Sampling is as easy as keeping the consistency: Convergence guarantee for Consistency Models

Junlong Lyu [1] **Zhitang Chen** [2] **Shoubo Feng** [3]

Abstract

We provide the first convergence guarantee for the Consistency Models (CMs), a newly emerging type of one-step generative models that is capable of generating comparable samples to those sampled from state-of-the-art Diffusion Models. Our main result is that, under the basic assumptions on score-matching errors, consistency errors, and smoothness of the data distribution, CMs can efficiently generate samples in one step with small W_2 error to any real data distribution. Our results (1) hold for L^2-accurate assumptions on both score and consistency functions (rather than L^∞-accurate assumptions); (2) do not require strong assumptions on the data distribution such as log-sobolev conditions; (3) scale polynomially in all parameters; and (4) match the state-of-the-art convergence guarantee for score-based generative models. We also show that the Multi-step Consistency Sampling procedure can further reduce the error comparing to one step sampling, which supports the original statement of (Song et al., 2023). Our result can be generalized to arbitrary bounded data distributions that may be supported on some low-dimensional sub-manifolds. Our results further imply TV error guarantees when making some Langevin-based modifications to the output distributions.

1. Introduction

Score-based generative models (SGMs), also known as diffusion models ((Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Dhariwal & Nichol, 2021; Song et al., 2021; 2020)), are a family of generative models which achieve unprecedented success across multiple fields like image generation ((Dhariwal & Nichol, 2021; Nichol et al., 2021; Ramesh et al., 2022; Saharia et al., 2022)), audio synthesis ((Kong et al., 2020; Chen et al., 2020; Popov et al., 2021)) and video generation ((Ho et al., 2022a;b)); see, e.g., the recent surveys ((Cao et al., 2022; Croitoru et al., 2022; Yang et al., 2022)). A key point of diffusion models is the iterative sampling process which gradually reduces noise from random initial vectors which provides a flexible trade-off of compute and sample quality, as using extra compute for more iterations usually yields samples of better quality. However, compared to single-step generative models like GANs, VAEs, or normalizing flows, the generation process of diffusion models requires 10-2000 times more, limiting it to a small number of real-time applications.

To overcome this limitation, (Song et al., 2023) proposed Consistency Models (CMs) that can directly map noise to data, which can be seen as an extension of SGMs. CMs support fast one-step generation by design, while still allowing multistep sampling to trade compute for sampling quality. CMs can be trained either by distilling pre-trained diffusion models or as stand-alone generative models altogether. (Song et al., 2023) demonstrates its superiority through extensive experiments outperforming existing distillation techniques for diffusion models, and when trained in isolation, CMs outperform existing one-step, non-adversarial generative models on standard benchmarks.

Besides the achievements of CMs in saving the the generation costs as well as keeping the sampling quality, it is a pressing question of both practical and theoretical concern to understand the mathematical underpinnings that explain their startling successes. The theoretical guarantee for SGMs has been extensively studied and well established (Chen et al., 2022; Lee et al., 2022a;b). Despite the theoretical successes of SGMs, one would wonder if CMs can inherit the good points from SGMs, as they are inextricably linked in their underlying mathematical process.

Providing a convergence analysis for CMs and exploring the benefits for CMs compared to SGMs is a pressing first step towards theoretically understanding why CMs work in practice.

[1]Theory Lab, Huawei [2]Noah's Ark Lab, Huawei [3]Noah's Ark Lab, Huawei. Correspondence to: Junlong Lyu <lyujunlong@huawei.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1.1. Our contributions

In this work, we take a step towards connecting theory and practice by providing a convergence guarantee for CMs, under minimal assumptions that coincide with our intrinsic. For the underlying SGMs, we make no more assumptions than the state-of-the-art works:

A1 The score function of the forward process is L_s-Lipschitz.
A2 The second moment of the data distribution p_{data} is bounded.

Note that these two assumptions are standard, no more than what is needed in prior works. The crucial point to these two assumptions is that they do not need log-concavity, a log-sobolev inequality, or dissipativity, which cover arbitrarily non-log-concave data distributions. Our main result is summarized informally as follows.

Theorem 1.1. *Under Assumptions* **A1** *and* **A2***, and in addition if the consistency model is Lipschitz and the consistency error, score estimation error in L^2 are at most $O(\varepsilon)$ with an appropriate choice of step size in the training procedure, then, the CM outputs a measure which is ε-close in Wasserstein-2 (W_2) distance to p_{data} in single step.*

We find Theorem 1.1 surprising, because it shows that CMs can output a distribution arbitrary close to the data distribution in W_2 distance with a single step. The error of CMs is just the same order as what SGMs achieved, under the assumption that the consistency error is small enough, which coincides with the incredible success of CMs in many benchmarks. In the fields of neural networks, our result implies that so long as the neural network succeeds at the score and consistency function estimation tasks, the remaining part of the CM algorithm is almost understood, as it admits a strong theoretical justification.

However, learning the score function and consistency function is also difficult in general. Nevertheless, our result still leads the way to further investigations, such as: do consistency function for real-life data have intrinsic structure which can be well-explored by neural networks? If the answer is true, this would then provide an end-to-end guarantee for CMs.

Better error bounds by multistep consistency sampling. Beyond the one step CM sampling, (Song et al., 2023) suggests a way for multistep sampling to trade compute for sampling quality. However in their original work, they did not make any analysis on the positive effect of multistep sampling comparing to one-step sampling. In our work, we analysis the error bound for each middle state of multistep sampling, showing an asymptotically convergent error bound that is greatly smaller than the error bound for one-step sampling. Our analysis reveals the fact that with a suitable choice of middle time points, the multistep consistency sampling takes a few more steps to achieve the near-best performance.

Bounding the W_2 error for general bounded data distribution. The foregoing results are established on the assumption **A1**, which only holds for L_s-smooth data distribution. When only assuming the data distribution to be bounded supported, which includes the situation when q is supported on a lower-dimensional submanifold of $\mathbb{R}^d$, we can still guarantee polynomial convergence in the Wasserstein metric by early stopping. As the methodology is the same as in (Chen et al., 2022) and (Lee et al., 2022b), we do not claim the originality, but just include this part for completeness.

Bounding the Total Variation error. As the mathematical foundation of CMs established on the reverse probability flow ODE, they share the same shortcoming compared to the probability flow SDE: they can not get an error bound in Total Variation (TV) distance or Kullback-Leibler divergence by only controlling the score-matching objective, thus may even fail to estimate the likelihood of very simple data distributions (Fig.1 in (Lu et al., 2022)). To solve this potential problem, we offered two modification processes to control the TV error: we can take an OU-type smoothing which takes no more evaluation costs to get a relatively larger TV error bound; we can also apply a Langevin dynamics for correcting purpose with the score model to get a smaller TV error bound, while it needs additional $O(\varepsilon^{-1}d^{1/2})$ evaluation steps.

1.2. Prior works

As far as we know, this is the first work to establish a systematical analysis of the convergence property of CMs. As the CMs and SGMs share a similar mathematical essence in the asymptotic situation, our result can be compared to a vast list of literature on the convergence of SGMs.

SDE-type SGMs. The Langevin Monte Carlo (LMC) algorithm ((Rossky et al., 1978)) can be seen as the predecessor to the SDE-type SGMs, and literature on the convergence of LMCs is extensive, such as (Durmus & Moulines, 2015; Cheng & Bartlett, 2017; Cheng et al., 2017). However, these works mainly consider the case of exact or stochastic gradients. By the structure of the score-matching loss function, only an L^2-accurate gradient can be guaranteed for SDE-type SGMs. [(Lee et al., 2022a)] is the first to give a polynomial convergence guarantee in TV distance under an L^2-accurate score. However, they rely on the data distribution satisfying smoothness conditions and a log-Sobolev inequality, which essentially limits the guarantees to unimodal distributions.

(Bortoli, 2022) instead only make minimal data assumptions, giving convergence in Wasserstein distance for distributions

with bounded support $\mathcal{M}$. In particular, this covers the case of distributions supported on lower-dimensional manifolds, where guarantees in TV distance are unattainable. However, their guarantees have exponential dependence on the diameter of $\mathcal{M}$ or other parameters such as the Lipstchitz constant of score function.

Recently, (Chen et al., 2022) and (Lee et al., 2022b) concurrently obtained theoretical guarantees for SGMs under similar general assumptions on the data distribution. They give Wasserstein bounds for any distribution of bounded support (or sufficiently decaying tails), and TV bounds for distributions under minimal smoothness assumptions, that are polynomial in all parameters. This gives theoretical grounding to the success of SGM of data distribution that is often non-smooth and multimodal.

ODE-type SGMs. Instead of implementing the time-reversed diffusion as an SDE, it is also possible to implement it as an ordinary differential equation (ODE). However, current analyses of SGMs cannot provide a polynomial-complexity TV bound of the probability flow ODE under minimal assumption on data distribution. (Lu et al., 2022) first bounded the KL divergence gap (and thus TV error) by higher-order gradients of the score function, and thus suggested controlling this bound by minimizing the higher order score-matching objectives, which causes much more difficulties in training the score model.

Instead of changing the training procedure, (Chen et al., 2023b) obtained a discretization analysis for the probability flow ODE in KL divergence, though their bounds have a large dependency on d, exponential in the Lipschitz constant of the score integrated over time, which rely on higher order regularities of the log-data density.

To overcome the difficulty on strong data density regularity assumptions, (Chen et al., 2023a) suggest to interleave steps of the discretized probability flow ODEs with Langevin diffusion correctors using the estimated score, and get a better convergence guarantee than SDEs thanks to the $\mathcal{C}^1$ trajectory for ODEs comparing to the $\mathcal{C}^{\frac{1}{2}-}$ trajectory for SDEs. This approach only need to assume the data density to be L-smooth.

2. Preliminary

2.1. Diffusion Models

Consistency models heavily rely on the denoising diffusion probabilistic modeling (DDPM). We start with a forward process defined in $\mathbb{R}^d$, which is expressed as a stochastic differential equation

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{\mu}(\boldsymbol{x}_t, t)\mathrm{d}t + \sigma(t)\mathrm{d}\boldsymbol{w}_t \tag{1}$$

where $t \in [0, T], T > 0$ is a fixed constant, $\boldsymbol{\mu}(\cdot,\cdot)$ and $\sigma(\cdot)$ are the drift and diffusion coefficients respectively, and $\{\boldsymbol{w}_t\}_{t\in[0,T]}$ denotes the d-dimensional standard Brownian motion. Denote the disbribution of $\boldsymbol{x}_t$ as $p_t(\boldsymbol{x})$, therefore $p_0(\boldsymbol{x}) = p_{\mathrm{data}}(\boldsymbol{x})$. A remarkable property is the existence of an ordinary differential equation dubbed the *probability flow ODE*, whose solution trajectories sampled at t are distributed according to $p_t(\boldsymbol{x})$:

$$\mathrm{d}\boldsymbol{x}_t = \left[\boldsymbol{\mu}(\boldsymbol{x}_t, t) - \frac{1}{2}\sigma(t)^2\nabla \log p_t(\boldsymbol{x}_t)\right]\mathrm{d}t \tag{2}$$

here $\nabla \log p_t(\boldsymbol{x})$ is the *score function* of $p_t(\boldsymbol{x})$.

For clarity, we consider the simplest possible noise schedule choice, which is the Ornstein-Uhlenbeck (OU) process as in (Chen et al., 2023a), where $\mu(\boldsymbol{x}, t) = -\boldsymbol{x}$ and $\sigma(t) \equiv \sqrt{2}$,

$$\mathrm{d}\boldsymbol{x}_t = -\boldsymbol{x}_t\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{w}_t, \boldsymbol{x}_0 \sim p_{\mathrm{data}}, \tag{3}$$

The corresponding backward ODE is

$$\mathrm{d}\boldsymbol{x}_t = (-\boldsymbol{x}_t - \nabla \log p_t(\boldsymbol{x}_t))\mathrm{d}t. \tag{4}$$

In this case we have

$$p_t(\boldsymbol{x}) = e^{dt}p_{\mathrm{data}}(e^{-t}\boldsymbol{x}) * \mathcal{N}(\mathbf{0}, (1-e^{-2t})\boldsymbol{I}_d), \tag{5}$$

where $*$ denotes the convolution operator. We take $\pi(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, which is a tractable Gaussian distribution close to $p_T(\boldsymbol{x})$. For sampling, we first train a score model $\boldsymbol{s}_{\boldsymbol{\phi}} \approx \nabla \log p_t(\boldsymbol{x})$ via score matching ((Hyvärinen, 2005; Song & Ermon, 2019; Ho et al., 2020)), then plug into equation 4 to get the empirical estimation of the PF ODE, which takes the form of

$$\mathrm{d}\boldsymbol{x}_t = (-\boldsymbol{x}_t - \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t))\mathrm{d}t. \tag{6}$$

We call equation 6 the *empirical PF ODE*. Denote the distribution of $\boldsymbol{x}_t$ in equation 6 as $q_t(\boldsymbol{x})$. Empirical PF ODE gradually transforms $q_T(\boldsymbol{x}) = \pi(\boldsymbol{x})$ into $q_0(\boldsymbol{x})$, which can be viewed as an approximation of $p_{\mathrm{data}}(\boldsymbol{x})$.

2.2. Consistency Models

For any ordinary differential equation defined on $\mathbb{R}^d$ with vector field $\boldsymbol{v} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$,

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{v}(\boldsymbol{x}_t, t)\mathrm{d}t,$$

we may define the associate backward mapping $\boldsymbol{f}^{\boldsymbol{v}} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$ such that

$$\boldsymbol{f}^{\boldsymbol{v}}(\boldsymbol{x}_t, t) = \boldsymbol{x}_\delta. \tag{7}$$

with an early-stopping time $\delta > 0$. Under mild conditions on $\boldsymbol{v}$, such a mapping $\boldsymbol{f}^{\boldsymbol{v}}$ exists for any $t \in \mathbb{R}^+$, and is

smoothly relied on $\boldsymbol{x}$ and t. Note that equation 7 is equivalent to the following conditions

$$\boldsymbol{f}^{\boldsymbol{v}}(\boldsymbol{x}_t, t) = \boldsymbol{f}^{\boldsymbol{v}}(\boldsymbol{x}_s, \tau),\ \forall\, 0 \le \tau, t \le T, \text{ and}$$
$$\boldsymbol{f}^{\boldsymbol{v}}(\boldsymbol{x}, \delta) = \boldsymbol{x},\ \forall \boldsymbol{x} \in \mathbb{R}^d$$

which playing the essential role in constructing consistency loss.

Now let us take $\boldsymbol{v}^{\text{ex}}(\boldsymbol{x}, t) = -\boldsymbol{x} - \nabla \log p_t(\boldsymbol{x})$ and $\boldsymbol{v}^{\text{em}}(\boldsymbol{x}, t) = -\boldsymbol{x} - \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}, t)$, and denote the corresponding backward mapping function as $\boldsymbol{f}^{\text{ex}}$, for exact vector field $\boldsymbol{v}^{\text{ex}}$, and $\boldsymbol{f}^{\text{em}}$, for empirical vector field $\boldsymbol{v}^{\text{em}}$, respectively. We aim to construct a parametric model $\boldsymbol{f}_{\boldsymbol{\theta}}$ to approximate $\boldsymbol{f}^{\text{ex}}$. (Song et al., 2023) first implement the boundary condition using the skip connection,

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) = c_{\text{skip}}(t)\boldsymbol{x} + c_{\text{out}}(t) F_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$$

with differentiable $c_{\text{skip}}(t), c_{\text{out}}(t)$ such that $c_{\text{skip}}(\delta) = 1, c_{\text{out}}(\delta) = 0$, and then define the following Consistency Distillation (CD) object with a additional parameter $\boldsymbol{\theta}^-$ that shall be given later:

$$\begin{aligned}&\mathcal{L}_{\text{CD}}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \boldsymbol{\phi})\\ &:= \mathbb{E}[\lambda(t_n) \| \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}) - \boldsymbol{f}_{\boldsymbol{\theta}^-}(\hat{\boldsymbol{x}}_{t_n}^{\boldsymbol{\phi}}, t_n)\|_2^2], \quad (8)\end{aligned}$$

where $0 < t_1 = \delta < t_2 \cdots < t_N = T$, n uniformly distributed over $\{1, 2, \cdots, N-1\}$, $\boldsymbol{x} \sim p_{\text{data}}$, $\boldsymbol{x}_{t_{n+1}} = e^{-t}\boldsymbol{x} + \sqrt{1 - e^{-2t}}\boldsymbol{\xi}, \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Here $\hat{\boldsymbol{x}}_{t_n}^{\boldsymbol{\phi}}$ is calculated by

$$\hat{\boldsymbol{x}}_{t_n}^{\boldsymbol{\phi}} := \Phi(\boldsymbol{x}_{t_{n+1}}, t_{n+1}, t_n; \boldsymbol{\phi}) \quad (9)$$

where $\Phi(\cdots; \boldsymbol{\phi})$ represents the update function of a ODE solver applied to the empirical PF ODE 6. In our noise scheduler 3, We may use the exponential integrator (i.e., exactly integrating the linear part),

$$\hat{\boldsymbol{x}}_{t_n}^{\boldsymbol{\phi}} = e^{t_{n+1} - t_n}\boldsymbol{x}_{t_{n+1}} + (e^{t_{n+1} - t_n} - 1)\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}). \quad (10)$$

For simplicity, we assume $\lambda(t_n) \equiv 1$, and only consider the square of l_2 distance to build the loss metric, (Song et al., 2023) also considered other distance metric such as l_1 distance $\|\boldsymbol{x} - \boldsymbol{y}\|_1$, and the Learned Perceptual Image Patch Similarity (LPIPS, (Zhang et al., 2018)).

To stabilize the training process, (Song et al., 2023) introduce an additional parameter $\boldsymbol{\theta}^-$ and update it by an exponential moving average (EMA) strategy. That is, given a decay rate $0 \le \mu < 1$, the author perform the following update after each optimizaiton step: $\boldsymbol{\theta}^- = \text{stopgrad}(\mu\boldsymbol{\theta}^- + (1-\mu)\boldsymbol{\theta})$.

Besides the distillation strategy that needs an existing score model $\boldsymbol{s}_{\boldsymbol{\phi}}$, (Song et al., 2023) also introduced a way to train without any pre-trained score models called the Consistency Training (CT) objective. In their following work (Song & Dhariwal, 2023), they showed the Consistency Training has similar performance comparing to CD. We refer Lemma C.3 for the expression of CT objective under the OU scheduler 5 and the exponential integrator 9. However, as the CT objective do not directly match a loss function, it is hard to character it with some assumptions like 3.4. Thus we only focus on CD situations in this work.

(Song et al., 2023) gave an asymptotic analysis on the approximation error in their original work. If $\mathcal{L}_{\text{CD}}^N(\boldsymbol{\theta}, \boldsymbol{\theta}; \boldsymbol{\phi}) = 0$, we have $\sup_{n,\boldsymbol{x}} \|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}, t_n) - \boldsymbol{f}^{\text{em}}(\boldsymbol{x}, t_n)\|_2 = O((\Delta t)^p)$ when the numerical integrator has local error uniformly bounded by $O((\Delta t)^{p+1})$ with $p \ge 1$. However, when $\mathcal{L}_{\text{CD}}^N \ne 0$, we also need a quantitative analysis on how far between $\boldsymbol{f}_{\boldsymbol{\theta}}$ and $\boldsymbol{f}^{\text{em}}$, and further, a quantitative analysis between $\boldsymbol{f}_{\boldsymbol{\theta}}$ and $\boldsymbol{f}^{\text{ex}}$, and thus the distance between generated distribution and true data distribution.

3. Main Results

In this section, we formally state our assumptions and our main results. We denote $\boldsymbol{f}_{\boldsymbol{\theta},t}(\boldsymbol{x}) = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ to emphasize the mapping over $\boldsymbol{x}$ at time t. We summarized definition of some notations in the Appendix A.

3.1. Assumptions

We assume the following mild conditions on the data distribution p_{data}.

Assumption 3.1. *The data distribution has finite 2nd moment, that is, $\mathbb{E}_{\boldsymbol{x}_0 \sim p_{\text{data}}}[\|\boldsymbol{x}_0\|_2^2] = \mathfrak{m}^2 < \infty$.*

Assumption 3.2. *The score function $\nabla \log p_t(\boldsymbol{x})$ is Lipschitz on the variable $\boldsymbol{x}$ with Lipschitz constant $L_s \ge 1$, $\forall t \in [0, T]$.*

This two assumptions are standard and has been used in prior works (Block et al., 2020; Lee et al., 2022a;b; Chen et al., 2022). As (Lee et al., 2022b; Chen et al., 2022), we do not assume Lipschitzness of the score estimate; unlike (Block et al., 2020; Bortoli et al., 2021), we do not assume any convexity or dissipativity assumptions on the potential $U = -\log(p_{\text{data}})$, and unlike (Lee et al., 2022a) we do not assume p_{data} satisfies a log-Sobolev inequality. Thus our assumptions are general enough to cover the highly non-log-concave data distributions. Our assumption could be further weakened to only be compactly supported, which will be further discussed in 3.4.

We also assume bounds on the score estimation error and consistency error.

Assumption 3.3. *Assume*

$$\mathbb{E}_{p_{t_n}}[\|\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_n}, t_n) - \nabla \log p_{t_n}(\boldsymbol{x}_{t_n})\|_2^2] \le \varepsilon_{sc}^2, \forall n \in [\![1, N]\!].$$

Assumption 3.4. *Assume*

$$\mathbb{E}_{\boldsymbol{x}_{t_{n+1}}\sim p_{t_n}}[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}},t_{n+1})-\boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}^{\phi}_{t_n},t_n)\|_2^2] \leq \varepsilon_{cm}^2(t_{n+1}-t_n)^2, \forall n\in[\![1,N-1]\!],$$

where $\hat{\boldsymbol{x}}^{\phi}_{t_n}$ is the exponential integrator defined as in equation 10.

The score estimation error is the same as in (Lee et al., 2022a; Chen et al., 2022). As discussed in Section 2, these two assumption are nature and realistic in light of the derivation of the score matching objective and consistency distillation object.

The following Lipschitz-like condition for the consistency model is natural but a little stronger under 3.3 and 3.4 as it requires a L^∞ assumption. Note that originally in ((Song et al., 2023), Theorem 1), they assumed a Lipschitz condition that is essentially stronger than ours.

Assumption 3.5. *The consistency model $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x},t_n)$ satisfies the following Lipschitz-like condition on the variable $\boldsymbol{x}$: for any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d, n\in[\![1,N]\!]$,*

$$\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x},t_n)-\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y},t_n)\|_2 \lesssim t_n(L_f\varepsilon_{sc}+\varepsilon_{cm})+L_f\|\boldsymbol{x}-\boldsymbol{y}\|_2.$$

We will verify this assumption in C.1.

To control the discretization error near t = 0, we divide our discretization schedule into two stages: in the first stage, which lasts from T to h, we keep the step size equal to h; in the second stage, which lasts from h to δ, we take a geometric reducing sequence $2^{-1}h, 2^{-2}h, \cdots$ until $2^{-l}h \leq \delta$ for some $l\geq 1$. This can help establishing equation 25, and theorem 3.7.

Assumption 3.6. *Assume the discretization schedule $0<\delta=t_1<t_2<\cdots<t_N=T$, $h_k=t_{k+1}-t_k$ to equation 6 is divided into two stages:*

1. *$h_k\equiv h$ for all $k\in[\![N_1,N-1]\!]$, and $(N-N_1-1)h<T\leq(N-N_1)h$;*
2. *$h_k=2^{-(N_1-k)}h=\frac{h_{k+1}}{2}$ for $k\in[\![1,N_1-1]\!]$, N_1 satisfies $h_2=2^{-(N_1-2)}h\leq 2\delta$.*

note that in this case, we have $h_1=T-\sum_{k=2}^N h_k-\delta\leq h-(1-2^{-(N_1-2)})h-\delta\leq\delta$, and $t_{N_1}\leq h$

3.2. W_2 error guarantee for one-step Consistency Generating

In this section we introduce our main results. The first result bounds the Consistency Model estimation error in an expectation mean.

Figure 1. Illustration of the discretization schedule in Assumption 3.6

Theorem 3.7 (see Section B.1 of Appendix)**.** *Under Assumptions 3.1-3.5, assume we choose Φ as the exponential integrator, and assume the timestep schedule satisfies assumption 3.6, for $1\leq n\leq N-1$, $\boldsymbol{x}_{t_n}\sim p_{t_n}$:*

$$\left(\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n},t_n)-\boldsymbol{f}^{ex}(\boldsymbol{x}_{t_n},t_n)\|_2^2]\right)^{1/2} \lesssim t_n^2(\varepsilon_{cm}+L_f\varepsilon_{sc}) +t_n(\varepsilon_{cm}+L_f\varepsilon_{sc}+L_fL_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+t_n^{\frac{1}{2}}L_fL_sd^{\frac{1}{2}}h.$$

Now we can get our first theorem that analysis the W_2 distance after Consistency Model mapping.

Theorem 3.8 (see Section B.2 of Appendix)**.** *Under Assumptions 3.1-3.6, let $\mu(\boldsymbol{x})$ be any probability density, $p_t(\boldsymbol{x})=e^{dt}p_{\text{data}}(e^t\boldsymbol{x})*\mathcal{N}(\boldsymbol{0},(1-e^{-2t})\boldsymbol{I}_d)$, then the following estimation holds,*

$$W_2(\boldsymbol{f}_{\boldsymbol{\theta},t_n}\sharp\mu,p_\delta)\lesssim L_fW_2(\mu,p_{t_n})+t_n^2(\varepsilon_{cm}+L_f\varepsilon_{sc}) +t_n(\varepsilon_{cm}+L_f\varepsilon_{sc}+L_fL_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+t_n^{\frac{1}{2}}L_fL_sd^{\frac{1}{2}}h.$$

As a consequence, we directly get the one-step generation error.

Corollary 3.9 (see Section B.2 of Appendix)**.** *Under Assumptions 3.1-3.6, when $T>L_s^{-1}$, let $\mu\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)$ be the standard normal distribution, then the one-step generating error is bounded as follows,*

$$W_2(\boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mu,p_{\text{data}})\lesssim(d^{\frac{1}{2}}\vee\mathfrak{m})L_fe^{-T}+T^2(\varepsilon_{cm}+L_f\varepsilon_{sc}) +T(\varepsilon_{cm}+L_f\varepsilon_{sc}+L_fL_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+(d^{\frac{1}{2}}\vee\mathfrak{m})\delta^{\frac{1}{2}}.$$

In particular, for any $\varepsilon>0$, if we set $\delta\asymp\frac{\varepsilon^2}{d\vee\mathfrak{m}^2}$, $T\geq O(\log(\frac{L_f(\sqrt{d}\vee R)}{\varepsilon}))$, step size $h=O(\frac{\varepsilon}{TL_fL_s^{3/2}d^{1/2}})$, $\varepsilon_{cm}=O(\frac{\varepsilon}{T^2}),\varepsilon_{sc}=O(\frac{\varepsilon}{L_fT^2})$, we can guarantee $W_2(\boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mu,p_{\text{data}})\lesssim\varepsilon$.

We remark that our discretization complexity $N=O(\frac{T}{h}\log(\frac{1}{h}))=O(\frac{L_fL_s^{3/2}d^{1/2}}{\varepsilon})$ matches state-of-the-art complexity for ODE-type SGMs (Chen et al., 2023a;b). This provides some evidence that our descretization bounds are of the correct order.

Note that the error bound in 11 relied on the final time T, consistency error ε_{cm}, score error ε_{sc} and step size h. Actually we can refine the error and reduce the linear dependency on T to log dependency by Multistep Consistency Sampling that will be introduced in the next section.

3.3. Multistep Consistency Sampling can reduce the W_2 error

Now let us analysis the effect of Multistep Consistency Sampling introduced in original CM work, Algorithm 1, (Song et al., 2023), which has been introduced to improve the sample quality by alternating denoising and noise injection steps. Given a sequence of time points $T = t_{n_1} \geq t_{n_2} \geq \cdots$, adapted to the OU noise scheduler 3 , the generating procedure can be written as

$$\begin{aligned} \boldsymbol{z}_1 &:= \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{\xi}_1, T), \\ \boldsymbol{u}_k &:= e^{-(t_{n_k}-\delta)}\boldsymbol{z}_{k-1} + \sqrt{1-e^{-2(t_{n_k}-\delta)}}\boldsymbol{\xi}_k, \\ \boldsymbol{z}_k &:= \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{u}_k, t_{n_k}), \end{aligned} \tag{11}$$

with $\boldsymbol{\xi}_i$ i.i.d. $N(\mathbf{0}, \boldsymbol{I}_d)$ distributed, and $q_k := \text{law}(\boldsymbol{z}_k)$ satisfies the following relationship:

$$q_1 = \boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), \tag{12}$$

$$\begin{aligned} \mu_k := &\left(e^{d(t_{n_k}-\delta)}q_{k-1}(e^{(t_{n_k}-\delta)}\boldsymbol{x})\right) \\ &* \mathcal{N}(\mathbf{0}, (1-e^{-2(t_{n_k}-\delta)})\boldsymbol{I}_d). \end{aligned}$$

$$q_k = \boldsymbol{f}_{\boldsymbol{\theta},t_{n_k}}\sharp\mu_k, \tag{13}$$

We thus have the following upper bound of the W_2 distance between q_k and p_δ.

Corollary 3.10 (see Section B.3 of Appendix)**.** *Under Assumptions 3.1-3.6, when $T > L_s^{-1}$, the W_2 distance between q_k and p_δ can be controlled by q_{k-1} and p_δ as follows,*

$$\begin{aligned} W_2(q_k, p_\delta) &\lesssim L_f e^{-t_{n_k}} W_2(q_{k-1}, p_\delta) \\ &+ t_{n_k}^2(L_f\varepsilon_{sc} + \varepsilon_{cm}) + t_{n_k}(\varepsilon_{cm} + L_f\varepsilon_{sc} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h). \end{aligned} \tag{14}$$

Note that in equation 14, we have an exponentially small multiplier $e^{-t_{n_k}}$ gradually reduce the error introduced from the previous steps, and a t_{n_k}-polynomial term representing the error introduced from the current step. By choosing a suitable time schedule $\{t_{n_k}\}_{k\geq 1}$, we can get a finer bound in W_2.

Corollary 3.11 (see Section B.3 of Appendix)**.** *Under Assumptions 3.1-3.6, there exists $\hat{n} \in [\![1, N]\!]$, $T \geq t_{\hat{n}} \geq \max(\log(2L_f) + \delta, L_s^{-1})$, such that when taking $n_k \equiv \hat{n}$ for all k,*

$$\begin{aligned} W_2(q_k, p_{\text{data}}) \lesssim &(\log(L_f)^2 + 2^{-k}T^2)(\varepsilon_{cm} + L_f\varepsilon_{sc}) \\ &+ (\log(L_f) + 2^{-k}T)(\varepsilon_{cm} + L_f\varepsilon_{sc} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) \\ &+ 2^{-k}(d^{\frac{1}{2}} \vee \mathfrak{m})L_f e^{-T} + (d^{\frac{1}{2}} \vee \mathfrak{m})\delta^{\frac{1}{2}}. \end{aligned}$$

Thus, for any $\varepsilon > 0$, if we set $\delta \asymp \frac{\varepsilon^2}{d\vee\mathfrak{m}^2}$, $k = O(\log(T\vee(\frac{(d\vee\mathfrak{m}^2)L_f}{\varepsilon})))$, $h = O(\frac{\varepsilon}{\log(L_f)L_f L_s^{3/2} d^{1/2}})$, $\varepsilon_{cm} = O(\frac{\varepsilon}{\log(L_f)^2})$, $\varepsilon_{sc} = O(\frac{\varepsilon}{L_f\log(L_f)^2})$, we can guarantee $W_2(q_k, p_{\text{data}}) \lesssim \varepsilon$.

Remark 3.12. *Comparing the result between multistep sampling error 3.11 and one step sampling error 3.9, the main improvement of multistep sampling is getting rid of the linear dependency from T. In one step sampling, one should take the step size smaller, and train the consistency model and score model better. Besides, multistep sampling 3.11 only requires $T \geq \max(\log(2L_f) + \delta, L_s^{-1})$, while one step sampling 3.9 requires $T \geq O(\log(\frac{L_f(\sqrt{d}\vee R)}{\varepsilon}))$, which is an added benefit that multistep sampling requires lower training complexity comparing to one step sampling.*

3.4. W_2 convergence guarantee for arbitrarily data distributions with bounded support

In this section, we consider a much more general case: in fact for any compactly supported distribution p_{data}, supp $p_{\text{data}} \subseteq B(\mathbf{0}, R)$, for any t_0, we can get a positive $L_s(t_0)$, such that Assumption 3.2 is satisfied for any $t > t_0$ with Lipschitz constant $L_s(t_0)$. This include a wide range of situations even when p do not have smooth density w.r.t. Lebesgue measure such as when p supported on a lower-dimensional submanifold of $\mathbb{R}^d$, which recently investigated in (Bortoli, 2022; Lee et al., 2022b; Chen et al., 2022).

Namely, based on the following lemma, we can conduct regularity properties for the score functions.

Lemma 3.13 (see Section B.4 of Appendix)**.** *Suppose that supp $p_{\text{data}} \subseteq B(\mathbf{0}, R)$ where $R \geq 1$, and let p_t denote the law of the OU process at time t, started at p: that is, $p_t(\boldsymbol{x}) = e^{dt}p_{\text{data}}(e^t\boldsymbol{x}) * \mathcal{N}(\mathbf{0}, (1-e^{-2t})\boldsymbol{I}_d)$. Then the Hessian of the score function satisfies:*

$$\|\nabla^2 \log p_t(\boldsymbol{x})\|_{op} \leq \frac{e^{-2t}R^2}{(1-e^{-2t})^2} + \frac{1}{1-e^{-2t}}.$$

Note that in our proof of Corollary 3.9 and 3.11, we only use the Assumption 3.2 over $t \in [\delta, T]$. Combining Lemma 3.13, we immediately get the following corollary.

Corollary 3.14 (see Section B.4 of Appendix)**.** *Under Assumptions 3.3-3.6, suppose that supp $p_{\text{data}} \subseteq B(\mathbf{0}, R)$ where $R \geq 1$. Let $\delta \asymp \frac{\varepsilon^2}{R^2\vee d}$, then*

1. *the one-step generating error satisfies $W_2(q_1, p_{\text{data}}) \lesssim \varepsilon$, provided that $T = O(\log(\frac{L_f(\sqrt{d}\vee R)}{\varepsilon}))$, $h = O(\frac{\varepsilon^7}{d^{1/2}R^3(R^6\vee d^3)L_f T})$, $\varepsilon_{cm} = O(\frac{\varepsilon}{T^2})$, $\varepsilon_{sc} = O(\frac{\varepsilon}{L_f T^2})$;*
2. *the multi-step generating error satisfies $W_2(q_k, p_{\text{data}}) \lesssim \varepsilon$, provided that $k = O(\log(T \vee (\frac{(d\vee\mathfrak{m}^2)L_f L_s}{\varepsilon})))$, $T = O(\max(\log(2L_f) + \delta, L_s^{-1}))$,*

$h = O(\frac{\varepsilon^7}{d^{1/2}R^3(R^6\vee d^3)L_f\log(L_f)})$, $\varepsilon_{cm} = O(\frac{\varepsilon}{\log(L_f)^2})$, $\varepsilon_{sc} = O(\frac{\varepsilon}{L_f\log(L_f)^2})$.

3.5. Bounding the TV error

In the sections before, we have showed that the generated distribution of Consistency Models are close to the true data distribution in the metric of Wasserstein-2 distance. When we turn to the Total Variation (TV) distance, however, the error bound is deficient as the situation for the probability flow ODEs, in contrast to the situation for the probability flow SDEs. Here we introduce two operations that can further bound the TV error.

3.5.1. Bounding the TV error by forward OU process

Let the forward OU process be

$$\mathrm{d}\boldsymbol{x}_t = -\boldsymbol{x}_t\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{w}_t, \tag{15}$$

and denote the associate Markov kernel as P^s_{OU}, that is, if $\boldsymbol{x}_t \sim p$, $\boldsymbol{x}_{t+s} \sim pP^s_{\mathrm{OU}}$. Let q be the output of our Consistency Models, either the one step consistency sampling result, or the k-th multistep consistency sampling result. To control the TV error, we smooth the generated sample by the forward OU process with a small time that is the same as the early stopping time δ, and then we can get the TV distence between qP^δ_{OU} and p_{data}:

Corollary 3.15 (see Section B.5 of Appendix)**.** *Under Assumptions 3.1-3.6, suppose q is: (1) the one step consistency sampling result, $q = q_1 = \boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$; (2) the k-th multistep consistency sampling result, $q = q_k$ defined as in equation 13 with multistep schedule as in Corollary 3.11. Choose the early stopping time $\delta \asymp \frac{\varepsilon^2}{L_s^2(d\vee\mathfrak{m}^2)}$ for some $\varepsilon > 0$, then if $T \geq \max(\log(2L_f) + \delta, L_s^{-1})$, denote*

$$D := \varepsilon_{cm} + L_f\varepsilon_{sc} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h,$$

$$\begin{aligned} TV(q_1P^\delta_{OU}, p_{\mathrm{data}}) \lesssim & \frac{L_sL_f(d\vee\mathfrak{m}^2)}{\varepsilon}e^{-T} + \frac{L_s(d^{\frac{1}{2}}\vee\mathfrak{m})}{\varepsilon}TD \\ & + \frac{L_s(d^{\frac{1}{2}}\vee\mathfrak{m})}{\varepsilon}T^2(\varepsilon_{cm} + L_f\varepsilon_{sc}) + \varepsilon, \end{aligned}$$

$$\begin{aligned} & TV(q_kP^\delta_{OU}, p_{\mathrm{data}}) \\ & \lesssim \frac{L_s(d^{\frac{1}{2}}\vee\mathfrak{m})}{\varepsilon}\Big((\log L_f + \frac{T}{2^k})D + \frac{(d^{\frac{1}{2}}\vee\mathfrak{m})L_f}{2^k e^T} \\ & + ((\log L_f)^2 + \frac{T^2}{2^k})(\varepsilon_{cm} + L_f\varepsilon_{sc})\Big) + \varepsilon. \end{aligned}$$

In particular,

1. *If we set $T = O(\log(\frac{(d\vee\mathfrak{m}^2)L_fL_s}{\varepsilon^4}))$, $h = O(\frac{\varepsilon^2}{L_fL_s^{5/2}(d\vee\mathfrak{m}^2)T})$, $\varepsilon_{sc} \leq O(\frac{\varepsilon^2}{L_fT^2L_s(d^{1/2}\vee\mathfrak{m})})$, $\varepsilon_{cm} \leq O(\frac{\varepsilon^2}{T^2L_s(d^{1/2}\vee\mathfrak{m})})$, then we can guarantee TV error $O(\varepsilon)$ with one step prediction and one additional OU correction (with no NN evaluation);*

2. *If we set $k = O(\log(T \vee (\frac{(d\vee\mathfrak{m}^2)L_fL_s}{\varepsilon})))$, $h = O(\frac{\varepsilon^2}{\log(L_f)L_fL_s^{5/2}(d\vee\mathfrak{m}^2)})$, then if we have $\varepsilon_{sc} \leq O(\frac{\varepsilon^2}{\log(L_f)^2L_fL_s(d^{1/2}\vee\mathfrak{m})})$, $\varepsilon_{cm} \leq O(\frac{\varepsilon^2}{\log(L_f)L_s(d^{1/2}\vee\mathfrak{m})})$, then we can guarantee TV error $O(\varepsilon)$ with k steps prediction and one additional OU correction (with no NN evaluation).*

3.5.2. Bounding the TV error by Underdamped Langevin Corrector

We may adopt the idea from (Chen et al., 2023a), who introduce the Langevin-correcting procedure into the probability flow ODEs to get a TV error guarantee.

The Langevin dynamics for correcting purpose is defined as follows: let p be a distribution over $\mathbb{R}^d$, and write U as a shorthand for the potential $-\log p$.

Given a friction parameter $\gamma > 0$, consider the following discretized process with step size τ, where $-\nabla U$ is replaced by a score estimator $\boldsymbol{s}$. Let $(\hat{z}_t, \hat{v}_t)_{t\geq 0}$ over $\mathbb{R}^d\otimes\mathbb{R}^d$ be given by

$$\begin{aligned} \mathrm{d}\hat{\boldsymbol{z}}_t &= \hat{\boldsymbol{v}}_t\mathrm{d}t, \\ \mathrm{d}\hat{\boldsymbol{v}}_t &= (\boldsymbol{s}(\hat{\boldsymbol{z}}_{\lfloor t/\tau\rfloor\tau}) - \gamma\hat{\boldsymbol{v}}_t)\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}\boldsymbol{w}_t. \end{aligned} \tag{16}$$

Denote the Markov kernel $\hat{P}_{\mathrm{ULMC}}$ to be defined by the equation 16, that is, if $(\hat{\boldsymbol{z}}_{k\tau}, \hat{\boldsymbol{v}}_{k\tau}) \sim \boldsymbol{\mu}$ for some $\boldsymbol{\mu}$ be a distribution over $\mathbb{R}^{d\times d}$, $(\hat{\boldsymbol{z}}_{(k+1)\tau}, \hat{\boldsymbol{v}}_{(k+1)\tau}) \sim \boldsymbol{\mu}\hat{P}_{\mathrm{ULMC}}$. We denote the k-th composition $\hat{P}^k_{\mathrm{ULMC}} = \hat{P}_{\mathrm{ULMC}} \circ \hat{P}^{k-1}_{\mathrm{ULMC}}$, and $\boldsymbol{q} = q \otimes \mathcal{N}(0, \boldsymbol{I}_d), \boldsymbol{p} = p \otimes \mathcal{N}(0, \boldsymbol{I}_d)$. In what follows, we abuse the notation as follows. Given a distribution q on $\mathbb{R}^d$, we write $q\hat{P}_{\mathrm{ULMC}}$ to denote the projection onto the $\boldsymbol{z}-$coordinates of $\boldsymbol{q}\hat{P}_{\mathrm{ULMC}}$. It's obvious that

$$\mathrm{TV}(q\hat{P}^N_{\mathrm{ULMC}}, p) \leq \mathrm{TV}(\boldsymbol{q}\hat{P}^N_{\mathrm{ULMC}}, \boldsymbol{p}).$$

Now let's take $p = p_{\mathrm{data}}$ as the data distribution, and $q = q_k$ for some $k \geq 1$ as the output distribution of k-th multistep consistency sampling defined as in equation 13. We can use the score model $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}, \delta) \approx \nabla \log p_\delta(\boldsymbol{x})$ to do corrector steps. The end-to-end error now can be written as follows:

Corollary 3.16 (see Section B.6 of Appendix)**.** *Under Assumptions 3.1-3.6, suppose q is: (1) the one step consistency sampling result, $q = q_1 = \boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$; (2) the k-th multistep consistency sampling result, $q = q_k$ defined as in equation 13 with multistep schedule as in Corollary 3.11. Choose $\gamma \asymp L_s$, and $\delta \asymp \frac{\varepsilon^2}{L_s^2(d\vee\mathfrak{m}^2)}$ for some $\varepsilon > 0$, then if*

$T \geq \max(\log(2L_f) + \delta, L_s^{-1}), N\tau \asymp \frac{1}{\sqrt{L_s}}$, *denote*

$$D_1 := \varepsilon_{cm} + L_f \varepsilon_{sc},$$
$$D_2 := \varepsilon_{cm} + L_f \varepsilon_{sc} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h,$$
$$D_3 := L_s^{-\frac{1}{2}} \varepsilon_{sc} + L_s^{\frac{1}{2}} d^{\frac{1}{2}} \tau + \varepsilon,$$

then

$$\begin{aligned}
&TV(q_1 \hat{P}^N_{ULMC}, p_{\text{data}}) \\
\lesssim &(d^{\frac{1}{2}} \vee \mathfrak{m}) L_f L_s^{\frac{1}{2}} e^{-T} + T^2 L_s^{\frac{1}{2}} D_1 + T L_s^{\frac{1}{2}} D_2 + D_3, \\
&TV(q_k \hat{P}^N_{ULMC}, p_{\text{data}}) \\
\lesssim &(\log(L_f)^2 + \frac{T^2}{2^k}) L_s^{\frac{1}{2}} D_1 + (\log(L_f) + \frac{T}{2^k}) L_s^{\frac{1}{2}} D_2 \\
&+ \frac{(d^{\frac{1}{2}} \vee \mathfrak{m}) L_s^{\frac{1}{2}} L_f}{2^k e^T} + D_3.
\end{aligned}$$

In particular,

1. *if we set* $T = O(\log(\frac{(d \vee \mathfrak{m}^2) L_f^2 L_s}{\varepsilon^2}))$, $h = O(\frac{\varepsilon}{L_f L_s^2 d^{1/2} T})$, $\tau = O(\frac{\varepsilon}{L_s^{1/2} d^{1/2}})$, *and if* $\varepsilon_{sc} \leq O(\frac{\varepsilon}{L_f T^2 L_s^{1/2}})$, $\varepsilon_{cm} \leq O(\frac{\varepsilon}{T^2 L_s^{1/2}})$, *then we can obtain TV error* $O(\varepsilon)$ *with one step consistency prediction and* $O(\frac{\sqrt{d}}{\varepsilon})$ *steps correcting;*

2. *if we set* $k = O(\log(T \vee (\frac{(d \vee \mathfrak{m}^2) L_f L_s}{\varepsilon})))$, $h = O(\frac{\varepsilon}{\log(L_f) L_f L_s^2 d^{1/2}})$, $\tau = O(\frac{\varepsilon}{L_s^{1/2} d^{1/2}})$ *and if* $\varepsilon_{sc} \leq O(\frac{\varepsilon}{\log(L_f)^2 L_f L_s^{1/2}})$, $\varepsilon_{cm} \leq O(\frac{\varepsilon}{\log(L_f)^2 L_s^{1/2}})$, *then we can obtain TV error* $O(\varepsilon)$ *with* k *steps consistency prediction and* $O(\frac{\sqrt{d}}{\varepsilon})$ *steps correcting.*

Remark 3.17. *In case of missing the score model* $\boldsymbol{s_\phi}$ *but only remain the consistency model* $\boldsymbol{f_\theta}$*, such as training the consistency model by CT objective 54 without a pretrained score model, we can also recover a score model* $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}$ *that approx* $\nabla \log p_{t_2}(\boldsymbol{x})$*: in fact* $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{\boldsymbol{f_\theta}(\boldsymbol{x}, t_2) - e^{h_1} \boldsymbol{x}}{e^{h_1} - 1}$ *is a score model that approximate* $\nabla \log p_{t_2}(\boldsymbol{x})$ *with* L^2 *error* $\varepsilon_{cm} + \varepsilon_{sc}$ *(proved in Appendix, Lemma C.2). Hence, we may run similar procedure as in Theorem 3.16, where the output distribution* q_k *of consistency model should firstly be transformed with forward OU process equation 3 under a small time* $h_1 = t_2 - \delta < \delta$*, then apply the Underdamped Langevin Corrector Operator with* $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{x})$.

4. Conclusions and Limitations

In this work, we provided a first convergence guarantee for CMs which holds true under realistic assumptions (L^2-accurate score and consistency function surrogates; arbitrarily data distributions with smooth densities respect to Lebesgue measure, or bounded distributions) and which scale at most polynomially in all relavent parameters. Our results take a step towards explaining the success of CMs. We also provide theoretical evidence that multistep CM sampling technique can further reduce the error comparing to one step CM sampling .

There are mainly three shortcomings in our work. Firstly, our proofs relied on the lipschitz conditions of the surrogate model $\boldsymbol{f_\theta}$, which is somehow unrealistic. We will keep in improving our results by replacing this condition on the exact consistency function $\boldsymbol{f}^{\text{ex}}$ in our future works. Secondly, to bound the TV error, we introduced an additional smoothing procedure after the original CM sampling steps. It would be better to remove this unnatural procedure by introduce new assumptions or techniques. Lastly, we did not address the question of when the score and consistency function can be learned well enough. We believe that the resolution of these problem would shed considerable light on Consistency Models.

Impact Statement

This work aims to establish a foundational theoretical framework for consistency models when applied to Ornstein-Uhlenbeck (OU) processes. Under appropriate conditions, the authors provide theoretical guarantees for this class of models. Moreover, this work shows that employing a multistep consistency model can further enhance this convergence.

This work represents an initial step toward a more comprehensive theoretical analysis of consistency models

References

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling. *ArXiv*, abs/2002.00107, 2020. URL https://api.semanticscholar.org/CorpusID:211010797.

Bortoli, V. D. Convergence of denoising diffusion models under the manifold hypothesis. *ArXiv*, abs/2208.05314, 2022.

Bortoli, V. D., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235294278.

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.-A., and Li, S. Z. A survey on generative diffusion model. *ArXiv*, abs/2209.02646,

2022. URL https://api.semanticscholar.org/CorpusID:252090040.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. *ArXiv*, abs/2009.00713, 2020. URL https://api.semanticscholar.org/CorpusID:221447287.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *ArXiv*, abs/2209.11215, 2022.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. The probability flow ode is provably fast. *ArXiv*, abs/2305.11798, 2023a.

Chen, S., Daras, G., and Dimakis, A. G. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *ArXiv*, abs/2303.03384, 2023b.

Cheng, X. and Bartlett, P. L. Convergence of langevin mcmc in kl-divergence. In *International Conference on Algorithmic Learning Theory*, 2017.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped langevin mcmc: A non-asymptotic analysis. *ArXiv*, abs/1707.03663, 2017.

Clark, D. S. Short proof of a discrete gronwall inequality. *Discret. Appl. Math.*, 16:279–281, 1987. URL https://api.semanticscholar.org/CorpusID:205054476.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10850–10869, 2022. URL https://api.semanticscholar.org/CorpusID:252199918.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. URL https://api.semanticscholar.org/CorpusID:234357997.

Durmus, A. and Moulines, É. Non-asymptotic convergence analysis for the unadjusted langevin algorithm. *arXiv: Statistics Theory*, 2015.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A. A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., and Salimans, T. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022a. URL https://api.semanticscholar.org/CorpusID:252715883.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *ArXiv*, abs/2204.03458, 2022b. URL https://api.semanticscholar.org/CorpusID:248006185.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *ArXiv*, abs/2009.09761, 2020. URL https://api.semanticscholar.org/CorpusID:221818900.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *ArXiv*, abs/2206.06227, 2022a.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, 2022b.

Lu, C., Zheng, K., Bao, F., Chen, J., Li, C., and Zhu, J. Maximum likelihood training for score-based diffusion odes by high-order denoising score matching. In *International Conference on Machine Learning*, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:245335086.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. A. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:234483016.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. URL https://api.semanticscholar.org/CorpusID:248097655.

Rossky, P. J., Doll, J. D., and Friedman, H. L. Brownian dynamics as smart monte carlo simulation. *Journal of Chemical Physics*, 69:4628–4633, 1978.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. URL https://api.semanticscholar.org/CorpusID:248986576.

Sohl-Dickstein, J. N., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585, 2015. URL https://api.semanticscholar.org/CorpusID:14888175.

Song, Y. and Dhariwal, P. Improved techniques for training consistency models. *ArXiv*, abs/2310.14189, 2023. URL https://api.semanticscholar.org/CorpusID:264426451.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *ArXiv*, abs/1907.05600, 2019.

Song, Y., Sohl-Dickstein, J. N., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020. URL https://api.semanticscholar.org/CorpusID:227209335.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235352469.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *ArXiv*, abs/2303.01469, 2023.

Yang, L., Zhang, Z., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Yang, M.-H., and Cui, B. Diffusion models: A comprehensive survey of methods and applications. *ArXiv*, abs/2209.00796, 2022. URL https://api.semanticscholar.org/CorpusID:252070859.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A. Common Symbols

- $*$: the convolution operator defined for two functions $f, g \in L^2(\mathbb{R}^d)$.
  $f * g(\boldsymbol{x}) := \int_{\mathbb{R}^d} f(\boldsymbol{x} - \boldsymbol{u}) g(\boldsymbol{u}) d\boldsymbol{u}$.

- $\lesssim$: less or similar to. If $a \lesssim b$, it means $a \leq Cb$ for some constant C.

- $\sharp$ the push-forward operator associated with a measurable map $f : \mathcal{M} \to \mathcal{N}$. For any measure μ over $\mathcal{M}$, we may define the push-forward measure $f\sharp\mu$ over $\mathcal{N}$ by: $f\sharp\mu(A) = \mu(f^{-1}(A))$, for any A be measurable set in $\mathcal{N}$.

- $\vee$: take the larger one. $a \vee b = \max(a, b)$;
  $\wedge$: take the smaller one. $a \wedge b = \min(a, b)$.

- $\asymp$: asymptotic to. If $a_n \asymp b_n$, it means $\lim_{n\to\infty} a_n/b_n = C$ for some constant C.

- $[\![a, b]\!] := [a, b] \cap \mathbb{Z}$. For example, $[\![1, N]\!] = \{1, 2, 3, \cdots, N\}$, for any $N \in \mathbb{Z}^+$

B. Proofs

B.1. Proofs for Theorem 2

Before we proof our main theorem, we introduce a score perturbation lemma which comes from Lemma 1 in (Chen et al., 2023a).

Lemma B.1 (Lemma 3 in (Chen et al., 2023a)). *(Score perturbation for $\boldsymbol{x}_t$). Suppose $p_t(\boldsymbol{x}) = e^{dt} p_{\text{data}}(e^t \boldsymbol{x}) * \mathcal{N}(\boldsymbol{0}, (1 - e^{-2t})\boldsymbol{I}_d)$ started at p_0, and $\boldsymbol{x}_0 \sim p_0$, $d\boldsymbol{x}_t = -\boldsymbol{x}_t - \nabla \log p_t(\boldsymbol{x})$. Suppose that $\|\nabla^2 \log p_t(\boldsymbol{x})\|_{op} \leq L$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $t \in [0, T]$, where $L \geq 1$. Then,*

$$\mathbb{E}\left[\left\|\frac{\partial}{\partial_t} \nabla \log p_t(\boldsymbol{x}_t)\right\|_2^2\right] \lesssim L^2 d(L + \frac{1}{t})$$

Proof of Theorem 3.7. We divide the left-hand side by Cauchy-Schwarz inequality as follows: let $\boldsymbol{x}_t$ be the solution to the probability flow ODE 4. Notice that $\boldsymbol{f}^{\text{ex}}(\boldsymbol{x}_{t_n}, t_n) = \boldsymbol{x}_\delta = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_1}, t_1)$

$$\begin{aligned}
&\left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}}[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n}, t_n) - \boldsymbol{f}^{\text{ex}}(\boldsymbol{x}_{t_n}, t_n)\|_2^2]\right)^{1/2} \\
=& \left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}}\left[\left\|\sum_{k=1}^{n-1}(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k}, t_k))\right\|_2^2\right]\right)^{1/2} \\
\leq& \sum_{k=1}^{n-1}\left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}}\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k}, t_k)\|_2^2\right]\right)^{1/2} \\
=& \sum_{k=1}^{n-1}\left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}}\left[\left\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_k}^{\boldsymbol{\phi}}, t_k) + \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_k}^{\boldsymbol{\phi}}, t_k) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k}, t_k)\right\|_2^2\right]\right)^{1/2} \\
\leq& \sum_{k=1}^{n-1}\left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}}\left[\left\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_k}^{\boldsymbol{\phi}}, t_k)\right\|_2^2\right]\right)^{1/2} \\
&+ \sum_{k=1}^{n-1}\left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}}\left[\left\|\boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_k}^{\boldsymbol{\phi}}, t_k) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k}, t_k)\right\|_2^2\right]\right)^{1/2} \\
:=& E_1 + E_2
\end{aligned} \tag{17}$$

We can bound E_1 by assumption 3.4: note that

$$\begin{aligned}
E_1 =& \sum_{k=1}^{n-1} \left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}} \left[\left\| \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k}, t_k) \right\|_2^2 \right] \right)^{1/2} \\
=& \sum_{k=1}^{n-1} \left(\mathbb{E}_{\boldsymbol{x}_{t_{k+1}} \sim p_{t_{k+1}}} \left[\left\| \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k}, t_k) \right\|_2^2 \right] \right)^{1/2} \\
\leq& \varepsilon_{\text{cm}} \sum_{k=1}^{n-1} h_k = \varepsilon_{\text{cm}}(t_n - t_1),
\end{aligned} \tag{18}$$

where the last equality we use the fact that when $\boldsymbol{x}_t$ satisfies equation 4 and $\boldsymbol{x}_{t_n} \sim p_{t_n}, \boldsymbol{x}_{t_k} \sim p_{t_k}$ for all $k \leq N$.

Now we turn to bounding the second term. We notice that by Lipschitz assumption 3.5,

$$\begin{aligned}
E_2 =& \sum_{k=1}^{n-1} \left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}} \left[\left\| \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k}, t_k) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k}, t_k) \right\|_2^2 \right] \right)^{1/2} \\
\leq& \sum_{k=1}^{n-1} L_f \left(\mathbb{E}_{\boldsymbol{x}_{t_n} \sim p_{t_n}} \left[\|\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k} - \boldsymbol{x}_{t_k}\|_2^2 \right] \right)^{1/2} + (\varepsilon_{\text{cm}} + \varepsilon_{\text{sc}}) \sum_{k=1}^{n-1} t_k \\
=& \sum_{k=1}^{n-1} L_f \left(\mathbb{E}_{\boldsymbol{x}_{t_{k+1}} \sim p_{t_{k+1}}} \left[\|\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k} - \boldsymbol{x}_{t_k}\|_2^2 \right] \right)^{1/2} + (\varepsilon_{\text{cm}} + \varepsilon_{\text{sc}}) \sum_{k=1}^{n-1} t_k.
\end{aligned} \tag{19}$$

Now let us bound the term $\|\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k} - \boldsymbol{x}_{t_k}\|_2^2$. Note that $\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k}$ is the exponential integrator solution to the ODE 6, we have

$$\begin{aligned}
\mathrm{d}\boldsymbol{x}_t &= -(\boldsymbol{x}_t + \nabla \log p_t(\boldsymbol{x}_t))\mathrm{d}t, \\
\mathrm{d}\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t &= -(\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t + \boldsymbol{s}(\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_{k+1}}, t_{k+1}))\mathrm{d}t.
\end{aligned} \tag{20}$$

for $t_k \leq t \leq t_{k+1}$ with $\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_{k+1}} = \boldsymbol{x}_{t_{k+1}}$. Denote $h_k = t_{k+1} - t_k$, then,

$$\begin{aligned}
\frac{\partial}{\partial_t} \|\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t - \boldsymbol{x}_t\|_2^2 &= 2\langle \hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t - \boldsymbol{x}_t, \frac{\partial}{\partial_t}(\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t - \boldsymbol{x}_t)\rangle \\
&= 2\left(\|\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t - \boldsymbol{x}_t\|_2^2 + \langle \hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t - \boldsymbol{x}_t, \boldsymbol{s}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \nabla \log p_t(\boldsymbol{x}_t)\rangle \right) \\
&\leq (2 + \frac{1}{h_k})\|\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_t - \boldsymbol{x}_t\|_2^2 + h_k \|\boldsymbol{s}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \nabla \log p_t(\boldsymbol{x}_t)\|_2^2.
\end{aligned} \tag{21}$$

By Grönwall's inequality,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_{t_{k+1}} \sim p_{t_{k+1}}} \left[\|\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_k} - \boldsymbol{x}_{t_k}\|_2^2 \right] \\
&\leq \exp((2 + \frac{1}{h_k})h_k) \int_{t_k}^{t_{k+1}} h_k \mathbb{E}_{\boldsymbol{x}_{t_{k+1}} \sim p_{t_{k+1}}} \left[\|\boldsymbol{s}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \nabla \log p_t(\boldsymbol{x}_t)\|_2^2 \right] \mathrm{d}t \\
&\lesssim h_k \int_{t_k}^{t_{k+1}} \mathbb{E}_{\boldsymbol{x}_{t_{k+1}} \sim p_{t_{k+1}}} \left[\|\boldsymbol{s}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \nabla \log p_t(\boldsymbol{x}_t)\|_2^2 \right] \mathrm{d}t.
\end{aligned} \tag{22}$$

We split up the error term as

$$\begin{aligned}
&\|\boldsymbol{s}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \nabla \log p_t(\boldsymbol{x}_t)\|_2^2 \\
&\lesssim \|\boldsymbol{s}(\boldsymbol{x}_{t_{k+1}}, t_{k+1}) - \nabla \log p_{t_{k+1}}(\boldsymbol{x}_{t_{k+1}})\|_2^2 + \|\nabla \log p_{t_{k+1}}(\boldsymbol{x}_{t_{k+1}}) - \nabla \log p_t(\boldsymbol{x}_t)\|_2^2.
\end{aligned} \tag{23}$$

By assumption 3.3, the first term is bounded in expectation by $\varepsilon_{\text{sc}}^2$. By Lemma B.1 and $t_k \leq t \leq t_{k+1}$, the second term is

bounded by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_{t_{k+1}}\sim p_{t_{k+1}}}\left[\|\nabla\log p_{t_{k+1}}(\boldsymbol{x}_{t_{k+1}})-\nabla\log p_t(\boldsymbol{x}_t)\|_2^2\right] &= \mathbb{E}_{\boldsymbol{x}_{t_{k+1}}\sim p_{t_{k+1}}}\left[\left\|\int_t^{t_{k+1}}\frac{\partial}{\partial_u}\nabla\log p_u(\boldsymbol{x}_u)\mathrm{d}u\right\|_2^2\right]\\ &\le (t_{k+1}-t)\int_t^{t_{k+1}}\mathbb{E}\left[\left\|\frac{\partial}{\partial_u}\nabla\log p_u(\boldsymbol{x}_u)\right\|_2^2\right]\mathrm{d}u\\ &\le h_k\int_t^{t_{k+1}}L_s^2 d(L_s+\frac{1}{u})du\\ &\lesssim L_s^2 d h_k^2(L_s+\frac{1}{t_k}),\end{aligned}\tag{24}$$

thus

$$\mathbb{E}_{\boldsymbol{x}_{t_{k+1}}\sim p_{t_{k+1}}}\left[\|\hat{\boldsymbol{x}}_{t_k}^{\boldsymbol{\phi}}-\boldsymbol{x}_{t_k}\|_2^2\right]\le h_k^2(L_s^2 d h_k^2(L_s+\frac{1}{t_k})+\varepsilon_{\text{sc}}^2),$$

Now take it back to equation 19

$$\begin{aligned}E_2 &\lesssim L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}}\sum_{k=1}^n h_k^2 + L_f L_s d^{\frac{1}{2}}\sum_{k=1}^n\frac{h_k^2}{t_k^{\frac{1}{2}}}+L_f\varepsilon_{\text{sc}}\sum_{k=1}^n h_k+(\varepsilon_{\text{sc}}+\varepsilon_{\text{cm}})\sum_{k=1}^n\\ &\lesssim L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h(t_n-t_1)+L_f L_s d^{\frac{1}{2}} h(t_n-t_1)^{\frac{1}{2}}+L_f\varepsilon_{\text{sc}}(t_n-t_1)+(\varepsilon_{\text{sc}}+\varepsilon_{\text{cm}})(t_n-t_1)^2\end{aligned}\tag{25}$$

as we specially designed h_k in assumption 3.6 for $k\in[\![1,N_1]\!]$ such that $h_k\le\frac{t_{k+1}}{2}$, which implies $\sum_{k=1}^n\frac{h_k}{t_k^{\frac{1}{2}}}$ is a constant-factor approximation of the integral $\int_{t_1}^{t_n}\frac{1}{t^{\frac{1}{2}}}\mathrm{d}t\lesssim\sqrt{t_n-t_1}$.

Combining equations 18 and 25, we immediately get the result.

□

B.2. Proof of Theorem 3.8 and Corollary 3.9

Proof of Theorem 3.8. Take a couple of $(\boldsymbol{Y},\boldsymbol{Z})\sim\gamma(\boldsymbol{y},\boldsymbol{z})$ where $\gamma\in\Gamma(\mu,p_{t_n})$ is a coupling between μ and p_{t_n}, that is,

$$\int_{\mathbb{R}^d}\gamma(\boldsymbol{y},\boldsymbol{z})\mathrm{d}\boldsymbol{z}=\mu(\boldsymbol{y}),\tag{26}$$

$$\int_{\mathbb{R}^d}\gamma(\boldsymbol{y},\boldsymbol{z})\mathrm{d}\boldsymbol{y}=p_{t_n}(\boldsymbol{z}).\tag{27}$$

then we have $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{Y},t_n)\sim\boldsymbol{f}_{\boldsymbol{\theta},t_n}\sharp\mu$, $\boldsymbol{f}^{\text{ex}}(\boldsymbol{Z},t_n)\sim p_\delta$, thus

$$\begin{aligned}W_2(\boldsymbol{f}_{\boldsymbol{\theta},t_n}\sharp\mu,p_\delta) &\le\left(\mathbb{E}_\gamma\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{Y},t_n)-\boldsymbol{f}^{\text{ex}}(\boldsymbol{Z},t_n)\|_2^2\right]\right)^{1/2}\\ &\le\left(\mathbb{E}_\gamma\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{Y},t_n)-\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{Z},t_n)\|_2^2\right]\right)^{1/2}+\left(\mathbb{E}_\gamma\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{Z},t_n)-\boldsymbol{f}^{\text{ex}}(\boldsymbol{Z},t_n)\|_2^2\right]\right)^{1/2}\\ &\lesssim L_f\left(\mathbb{E}_\gamma\left[\|\boldsymbol{Y}-\boldsymbol{Z}\|_2^2\right]\right)^{1/2}+t_n(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}})+t_n^2(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}})+t_n(\varepsilon_{\text{cm}}+L_f\varepsilon_{\text{sc}}+L_f L_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+t_n^{\frac{1}{2}}L_f L_s d^{\frac{1}{2}}h\end{aligned}\tag{28}$$

$$\lesssim L_f\left(\mathbb{E}_\gamma\left[\|\boldsymbol{Y}-\boldsymbol{Z}\|_2^2\right]\right)^{1/2}+t_n^2(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}})+t_n(\varepsilon_{\text{cm}}+L_f\varepsilon_{\text{sc}}+L_f L_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+t_n^{\frac{1}{2}}L_f L_s d^{\frac{1}{2}}h.\tag{29}$$

Note that γ can be any coupling between μ and p_{t_n}, this implies

$$W_2(\boldsymbol{f}_{\boldsymbol{\theta},t_n}\sharp\mu,p_\delta)\lesssim L_f W_2(\mu,p_{t_n})+t_n^2(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}})+t_n(\varepsilon_{\text{cm}}+L_f\varepsilon_{\text{sc}}+L_f L_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+t_n^{\frac{1}{2}}L_f L_s d^{\frac{1}{2}}h$$

□

Proof of Corollary 3.9. We first proof that

$$W_2(\boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mathcal{N}(\mathbf{0},\boldsymbol{I}_d),p_\delta) \lesssim (d^{\frac{1}{2}} \vee \mathfrak{m})L_f e^{-T} + T(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + T^2(\varepsilon_{\text{cm}} + \varepsilon_{\text{sc}}). \tag{30}$$

This follows directly from the fact that $e^{-T}\boldsymbol{x}_0 + \sqrt{1-e^{-2T}}\boldsymbol{\xi} \sim p_T(\boldsymbol{x})$, if $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\boldsymbol{I}_d)$, thus

$$W_2(\mathcal{N}(\mathbf{0},\boldsymbol{I}_d),p_T) \le \left(\mathbb{E}_{p_{\text{data}}}[\|e^{-T}\boldsymbol{x}_0 + (\sqrt{1-e^{-2T}}-1)\xi\|_2^2]\right)^{1/2} \lesssim (\sqrt{d}\vee\mathfrak{m})e^{-T},$$

and Theorem 3.8, where we taking $\mu = \mathcal{N}(\mathbf{0},\boldsymbol{I}_d)$.

The corollary then follow from a simple triangular inequality and the fact that

$$\begin{aligned} W_2(p_\delta,p_0) &\le \left(\mathbb{E}_{p_{\text{data}}}[\|(1-e^{-\delta})\boldsymbol{x}_0 + (\sqrt{1-e^{-2\delta}})\xi\|_2^2]\right)^{1/2} \\ &\le ((1-e^{-\delta})^2\mathfrak{m}^2 + (1-e^{-2\delta})d)^{1/2} \\ &\lesssim (\sqrt{d}\vee\mathfrak{m})\sqrt{\delta}. \end{aligned} \tag{31}$$

□

B.3. Proofs for Corollary 3.10 and 3.11

Proof of Corollary 3.10. Take a couple of $(\boldsymbol{Y},\boldsymbol{Z}) \sim \gamma(\boldsymbol{y},\boldsymbol{z})$ where $\gamma \in \Gamma(q_{k-1},p_\delta)$, take $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\boldsymbol{I}_d)$, then we have

$$\begin{aligned} \hat{\boldsymbol{Y}} &= e^{-(t_{k_n}-\delta)}\boldsymbol{Y} + \sqrt{1-e^{-2(t_{k_n}-\delta)}}\boldsymbol{\xi} \sim \mu_k, \\ \hat{\boldsymbol{Z}} &= e^{-(t_{k_n}-\delta)}\boldsymbol{Z} + \sqrt{1-e^{-2(t_{k_n}-\delta)}}\boldsymbol{\xi} \sim p_{t_{n_k}}, \end{aligned} \tag{32}$$

The statement follows from the fact that

$$\begin{aligned} W_2(\boldsymbol{f}_{\boldsymbol{\theta},t_{n_k}}\sharp\mu_k,p_\delta) &\lesssim L_f W_2(\mu_k,p_{t_{n_k}}) + t_n^2(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}}) + t_n(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + t_n^{\frac{1}{2}} L_f L_s d^{\frac{1}{2}} h \\ &\le L_f(\mathbb{E}_\gamma\|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Z}}\|_2^2)^{1/2} + t_n^2(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}}) + t_n(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + t_n^{\frac{1}{2}} L_f L_s d^{\frac{1}{2}} h \\ &= L_f e^{-(t_{k_n}-\delta)}(\mathbb{E}_\gamma\|\boldsymbol{Y}-\boldsymbol{Z}\|_2^2)^{1/2} + t_n^2(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}}) + t_n(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + t_n^{\frac{1}{2}} L_f L_s d^{\frac{1}{2}} h, \end{aligned} \tag{33}$$

and γ can be arbitrary coupling between q_{k-1} and p_δ, which means

$$W_2(q_k,p_\delta) \lesssim L_f e^{-(t_{n_k}-\delta)} W_2(q_{k-1},p_\delta) + t_{n_k}^2(\varepsilon_{\text{cm}}+\varepsilon_{\text{sc}}) + t_{n_k}(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + t_{n_k}^{\frac{1}{2}} L_f L_s d^{\frac{1}{2}} h.$$

□

Proof of Corollary 3.11. For the first statement, fix $n_k \equiv \hat{n}$, note that $L_s \ge 1, t_{\hat{n}} \ge \log(2L_f) + \delta$, according to the proof of Corollary 3.10, we may assume C is the constant factor such that

$$W_2(q_k,p_\delta) \le L_f e^{-(t_{\hat{n}}-\delta)} W_2(q_{k-1},p_\delta) + C_1 t_{\hat{n}}(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + C_2 t_{\hat{n}}^2(\varepsilon_1+\varepsilon_2).$$

where we omit the term with $t_{\hat{n}}^{\frac{1}{2}}$ as it is controlled by the terms with higher orders. Denote

$$D_1 = C_1(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h), D_2 = C_2(\varepsilon_1+\varepsilon_2) \text{ and } \mathcal{E}_k = W_2(q_k,p_{\text{data}})$$

for short, we have

$$\mathcal{E}_k \le L_f e^{-(t_{\hat{n}}-\delta)}\mathcal{E}_{k-1} + t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2,$$

which means

$$\mathcal{E}_k - \frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}}-\delta)}} \le L_f e^{-(t_{\hat{n}}-\delta)}\left(\mathcal{E}_{k-1} - \frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}}-\delta)}}\right),$$

thus,

1. if $\mathcal{E}_{k-1} \le \frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}} - \delta)}}$, $\mathcal{E}_K \le \frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}} - \delta)}}$ for all $K \ge k$;
2. if $\mathcal{E}_{k-1} > \frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}} - \delta)}}$,

$$\mathcal{E}_k \le \frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}} - \delta)}} + \left(L_f e^{-(t_{\hat{n}} - \delta)}\right)^{k-1} \left(\mathcal{E}_1 - \frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}} - \delta)}}\right). \tag{34}$$

These observation shows that $\mathcal{E}_k$ is exponentially upper-bounded by $\frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}} - \delta)}}$, and such an upper bound can be further minimized over $\hat{n} \in [\![1, N]\!]$: in fact if we take $t_{\hat{n}} \approx \log(2L_f) + \delta$, then

$$\frac{t_{\hat{n}} D_1 + t_{\hat{n}}^2 D_2}{1 - L_f e^{-(t_{\hat{n}} - \delta)}} \approx 2\log(2L_f) D_1 1 + 2\log(2L_f)^2 D_2 = O(\log(L_f) D).$$

together with $L_f e^{-(t_{\hat{n}} - \delta)} \approx \frac{1}{2}$, $\mathcal{E}_1 = O((d^{\frac{1}{2}} \vee \mathfrak{m}) L_f e^{-T} + T(\varepsilon_{\text{cm}} + L_f \varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h)) + T^2(\varepsilon_{\text{cm}} + \varepsilon_{\text{sc}})$, this completes the proof of the first statement.

The second statement follows from the fact that

$$\begin{aligned} W_2(p_\delta, p_0) &\le \left(\mathbb{E}_{p_{\text{data}}}[\|(1 - e^{-\delta})\boldsymbol{x}_0 + (\sqrt{1 - e^{-2\delta}})\xi\|_2^2]\right)^{1/2} \\ &\le ((1 - e^{-\delta})^2 \mathfrak{m}^2 + (1 - e^{-2\delta}) d)^{1/2} \\ &\lesssim (\sqrt{d} \vee \mathfrak{m})\sqrt{\delta}. \end{aligned} \tag{35}$$

and a simple use of triangular inequality: $W_2(q_k, p_0) \le W_2(q_k, p_\delta) + W_2(p_\delta, p_0)$. □

B.4. Proof of Lemma 3.13 and Corollary 3.14

Proof of Lemma 3.13. Let $\mu_{\boldsymbol{x},\sigma^2}$ be the density $\mu(\mathrm{d}\boldsymbol{u})$ weighted with the gaussian $\psi_{\sigma^2}(\boldsymbol{u} - \boldsymbol{x}) \sim e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}}$, that is,

$$\mu_{\boldsymbol{x},\sigma^2}(\mathrm{d}\boldsymbol{u}) = \frac{e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}{\int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{u}}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\tilde{\boldsymbol{u}})}.$$

Note that

$$\nabla \log(\mu * \psi_{\sigma^2}(\boldsymbol{x})) = \frac{\nabla \int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}{\int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})} = \frac{\int_{\mathbb{R}^d} -\frac{\boldsymbol{x} - \boldsymbol{u}}{\sigma^2} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}{\int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})} = -\frac{1}{\sigma^2} \mathbb{E}_{\mu_{\boldsymbol{x},\sigma^2}}[\boldsymbol{x} - \boldsymbol{u}],$$

$$\begin{aligned} \nabla^2 \log(\mu * \psi_{\sigma^2}(\boldsymbol{x})) &= \frac{\nabla \otimes \int_{\mathbb{R}^d} -\frac{\boldsymbol{x} - \boldsymbol{u}}{\sigma^2} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}{\int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})} - \left(\frac{\int_{\mathbb{R}^d} -\frac{\boldsymbol{x} - \boldsymbol{u}}{\sigma^2} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}{\int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}\right)^{\otimes 2} \\ &= -\frac{1}{\sigma^2} \boldsymbol{I}_d + \frac{\int_{\mathbb{R}^d} \left(\frac{\boldsymbol{x} - \boldsymbol{u}}{\sigma^2}\right)^{\otimes 2} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}{\int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})} - \left(\frac{\int_{\mathbb{R}^d} -\frac{\boldsymbol{x} - \boldsymbol{u}}{\sigma^2} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}{\int_{\mathbb{R}^d} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{u}\|_2^2}{2\sigma^2}} \mu(\mathrm{d}\boldsymbol{u})}\right)^{\otimes 2} \\ &= \frac{1}{\sigma^4} \int_{\mathbb{R}^d} (\boldsymbol{u} - \mathbb{E}_{\mu_{\boldsymbol{x},\sigma^2}}[\boldsymbol{u}]) \otimes (\boldsymbol{u} - \mathbb{E}_{\mu_{\boldsymbol{x},\sigma^2}}[\boldsymbol{u}]) \mu_{\boldsymbol{x},\sigma^2}(\mathrm{d}\boldsymbol{u}) - \frac{1}{\sigma^2} \boldsymbol{I}_d \end{aligned} \tag{36}$$

where for any vector $\boldsymbol{y} \in \mathbb{R}^d$, we denote $\boldsymbol{y}^{\otimes 2} = \boldsymbol{y} \otimes \boldsymbol{y} = \boldsymbol{y}\boldsymbol{y}^T \in \mathbb{R}^{d \times d}$ as a matrix and denote $\nabla \otimes \boldsymbol{f}(\boldsymbol{x})$ as the Jacobbian matrix $[\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}]_{1 \le i,j \le d}$

Note that if μ is bounded on a set of radius R, so as $\mu_{\boldsymbol{x},\sigma^2}$, then the covariance of $\mu_{\boldsymbol{x},\sigma^2}$ is bounded by R^2 in operator norm.

Now we take $\mu(\boldsymbol{u}) = e^{dt}p_0(e^t\boldsymbol{u})$, which is bounded on a set of radius $e^{-t}R$. Take $\sigma^2 = 1 - e^{-2t}$, we have $\mu * \psi_{\sigma^2}(\boldsymbol{x}) = p_t(\boldsymbol{x})$, thus

$$\|\nabla^2 \log p_t(\boldsymbol{x})\|_{\text{op}} \leq \frac{e^{-2t}R^2}{(1-e^{-2t})^2} + \frac{1}{1-e^{-2t}}.$$

□

Proof of Corollary 3.14. Note that supp $p_{\text{data}} \subseteq B(\mathbf{0}, R)$ implies $\mathbb{E}_{\boldsymbol{x}_0 \sim p_{\text{data}}}[\|\boldsymbol{x}_0\|_2^2] \lesssim R^2$. Taking $\delta \asymp \frac{\varepsilon^2}{R^2 \vee d}$, according to Lemma 3.13, we have

$$\|\nabla^2 \log p_t(\boldsymbol{x})\|_{\text{op}} \leq \frac{e^{-2t}R^2}{(1-e^{-2t})^2} + \frac{1}{1-e^{-2t}} \lesssim \frac{1}{t} \vee \frac{R^2}{t^2} \lesssim \frac{R^2(R^2 \vee d)^2}{\varepsilon^4}, \forall t \geq \delta$$

and thus $\nabla \log p_t(\boldsymbol{x})$ satisfies Assumption 3.2 with $L_s \asymp \frac{R^2(R^2 \vee d)^2}{\varepsilon^4}$ for $t \geq \delta$. Now according to Corollary 3.9, we have that

$$W_2(\boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), p_{\text{data}}) \lesssim (d^{\frac{1}{2}} \vee R)L_f e^{-T} + T(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + \frac{L_f d^{\frac{1}{2}} R^3 (R^2 \vee d)^3 h}{\varepsilon^6}) + \varepsilon,$$

thus if we take $h = O(\frac{\varepsilon^7}{d^{1/2}R^3(R^6 \vee d^3)L_f T})$, $\varepsilon_{\text{cm}} = O(\frac{\varepsilon}{T}), \varepsilon_{\text{sc}} = O(\frac{\varepsilon}{L_f T})$, $T = O(\log(\frac{L_f(\sqrt{d} \vee R)}{\varepsilon}))$, we can guarantee $W_2(\boldsymbol{f}_{\boldsymbol{\theta},T}\sharp\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), p_{\text{data}}) \lesssim \varepsilon$.

Similarly, according to Corollary 3.11, we have

$$W_2(q_k, p_{\text{data}}) \lesssim (\log(L_f) + 2^{-k}T)(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + \frac{L_f d^{\frac{1}{2}} R^3 (R^2 \vee d)^3 h}{\varepsilon^6}) + 2^{-k}(d^{\frac{1}{2}} \vee \mathfrak{m})L_f e^{-T} + \varepsilon,$$

thus if we take $h = O(\frac{\varepsilon^7}{d^{1/2}R^3(R^6 \vee d^3)L_f \log(L_f)})$, $\varepsilon_{\text{cm}} = O(\frac{\varepsilon}{\log(L_f)}), \varepsilon_{\text{sc}} = O(\frac{\varepsilon}{L_f \log(L_f)})$, $k = O(\log(T \vee (\frac{(d \vee \mathfrak{m}^2)L_f L_s}{\varepsilon})))$, we can guarantee $W_2(q_k, p_{\text{data}}) \lesssim \varepsilon$. □

B.5. Proof of Corollary 3.15

Before we prove the corollary 3.15, we first prove the following lemma, which shows that TV error can be bounded after a small time OU regularization.

Lemma B.2. *For any two distribution p and q, running the OU process 15 for p, q individually with time $\tau > 0$, the following TV distance bound holds,*

$$TV(pP_{OU}^{\tau}, qP_{OU}^{\tau}) \lesssim \frac{1}{\sqrt{\tau}} W_1(p, q) \leq \frac{1}{\sqrt{\tau}} W_2(p, q)$$

Proof. Denote $\psi_{\sigma^2}(\boldsymbol{y})$ as the density function to the normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_d)$. We write the $\text{TV}(pP_{\text{OU}}^{\tau}, qP_{\text{OU}}^{\tau})$ into integral form as:

$$\begin{aligned}\text{TV}(pP_{\text{OU}}^{\tau}, qP_{\text{OU}}^{\tau}) &= \frac{1}{2}\int_{\mathbb{R}^d} |(pP_{\text{OU}}^{\tau})(\boldsymbol{x}) - (qP_{\text{OU}}^{\tau})(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} p(\boldsymbol{y})\psi_{1-e^{-2\tau}}(\boldsymbol{x} - e^{-\tau}\boldsymbol{y})\mathrm{d}\boldsymbol{y} - \int_{\mathbb{R}^d} q(\boldsymbol{z})\psi_{1-e^{-2\tau}}(\boldsymbol{x} - e^{-\tau}\boldsymbol{z})\mathrm{d}\boldsymbol{z} \right| \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{37}$$

Taking a coupling $\gamma \in \Gamma(p, q)$,

$$\begin{aligned}\int_{\mathbb{R}^d} \gamma(\boldsymbol{y}, \boldsymbol{z})\mathrm{d}\boldsymbol{z} &= p(\boldsymbol{y}), \\ \int_{\mathbb{R}^d} \gamma(\boldsymbol{y}, \boldsymbol{z})\mathrm{d}\boldsymbol{y} &= q(\boldsymbol{z}),\end{aligned} \tag{38}$$

we have

$$\begin{aligned}
\mathrm{TV}(pP_{\mathrm{OU}}^{\tau}, qP_{\mathrm{OU}}^{\tau}) &= \frac{1}{2}\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^{\mathrm{d}\times d}}\gamma(\boldsymbol{y},\boldsymbol{z})[\psi_{1-e^{-2\tau}}(\boldsymbol{x}-e^{-\tau}\boldsymbol{y})-\psi_{1-e^{-2\tau}}(\boldsymbol{x}-e^{-\tau}\boldsymbol{z})]\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\right|\mathrm{d}\boldsymbol{x}.\\
&\le \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^{d\times d}}\gamma(\boldsymbol{y},\boldsymbol{z})\left|\psi_{1-e^{-2\tau}}(\boldsymbol{x}-e^{-\tau}\boldsymbol{y})-\psi_{1-e^{-2\tau}}(\boldsymbol{x}-e^{-\tau}\boldsymbol{z})\right|\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x}\\
&= \int_{\mathbb{R}^{d\times d}}\gamma(\boldsymbol{y},\boldsymbol{z})\left(\frac{1}{2}\int_{\mathbb{R}^d}\left|\psi_{1-e^{-2\tau}}(\boldsymbol{x}-e^{-\tau}\boldsymbol{y})-\psi_{1-e^{-2\tau}}(\boldsymbol{x}-e^{-\tau}\boldsymbol{z})\right|\mathrm{d}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\\
&= \int_{\mathbb{R}^{d\times d}}\gamma(\boldsymbol{y},\boldsymbol{z})\mathrm{TV}\Big(\psi_{1-e^{-2\tau}}(\cdot-e^{-\tau}\boldsymbol{y}),\psi_{1-e^{-2\tau}}(\cdot-e^{-\tau}\boldsymbol{z})\Big)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\\
&\le \int_{\mathbb{R}^{d\times d}}\gamma(\boldsymbol{y},\boldsymbol{z})\sqrt{\frac{1}{2}\mathrm{KL}\Big(\psi_{1-e^{-2\tau}}(\cdot-e^{-\tau}\boldsymbol{y})\Big\|\psi_{1-e^{-2\tau}}(\cdot-e^{-\tau}\boldsymbol{z})\Big)}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\\
&= \int_{\mathbb{R}^{d\times d}}\gamma(\boldsymbol{y},\boldsymbol{z})\frac{1}{2}\sqrt{\frac{e^{-2\tau}}{1-e^{-2\tau}}\|\boldsymbol{y}-\boldsymbol{z}\|_2^2}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\\
&= \frac{1}{2\sqrt{e^{2\tau}-1}}\int_{\mathbb{R}^{d\times d}}\gamma(\boldsymbol{y},\boldsymbol{z})\|\boldsymbol{y}-\boldsymbol{z}\|_2\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z},
\end{aligned}\tag{39}$$

where for the second inequality we use the fact that $\mathrm{TV}(u,v)\le\sqrt{\frac{1}{2}\mathrm{KL}(u\|v)}$, and the next equality we use the formula of KL divergence between two Gaussian distribution. Noting $\frac{1}{2\sqrt{e^{2\tau}-1}}\le\frac{1}{2\sqrt{2\tau}}$, and taking γ over all coupling $\Gamma(p,q)$, we have

$$\mathrm{TV}(pP_{\mathrm{OU}}^{\tau}, qP_{\mathrm{OU}}^{\tau})\lesssim\frac{1}{\sqrt{\tau}}W_1(p,q)\le\frac{1}{\sqrt{\tau}}W_2(p,q)$$

□

Proof of Corollary 3.15. According to the triangular inequality, Lemma B.2 and equation 11,

$$\begin{aligned}
\mathrm{TV}(q_1P_{\mathrm{OU}}^{\delta}, p_{\mathrm{data}}) &\le \mathrm{TV}(q_1P_{\mathrm{OU}}^{\delta}, p_\delta P_{\mathrm{OU}}^{\delta})+\mathrm{TV}(p_\delta P_{\mathrm{OU}}^{\delta}, p_{\mathrm{data}})\\
&\lesssim \frac{1}{\sqrt{\delta}}W_2(q_1,p_\delta)+\mathrm{TV}(p_{2\delta},p_{\mathrm{data}})\\
&\lesssim \frac{1}{\sqrt{\delta}}\left((d^{\frac{1}{2}}\vee\mathfrak{m})L_fe^{-T}+T(\varepsilon_{\mathrm{cm}}+L_f\varepsilon_{\mathrm{sc}}+L_fL_s^{\frac{3}{2}}d^{\frac{1}{2}}h)\right)+\mathrm{TV}(p_{2\delta},p_{\mathrm{data}}).
\end{aligned}\tag{40}$$

Note that if we take $\delta\asymp\frac{\varepsilon^2}{L_s^2(d\vee\mathfrak{m}^2)}$, then by Lemma 6.4, (Lee et al., 2022b), $\mathrm{TV}(p_{2\delta},p_{\mathrm{data}})\le\varepsilon$, this concludes that

$$\mathrm{TV}(q_1P_{\mathrm{OU}}^{\delta}, p_{\mathrm{data}})\lesssim\frac{L_s(d^{\frac{1}{2}}\vee\mathfrak{m})}{\varepsilon}[(\log L_f+\frac{T}{2^k})(\varepsilon_{\mathrm{cm}}+L_f\varepsilon_{\mathrm{sc}}+L_fL_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+\frac{(d^{\frac{1}{2}}\vee\mathfrak{m})L_f}{2^ke^T}]+\varepsilon.$$

Similarly for q_k, if we take $\delta\asymp\frac{\varepsilon^2}{L_s^2(d\vee\mathfrak{m}^2)}$

$$\mathrm{TV}(q_kP_{\mathrm{OU}}^{\delta}, p_{\mathrm{data}})\tag{41}$$

$$\begin{aligned}
&\le\mathrm{TV}(q_kP_{\mathrm{OU}}^{\delta}, p_\delta P_{\mathrm{OU}}^{\delta})+\mathrm{TV}(p_\delta P_{\mathrm{OU}}^{\delta}, p_{\mathrm{data}})\\
&\lesssim\frac{L_s(d^{\frac{1}{2}}\vee\mathfrak{m})}{\varepsilon}[(\log L_f+\frac{T}{2^k})(\varepsilon_{\mathrm{cm}}+L_f\varepsilon_{\mathrm{sc}}+L_fL_s^{\frac{3}{2}}d^{\frac{1}{2}}h)+\frac{(d^{\frac{1}{2}}\vee\mathfrak{m})L_f}{2^ke^T}]+\varepsilon.
\end{aligned}\tag{42}$$

□

B.6. Proof of Corollary 3.16

We first introduce the following lemma which originally comes from Theorem 5 in (Chen et al., 2023a).

Lemma B.3 (Underdamped corrector, Theorem 5 in (Chen et al., 2023a)). *Denote the Markov kernel $\hat{P}_{ULMC}$ to be defined by the equation 16. Suppose $\gamma \asymp L_s$, ∇U is L_s-Lipschitz, $p \propto \exp(-U)$, and $\mathbb{E}_{\boldsymbol{x}\sim q}\|\boldsymbol{s}(\boldsymbol{x}) - (-\nabla U(\boldsymbol{x}))\|_2^2 \le \varepsilon_{sc}^2$. Denote $\boldsymbol{p} := p \otimes \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, and $\boldsymbol{q} := q \otimes \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. For any $T_{corr} := N\tau \lesssim 1/\sqrt{L_s}$,*

$$TV(\boldsymbol{q}\hat{P}^N_{ULMC}, \boldsymbol{p}) \lesssim \frac{W_2(q,p)}{L_s^{1/4}T_{corr}^{3/2}} + \frac{\varepsilon_{sc}T_{corr}^{1/2}}{L_s^{1/4}} + L_s^{3/4}T_{corr}^{1/2}d^{1/2}\tau.$$

In particular, for $T_{corr} \asymp 1/\sqrt{L_s}$,

$$TV(\boldsymbol{q}\hat{P}^N_{ULMC}, \boldsymbol{p}) \lesssim \sqrt{L_s}W_2(q,p) + \varepsilon_{sc}/\sqrt{L_s} + \sqrt{L_s d}\tau.$$

Under the Lemma B.3, we can immediately get the following result,

Proof of Corollary 3.16. Given any distribution q on $\mathbb{R}^d$, we write $q\hat{P}_{\text{ULMC}}$ to denote the projection onto the $\boldsymbol{z}-$coordinates of $\boldsymbol{q}\hat{P}_{\text{ULMC}}$. Now according to B.3, we have

$$\begin{aligned}\text{TV}(q_1\hat{P}^N_{\text{ULMC}}, p_{\text{data}}) &\le \text{TV}(q_1\hat{P}^N_{\text{ULMC}}, p_\delta) + \text{TV}(p_\delta, p_{\text{data}}) \\ &\lesssim \sqrt{L_s}W_2(q_1, p_\delta) + \varepsilon_{\text{sc}}/\sqrt{L_s} + \sqrt{L_s d}\tau + \text{TV}(p_\delta, p_{\text{data}}).\end{aligned} \tag{43}$$

According to equation 11, we have

$$\begin{aligned}&\text{TV}(q_1\hat{P}^N_{\text{ULMC}}, p_{\text{data}}) \\ &\lesssim (d^{\frac{1}{2}} \vee \mathfrak{m})L_f L_s^{\frac{1}{2}} e^{-T} + TL_s^{\frac{1}{2}}(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + L_s^{-\frac{1}{2}}\varepsilon_{\text{sc}} + L_s^{\frac{1}{2}} d^{\frac{1}{2}}\tau + \varepsilon.\end{aligned} \tag{44}$$

Similarly, we have

$$\begin{aligned}&\text{TV}(q_k\hat{P}^N_{\text{ULMC}}, p_{\text{data}}) \\ &\lesssim (\log(L_f) + \frac{T}{2^k})L_s^{\frac{1}{2}}(\varepsilon_{\text{cm}} + L_f\varepsilon_{\text{sc}} + L_f L_s^{\frac{3}{2}} d^{\frac{1}{2}} h) + \frac{(d^{\frac{1}{2}} \vee \mathfrak{m})L_s^{\frac{1}{2}} L_f}{2^k e^T} + L_s^{-\frac{1}{2}}\varepsilon_{\text{sc}} + L_s^{\frac{1}{2}} d^{\frac{1}{2}}\tau + \varepsilon.\end{aligned} \tag{45}$$

□

C. Additional proofs

Theorem C.1. *[Validation of assumption 3.5] If we strength the assumption 3.3, 3.4 to L^∞ norm: that is,*

$$\begin{aligned}\|\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}, t_n) - \nabla \log p_{t_n}(\boldsymbol{x})\|_2 &\le \varepsilon_{sc}, \forall n \in [\![1, N]\!], \forall \boldsymbol{x} \in \mathbb{R}^d, \\ \|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_n}, t_n)\|_2 &\le \varepsilon_{cm}(t_{n+1} - t_n), \forall n \in [\![1, N-1]\!], \boldsymbol{x}_{t_{n+1}} \in \mathbb{R}^d, \\ \text{where } \hat{\boldsymbol{x}}^{\boldsymbol{\phi}}_{t_n} &= e^{t_{n+1}-t_n}\boldsymbol{x}_{t_{n+1}} + (e^{t_{n+1}-t_n} - 1)\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}),\end{aligned}$$

and if we assume p_{data} satisfy assumption 3.2. In addition we assume p_{data} is a bounded distribution with radius R, then we have the following statement: There exists $L_f := e^{C_2R^2}$ for some $C_2 > 0$, such that

$$\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n}, t_n) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t_n}, t_n)\|_2 \lesssim (t_n - t_1)(\varepsilon_{cm} + L_f\varepsilon_{sc}) + L_f\|\boldsymbol{x}_{t_n} - \boldsymbol{y}_{t_n}\|_2 \tag{46}$$

Proof. For any $\boldsymbol{x}_{t_n}, \boldsymbol{y}_{t_n}$ in $\mathbb{R}^d$, $n \in [\![2, N]\!]$, denote the solution path of the ODE solver with exact score function $\nabla \log p_t(\boldsymbol{x})$ is $\{\boldsymbol{x}_{t_i}\}_{i=1}^n$, $\{\boldsymbol{y}_{t_i}\}_{i=1}^n$, that is,

$$\begin{aligned}\boldsymbol{x}_{t_i} &= e^{h_i}\boldsymbol{x}_{t_{i+1}} + (e^{h_i} - 1)\nabla \log p_{t_i}(\boldsymbol{x}_{t_{i+1}}), \\ \boldsymbol{y}_{t_i} &= e^{h_i}\boldsymbol{y}_{t_{i+1}} + (e^{h_i} - 1)\nabla \log p_{t_i}(\boldsymbol{y}_{t_{i+1}}).\end{aligned}$$

We also denote the solution path of the ODE solver with estimate score function $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}, t_n)$ is $\{\hat{\boldsymbol{x}}_{t_i}\}_{i=1}^n$,$\{\hat{\boldsymbol{y}}_{t_i}\}_{i=1}^n$, that is,

$$\hat{\boldsymbol{x}}_{t_i} = e^{h_i}\hat{\boldsymbol{x}}_{t_{i+1}} + (e^{h_i} - 1)\boldsymbol{s}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_{t_{i+1}}, t_{i+1}),$$

$$\hat{\boldsymbol{y}}_{t_i} = e^{h_i}\hat{\boldsymbol{y}}_{t_{i+1}} + (e^{h_i} - 1)\boldsymbol{s}_{\boldsymbol{\phi}}(\hat{\boldsymbol{y}}_{t_{i+1}}, t_{i+1}).$$

and $\hat{\boldsymbol{x}}_{t_n} = \boldsymbol{x}_{t_n}, \hat{\boldsymbol{y}}_{t_n} = \boldsymbol{y}_{t_n}$. Now we have:

$$\begin{aligned}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n}, t_n) &= \sum_{i=2}^{n}[\boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_i}, t_i) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_{i-1}}, t_{i-1})] + \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_1}, t_1) \\ &= \sum_{i=2}^{n}[\boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_i}, t_i) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_{i-1}}, t_{i-1})] + \hat{\boldsymbol{x}}_{t_1} - \boldsymbol{x}_{t_1} + \boldsymbol{x}_{t_1},\end{aligned}$$

thus

$$\begin{aligned}&\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n}, t_n) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t_n}, t_n)\|_2 \\ &\leq \sum_{i=2}^{n}\|\boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_i}, t_i) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t_{i-1}}, t_{i-1})\|_2 + \sum_{i=2}^{n}\|\boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}_{t_i}, t_i) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}_{t_{i-1}}, t_{i-1})\|_2 \qquad (47) \\ &+ \|\hat{\boldsymbol{x}}_{t_1} - \boldsymbol{x}_{t_1}\|_2 + \|\hat{\boldsymbol{y}}_{t_1} - \boldsymbol{y}_{t_1}\|_2 + \|\boldsymbol{x}_{t_1} - \boldsymbol{y}_{t_1}\|_2 \\ &= 2(t_n - t_1)\varepsilon_{\text{cm}} + \|\hat{\boldsymbol{x}}_{t_1} - \boldsymbol{x}_{t_1}\|_2 + \|\hat{\boldsymbol{y}}_{t_1} - \boldsymbol{y}_{t_1}\|_2 + \|\boldsymbol{x}_{t_1} - \boldsymbol{y}_{t_1}\|_2. \qquad (48)\end{aligned}$$

Now let us analysis $\|\hat{\boldsymbol{x}}_{t_1} - \boldsymbol{x}_{t_1}\|_2$: we denote $\boldsymbol{v}_t(\boldsymbol{x}) = x + \nabla \log p_t(\boldsymbol{x})$, $\hat{\boldsymbol{v}}_t(\boldsymbol{x}) = x + \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}, t)$. Then

$$\boldsymbol{x}_{t_{i-1}} = \boldsymbol{x}_{t_i} + (e^{h_i} - 1)(\boldsymbol{x}_{t_i} + \nabla \log p_{t_i}(\boldsymbol{x}_{t_i})) = \boldsymbol{x}_{t_i} + (e^{h_i} - 1)\boldsymbol{v}_{t_i}(\boldsymbol{x}_{t_i})$$

$$\hat{\boldsymbol{x}}_{t_{i-1}} = \hat{\boldsymbol{x}}_{t_i} + (e^{h_i} - 1)(\hat{\boldsymbol{x}}_{t_i} + \boldsymbol{s}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_{t_i}, t_i)) = \hat{\boldsymbol{x}}_{t_i} + (e^{h_i} - 1)\hat{\boldsymbol{v}}_{t_i}(\hat{\boldsymbol{x}}_{t_i})$$

$$\begin{aligned}\hat{\boldsymbol{x}}_{t_1} - \boldsymbol{x}_{t_1} &= \hat{\boldsymbol{x}}_{t_2} - \boldsymbol{x}_{t_2} + (e^{h_1} - 1)(\boldsymbol{v}_{t_2}(\boldsymbol{x}_{t_2}) - \hat{\boldsymbol{v}}_{t_2}(\hat{\boldsymbol{x}}_{t_2})) \\ &= \hat{\boldsymbol{x}}_{t_2} - \boldsymbol{x}_{t_2} + (e^{h_1} - 1)\left(\boldsymbol{v}_{t_2}(\boldsymbol{x}_{t_2}) - \boldsymbol{v}_{t_2}(\hat{\boldsymbol{x}}_{t_2}) + \boldsymbol{v}_{t_2}(\hat{\boldsymbol{x}}_{t_2}) - \hat{\boldsymbol{v}}_{t_2}(\hat{\boldsymbol{x}}_{t_2})\right) \\ &= \sum_{i=2}^{n}(e^{h_{i-1}} - 1)(\boldsymbol{v}_{t_i}(\boldsymbol{x}_{t_i}) - \boldsymbol{v}_{t_i}(\hat{\boldsymbol{x}}_{t_i})) + \sum_{i=2}^{n}(\boldsymbol{v}_{t_i}(\hat{\boldsymbol{x}}_{t_i}) - \hat{\boldsymbol{v}}_{t_i}(\hat{\boldsymbol{x}}_{t_i})) \qquad (49)\end{aligned}$$

We can immediately get the bound for the second term: $\|\boldsymbol{v}_{t_i}(\hat{\boldsymbol{x}}_{t_i}) - \hat{\boldsymbol{v}}_{t_i}(\hat{\boldsymbol{x}}_{t_i})\|_2 = \|\nabla \log p_{t_i}(\hat{\boldsymbol{x}}_{t_i})) - \boldsymbol{s}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_{t_i}, t_i))\|_2 \leq \varepsilon_{sc}$. For the first term, we have the following statement of $\boldsymbol{v}_t(\boldsymbol{x}) = x + \nabla \log p_t(\boldsymbol{x})$:

$$\|\nabla \boldsymbol{v}_t(\boldsymbol{x})\|_{op} \lesssim R^2 e^{-2t}.$$

Actually according to Lemma 3.13,

$$\|\nabla \boldsymbol{v}_t(\boldsymbol{x})\|_{op} = \|\boldsymbol{I}_d + \nabla^2 \log p_t(\boldsymbol{x})\|_{op} \leq e^{-2t}\left(\frac{R^2}{(1 - e^{-2t})^2} + \frac{1}{1 - e^{-2t}}\right).$$

Now we return to 49, we have

$$\begin{aligned}\|\hat{\boldsymbol{x}}_{t_1} - \boldsymbol{x}_{t_1}\|_2 \leq& \sum_{i=2}^{n}(e^{h_{i-1}} - 1)R^2 e^{-2t_i}\|\hat{\boldsymbol{x}}_{t_i} - \boldsymbol{x}_{t_i}\|_2 + \varepsilon_{\text{sc}}\sum_{i=2}^{n}(e^{h_{i-1}} - 1) \\ \lesssim& R^2\sum_{i=2}^{n}h_{i-1}e^{-2t_i}\|\hat{\boldsymbol{x}}_{t_i} - \boldsymbol{x}_{t_i}\|_2 + (t_n - t_1)\varepsilon_{\text{sc}}\end{aligned}$$

and by the discrete Gronwall's inequality (Clark, 1987),

$$\begin{aligned}\|\hat{\boldsymbol{x}}_{t_1} - \boldsymbol{x}_{t_1}\|_2 &\leq (t_n - t_1)\varepsilon_{\text{sc}} + \varepsilon_{\text{sc}}R^2\sum_{i=2}^{n}(t_i - t_1)h_{i-1}e^{-2t_i}\exp\left(C_1R^2\sum_{j=2}^{i}h_{j-1}e^{-2t_j}\right) \\ &\leq (t_n - t_1)\varepsilon_{\text{sc}} + \varepsilon_{\text{sc}}R^2(t_n - t_1)\sum_{i=2}^{n}h_{i-1}e^{-2t_i}\exp\left(C_2R^2(1 - e^{-2t_i})\right) \\ &\lesssim (t_n - t_1)\varepsilon_{\text{sc}} + \varepsilon_{\text{sc}}(t_n - t_1)(\exp\left(C_2R^2(1 - e^{-2t_n})\right) - 1) \\ &= (t_n - t_1)\varepsilon_{\text{sc}}e^{C_2R^2}\end{aligned}$$

Thus the equation 48 now becomes

$$\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n}, t_n) - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t_n}, t_n)\|_2 \leq 2(t_n - t_1)\varepsilon_{\text{cm}} + 2(t_n - t_1)\varepsilon_{\text{sc}} e^{C_2 R^2} + \|\boldsymbol{x}_{t_1} - \boldsymbol{y}_{t_1}\|_2, \tag{50}$$

and similar to 49 the term $\|\boldsymbol{x}_{t_1} - \boldsymbol{y}_{t_1}\|_2$ satisfies

$$\|\boldsymbol{x}_{t_1} - \boldsymbol{y}_{t_1}\|_2 \leq \sum_{i=2}^{n} (e^{h_{i-1}} - 1)\|\boldsymbol{v}_{t_i}(\boldsymbol{x}_{t_i}) - \boldsymbol{v}_{t_i}(\boldsymbol{y}_{t_i})\|_2 \tag{51}$$

$$\lesssim R^2 \sum_{i=2}^{n} h_{i-1} e^{-2t_i} \|\boldsymbol{x}_{t_i} - \boldsymbol{y}_{t_i}\|_2, \tag{52}$$

by the same discrete Gronwall's inequality, we have

$$\|\boldsymbol{x}_{t_1} - \boldsymbol{y}_{t_1}\|_2 \lesssim e^{C_2 R^2} \|\boldsymbol{x}_{t_n} - \boldsymbol{y}_{t_n}\|_2. \tag{53}$$

Now combine 53, 49, 48, we immediately get the result. □

Lemma C.2. *Assuming 3.3 and 3.4, then* $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}, t_2) - e^{h_1}\boldsymbol{x}}{e^{h_1} - 1}$ *is a score model approximating* $\nabla \log p_{t_2}(\boldsymbol{x})$ *with* L^2 *error* $\varepsilon_{cm} + \varepsilon_{sc}$.

Proof. Letting $h_1 = t_2 - t_1 = t_2 - \delta$, the consistency loss assumption 3.4 ensures

$$\mathbb{E}_{\boldsymbol{x}_{t_2} \sim p_{t_2}} \left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_2}, t_2) - \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}^{\phi}_{t_1}, t_1)\|_2^2 \right] \leq \varepsilon_{\text{cm}}^2 h_1^2,$$

where we have $\boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}^{\phi}_{t_1}, t_1) = \hat{\boldsymbol{x}}^{\phi}_{t_1} = e^{h_1}\boldsymbol{x}_{t_2} + (e^{h_1} - 1)\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_2}, t_2)$, thus

$$\mathbb{E}_{\boldsymbol{x}_{t_2} \sim p_{t_2}} \left[\left\| \frac{\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_2}, t_2) - e^{h_1}\boldsymbol{x}_{t_2}}{e^{h_1} - 1} - \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_2}, t_2) \right\|_2^2 \right] \leq \varepsilon_{\text{cm}}^2 \frac{h_1^2}{(e^{h_1} - 1)^2} \leq \varepsilon_{\text{cm}}^2,$$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_{t_2} \sim p_{t_2}} \left[\left\| \frac{\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_2}, t_2) - e^{h_1}\boldsymbol{x}_{t_2}}{e^{h_1} - 1} - \nabla \log p_{t_2}(\boldsymbol{x}) \right\|_2^2 \right] \\
\leq &\mathbb{E}_{\boldsymbol{x}_{t_2} \sim p_{t_2}} \left[\left(\left\| \frac{\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_2}, t_2) - e^{h_1}\boldsymbol{x}_{t_2}}{e^{h_1} - 1} - \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_2}, t_2) \right\|_2 + \left\| \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_2}, t_2) - \nabla \log p_{t_2}(\boldsymbol{x}) \right\|_2 \right)^2 \right] \\
\leq &(\varepsilon_{\text{cm}} + \varepsilon_{\text{sc}})^2.
\end{aligned}$$

□

Lemma C.3. *Under the OU scheduler 5, let* $\Delta t := \max_{1 \leq n \leq N-1}(t_{n+1} - t_n)$. *Assume* $\boldsymbol{f}_{\boldsymbol{\theta}^-}$ *is twice continuously differentiable with bounded second derivatives, and* $\mathbb{E}[\|\nabla \log p_{t_n}(\boldsymbol{x}_{t_n})\|_2^2] < \infty$. *Assume the CD objective 8 is defined with the exponential integrator 9 and exact score model* $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}, t) = \nabla \log p_t(\boldsymbol{x})$, *then we can define the CT objective as*

$$\mathcal{L}_{CT}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-) := \mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\theta}}(e^{-t_{n+1}}\boldsymbol{x}_0 + \sqrt{1 - e^{-2t_{n+1}}}\boldsymbol{z}, t_{n+1}) - \boldsymbol{f}_{\boldsymbol{\theta}^-}(e^{-t_n}\boldsymbol{x}_0 + \frac{1 - e^{-(t_n + t_{n+1})}}{\sqrt{1 - e^{-2t_{n+1}}}}\boldsymbol{z}, t_n)\|_2^2], \tag{54}$$

where $\boldsymbol{x}_0 \sim p_{\text{data}}$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. *Then we have*

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_{CT}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_{CD}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-) + O((\Delta t)^2)$$

Proof. We first prove that, if $\boldsymbol{x}_0 \sim p_{\text{data}}, \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), \boldsymbol{x}_t = e^{-t}\boldsymbol{x}_0 + \sqrt{1-e^{-2t}}\boldsymbol{z}$, then we have $\nabla \log p_t(\boldsymbol{x}) = -\mathbb{E}[\frac{\boldsymbol{z}}{\sqrt{1-e^{-2t}}}|\boldsymbol{x}_t]$. This comes from the fact that

$$\begin{aligned}
\nabla \log p_t(\boldsymbol{x}_t) &= \frac{\int_{\mathbb{R}^d} p_{\text{data}}(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t} p(\boldsymbol{x}_t|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0}{\int_{\mathbb{R}^d} p_{\text{data}}(\boldsymbol{x}_0)p(\boldsymbol{x}_t|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0} \\
&= \frac{\int_{\mathbb{R}^d} p_{\text{data}}(\boldsymbol{x}_0)p(\boldsymbol{x}_t|\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0}{p_t(\boldsymbol{x}_t)} \\
&= \int_{\mathbb{R}^d} \frac{p_{\text{data}}(\boldsymbol{x}_0)p(\boldsymbol{x}_t|\boldsymbol{x}_0)}{p_t(\boldsymbol{x}_t)} \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0 \\
&= \int_{\mathbb{R}^d} p(\boldsymbol{x}_0|\boldsymbol{x}_t)\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0 \\
&= \mathbb{E}[\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)|\boldsymbol{x}_t],
\end{aligned}$$

and $p(\boldsymbol{x}_t|\boldsymbol{x}_0) \sim e^{-\frac{\|\boldsymbol{x}_t - e^{-t}\boldsymbol{x}_0\|^2}{2(1-e^{-2t})}}$, which means

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0) = -\frac{\boldsymbol{x}_t - e^{-t}\boldsymbol{x}_0}{1-e^{-2t}} = -\frac{\boldsymbol{z}}{\sqrt{1-e^{-2t}}}.$$

Now we may rewrite $\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}_{\text{CD}}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$ as

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}_{\text{CD}}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-) = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}) - \boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{(t_{n+1}-t_n)}\boldsymbol{x}_{t_{n+1}} + (e^{(t_{n+1}-t_n)}-1)\nabla \log p_{t_{n+1}}(\boldsymbol{x}_{t_{n+1}}), t_n\right)\|_2^2] \\
=&\frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1})\|_2^2 - 2\left\langle \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}), \boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{(t_{n+1}-t_n)}\boldsymbol{x}_{t_{n+1}} + (e^{(t_{n+1}-t_n)}-1)\nabla \log p_{t_{n+1}}(\boldsymbol{x}_{t_{n+1}}), t_n\right)\right\rangle\right],
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}_{\text{CT}}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-) = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\left[\left\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}) - \boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{-t_n}\boldsymbol{x}_0 + \frac{1-e^{-(t_n+t_{n+1})}}{\sqrt{1-e^{-2t_{n+1}}}}\boldsymbol{z}, t_n\right)\right\|_2^2\right] \\
=&\frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1})\|_2^2 - 2\left\langle \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}), \boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{-t_n}\boldsymbol{x}_0 + \frac{1-e^{-(t_n+t_{n+1})}}{\sqrt{1-e^{-2t_{n+1}}}}\boldsymbol{z}, t_n\right)\right\rangle\right],
\end{aligned}$$

thus

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}_{\text{CD}}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-) - \frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}_{\text{CT}}^N(\boldsymbol{\theta}, \boldsymbol{\theta}^-) \\
=&2\mathbb{E}\Bigg[\Bigg\langle \frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}), \\
&\boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{-t_n}\boldsymbol{x}_0 + \frac{1-e^{-(t_n+t_{n+1})}}{\sqrt{1-e^{-2t_{n+1}}}}\boldsymbol{z}, t_n\right) - \boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{(t_{n+1}-t_n)}\boldsymbol{x}_{t_{n+1}} + (e^{(t_{n+1}-t_n)}-1)\nabla \log p_{t_{n+1}}(\boldsymbol{x}_{t_{n+1}}), t_n\right)\Bigg\rangle\Bigg].
\end{aligned}$$

Notice that

$$\begin{aligned}
&\boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{-t_n}\boldsymbol{x}_0 + \frac{1-e^{-(t_n+t_{n+1})}}{\sqrt{1-e^{-2t_{n+1}}}}\boldsymbol{z}, t_n\right) - \boldsymbol{f}_{\boldsymbol{\theta}^-}\left(e^{(t_{n+1}-t_n)}\boldsymbol{x}_{t_{n+1}} + (e^{(t_{n+1}-t_n)}-1)\nabla \log p_{t_{n+1}}(\boldsymbol{x}_{t_{n+1}}), t_n\right) \\
=&\nabla_{\boldsymbol{x}}\boldsymbol{f}_{\boldsymbol{\theta}^-}(e^{(t_{n+1}-t_n)}\boldsymbol{x}_{t_{n+1}} + (e^{(t_{n+1}-t_n)}-1)\nabla \log p_{t_{n+1}}(\boldsymbol{x}_{t_{n+1}}), t_n)\cdot \\
&\left(\frac{1-e^{-(t_n+t_{n+1})}}{\sqrt{1-e^{-2t_{n+1}}}}\boldsymbol{z} - e^{(t_{n+1}-t_n)}\sqrt{1-e^{-2t_n}}\boldsymbol{z} + (e^{(t_{n+1}-t_n)}-1)\nabla \log p_{t_{n+1}}(\boldsymbol{x}_{t_{n+1}})\right) + O((t_{n+1}-t_n)^2) \\
=&\nabla_{\boldsymbol{x}}\boldsymbol{f}_{\boldsymbol{\theta}^-}(e^{(t_{n+1}-t_n)}\boldsymbol{x}_{t_{n+1}} + (e^{t_{n+1}-t_n}-1)\nabla \log p_{t_{n+1}}(\boldsymbol{x}_{t_{n+1}}), t_n)\cdot \\
&(e^{(t_{n+1}-t_n)}-1)\left(\frac{\boldsymbol{z}}{\sqrt{1-e^{-2t_{n+1}}}} - \mathbb{E}\left[\frac{\boldsymbol{z}}{\sqrt{1-e^{-2t_{n+1}}}}\middle|\boldsymbol{x}_{t_{n+1}}\right]\right) + O((t_{n+1}-t_n)^2),
\end{aligned}$$

we have (writing the terms related of $\boldsymbol{x}_{t_{n+1}}, t_{n+1}$ with $C(\boldsymbol{x}_{t_{n+1}}, t_{n+1})$ for simplicity)

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}_{\mathrm{CD}}^{N}(\boldsymbol{\theta}, \boldsymbol{\theta}^{-}) - \frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}_{\mathrm{CT}}^{N}(\boldsymbol{\theta}, \boldsymbol{\theta}^{-}) \\
=&\mathbb{E}[C(\boldsymbol{x}_{t_{n+1}}, t_{n+1})(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z}|\boldsymbol{x}_{t_{n+1}}])] + O((t_{n+1} - t_n)^2) \\
=&\mathbb{E}[C(\boldsymbol{x}_{t_{n+1}}, t_{n+1})\boldsymbol{z}] - \mathbb{E}[\mathbb{E}[C(\boldsymbol{x}_{t_{n+1}}, t_{n+1})\boldsymbol{z}|\boldsymbol{x}_{t_{n+1}}]] + O((t_{n+1} - t_n)^2) \\
=&O((t_{n+1} - t_n)^2)
\end{aligned}$$

which finishes the proof. □