have, for all $x, \bar{x} \in \mathbb{R}$

$$\begin{aligned} \langle h(x) - h(\bar{x}), x - \bar{x} \rangle &= 2\left(\xi|x - \bar{x}|^2 + (x - \bar{x})\mathbb{1}_{x>0,\ \bar{x}<0} - (x - \bar{x})\mathbb{1}_{x<0,\ \bar{x}>0}\right) \\ &\geq 2\xi|x - \bar{x}|^2, \end{aligned} \tag{28}$$

which shows that Assumption 2-(ii) is verified for any $K \geq 0$, and Assumption 2-(iii) is verified for $\mu = 2\xi$. Therefore, we can conclude that 27 satisfies Assumption 2.

3.4.2 Multidimensional Gaussian mixture distribution

We consider a multidimensional Gaussian mixture data distribution with unknown mean and variance, i.e.,

$$\pi_{\mathsf{D}}(\mathrm{d}x) = \sum_{j=1}^{J} \widetilde{\xi}_j \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left(-\frac{|x - \eta_j|^2}{2\sigma_j^2}\right) \mathrm{d}x, \quad x \in \mathbb{R}^d, \tag{29}$$

with $\sigma_j > 0$, $\eta_j \in \mathbb{R}^d$, and $\widetilde{\xi}_j \in [0, 1]$ for $j \in \{1, \ldots, J\}$ such that $\sum_{j=1}^{J} \widetilde{\xi}_j = 1$. The authors in Gentiloni-Silveri & Ocello (2025, Appendix A) show that the score function of 29 is Lipschitz continuous and $-\log \pi_{\mathsf{D}}$ is weakly convex. Therefore, Assumption 2 is satisfied. In addition, the distribution 29 covers also case of the double-well potential:

$$U(x) = x^4 - |x|^2, \quad x \in \mathbb{R}^d, \tag{30}$$

which is 2-semiconvex and strongly convex at infinity.

3.4.3 Multi-dimensional Potentials

Similarly as in Section 3.4.1, one can proves that the elastic net potential:

$$U(x) = |x|^2 + \sum_{i=1}^{d} |x_i|, \quad x \in \mathbb{R}^d, \tag{31}$$

satisfies Assumption 2. Moreover, the following potential

$$U(x) = \max\left\{|x|, |x|^2\right\}, \qquad x \in \mathbb{R}^d, \tag{32}$$

verifies Assumption 2 with $K = 0$, $R = 1$, and $\mu = 2$ as well as the following non-convex potential presented in Johnston et al. (2025, Example 4.2):

$$U(x) = \max\left\{|x|, |x|^2\right\} - \frac{1}{2}|x|^2, \qquad x \in \mathbb{R}^d. \tag{33}$$

4 Related Work and Comparison

In recent years, there has been a rapidly expanding body of research on the convergence theory of Score-based Generative Models. Existing works for convergence bounds can be divided into two main approaches, depending on the divergence or distance used.

The first approach focuses on α-divergences, particularly the Kullback–Leibler (KL) divergence and Total Variation (TV) distance (e.g., Benton et al. (2024); Conforti et al. (2025); Yang & Wibisono (2023); Li & Cai (2024); Block et al. (2020); De Bortoli et al. (2021); Lee et al. (2022); Li et al. (2024); Lee et al. (2023); Chen et al. (2023a;b); Oko et al. (2023); Liang et al. (2025); Yang & Wibisono (2022)), which are the vast majority of the results available in the literature. Crucially, bounds on KL divergence imply bounds on TV distance via Pinsker's inequality, strengthening their wide applicability. We provide a brief and selective overview of some of the findings following this first approach. The results in TV distance in Lee et al. (2022)

and in KL divergence Yang & Wibisono (2023) established convergence bounds characterized by polynomial complexity under the assumption that the data distribution satisfies a logarithmic Sobolev inequality and that the score function is Lipschitz continuous. By replacing the requirement that the data distribution satisfies a functional inequality with the assumption that π_{D} has finite KL divergence with respect to the standard Gaussian and by assuming that the score function for the forward process is Lipschitz, the authors in Chen et al. (2023b) managed to derive bounds in TV distance which scale polynomially in all the problem parameters. By requiring only the Lipschitzness of the score at the initial time rather than along the full trajectory, the authors in Chen et al. (2023a, Theorem 2.5) managed to establish, using an exponentially decreasing then linear step size, convergence bounds in KL divergence with quadratic dimensional dependence and logarithmic complexity in the Lipschitz constant. Later, Benton et al. (2024) provided KL convergence bounds that are linear in the data dimension, up to logarithmic factors, by assuming finite second moments of the data distribution and employing early stopping. However, both the results of Chen et al. (2023a, Theorem 2.5) and Benton et al. (2024, Theorem 1 and Corollary 1) still require the uniqueness of solutions for the backward SDE 3, and therefore additional assumptions on the score function are needed. For further discussion on this point, we refer the reader to Bruno et al. (2025, Section 4.2). Assuming finite second moments and using an exponential integrator (EI) scheme with both constant and exponentially decaying step sizes, the authors in Conforti et al. (2025, Corollary 2.4) derive a KL divergence bound with early stopping, which scales linearly in the data dimension up to logarithmic factors. Bounds in KL without early stopping have been derived in Conforti et al. (2025) for data distributions with finite Fisher information with respect to the standard Gaussian distribution. We note that this condition on π_{D} stated in Conforti et al. (2025, Assumption H2) still requires that the potential $U \in C^1(\mathbb{R}^d)$. The KL bounds provided in Conforti et al. (2025, Theorem 2.1 and 2.2) scale linearly in the Fisher information when an EI discretization scheme with constant step size is used, and logarithmically in the Fisher information when an exponential-then-constant step size Conforti et al. (2025, Theorem 2.3) is employed.

The second approach focuses on convergence bounds in Wasserstein distance, a metric which is often considered more practical and informative for estimation tasks. We can relate results following this approach with the results of the first approach only when π_{D} is a strongly log-concave distribution. In this case, W_2-bounds in terms of KL divergence follow from an extension of Talagrand's inequality (Gozlan & Léonard, 2010, Corollary 7.2). However, for two general data distributions, there is no known relationship between their KL divergence and their W_2. Therefore, we cannot compare our findings in Theorem 13 and Theorem 15 with the results derived following the first approach. One line of work within the second approach assumes (at least) strong log-concavity of the data distribution (Strasman et al., 2025; Gao et al., 2025; Bruno et al., 2025; Tang & Zhao, 2024; Yu & Yu, 2025). Under this (strict) assumption, Bruno et al. (2025, Remark 12) achieved optimal data dimensional dependence, i.e., reaching $O(\sqrt{d})$. The recent bound in Gentiloni-Silveri & Ocello (2025, Theorem D.1) scales linearly in the data dimension while relaxing the strong log-concavity assumption on π_{D} to weakly log-concavity, but still requiring that the potential $\nabla^2 U$ exists (see, e.g., Gentiloni-Silveri & Ocello (2025, Proof of Proposition B.1 and B.2)). Our Assumption 2 is much weaker than this requirement and it allows to consider the case of potentials with discontinuous gradients covering a wider range of distributions as outlined in Section 3.4. Another line of work following this approach focuses on specific structural assumptions of the data distribution. For instance, convergence bounds in Wasserstein distance of order one with exponential dependence on the problem parameters have been obtained in Bortoli (2022) under the so-called manifold hypothesis, namely assuming that the target distribution is supported on a lower-dimensional manifold or is given by some empirical distribution. Under the same metric, the authors in Mimikos-Stamatopoulos et al. (2024) provide a convergence analysis when the data distribution is defined on a torus. We summarize in Table 1 and the best results obtained in W_2, i.e., Bruno et al. (2025); Gentiloni-Silveri & Ocello (2025) and compare with our best result, which scale polynomially in the data dimension, i.e. $O(\sqrt{d})$ in Theorem 13.

We close this section by briefly commenting on the choice of deriving our results in Wasserstein distance of order two. Beyond its theoretical relevance, this choice is motivated by practical considerations in generative modeling. First, the Wasserstein distance is often regarded as a more informative and robust metric for estimation tasks. Second, a widely used performance metric for evaluating the quality of images produced by generative models is the Fréchet Inception Distance (FID) Heusel et al. (2017), which measures the Fréchet distance between the distributions of generated and real samples, assuming Gaussian distributions.

In particular, this Fréchet distance is equivalent to the Wasserstein-2 distance. Thus, providing convergence results under the Wasserstein-2 metric enhances the practical relevance of our theoretical findings.

Table 1: Summary of previous bounds for $W_2(\mathcal{L}(\widehat{Y}_K^{\mathrm{EM}}), \pi_{\mathsf{D}})$ and our result in Theorem 13. All the bounds assume that $\pi_{\mathsf{D}}(\mathrm{d}x) \propto e^{-U(x)}\mathrm{d}x$ has finite second moments.

Assumption on π_{D}	Error bound	Reference
U strongly convex, $\nabla \log p_t(0) \in L^2([\epsilon, T])$, and Assumption 4	$O(\sqrt{d})\sqrt{\epsilon} + O(\sqrt{d})e^{-2\widehat{L}_{\mathrm{MO}}(T-\epsilon)-\epsilon} + O(e^{(1+\zeta-2\widehat{L}_{\mathrm{MO}})(T-\epsilon)})\sqrt{\varepsilon_{\mathrm{SN}}} + O(\sqrt{d}e^{T^{2\alpha+1}}T^{2\alpha+1}\widetilde{\varepsilon}_{\mathrm{AL}}^{1/2})\gamma^{1/2}$, with $\widehat{L}_{\mathrm{MO}} > 0$ lower bound of the strongly convex constant of U, see e.g., Bruno et al. (2025, Remark 4).	(Bruno et al., 2025, Remark 12)
$U \in C^2(\mathbb{R}^d)$, weakly convex, and Assumption 4	$e^{(2L_U+5)\eta(\beta,L,\gamma)}[e^{-T}W_2(\pi_{\mathsf{D}}, \pi_\infty) + 4\varepsilon_{\mathrm{SN}}(T-\eta(\beta,L,0)) + \sqrt{2\gamma}(4L_U d + 6d + \sqrt{d+\mathbb{E}[\lvert X_0\rvert^2]})(T-\eta(\beta,L,0))]$, with $L_U \geq 0$ one-sided Lipschitz constant for ∇U, see e.g., Gentiloni-Silveri & Ocello (2025, Assumption H1), $\eta(\beta,L,\gamma)$ defined in (Gentiloni-Silveri & Ocello, 2025, equation (29)), and $\gamma < 2/(2L_U+5)^2$.	Gentiloni-Silveri & Ocello (2025, Theorem D.1)
Assumption 2 and Assumption 4	$O(\sqrt{d})\sqrt{\epsilon} + O(\sqrt{d})e^{-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu}\,\mathrm{d}t-\epsilon} + O(e^{(1+\zeta)(T-\epsilon)-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu}\,\mathrm{d}t})\sqrt{\varepsilon_{\mathrm{SN}}} + O(\sqrt{d}e^{T^{2\alpha+1}}T^{3\alpha+1}\widetilde{\varepsilon}_{\mathrm{AL}}^{1/2})\gamma^{1/2}$.	Theorem 13

References

Juan Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18187–18197, 2021.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 843–852, 2023.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Stefano Bruno, Ying Zhang, Dongyoung Lim, Omer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, volume 59, pp. 1844–1881, 2023.

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

Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024.

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

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.

Hyungjin Chung and Jong-Chul Ye. Score-based diffusion models for accelerated MRI. *Medical image analysis*, 80:102479, 2021.

Giovanni Conforti, Daniel Lacker, and Soumik Pal. Projected Langevin dynamics and a gradient flow for entropic optimal transport. *arXiv preprint arXiv:2309.08598*, 2023.

Giovanni Conforti, Alain Durmus, and Marta G Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025.

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.

Zihan Ding and Chi Jin. Consistency models as a rich and efficient policy class for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Charles Fox. The asymptotic expansion of generalized hypergeometric functions. *Proceedings of the London Mathematical Society*, 2(1):389–400, 1928.

Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54, 2025.

Marta Gentiloni-Silveri and Antonio Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in W2-distance. *arXiv preprint arXiv:2501.02298*, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Nathael Gozlan and Christian Léonard. Transport inequalities. A survey. *Markov Processes and Related Fields*, 16:635–736, 2010.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. IDQL: Implicit Q-learning as an actor-critic method with diffusion policies. *ArXiv*, abs/2304.10573, 2023.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pp. 1188–1205, 1986.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Arnulf Jentzen, Benno Kuckuck, Ariel Neufeld, and Philippe von Wurstemberger. Strong error analysis for stochastic gradient descent optimization algorithms. *IMA Journal of Numerical Analysis*, 41(1):455–492, 2021.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi S. Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems*, 2022.

Tim Johnston, Iosif Lytras, Nikolaos Makras, and Sotirios Sabanis. The performance of the unadjusted Langevin algorithm without smoothness assumptions. *arXiv preprint arXiv:2502.03458*, 2025.

Gwanghyun Kim, Taesung Kwon, and Jong-Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2416–2425, 2021.

Diederik Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *The Eleventh International Conference on Learning Representations*, 2014.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2020.

Chaman Kumar and Sotirios Sabanis. On Milstein approximations with varying coefficients: The case of super-linear diffusion coefficients. *BIT Numerical Mathematics*, 59(4):929–968, 2019.

Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the wasserstein distance. *Advances in Neural Information Processing Systems*, 35:20205–20217, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

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

Gen Li and Changxiao Cai. Provable acceleration for diffusion models under minimal assumptions. *arXiv preprint arXiv:2410.23285*, 2024.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in Total Variation. *arXiv preprint arXiv:2501.12982*, 2025.

Dong-Young Lim and Sotirios Sabanis. Polygonal unadjusted Langevin algorithms: Creating stable and efficient adaptive algorithms for neural networks. *Journal of Machine Learning Research*, 25(53):1–52, 2024.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Seo Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Nikiforos Mimikos-Stamatopoulos, Benjamin Zhang, and Markos Katsoulakis. Score-based generative models are provably robust: An uncertainty quantification perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 26517–26582, 2023.

Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, and Sam Devlin. Imitating human behaviour with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.

Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric R. Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, C. Karen Liu, Lingjie Liu, Ben Mildenhall, Matthias Nießner, Bjorn Ommer, Christian Theobalt, Peter Wonka, and Gordon Wetzstein. State of the art on diffusion models for visual computing. *Computer Graphics Forum*, 43, 2023.

Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. In *Robotics: Science and Systems*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022.

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, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 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*, 2021.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022.

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, and Vincent Lemaire. An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Jingchao Sun, Maiying Kong, and Subhadip Pal. The Modified-Half-Normal distribution: Properties and an efficient sampling scheme. *Communications in Statistics-Theory and Methods*, 52(5):1591–1613, 2023.

Wenpin Tang and Hanyang Zhao. Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*, 2024.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Neural Information Processing Systems*, 2021. URL `https://api.semanticscholar.org/CorpusID:235765577`.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, T. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089 – 1100, 2023.

Tomer Weiss, Eduardo Mayo Yanes, Sabyasachi Chakraborty, Luca Cosmo, Alex M. Bronstein, and Renana Gershoni-Poranne. Guided diffusion for inverse molecular design. *Nature Computational Science*, 3:873 – 882, 2023.

E Maitland Wright. The asymptotic expansion of the generalized hypergeometric function. *Journal of the London Mathematical Society*, 1(4):286–293, 1935.

Kaylee Yingxi Yang and Andre Wibisono. Convergence in KL and Rényi divergence of the Unadjusted Langevin Algorithm using estimated score. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Kaylee Yingxi Yang and Andre Wibisono. Convergence of the inexact Langevin algorithm and score-based generative models in KL divergence. *arXiv preprint arXiv:2211.01512*, 2023.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruigi Gao, Yixin Zhu, Song-Chun Zhu, and Ying Nian Wu. Latent Diffusion energy-based model for interpretable text modeling. In *International Conference on Machine Learning*, 2022.

Yifeng Yu and Lu Yu. Advancing Wasserstein convergence analysis of Score-based Models: Insights from discretization and second-order acceleration. *arXiv preprint arXiv:2502.04849*, 2025.

Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In-So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative AI. *ArXiv*, abs/2303.13336, 2023.

Zhengbang Zhu, Hanye Zhao, Haoran He, Yichao Zhong, Shenyu Zhang, Yong Yu, and Weinan Zhang. Diffusion models for reinforcement learning: A survey. *ArXiv*, abs/2311.01223, 2023.

Appendix

A Regularity of the Score Function

We recall the following result to justify the smoothness of the map

$$(0,T] \times \mathbb{R}^d \ni (t,x) \mapsto p_t(x) \in \mathbb{R}_+, \tag{34}$$

where p_t density of the forward process defined in Section 2.

Proposition 17. *(Conforti et al., 2025, Proposition 3.1) Let π_{D} be absolutely continuous with respect to the Lebesgue measure, and denote its density by p_0. The map defined in 34 is positive and solution of the following Fokker–Planck equation on $(0,T] \times \mathbb{R}^d$:*

$$\partial_t p_t(x) - div(x\ p_t) - \Delta p_t(x) = 0, \quad \textit{for } (t,x) \in (0,T] \times \mathbb{R}^d.$$

Moreover, it belongs to $C^{1,2}((0,T]\times\mathbb{R}^d)$; i.e. for any $t \in (0,T)$, $x \mapsto p_t(x)$ is twice continuously differentiable, and for any $x \in \mathbb{R}^d$, $t \mapsto p_t(x)$ is continuously differentiable on $(0,T]$.

B Further Details on Assumption 2 and Weak Convexity of the Data Distribution

We provide the proofs of Section 3.2.

Proof of Proposition 8. We begin by considering that π_{D} satisfies Assumption 2. Recall that f_L is defined as in 13. Note that $r \mapsto r^{-1} f_L(r)$ is non-increasing on $(0, \infty)$ and $f_L^{'}(0) = L > r^{-1} f_L(r)$ for $r \in (0, R]$. We look for $L > 0$ satisfying

$$\inf_{r \in (0,R]} r^{-1} f_L(r) = R^{-1} f_L(R) = 2R^{-1} L^{1/2} \tanh((RL^{1/2})/2) = K + \mu. \tag{35}$$

Equivalently, we look for $x = L^{1/2} R/2 > 0$ such that

$$x \tanh(x) = \frac{K+\mu}{4} R^2, \quad \text{subject to} \quad x > \frac{\sqrt{K+\mu}}{2} R, \tag{36}$$

so as $L > K + \mu$. Note that $\tanh(x) \leq x$ for all $x \geq 0$. Therefore, if we choose $x = \frac{\sqrt{K+\mu}}{2} R$, then

$$\frac{\sqrt{K+\mu}}{2} R \tanh\left(\frac{\sqrt{K+\mu}}{2} R\right) \leq \frac{K+\mu}{4} R^2. \tag{37}$$

Using 37 and $\lim_{x \uparrow \infty} x \tanh(x) = \infty$, we deduce that there exists $x^\star > 0$ such that

$$x^\star \tanh(x^\star) = \frac{K+\mu}{4} R^2, \tag{38}$$

with $x^\star > \frac{\sqrt{K+\mu}}{2} R$, since $x \mapsto x \tanh(x)$ is non-decreasing on $(0, \infty)$. By Assumption 2 and 35, we have

$$\begin{aligned} k_U(r) &\geq \mu - (K + \mu) \\ &\geq \mu - r^{-1} f_L(r), \qquad \text{for } r \leq R. \end{aligned} \tag{39}$$

Moreover,

$$\begin{aligned} k_U(r) &\geq \mu \\ &\geq \mu - r^{-1} f_L(r), \qquad \text{for } r > R, \end{aligned}$$

where it is used that $r^{-1} f_L(r) > 0$ for all $r > 0$. This proves the first part of the statement in Proposition 8, i.e. the lower bound 16.

Conversely, assume that U is weakly convex as in Definition 6 with lower bound 16 for some known constants μ and $L > 0$. We look for R such that

$$\begin{aligned} \kappa_U(r) &\geq \mu - r^{-1} f_L(r) \\ &\geq \mu - R^{-1} f_L(R) \\ &> 0, \qquad\qquad \forall\, r > R, \end{aligned} \tag{40}$$

where it is used that $r^{-1} f_L(r)$ is decreasing on $(0, \infty)$. Let $\widetilde{\mu} := \mu - R^{-1} f_L(R)$, so 40 becomes $\kappa_U(r) \geq \widetilde{\mu} > 0$, for all $r > R$. One notes that

$$\widetilde{\mu} = \mu - L \frac{\tanh((RL^{1/2})/2)}{(RL^{1/2})/2} > 0. \tag{41}$$

If $\mu > L$, 41 is satisfied for all $R > 0$. If $\mu \leq L$, 41 holds for $R \geq R_0$, where R_0 is the unique solution to

$$\mu = \frac{2L^{1/2}}{R} \tanh\left(\frac{RL^{1/2}}{2}\right). \tag{42}$$

Let $z = \frac{RL^{1/2}}{2}$, then $R_0 = \frac{2z_0}{L^{1/2}}$, where z_0 solves

$$\frac{\tanh(z)}{z} = \frac{\mu}{L}. \tag{43}$$

Since $\frac{\tanh(z)}{z}$ monotonically decreases from 1 to 0 as z increases, a unique $z_0 > 0$ solving 43 exists for $\mu < L$. Therefore, 40 is satisfied for $R \geq R_0 = \frac{2z_0}{L^{1/2}}$. This proves that U is $\widetilde{\mu}$-strongly convex at infinity, and therefore 17. Using the assumption that U is weakly convex as in Definition 6, one obtains that

$$\begin{aligned} \kappa_U(r) &\geq \mu - r^{-1} f_L(r) \\ &\geq \mu - L, \qquad\qquad \text{for} \quad r \leq R. \end{aligned} \tag{44}$$

We distinguish two cases for the lower bound in 44. If $\mu > L$, then $\kappa_U(r) \geq -K$ for $r \leq R$ for all $R > 0$ and $K \geq 0$. If $\mu \leq L$, then, by setting $K = L - \mu$ in 43, we have $\kappa_U(r) \geq -K$ for $r \leq R$ for all $R > 0$. This proves that U is K-semiconvex, and therefore 18. This concludes the proof for the second part of the statement in Proposition 8.

□

Proof of Proposition 11. We look for $t^\star$ satisfying

$$B(t^\star, 0, \mu, K) = \frac{1}{2}\left[\log\left(\mu(e^{2t^\star} - 1) + 1\right) + \left(\frac{K}{\mu} + 1\right)\left(\frac{1}{\mu(e^{2t^\star} - 1) + 1} - 1\right)\right] > 0. \tag{45}$$

Equivalently, we look for $x := e^{2t^\star} - 1$ such that

$$g(\mu x + 1) = \log(\mu x + 1) - \left(\frac{K}{\mu} + 1\right)\frac{\mu x}{\mu x + 1} > 0. \tag{46}$$

Note that 46 is satisfied for all $x > 0$ when $K = 0$. In addition, we have

$$\begin{aligned} \lim_{x \to 0+} g(\mu x + 1) &= 0, \\ \lim_{x \to +\infty} g(\mu x + 1) &= \infty, \end{aligned} \tag{47}$$

and

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} g(\mu x + 1) &= \frac{\mu}{\mu x + 1} - \frac{K + \mu}{(\mu x + 1)^2} \geq 0 \quad \text{when} \quad x \geq \frac{K}{\mu^2}. \\ \frac{\mathrm{d}^2}{\mathrm{d}x^2} g(\mu x + 1) &= -\frac{\mu^2}{(\mu x + 1)^2} + \frac{2(K + \mu)\mu}{(\mu x + 1)^3} \geq 0 \quad \text{when} \quad x \leq \frac{2K}{\mu^2} + \frac{1}{\mu}. \end{aligned} \tag{48}$$

By 48, the function g in 46 has a minimum at $\frac{K}{\mu^2}$ and

$$g\left(\frac{K}{\mu} + 1\right) = \log\left(\frac{K}{\mu} + 1\right) - \frac{K}{\mu} < 0,$$

for all $K, \mu > 0$. By 47 and 48, there exists $x > \frac{K}{\mu^2}$ such that 46 is strictly positive. Therefore, there exists $t^\star > \ln\left(\sqrt{1 + \frac{K}{\mu^2}}\right)$ such that 45 holds. □

C Proof of the Main Results

In this section, we present the proofs of Theorem 13 and Theorem 15. We begin by recalling an upper bound on the moments of the process $(\widehat{Y}_t^{\mathrm{EM}})_{t \in [0,T]}$ defined in 8, along with an estimate for its one-step discretization error. These results will be instrumental in the subsequent proofs.

Lemma 18. *(Bruno et al., 2025, Lemma 20) Let Assumptions 1 and 3.a hold, and suppose that* $\mathbb{E}[|\hat{\theta}|^p] < \infty$ *for any* $p \in [2,4]$*. Then, for any* $t \in [0, T-\epsilon]$*,*

$$\sup_{0 \le s \le t} \mathbb{E}\left[|\widehat{Y}_s^{EM}|^p\right] \le C_{\mathsf{EM},p}(t),$$

where

$$\begin{aligned} C_{\mathsf{EM},p}(t) &:= e^{t(3p-1-\frac{2}{p}+2^{2p-1}\mathsf{K}^p_{Total}(1+T^{\alpha p}))} \\ &\quad \times \left(\mathbb{E}\left[|\widehat{Y}_0^{EM}|^p\right] + 2^{3p-2}\mathsf{K}^p_{Total} t(1+\mathbb{E}[|\hat{\theta}|^p])(1+T^{\alpha p}) + \frac{2}{p}(pd+p(p-2))^{\frac{p}{2}} t\right), \end{aligned}$$

and K_{Total} *is defined in Remark 4.*

Lemma 19. *(Bruno et al., 2025, Lemma 21) Let Assumptions 1 and 3.a hold, and suppose that* $\mathbb{E}[|\hat{\theta}|^p] < \infty$ *for any* $p \in [2,4]$*. Then, for any* $t \in [0, T-\epsilon]$*,*

$$\mathbb{E}\left[|\widehat{Y}_t^{EM} - \widehat{Y}^{EM}_{\lfloor t/\gamma \rfloor \gamma}|^p\right] \le \gamma^{\frac{p}{2}} C_{\mathsf{EMose},p},$$

where

$$\begin{aligned} C_{\mathsf{EMose},p} &:= 2^{p-1}(C_{\mathsf{EM},p}(T) + \mathsf{K}^p_{Total}(1+T^{\alpha p})(2^{3p-2}C_{\mathsf{EM},p}(T) + 2^{4p-3}(1+\mathbb{E}[|\hat{\theta}|^p]))) \\ &\quad + (dp(p-1))^{\frac{p}{2}}, \end{aligned}$$

$C_{\mathsf{EM},p}$ *and* K_{Total} *are defined in Lemma 18 and in Remark 4, respectively.*

Proof of Theorem 13. We derive the non-asymptotic estimate for $W_2(\mathcal{L}(Y_K^{\mathrm{EM}}), \pi_{\mathsf{D}})$ using the splitting

$$\begin{aligned} W_2(\mathcal{L}(Y_K^{\mathrm{EM}}), \pi_{\mathsf{D}}) &\le W_2(\pi_{\mathsf{D}}, \mathcal{L}(Y_{t_K})) + W_2(\mathcal{L}(Y_{t_K}), \mathcal{L}(\widetilde{Y}_{t_K})) \\ &\quad + W_2(\mathcal{L}(\widetilde{Y}_{t_K}), \mathcal{L}(Y^{\mathrm{aux}}_{t_K})) + W_2(\mathcal{L}(Y^{\mathrm{aux}}_{t_K}), \mathcal{L}(Y_K^{\mathrm{EM}})). \end{aligned} \tag{49}$$

We provide upper bounds on the error made by the early stopping, i.e. $W_2(\pi_{\mathsf{D}}, \mathcal{L}(Y_{t_K}))$, the error made by approximating the initial condition of the backward process $Y_0 \sim \mathcal{L}(X_T)$ with $\widetilde{Y}_0 \sim \pi_\infty$, i.e. $W_2(\mathcal{L}(Y_{t_K}), \mathcal{L}(\widetilde{Y}_{t_K}))$, the error made by approximating the score function with s, i.e. $W_2(\mathcal{L}(\widetilde{Y}_{t_K}), \mathcal{L}(Y^{\mathrm{aux}}_{t_K}))$, and the discretisation error, i.e. $W_2(\mathcal{L}(Y^{\mathrm{aux}}_{t_K}), \mathcal{L}(Y_K^{\mathrm{EM}}))$, separately.

Upper bound on $W_2(\pi_{\mathsf{D}}, \mathcal{L}(Y_{t_K}))$**.** This bound can be established by following the same argument as in (Bruno et al., 2025, Proof of Theorem 10), which relies on the representation of the OU process

$$X_t \overset{\text{a.s.}}{=} m_t X_0 + \sigma_t Z_t, \quad m_t = e^{-t}, \quad \sigma_t^2 = 1 - e^{-2t}, \quad Z_t \sim \mathcal{N}(0, I_d), \tag{50}$$

where $\overset{\text{a.s.}}{=}$ denotes almost sure equality. Therefore, we have

$$W_2(\pi_{\mathsf{D}}, \mathcal{L}(Y_{t_K})) \le 2\sqrt{\epsilon}(\sqrt{\mathbb{E}[|X_0|^2]} + \sqrt{d}), \tag{51}$$

where $t_K = T - \epsilon$.

Upper bound on $W_2(\mathcal{L}(Y_{t_K}), \mathcal{L}(\widetilde{Y}_{t_K}))$**.** Using Itô's formula, we have, for any $t \in [0, T-\epsilon]$,

$$\begin{aligned} \mathrm{d}|Y_t - \widetilde{Y}_t|^2 &= 2\langle Y_t - \widetilde{Y}_t, Y_t + 2\nabla \log p_{T-t}(Y_t) - \widetilde{Y}_t - 2\nabla \log p_{T-t}(\widetilde{Y}_t)\rangle \ \mathrm{d}t \\ &= 2|Y_t - \widetilde{Y}_t|^2 \ \mathrm{d}t + 4\langle Y_t - \widetilde{Y}_t, \nabla \log p_{T-t}(Y_t) - \nabla \log p_{T-t}(\widetilde{Y}_t)\rangle \ \mathrm{d}t. \end{aligned} \tag{52}$$

By integrating and taking on both sides in 52, we have

$$\begin{aligned} \mathbb{E}\left[|Y_{t_K} - \widetilde{Y}_{t_K}|^2\right] &= \mathbb{E}\left[|Y_0 - \widetilde{Y}_0|^2\right] + \int_0^{t_K} 2\mathbb{E}\left[|Y_t - \widetilde{Y}_t|^2\right] \ \mathrm{d}t \\ &\quad + \int_0^{t_K} 4\mathbb{E}\left[\langle Y_t - \widetilde{Y}_t, \nabla \log p_{T-t}(Y_t) - \nabla \log p_{T-t}(\widetilde{Y}_t)\rangle\right] \ \mathrm{d}t. \end{aligned} \tag{53}$$

By integrating, taking expectations on both sides in 53, using Corollary 9, the representation 50 with $Z_T \overset{\mathrm{d}}{=} \widetilde{Y}_0$ (where $\overset{\mathrm{d}}{=}$ denotes equality in distribution), the inequality $1 - \sigma_t \le m_t$ for any $t \in [0, T]$, we have

$$
\begin{aligned}
&\mathbb{E}\left[|Y_{t_K} - \widetilde{Y}_{t_K}|^2\right] \\
&\le \mathbb{E}\left[|Y_0 - \widetilde{Y}_0|^2\right] + 2\int_0^{t_K} \mathbb{E}\left[|Y_t - \widetilde{Y}_t|^2\right]\ \mathrm{d}t - 4\int_0^{t_K} \beta^{\mathrm{OS}}_{T-t}\mathbb{E}\left[|Y_t - \widetilde{Y}_t|^2\right]\mathrm{d}t \\
&\le \mathbb{E}[|Y_0 - \widetilde{Y}_0|^2] e^{2[t_K - 2\int_0^{t_K} \beta^{\mathrm{OS}}_{T-t}\ \mathrm{d}t]} \\
&= \mathbb{E}[|m_T X_0 + (\sigma_T - 1)\widetilde{Y}_0|^2] e^{2[t_K - 2\int_0^{t_K} \beta^{\mathrm{OS}}_{T-t}\ \mathrm{d}t]} \\
&\le 2\left(\mathbb{E}[|X_0|^2] + d\right)\ e^{2[t_K - 2\int_0^{t_K} \beta^{\mathrm{OS}}_{T-t}\ \mathrm{d}t] - 2T}.
\end{aligned}
\tag{54}
$$

Using 54, Remark 10, and and $t_K = T - \epsilon$, we have

$$
\begin{aligned}
W_2(\mathcal{L}(Y_{t_K}), \mathcal{L}(\widetilde{Y}_{t_K})) &\le \sqrt{\mathbb{E}[|Y_{t_K} - \widetilde{Y}_{t_K}|^2]} \\
&\le \sqrt{2}(\sqrt{\mathbb{E}[|X_0|^2]} + \sqrt{d}) e^{-2\int_\epsilon^T \beta^{\mathrm{OS},K,\mu}_t\ \mathrm{d}t - \epsilon}.
\end{aligned}
\tag{55}
$$

Upper bound on $W_2(\mathcal{L}(\widetilde{Y}_{t_K}), \mathcal{L}(Y^{\mathbf{aux}}_{t_K}))$. Using Itô's formula, we have, for $t \in [0, T-\epsilon]$,

$$
\begin{aligned}
\mathrm{d}|\widetilde{Y}_t - Y^{\mathrm{aux}}_t|^2 &= 2\langle \widetilde{Y}_t - Y^{\mathrm{aux}}_t, \widetilde{Y}_t + 2\ \nabla \log p_{T-t}(\widetilde{Y}_t) - Y^{\mathrm{aux}}_t - 2\ s(T-t, \hat{\theta}, Y^{\mathrm{aux}}_t)\rangle\ \mathrm{d}t \\
&= 2|\widetilde{Y}_t - Y^{\mathrm{aux}}_t|^2\ \mathrm{d}t + 4\ \langle \widetilde{Y}_t - Y^{\mathrm{aux}}_t, \nabla \log p_{T-t}(\widetilde{Y}_t) - \nabla \log p_{T-t}(Y^{\mathrm{aux}}_t)\rangle\ \mathrm{d}t \\
&\quad + 4\ \langle \widetilde{Y}_t - Y^{\mathrm{aux}}_t, \nabla \log p_{T-t}(Y^{\mathrm{aux}}_t) - s(T-t, \hat{\theta}, Y^{\mathrm{aux}}_t)\rangle\ \mathrm{d}t.
\end{aligned}
\tag{56}
$$

By integrating and taking the expectation on both sides in 56, using Corollary 9, Young's inequality with $\zeta \in (0, 1)$ and Assumption 4, we have

$$
\begin{aligned}
\mathbb{E}[|\widetilde{Y}_{T-\epsilon} - Y^{\mathrm{aux}}_{T-\epsilon}|^2] &= 2\int_0^{T-\epsilon} \mathbb{E}[|\widetilde{Y}_s - Y^{\mathrm{aux}}_s|^2]\ \mathrm{d}s \\
&\quad + 4\int_0^{T-\epsilon} \mathbb{E}[\langle \widetilde{Y}_s - Y^{\mathrm{aux}}_s, \nabla \log p_{T-s}(\widetilde{Y}_s) - \nabla \log p_{T-s}(Y^{\mathrm{aux}}_s)\rangle]\ \mathrm{d}s \\
&\quad + 4\int_0^{T-\epsilon} \mathbb{E}[\langle \widetilde{Y}_s - Y^{\mathrm{aux}}_s, \nabla \log p_{T-s}(Y^{\mathrm{aux}}_s) - s(T-s, \hat{\theta}, Y^{\mathrm{aux}}_s)\rangle]\ \mathrm{d}s \\
&\le \int_0^{T-\epsilon} 2(1+\zeta)\ \mathbb{E}[|\widetilde{Y}_s - Y^{\mathrm{aux}}_s|^2]\ \mathrm{d}s \\
&\quad - 4\int_0^{t_K} \beta^{\mathrm{OS}}_{T-s}\mathbb{E}\left[|\widetilde{Y}_s - Y^{\mathrm{aux}}_s|^2\right]\mathrm{d}t + 2\zeta^{-1}\varepsilon_{\mathrm{SN}} \\
&\le 2e^{2(1+\zeta)(T-\epsilon) - 4\int_0^{t_K} \beta^{\mathrm{OS}}_{T-t}\ \mathrm{d}t}\ \zeta^{-1}\varepsilon_{\mathrm{SN}}.
\end{aligned}
\tag{57}
$$

Using 57, Remark 10, and $t_K = T - \epsilon$, we have

$$
\begin{aligned}
W_2(\mathcal{L}(\widetilde{Y}_{t_K}), \mathcal{L}(Y^{\mathrm{aux}}_{t_K})) &\le \sqrt{\mathbb{E}[|\widetilde{Y}_{t_K} - Y^{\mathrm{aux}}_{t_K}|^2]} \\
&\le \sqrt{2\zeta^{-1}} e^{(1+\zeta)(T-\epsilon) - 2\int_\epsilon^T \beta^{\mathrm{OS},K,\mu}_t\ \mathrm{d}t} \sqrt{\varepsilon_{\mathrm{SN}}}.
\end{aligned}
\tag{58}
$$

Upper bound on $W_2(\mathcal{L}(Y^{\mathbf{aux}}_{t_K}), \mathcal{L}(\widehat{Y}^{\mathbf{EM}}_t))$. Using Itô's formula, we have, for $t \in [0, T-\epsilon]$,

$$
\begin{aligned}
&\mathrm{d}|Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t|^2 \\
&= 2\langle Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t, Y^{\mathrm{aux}}_t + 2\, s(T-t, \hat{\theta}, Y^{\mathrm{aux}}_t) - \widehat{Y}^{\mathrm{EM}}_{\lfloor t/\gamma \rfloor \gamma} - 2\, s(T - \lfloor t/\gamma \rfloor \gamma, \hat{\theta}, \widehat{Y}^{\mathrm{EM}}_{\lfloor t/\gamma \rfloor \gamma})\rangle \,\mathrm{d}t \\
&= 2|Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t|^2 \,\mathrm{d}t + 2\langle Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t, \widehat{Y}^{\mathrm{EM}}_t - \widehat{Y}^{\mathrm{EM}}_{\lfloor t/\gamma \rfloor \gamma}\rangle \,\mathrm{d}t \\
&\quad + 4\langle Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t, s(T-t, \hat{\theta}, Y^{\mathrm{aux}}_t) - s(T-t, \hat{\theta}, \widehat{Y}^{\mathrm{EM}}_t)\rangle \,\mathrm{d}t \\
&\quad + 4\langle Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t, s(T-t, \hat{\theta}, \widehat{Y}^{\mathrm{EM}}_t) - s(T - \lfloor t/\gamma \rfloor \gamma, \hat{\theta}, \widehat{Y}^{\mathrm{EM}}_{\lfloor t/\gamma \rfloor \gamma})\rangle \,\mathrm{d}t.
\end{aligned} \tag{59}
$$

Integrating and taking the expectation on both sides in 59, using Young's inequality for $\zeta \in (0,1)$, Cauchy Schwarz inequality, Assumption 3.a, Lemma 19, and Remark 1, we have

$$
\begin{aligned}
\mathbb{E}\left[|Y^{\mathrm{aux}}_{T-\epsilon} - \widehat{Y}^{\mathrm{EM}}_{T-\epsilon}|^2\right] &\le (2+3\zeta)\int_0^{T-\epsilon} \mathbb{E}[|Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t|^2] \,\mathrm{d}t + \zeta^{-1}\int_0^{T-\epsilon} \mathbb{E}[|\widehat{Y}^{\mathrm{EM}}_t - \widehat{Y}^{\mathrm{EM}}_{\lfloor t/\gamma \rfloor \gamma}|^2] \,\mathrm{d}t \\
&\quad + 4\mathsf{K}_3(1+2T^\alpha)\int_0^{T-\epsilon} \mathbb{E}[|Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t|^2]\mathrm{d}t \\
&\quad + 2\zeta^{-1}\int_0^{T-\epsilon} \mathbb{E}[|s(T-t, \hat{\theta}, \widehat{Y}^{\mathrm{EM}}_t) - s(T - \lfloor t/\gamma \rfloor \gamma, \hat{\theta}, \widehat{Y}^{\mathrm{EM}}_{\lfloor t/\gamma \rfloor \gamma}|^2]\mathrm{d}t \\
&\le (2+3\zeta+4\mathsf{K}_3(1+2T^\alpha))\int_0^{T-\epsilon} \mathbb{E}[|Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t|^2] \,\mathrm{d}t \\
&\quad + \zeta^{-1}\gamma(T-\epsilon)C_{\mathsf{EMose},2} + 8\zeta^{-1}\gamma^{2\alpha}(T-\epsilon)\mathsf{K}_1^2(1+4\mathbb{E}[|\hat{\theta}|^2]) \\
&\quad + 4\zeta^{-1}\mathsf{K}_3^2(1+2T^\alpha)^2\int_0^{T-\epsilon} \mathbb{E}[|\widehat{Y}^{\mathrm{EM}}_t - \widehat{Y}^{\mathrm{EM}}_{\lfloor t/\gamma \rfloor \gamma}|^2]\mathrm{d}t \\
&\le (2+3\zeta+4\mathsf{K}_3(1+2T^\alpha))\int_0^{T-\epsilon} \mathbb{E}[|Y^{\mathrm{aux}}_t - \widehat{Y}^{\mathrm{EM}}_t|^2] \,\mathrm{d}t \\
&\quad + \zeta^{-1}\gamma(T-\epsilon)C_{\mathsf{EMose},2}(1+4\mathsf{K}_3^2(1+2T^\alpha)^2) \\
&\quad + 8\zeta^{-1}\gamma^{2\alpha}(T-\epsilon)\mathsf{K}_1^2(1+8\widetilde{\varepsilon}_{\mathrm{AL}} + 8|\theta^*|^2) \\
&\le e^{(2+3\zeta+4\mathsf{K}_3(1+2T^\alpha))(T-\epsilon)} \\
&\quad \times \Bigg(\zeta^{-1}\gamma(T-\epsilon)C_{\mathsf{EMose},2}(1+4\mathsf{K}_3^2(1+2T^\alpha)^2) \\
&\qquad + 8\zeta^{-1}\gamma^{2\alpha}(T-\epsilon)\mathsf{K}_1^2(1+8\widetilde{\varepsilon}_{\mathrm{AL}} + 8|\theta^*|^2)\Bigg).
\end{aligned} \tag{60}
$$

Using 60 and $t_K = T - \epsilon$, we have

$$
\begin{aligned}
W_2(\mathcal{L}(Y^{\mathrm{aux}}_{T-\epsilon}), \mathcal{L}(\widehat{Y}^{\mathrm{EM}}_{T-\epsilon})) &\le \gamma^{1/2}\zeta^{-1/2}(T-\epsilon)^{1/2}e^{(1+(3/2)\zeta+2\mathsf{K}_3(1+2T^\alpha))(T-\epsilon)} \\
&\quad \times (C^{1/2}_{\mathsf{EMose},2}(1+2\mathsf{K}_3(1+2T^\alpha)) + 2\sqrt{2}\mathsf{K}_1(1+8\widetilde{\varepsilon}_{\mathrm{AL}}+8|\theta^*|^2)^{1/2}).
\end{aligned} \tag{61}
$$

Final upper bound on $W_2(\mathcal{L}(Y^{\mathbf{EM}}_K), \pi_{\mathsf{D}})$. Substituting 51, 55, 58, and 61 into 49, we have

$$
\begin{aligned}
W_2(\mathcal{L}(Y^{\mathrm{EM}}_K), \pi_{\mathsf{D}}) &\le (\sqrt{\mathbb{E}[|X_0|^2]} + \sqrt{d})2\sqrt{\epsilon} \\
&\quad + \sqrt{2}(\sqrt{\mathbb{E}[|X_0|^2]} + \sqrt{d})e^{-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu} \,\mathrm{d}t - \epsilon} \\
&\quad + \sqrt{2\zeta^{-1}}e^{(1+\zeta)(T-\epsilon)-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu} \,\mathrm{d}t}\sqrt{\varepsilon_{\mathrm{SN}}} \\
&\quad + \gamma^{1/2}\zeta^{-1/2}(T-\epsilon)^{1/2}e^{(1+(3/2)\zeta+2\mathsf{K}_3(1+2T^\alpha))(T-\epsilon)} \\
&\qquad \times (C^{1/2}_{\mathsf{EMose},2}(1+2\mathsf{K}_3(1+2T^\alpha)) + 2\sqrt{2}\mathsf{K}_1(1+8\widetilde{\varepsilon}_{\mathrm{AL}}+8|\theta^*|^2)^{1/2}).
\end{aligned} \tag{62}
$$

The bound for $W_2(\mathcal{L}(\widehat{Y}_K^{\mathrm{EM}}), \pi_{\mathsf{D}})$ in 62 can be made arbitrarily small by appropriately choosing parameters including $\epsilon, T, \varepsilon_{\mathrm{SN}}$ and γ. More precisely, for any $\delta > 0$, we first choose $0 < \epsilon < \epsilon_\delta$ with ϵ_δ given in Table 2 such that the first term on the right-hand side of 62 is

$$(\sqrt{\mathbb{E}[|X_0|^2]} + \sqrt{d})2\sqrt{\epsilon} < \delta/4. \tag{63}$$

Next, we choose $T > T_\delta$ with T_δ given in Table 2 such that the second term on the right-hand side of 62 is

$$\sqrt{2}(\sqrt{\mathbb{E}[|X_0|^2]} + \sqrt{d})e^{-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu}\ \mathrm{d}t - \epsilon} < \delta/4. \tag{64}$$

Next, we turn to the third term on the right-hand side of 62. We choose $0 < \varepsilon_{\mathrm{SN}} < \varepsilon_{\mathrm{SN},\delta}$ with $\varepsilon_{\mathrm{SN},\delta}$ given in Table 2 such that

$$\sqrt{2\zeta^{-1}}e^{(1+\zeta)(T-\epsilon)-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu}\ \mathrm{d}t}\sqrt{\varepsilon_{\mathrm{SN}}} < \delta/4. \tag{65}$$

Finally, we choose $0 < \gamma < \gamma_\delta$ with γ_δ given in Table 2 such that the fourth term on the right-hand side of 62 is

$$\begin{aligned} &\gamma^{1/2}\zeta^{-1/2}(T-\epsilon)^{1/2}e^{(1+(3/2)\zeta+2\mathsf{K}_3(1+2T^\alpha))(T-\epsilon)} \\ &\quad \times (C_{\mathsf{EMose},2}^{1/2}(1+2\mathsf{K}_3(1+2T^\alpha)) + 2\sqrt{2}\mathsf{K}_1(1+8\widetilde{\varepsilon}_{\mathrm{AL}}+8|\theta^*|^2)^{1/2}) < \delta/4. \end{aligned} \tag{66}$$

Using 63, 64, 65, and 66, we obtain $W_2(\mathcal{L}(\widehat{Y}_K^{\mathrm{EM}}), \pi_{\mathsf{D}}) < \delta$. □

Proof of Theorem 15. Using the splitting 49, the proof follows along the same lines of the Proof of Theorem 13 for the estimation of the error bounds of the terms $W_2(\pi_{\mathsf{D}}, \mathcal{L}(Y_{t_K}))$, $W_2(\mathcal{L}(Y_{t_K}), \mathcal{L}(\widetilde{Y}_{t_K}))$, and $W_2(\mathcal{L}(\widetilde{Y}_{t_K}), \mathcal{L}(Y_{t_K}^{\mathrm{aux}}))$. The error bound for $W_2(\mathcal{L}(Y_{t_K}^{\mathrm{aux}}), \mathcal{L}(Y_K^{\mathrm{EM}}))$ is derived along the same lines of Bruno et al. (2025, Proof of Theorem 10). Putting these four estimates together leads to 25 and 26. □

D Modified Half-Normal Distribution

In this section, we recall the probability density function of the modified half-normal distribution, see e.g., Sun et al. (2023), used in Section 3.4.1 and defined as

$$g(x) = \frac{2\xi^{\frac{\upsilon}{2}}x^{\upsilon-1}\exp\left(-\xi x^2 + \psi x\right)}{\Psi\left(\frac{\upsilon}{2}, \frac{\psi}{\sqrt{\xi}}\right)}, \qquad x \geq 0, \tag{67}$$

where $\upsilon, \xi > 0$, $\psi \in \mathbb{R}$, and the normalizing constant

$$\Psi\left(\frac{\upsilon}{2}, \frac{\psi}{\sqrt{\beta}}\right) := \sum_{n=0}^{\infty} \frac{\Gamma\left(\frac{\upsilon}{2}+\frac{n}{2}\right)}{\Gamma(n)} \frac{\psi^n \xi^{-n/2}}{n!},$$

is the Fox–Wright function (Fox, 1928; Wright, 1935). We point out that the half-normal distribution, truncated normal distribution, gamma distribution, and square root of the gamma distribution are all special cases of the modified Half-Normal distribution 67. The distribution 27 follows by taking the symmetric extension of 67, i.e. $g(|x|)/2$, and choosing $\upsilon = 1$ and $\psi = -1$.

E Table of Constants

Table 2 displays full expressions for constants which appear in Theorem 13 and Theorem 15.

Table 2: Explicit expressions for the constants in Theorem 13 and Theorem 15.

Constant	Dependency	Full Expression
C_1	$O(\sqrt{d})$	$2(\sqrt{\mathbb{E}[\|X_0\|^2]} + \sqrt{d})$
C_2	$O(\sqrt{d})$	$\sqrt{2}\left(\sqrt{\mathbb{E}[\|X_0\|^2]} + \sqrt{d}\right)$
$C_3(T,\epsilon)$	$O(e^{(1+\zeta)(T-\epsilon)-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu}\,\mathrm{d}t})$	$\sqrt{2\zeta^{-1}}e^{(1+\zeta)(T-\epsilon)-2\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu}\,\mathrm{d}t}$
$C_{\mathsf{EM},2}(T)$	$O(Me^{T^{2\alpha+1}}T^{2\alpha+1}\widetilde{\varepsilon}_{\mathrm{AL}})$	$e^{T(4+8\mathsf{K}_{\mathrm{Total}}^2(1+T^{2\alpha}))} \times (\mathbb{E}[\|\widehat{Y}_0^{\mathrm{EM}}\|^2] + 16\mathsf{K}_{\mathrm{Total}}^2 T(1+2\widetilde{\varepsilon}_{\mathrm{AL}}+2\|\theta^*\|^2)(1+T^{2\alpha}) + 2dT)$
$C_{\mathsf{EM},4}(T)$	$O(d^2e^{T^{4\alpha+1}}T^{4\alpha+1})$	$e^{T(\frac{21}{2}+128\mathsf{K}_{\mathrm{Total}}^4(1+T^{4\alpha}))} \times (\mathbb{E}[\|\widehat{Y}_0^{\mathrm{EM}}\|^4] + 1024\mathsf{K}_{\mathrm{Total}}^4 T(1+\mathbb{E}[\|\hat{\theta}\|^4])(1+T^{4\alpha}) + 8(d^2+4d+4)T)$
$C_{\mathsf{EMose},2}$	$O(de^{T^{2\alpha+1}}T^{4\alpha+1}\widetilde{\varepsilon}_{\mathrm{AL}})$	$2(C_{\mathsf{EM},2}(T) + \mathsf{K}_{\mathrm{Total}}^2(1+T^{2\alpha})(16C_{\mathsf{EM},2}(T) + 32(1+2\widetilde{\varepsilon}_{\mathrm{AL}}+2\|\theta^*\|^2))) + 2d$
$C_4(T,\epsilon)$	$O(\sqrt{d}e^{T^{2\alpha+1}}T^{3\alpha+1}\widetilde{\varepsilon}_{\mathrm{AL}}^{1/2})$	$\zeta^{-1/2}(T-\epsilon)^{1/2}e^{(1+(3/2)\zeta+2\mathsf{K}_3(1+2T^{\alpha}))(T-\epsilon)} \times (C_{\mathsf{EMose},2}^{1/2}(1+2\mathsf{K}_3(1+2T^{\alpha})) + 2\sqrt{2}\mathsf{K}_1(1+8\widetilde{\varepsilon}_{\mathrm{AL}}+8\|\theta^*\|^2)^{1/2})$
$C_{\mathsf{EMose},4}$	$O(d^2e^{T^{4\alpha+1}}T^{8\alpha+1})$	$8(C_{\mathsf{EM},4}(T) + \mathsf{K}_{\mathrm{Total}}^4(1+T^{4\alpha})(1024C_{\mathsf{EM},4}(T) + 8192(1+\mathbb{E}[\|\hat{\theta}\|^4]))) + 144d^2$
$\widetilde{C}_4(T,\epsilon)$	$O(de^{T^{4\alpha+1}}T^{4\alpha+1}\widetilde{\varepsilon}_{\mathrm{AL}}^{1/4})$	$\sqrt{2}e^{2(1+\zeta+\mathsf{K}_3(1+2T^{\alpha}+4\mathsf{K}_3(1+4T^{2\alpha})))(T-\epsilon)}\sqrt{T-\epsilon} \times \Big(\mathsf{K}_4^2\zeta^{-1}(1+4T^{2\alpha})C_{\mathsf{EMose},4} + 4d(1+8\mathsf{K}_3^2(1+4T^{2\alpha})) + 2\zeta^{-1}\mathsf{K}_1^2(1+8(\widetilde{\varepsilon}_{\mathrm{AL}}+\|\theta^*\|^2)) + 4\zeta^{-1}d(1+8\mathsf{K}_3^2(1+4T^{2\alpha})) \times [(1+16\mathsf{K}_{\mathrm{Total}}^2(1+T^{2\alpha}))C_{\mathsf{EM},2}(T) + 32\mathsf{K}_{\mathrm{Total}}^2(1+T^{2\alpha})(1+2\widetilde{\varepsilon}_{\mathrm{AL}}+2\|\theta^*\|^2)] + 2[(1+8\mathsf{K}_3^2(1+4T^{2\alpha}))^{1/2}C_{\mathsf{EMose},2}^{1/2} + 2\mathsf{K}_1(1+8\widetilde{\varepsilon}_{\mathrm{AL}}+8\|\theta^*\|^2)^{1/2}] \times [d\sqrt{2}(1+8\mathsf{K}_3^2(1+4T^{2\alpha}))^{1/2}]\Big)^{1/2}$
ϵ_δ	-	$\delta^2/(64(\sqrt{\mathbb{E}[\|X_0\|^2]} + \sqrt{d})^2)$
T_δ	-	Obtained solving $T > T_\delta$ using Proposition 11, i.e., $\ln(\mu(e^{2T}-1)+1) + (K/\mu+1)/(\mu(e^{2T}-1)+1) > \ln(4\sqrt{2}((\mathbb{E}[\|X_0\|^2])^{1/2} + \sqrt{d})/\delta) + 2\int_0^\epsilon \beta_t^{\mathrm{OS},K,\mu}\,\mathrm{d}t + K/\mu + 1 - \epsilon$
$\varepsilon_{\mathrm{SN},\delta}$	-	$(\delta^2\zeta/32)e^{-2(1+\zeta)(T-\epsilon)+4\int_\epsilon^T \beta_t^{\mathrm{OS},K,\mu}\,\mathrm{d}t}$
γ_δ	-	$(\delta^2\zeta/16)(T-\epsilon)^{-1}e^{-2(1+(3/2)\zeta+2\mathsf{K}_3(1+2T^{\alpha}))(T-\epsilon)} \times (C_{\mathsf{EMose},2}^{1/2}(1+2\mathsf{K}_3(1+2T^{\alpha})) + 2\sqrt{2}\mathsf{K}_1(1+8\widetilde{\varepsilon}_{\mathrm{AL}}+8\|\theta^*\|^2)^{1/2})^{-2}$
$\widetilde{\gamma}_\delta$	-	$\min\Big\{(\delta/(4\sqrt{2}))^{1/\alpha}(T-\epsilon)^{-1/(2\alpha)}e^{-(2/\alpha)(1+\zeta+\mathsf{K}_3(1+2T^{\alpha}+4\mathsf{K}_3(1+4T^{2\alpha})))(T-\epsilon)} \times \Big(\mathsf{K}_4^2\zeta^{-1}(1+4T^{2\alpha})C_{\mathsf{EMose},4} + 4d(1+8\mathsf{K}_3^2(1+4T^{2\alpha})) + 2\zeta^{-1}\mathsf{K}_1^2(1+8(\widetilde{\varepsilon}_{\mathrm{AL}}+\|\theta^*\|^2)) + 4\zeta^{-1}d(1+8\mathsf{K}_3^2(1+4T^{2\alpha})) \times [(1+16\mathsf{K}_{\mathrm{Total}}^2(1+T^{2\alpha}))C_{\mathsf{EM},2}(T) + 32\mathsf{K}_{\mathrm{Total}}^2(1+T^{2\alpha})(1+2\widetilde{\varepsilon}_{\mathrm{AL}}+2\|\theta^*\|^2)] + 2[(1+8\mathsf{K}_3^2(1+4T^{2\alpha}))^{1/2}C_{\mathsf{EMose},2}^{1/2} + 2\mathsf{K}_1(1+8\widetilde{\varepsilon}_{\mathrm{AL}}+8\|\theta^*\|^2)^{1/2}] \times [d\sqrt{2}(1+8\mathsf{K}_3^2(1+4T^{2\alpha}))^{1/2}]\Big)^{-1/(2\alpha)}, 1\Big\}$