Restoration-Degradation Beyond Linear Diffusions: A Non-Asymptotic Analysis For DDIM-type Samplers

Sitan Chen [1] Giannis Daras [2] Alexandros G. Dimakis [2]

Abstract

We develop a framework for non-asymptotic analysis of deterministic samplers used for diffusion generative modeling. Several recent works have analyzed *stochastic* samplers using tools like Girsanov's theorem and a chain rule variant of the interpolation argument. Unfortunately, these techniques give vacuous bounds when applied to deterministic samplers. We give a new operational interpretation for deterministic sampling by showing that one step along the probability flow ODE can be expressed as two steps: 1) a restoration step that runs gradient ascent on the conditional log-likelihood at some infinitesimally previous time, and 2) a degradation step that runs the forward process using noise pointing back towards the current iterate. This perspective allows us to extend denoising diffusion implicit models to general, non-linear forward processes. We then develop the first polynomial convergence bounds for these samplers under mild conditions on the data distribution.

1. Introduction

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019) have emerged as a powerful framework for generative modeling. One of the core components is corrupting samples at different scales, slowly molding the data into noise. The corruption process, also known as the *forward process*, can be fully described by the intermediate distributions, $\{q_t\}_{t\in[0,T]}$, it defines. Diffusion models learn to revert the forward Process by approximating the *score* function, i.e. the gradient of the log-likelihood, of the intermediate distributions q_t.

Once the score function has been learned, one can generate samples by running the reverse stochastic differential equation (SDE) associated with the forward Process (Anderson, 1982; Song et al., 2020). In practice however, one can only run a suitable discretization of the SDE, and due to the recursive nature of the sampling procedure, the discretization error from previous steps can accumulate, leading to sampling drift away from the true reverse process. Other sources of error come from the approximation error in estimating the score (Sehwag et al., 2022; Ho et al., 2020; Nichol & Dhariwal, 2021) and from the starting distribution. Controlling the propagation of errors in the reverse SDE has been studied in the recent works of Block et al. (2022); De Bortoli et al. (2021); De Bortoli (2022); Liu et al. (2022); Lee et al. (2022a); Pidstrigach (2022); Lee et al. (2022b); Chen et al. (2022a;b).

A second family of sampling methods is that of *deterministic* samplers. These samplers can be derived by deterministic ODE processes that satisfy the same Fokker-Planck equations (and hence have the same marginals) as the reverse SDE (Song et al., 2020). A different work, DDIM (Song et al., 2021), derives deterministic samplers by considering a non-Markovian diffusion process that leads to the same training objective, but a different reverse process. The two formulations turn out to be equivalent up to a reparametrization of the *probability flow ODE* (Song et al., 2020; Karras et al., 2022). DDIM samplers can be interpreted as iterating a combination of two steps: a restoration step that recovers some rough final reconstruction of the current iterate at time t, and a degradation step that corrupts this rough estimate to time $t+h$. This interpretation allowed the generalization of DDIM to general linear diffusions (Zhang et al., 2022; Daras et al., 2022b; Bansal et al., 2022; Zhang et al., 2022).

While stochastic samplers are typically state-of-the-art for image-generation, they require a large *number of function evaluations* which makes them impractical for many applications. The gap between sample quality for deterministic and stochastic samplers has been significantly narrowed in the recent work of Karras et al. (2022). Deterministic samplers are typically much faster (Song et al., 2021; Nichol & Dhariwal, 2021) and also useful for computing likelihoods (Ho et al., 2020; Song et al., 2020). Further, one of

[*] Equal contribution [1] John A. Paulson School of Engineering and Applied Sciences, Harvard University, MA, USA [2] Department of Computer Science, UT Austin, TX, USA. Correspondence to: Sitan Chen <sitan@seas.harvard.edu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

the most successful techniques for accelerating diffusion models, Progressive Distillation (Salimans & Ho, 2022), requires deterministic samplers. Finally, deterministic samplers allow the exploration of the semantic latent space of the trained network (Kwon et al., 2022).

Despite their significance, there is currently limited theoretical understanding for deterministic samplers. Specifically, there is no analysis for their non-asymptotic convergence behavior, in contrast to stochastic samplers. The ODE analysis is challenging because Girsanov's theorem– the main tool for bounding the propagation of errors when implementing the reverse SDE– and related techniques all yield vacuous bounds for deterministic samplers (see Section 5).

Our contributions are twofold. We first propose a new operational interpretation for the reverse ODE that generalizes DDIM sampling to arbitrary, non-linear forward processes.

Theorem 1.1 (Informal, see Section 3)**.** *Denote by h the infinitesimally small step size with which we discretize the probability flow ODE. Let $\ell \in \mathbb{N}$ be a parameter for which $\ell \to \infty$ and $\ell h \to 0$. For any forward process, running the probability flow ODE for time h is equivalent to running the following two steps: 1) restoring the current iterate to ℓh time steps in the past via a step of gradient ascent on conditional log-likelihood, 2) degrading this by $(\ell - 1)h$ steps by simulating the forward process with noise pointing in the direction of the current iterate.*

We then complement this new asymptotic result with a non-asymptotic proof that the sampler from this operational interpretation converges to the true process. This yields a deterministic sampling analogue of recent non-asymptotic analyses of stochastic samplers for diffusion models (Chen et al., 2022b; Lee et al., 2022b; Chen et al., 2022a):

Theorem 1.2 (Informal, see Theorem 4.3)**.** *Under mild assumptions on the smoothness of the data distribution (in particular, the distribution can be arbitrarily non-log-concave), the deterministic sampler arising from Theorem 1.1 generates samples for which the KL divergence with respect to the data distribution is small provided ℓh and ℓ^{-1} are polynomially small in the dimension and other problem-specific parameters.*

As a corollary, our techniques imply that the same bounds hold for the Euler discretization of the probability flow ODE, yielding, to our knowledge, the first non-asymptotic analysis of this sampler.

2. Preliminaries

In this work we consider a general forward process driven by a stochastic differential equation (SDE) of the form:

$$\mathrm{d}x_t = f_t(x_t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t, \qquad x_0 \sim q\,.$$

where (W_t) is a standard Brownian motion in $\mathbb{R}^d$. Let q_t denote the law of x_t, so that $q_0 = q$.

Suppose we run the forward process up to a terminal time $T > 0$. Under mild conditions on the diffusion (see e.g. (Anderson, 1982; Föllmer, 1985; Cattiaux et al., 2022)) which are satisfied by the processes we consider in this work, there is a suitable reverse process given by an SDE such that the marginal distribution at time t is given by q_{T-t}. For convenience, we will often refer to q_{T-t} as $q_t^{\leftarrow}$.

In fact, there is an entire family of SDEs with this property. For any $\lambda > 0$, consider the process $(x_t^{\leftarrow,\lambda})_{0\le t\le T}$ given by

$$\begin{aligned}\mathrm{d}x_t^{\leftarrow,\lambda} &= -\{f_{T-t}(x_t^{\leftarrow,\lambda}) \\ &-\frac{1+\lambda^2}{2}g(T-t)^2\nabla \ln q_t^{\leftarrow}(x_t^{\leftarrow,\lambda})\}\,\mathrm{d}t \\ &+\lambda g(T-t)\mathrm{d}W_t\,, \qquad x_0^{\leftarrow,\lambda} \sim q_0^{\leftarrow}\,.\end{aligned}$$

By checking the Fokker-Planck equations, one sees that the marginal distribution of $x_t^{\leftarrow,\lambda}$ is indeed given by $q_t^{\leftarrow}$.

One notable process in this family corresponds to the case of $\lambda = 0$. This is a *deterministic* process, denoted $(x_t^{\leftarrow})_{0\le t\le T}$, driven by the probability flow ODE (Song et al., 2020).

$$\mathrm{d}x_t^{\leftarrow} = -\{f_{T-t}(x_t^{\leftarrow}) - \frac{1}{2}g(T-t)^2\nabla \ln q_t^{\leftarrow}(x_t^{\leftarrow})\}\,\mathrm{d}t\,,$$

with $x_0^{\leftarrow} \sim q_0^{\leftarrow}$.

In the diffusion model literature, there are two popular choices of forward process: the *variance exploding (VE) SDE* (Song et al., 2020; Song & Ermon, 2019; 2020), which corresponds to $f_t(x_t) = 0,\ g(t) = \sqrt{\frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t}}$ for some increasing function σ_t^2; and the *variance preserving (VP) SDE* (Ho et al., 2020), which corresponds to $f_t(x_t) = -\frac{1}{2}\beta_t x_t,\ g(t) = \sqrt{\beta_t}$ for some variance schedule β_t. These two choices are used in state-of-the-art diffusion models (Dhariwal & Nichol, 2021; Kim et al., 2022) and form the backbone of systems like DALL·E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022).

3. Operational Interpretation for the probability flow ODE

3.1. Warmup: linear SDEs and DDIM

We begin by recalling the interpretation of the probability flow ODE associated to the variance exploding (VE) (Song et al., 2020) SDE as a *denoising diffusion implicit model* (DDIM) (Song et al., 2021). For simplicity of exposition, we specialize to the case of $\sigma_t^2 = t$, which corresponds to the forward process

$$\mathrm{d}x_t = \mathrm{d}W_t, \qquad x_0 \sim q\,.$$

According to (2), the associated probability flow ODE is:

$$\mathrm{d}x_t^{\leftarrow} = \frac{1}{2}\nabla \ln q_t^{\leftarrow}(x_t^{\leftarrow})\,\mathrm{d}t, \qquad x_t^{\leftarrow} \sim q_T,$$

so that the marginal distribution of $x_t^{\leftarrow}$ is $q_t^{\leftarrow}$ for any $0 \leq t \leq T$. The perspective of DDIM offers an interesting operational interpretation of (3.1). Fix some infinitesimally small step size h, and consider the following procedure for forming $x_{t+h}^{\leftarrow}$ given $x_t^{\leftarrow}$. We first produce an estimate for the *beginning* x_0 of the forward process. Note that

$$x_t^{\leftarrow} = x_{T-t} = x_0 + \varepsilon\sqrt{T-t}$$

for $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$, so by Tweedie's formula (Efron, 2011), the mean of the posterior distribution over x_0 given $x_t^{\leftarrow}$, i.e. $\mathbb{E}[x_0|x_t]$, is exactly:

$$z \triangleq E[x_0|x_t^{\leftarrow}] = x_t^{\leftarrow} + (T-t)\,\nabla \ln q_t^{\leftarrow}(x_t^{\leftarrow})\,.$$

Starting from z and degrading it along the forward process from time 0 to time $T-t$, we would end up with $z + \gamma\sqrt{T-t}$ for some Gaussian noise $\gamma \sim \mathcal{N}(0, \mathrm{Id})$.

Here is the key idea behind DDIMs: suppose we instead took γ to be the solution to

$$x_t^{\leftarrow} = z + \gamma\sqrt{T-t}\,,$$

i.e. suppose we took γ to be the "simulated noise" that would be needed to degrade z into $x_t^{\leftarrow}$, rather than fresh Gaussian noise.

Now imagine running the forward process to degrade z from time 0 to time $T-(t+h)$, but using this simulated noise $\gamma = \frac{x_t^{\leftarrow}-z}{\sqrt{T-t}}$ instead of Gaussian noise. It turns out that the resulting vector, which we will define $x_{t+h}^{\leftarrow}$ to be, is approximately what we would get by running the probability flow ODE for time h starting at $x_t^{\leftarrow}$!

Indeed, the result of degrading z in this fashion is

$$x_{t+h}^{\leftarrow} = z + \sqrt{T-(t+h)}\cdot\frac{x_t^{\leftarrow}-z}{\sqrt{T-t}}$$

$$= x_t^{\leftarrow} + (T-t)\cdot\left(1-\sqrt{1-\frac{h}{T-t}}\right)\cdot\nabla \ln q_{T-t}(x_t^{\leftarrow}).$$

As $h \to 0$, $x_{t+h}^{\leftarrow} \to x_t^{\leftarrow} + \frac{1}{2}\nabla \ln q_{T-t}(x_t^{\leftarrow})$, so the above interpretation indeed recovers the probability flow ODE (3.1). The above generalizes without much difficulty to any linear diffusion (Daras et al., 2022b; Bansal et al., 2022).

3.2. General diffusions

Let us now consider the setting where the forward process is given by an arbitrary diffusion as in Eq. (2) in Section 2, so that the associated probability flow ODE is given by Eq. (2). Unfortunately, as soon as we step away from the linear setting, the operational interpretation from the previous section breaks down. The key issue is that when forming our estimate z for the beginning of the forward process, there is no longer any simple expression for the posterior mean conditioned on $x_t^{\leftarrow}$.

Restoration operator. To get around this issue, our first insight is: instead of deriving an estimate for the beginning of the forward process, we instead derive one for the process *ℓh units of time in the past*, i.e. at time $T-t-\ell h$ of the forward process. In the previous section, we implicitly took $\ell = (T-t)/h$, but now ℓ is a parameter that needs to be tuned. Crucially, selecting ℓ such that $\ell h \to 0$ allows us to linearize around $T-t$. In analogy with (3.1), we get the approximate relation

$$\begin{aligned} x_t^{\leftarrow} &= x_{T-t} \\ &\approx x_{T-t-\ell h} + \ell h\, f_{T-t-\ell h}(x_{T-t-\ell h}) \\ &\quad + g(T-t-\ell h)\sqrt{\ell h}\cdot\varepsilon \\ &\approx x_{T-t-\ell h} + \ell h\, f_{T-t}(x_t^{\leftarrow}) + g(T-t)\sqrt{\ell h}\cdot\varepsilon \end{aligned}$$

for $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$, where the approximations hold up to $o(h)$ additive error. Rearranging, we see that $x_{T-t-\ell h}$ is simply $x_t^{\leftarrow} - \ell h f_{T-t}(x_t^{\leftarrow})$ plus some Gaussian noise of variance $\ell h g(T-t)^2$. So, again by Tweedie's formula, we find that the mean of the posterior distribution over $x_{T-t-\ell h}$ given $x_t^{\leftarrow}$ is approximately

$$z \triangleq x_t^{\leftarrow} - \ell h\left\{f_{T-t}(x_t^{\leftarrow}) - g(T-t)^2\nabla \ln q^{\leftarrow}(x_t^{\leftarrow})\right\}.$$

Borrowing terminology from (Bansal et al., 2022), we refer to the map from $x_t^{\leftarrow}$ to z as the *restoration operator*. Formally, for $t > s > 0$, define the restoration operator $R_{t\to s}(\cdot)$ by

$$R_{t\to s}(x) \triangleq x - (t-s)f_t(x) + (t-s)g(t)^2\nabla \ln q_t(x)$$

so that $z = R_{T-t\to T-t-\ell h}(x_t^{\leftarrow})$.

Restoration operator as gradient ascent. There turns out to be a different way of thinking about the restoration operator, namely as one step of gradient ascent.

Formally, given times $0 < t < s$, consider maximizing the conditional log-likelihood $\ln q_s^{\leftarrow}(\cdot \mid x_t^{\leftarrow})$. This is equivalent to maximizing

$$\ell_{x_t^{\leftarrow}}(x) \triangleq \ln q_t^{\leftarrow}(x_t^{\leftarrow} \mid x_s^{\leftarrow} = x) + \ln q_s^{\leftarrow}(x).$$

For s which is infinitesimally larger than t, the law of $x_t^{\leftarrow}$ conditioned on $x_s^{\leftarrow} = x$ is Gaussian with mean and covariance approximately $x + f_{T-s}(x_s^{\leftarrow})\,(t-s)$ and $g(T-t)^2(t-s)\,\mathrm{Id}$. We can thus compute the gradient of

(3.2) to get

$$\nabla \ell_{x_t^{\leftarrow}}(x) \approx \frac{1}{g(T-t)^2(t-s)}\big(\mathrm{Id} + (t-s)\,\nabla f_{T-s}(x)\big) \cdot\big(x_t^{\leftarrow} - x - f_{T-s}(x)\,(t-s)\big) + \nabla \ln q_s^{\leftarrow}(x)\,.$$

Now consider taking a single gradient step with learning rate η starting from $x_t^{\leftarrow}$ to get $x_t^{\leftarrow} + \eta \nabla \ell_{x_t^{\leftarrow}}(x_t^{\leftarrow})$. In Appendix A, we show that in the special case where q is Gaussian and the forward process is Ornstein-Uhlenbeck, the correct choice of learning rate to maximize the conditional log-likelihood with just one step of gradient ascent is

$$\eta \triangleq 2g(t)^2 \cdot (t-s)\,.$$

In this case, note that

$$\begin{aligned} x_t^{\leftarrow} + \eta\nabla \ell_{x_t^{\leftarrow}}(x_t^{\leftarrow}) &\approx x_t^{\leftarrow} - (t-s) f_{T-t}(x_t^{\leftarrow}) \\ &\quad -(t-s)^2(\nabla f_{T-s}(x_t^{\leftarrow}))f_{T-s}(x_t^{\leftarrow}) \\ &\quad +(t-s)g(T-t)^2\nabla \ln q_s^{\leftarrow}(x_t^{\leftarrow}) \\ &\approx x_t^{\leftarrow} - (t-s)f_{T-t}(x_t^{\leftarrow}) \\ &\quad +(t-s)g(T-t)^2\nabla \ln q_t^{\leftarrow}(x_t^{\leftarrow}), \end{aligned}$$

where in the second step we have dropped the second order term $(t-s)^2(\nabla f_{T-s}(x_t^{\leftarrow}))f_{T-s}(x_t^{\leftarrow})$ and approximated $(t-s)g(t)^2\nabla \ln q_s^{\leftarrow}(x_t^{\leftarrow})$ to first order by $(t-s)g(T-t)^2\nabla \ln q_t^{\leftarrow}(x_t^{\leftarrow})$. Observe now that for $s = t - \ell h$, the update rule of (3.2) is the same as the update rule of (3.2).

Degradation operator. The remainder of the derivation proceeds along similar lines to the previous section. Given noise vector $\gamma \in \mathbb{R}^d$, define the *degradation operator* $D^{\gamma}_{s,t}(\cdot)$ by

$$D^{\gamma}_{s\to t}(x) \triangleq x + f_s(x)(t-s) + g(s)\sqrt{t-s}\cdot \gamma\,.$$

This operator simply runs an Euler-Maruyama discretization of the forward process, starting at time s, for time $t-s$, with the noise taken to be γ.

Starting from z and degrading it along the forward process from $T-t-\ell h$ to time $T-t$, we would end up with $D^{\gamma}_{T-t-\ell h\to T-t}(z) = z + \ell h\, f_{T-t-\ell h}(z) + g(T-t-\ell h)\sqrt{\ell h}\cdot\gamma$ for some Gaussian noise $\gamma \sim \mathcal{N}(0,\mathrm{Id})$. As before, we instead take γ to be the simulated noise needed to degrade z into $x_t^{\leftarrow}$, which in this case is given by the solution to

$$x_t^{\leftarrow} = z + \ell h\, f_{T-t-\ell h}(z) + g(T-t-\ell h)\sqrt{\ell h}\cdot \gamma\,.$$

To produce the next iterate $x_{t+h}^{\leftarrow}$ of the reverse process, we use γ to degrade z from time $T-t-\ell h$ to $T-t-h$. The result is given by

$$\begin{aligned} x_{t+h}^{\leftarrow} &= z + (\ell-1)h\, f_{T-t-\ell h}(z) \\ &\quad + g(T-t-\ell h)\sqrt{(\ell-1)h}\cdot\frac{x_t^{\leftarrow} - z - \ell h\, f_{T-t-\ell h}(z)}{g(T-t-\ell h)\sqrt{\ell h}} \\ &\approx z + (\ell-1)hf_{T-t}(x_t^{\leftarrow}) \\ &\qquad + \sqrt{1-1/\ell}\cdot \ell h g(T-t)^2\nabla \ln q_{T-t}(x_t^{\leftarrow}) \\ &= x_t^{\leftarrow} - hf_{T-t}(x_t^{\leftarrow}) + \ell h\cdot\Big(1-\sqrt{1-1/\ell}\Big) \\ &\qquad \cdot g(T-t)^2\nabla \ln q_{T-t}(x_t^{\leftarrow})\,. \end{aligned}$$

where in the second step we approximated $f_{T-t-\ell h}(z)$ by $f_{T-t}(x_t^{\leftarrow})$ and dropped $o(h)$ terms. Finally as $\ell \to \infty$, the right-hand side converges to $x_t^{\leftarrow} - h\,\{f_{T-t}(x_t^{\leftarrow}) - \frac{1}{2}g(T-t)^2\nabla \ln q_{T-t}(x_t^{\leftarrow})\}$, which recovers the Euler discretization of the probability flow ODE. We note that this is the only place that requires taking $\ell \to \infty$. Finally, as we take $h \to 0$, the above recovers the probability flow ODE (2).

3.3. Extensions to other samplers

The operational interpretation framework that we developed to extend DDIM to non-linear forward processes can be adapted in a relatively straightforward way to describe more general samplers. For example, in Equation (2), we defined a more general family of reverse processes, each of which has the correct marginal law at time t. These can easily be described by a similar operational interpretation.

Specifically, we use the same restoration operator as before to arrive to z, which we then use to estimate the noise γ. The critical change to the framework is that now, to corrupt from z to $x_{t+h}^{\leftarrow}$, instead of just using the estimated noise, we use a linear combination of the estimated noise, γ and *fresh noise* ν. Specifically, in the degradation operator, we use a vector γ', defined as follows:

$$\gamma' = \sqrt{1-\frac{\lambda^2}{\ell-1}}\,\gamma + \frac{1}{\sqrt{\ell-1}}\lambda\nu, \quad \nu \sim \mathcal{N}(0,\mathrm{Id})\,.$$

The parameter λ here controls how close the update rule is to the deterministic sampler. Trivially, for $\lambda = 0$, we have a fully deterministic sampler, as before. For $\lambda = 1$, the sampler becomes the reverse SDE sampler of Song et al. (2020). The coefficients have been chosen such that if γ were actually a draw from $\mathcal{N}(0,\mathrm{Id})$ instead of simulated noise, then γ' would likewise be a draw from $\mathcal{N}(0,\mathrm{Id})$.

Note that

$$\begin{aligned} \widetilde{x}^{\lambda}_{(k-1)h} &= z + (\ell-1)h\, f_{(k-\ell)h}(z) \\ &\quad + g((k-\ell)h)\sqrt{(\ell-1)h}\cdot\gamma' \\ &\approx z + (\ell-1)h\, f_{kh}(\widetilde{x}^{\lambda}_{kh}) + g(kh)\sqrt{(\ell-1)h}\cdot\gamma' \end{aligned}$$

$$= z + (\ell-1)h\, f_{kh}(\widetilde{x}^{\lambda}_{kh}) + g(kh)\sqrt{(\ell-1)h} \cdot \Big($$
$$\sqrt{1-\frac{\lambda^2}{\ell-1}} \cdot \frac{\widetilde{x}^{\lambda}_{kh} - z - \ell h\, f_{(k-\ell)h}(z)}{g((k-\ell)h)\sqrt{\ell h}} + \frac{1}{\sqrt{\ell-1}}\lambda\nu\Big)$$
$$\approx z + (\ell-1)h\, f_{kh}(\widetilde{x}^{\lambda}_{kh}) + g(kh)\sqrt{(\ell-1)h} \cdot \Big($$
$$\sqrt{1-\frac{\lambda^2}{\ell-1}} \cdot \frac{\widetilde{x}^{\lambda}_{kh} - z - \ell h\, f_{kh}(z)}{g(kh)\sqrt{\ell h}} + \frac{1}{\sqrt{\ell-1}}\lambda\nu\Big)$$
$$= \widetilde{x}^{\lambda}_{kh} - h\, f_{kh}(\widetilde{x}^{\lambda}_{kh}) + \ell h\, g(kh)^2 \nabla \ln q_{kh}(\widetilde{x}^{\lambda}_{kh})$$
$$-\sqrt{1-\frac{1}{\ell}} \cdot \sqrt{1-\frac{\lambda^2}{\ell-1}} \cdot \ell h\, g(kh)^2 \nabla \ln q_{kh}(\widetilde{x}^{\lambda}_{kh})$$
$$+\lambda\sqrt{h}\, g(kh)^2\, \nu$$
$$_{(\ell\to\infty)} = \widetilde{x}^{\lambda}_{kh} - h\, \{f_{kh}(\widetilde{x}^{\lambda})$$
$$-\frac{1+\lambda^2}{2} g(kh)^2 \nabla \ln q_{kh}(\widetilde{x}^{\lambda})\} + \lambda\sqrt{h}\, g(kh)^2 \nu\,,$$

where in the second step we approximated $f_{(k-\ell)h}(z)$ and $g((k-\ell)h)$ by $f_{kh}(\widetilde{x}^{\lambda}_{kh})$ and $g(kh)$, dropping $o(h)$ terms; in the fourth step we approximated $f_{(k-\ell)h}(z)$ and $g((k-\ell)h)$ by $f_{kh}(z)$ by $g(kh)$; and in the last step we used that

$$\lim_{\ell\to\infty} \ell\Big(1-\sqrt{1-\frac{1}{\ell}} \cdot \sqrt{1-\frac{\lambda^2}{\ell-1}}\Big) = \frac{1+\lambda^2}{2}\,.$$

4. Discretization Analysis

In what follows, we provide a non-asymptotic convergence analysis for the DDIM-type samplers, i.e. for the update rule of (3.2). To the best of our knowledge, this constitutes the first KL convergence analysis for deterministic sampling with diffusion models. Since our sampler corresponds to the Euler discretization of the probability flow ODE plus some excess terms whose contribution we ultimately show is negligible, the following analysis also implies non-asymptotic convergence for the Euler discretization.

4.1. DDIM-type sampler

Motivated by the discussion in Section 3.2, our analysis will focus on the process $(\widetilde{x}_{kh})_{k\in\{0,\ldots,T/h\}}$ defined backwards in time as follows. The iterate $\widetilde{x}_T$ is sampled from $q_0^{\leftarrow}$. Given iterate $\widetilde{x}_{kh}$, the preceding iterate $\widetilde{x}_{(k-1)h}$ is defined as follows:

$$\widetilde{x}_{(k-1)h} = D^{\gamma}_{(k-\ell)h\to(k-1)h}(z)$$

for $z \triangleq R_{kh\to(k-\ell)h}(\widetilde{x}_{kh})$ and $\gamma \triangleq D^{\gamma}_{(k-\ell)\to kh}(z) = \widetilde{x}_{kh}$, where R and D were defined in (3.2) and (3.2) respectively. As z is the result of restoring the current iterate $\widetilde{x}_{kh}$,

$$z = \widetilde{x}_{kh} - \ell h\, f_{kh}(\widetilde{x}_{kh}) + \ell h\, g(kh)^2 \nabla \ln q_{kh}(\widetilde{x}_{kh}).$$

The next iterate $\widetilde{x}_{(k-1)h}$ is given by degrading z for time $(\ell-1)h$, with the noise vector taken to be the *simulated noise* γ. More precisely γ is the noise vector that one could have used to degrade z for time ℓh to obtain $\widetilde{x}_{kh}$. As γ is the solution to $D^{\gamma}_{(k-\ell)h\to kh}(z) = \widetilde{x}_{kh}$, an equivalent formulation is via

$$\gamma = \frac{\widetilde{x}_{kh} - z - \ell h\, f_{(k-\ell)h}(z)}{g((k-\ell)h)\sqrt{\ell h}}.$$

Note that (4.1) is not well-defined when $k < \ell$; in this case, we take the update according to the Euler-Maruyama discretization:

$$\widetilde{x}_{(k-1)h} = \widetilde{x}_{kh} - h(f_{kh}(\widetilde{x}_{kh}) - \frac{1}{2} g(kh)^2 \nabla \ln q_{kh}(\widetilde{x}_{kh}))\,.$$

It will be convenient to denote $\widetilde{x}^{\leftarrow}_{kh} \triangleq \widetilde{x}_{T-kh}$ in the sequel.

4.2. Statement of results

We make the following mild assumptions on the forward process (x_t) and the data distribution:

Assumption 4.1. For all $t \geq 0$, the following holds for parameters $L_{f;\mathsf{t}}, L_g, L_{f;\mathsf{x}}, L_{\mathsf{sc},t}, R, g_{\max}, \beta, M \geq 1, c > 0$:

1. $f_t(x)$ is $L_{f;\mathsf{t}}$-Lipschitz in t and $L_{f;\mathsf{x}}$-Lipschitz in x.
2. $g^2(t)$ is L_g-Lipschitz in t.
3. $\|f_t(0)\| \leq R$.
4. $g(t) \leq g_{\max}$.
5. $\nabla \ln q^{\leftarrow}_t(x)$ is $L_{\mathsf{sc},t}$-Lipschitz in x and satisfies
$$\|\nabla \ln \frac{q^{\leftarrow}_t}{q^{\leftarrow}_s}(x)\| \leq \beta|t-s|^c(1+\|x\|+\|\nabla q^{\leftarrow}_t(x)\|)$$
for all $s \geq 0$. Denote $\sup_{t\geq 0} L_{\mathsf{sc},t}$ by $L_{\mathsf{sc},*}$.
6. $\nabla f_t(x)$ and $\nabla^2 \ln q^{\leftarrow}_t$ are L_{high}-Lipschitz in operator norm.

Remark 4.2. We note that the first four Parts of Assumption 4.1, as well as the first half of Part 6, are quite mild and are satisfied by any reasonable choice of forward process. For instance, for the Ornstein-Uhlenbeck process $\mathrm{d}x_t = -x_t\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$, we can take $L_{f;\mathsf{t}} = 0, L_{f;\mathsf{x}} = 1, L_g = 0, g_{\max} = \sqrt{2}$, and $\nabla f_t(x) = -\mathrm{Id}$ for all x is thus clearly Lipschitz in operator norm. Part 5 ensures that the score functions $\nabla \ln q^{\leftarrow}_t$ do not change much when perturbed in space or time. The former is a standard assumption in the literature on discretization bounds for score-based generative modeling (Block et al., 2022; Chen et al., 2022b; Lee et al., 2022a;b; Chen et al., 2022a), and the latter holds for reasonable choices of forward process. For instance, for the Ornstein-Uhlenbeck process, we can take $c = 1/2$ and $\beta = \Theta(L\sqrt{d})$ (see e.g. Lemma C.12 from (Lee et al., 2022a)).

The most important distinction between Assumption 4.1 and the assumptions made in previous analyses for score-based generative models is the second half of Part 6 where we assume *higher-order smoothness* of $q^{\leftarrow}_t$. As we will see in Section C, this is essential to our analysis because third-order derivatives of $\ln q^{\leftarrow}_t$ naturally arise when one computes the time derivative of the Fisher information as described in Section 4.3. As discussed in that section, the need to compute such time derivatives is unique to the ODE setting, justifying why such an assumption was not needed in prior analysis of stochastic samplers.

Under these conditions, we show that our discretization procedure approximates the true reverse process to prescribed error ϵ provided ℓ and $(\ell h)^{-1}$ are larger than some quantities which are polynomially bounded in $1/\epsilon$ and all parameters from Assumption 4.1:

Theorem 4.3. *Let $\epsilon > 0$. Let $\widetilde{p}$ denote the law of the process $(\widetilde{x}^{\leftarrow}_{kh})$ at time T with this choice of learning rate. Suppose Assumption 4.1 holds and define*

$$\Lambda \triangleq \exp\left(\int_0^T (L_{f;\mathsf{x}}^2 + g_{\max}^2 L_{\mathsf{sc},t})\,\mathrm{d}t\right) \qquad \textit{and}$$

$$\Lambda' \triangleq \exp\left(\int_0^T (L_{f;\mathsf{x}}^2 + g_{\max}^2 L_{\mathsf{sc},\lfloor t/h\rfloor h})\,\mathrm{d}t\right).$$

Then there exist quantities $\mathfrak{C}_1$ and $\mathfrak{C}_2$ which are polynomially bounded in $L_{f;\mathsf{t}}$, $L_{f;\mathsf{x}}$, L_g, R, $g_{\max}$, β, $L_{\mathsf{sc},*}$, L_{high}, Λ, Λ', d, $\mathbb{E}\|x^{\leftarrow}_0\|^2$, and $1/\epsilon$ such that $\mathsf{KL}(\widetilde{p}\|q) \le \epsilon$ provided $\ell \ge \mathfrak{C}_1$ and $\ell h \le \mathfrak{C}_2^{-1}$.

Remark 4.4. We briefly remark on the quantities Λ, Λ' appearing in the above theorem. We typically think of $L_{f;\mathsf{x}}$ and $g_{\max}$ as of constant order, so Λ and Λ' scale polynomially with $\exp(\int_0^T L_{\mathsf{sc},t}\,\mathrm{d}t)$ and $\exp(\int_0^T L_{\mathsf{sc},\lfloor t/h\rfloor h}\,\mathrm{d}t)$. While this scales exponentially in T, the exponential convergence of reasonable forward processes like Ornstein-Uhlenbeck means we should think of T as scaling logarithmically in d/ϵ. And while naively one might suspect that Λ, Λ' scale exponentially with $L_{\mathsf{sc},*}$, we show in Example 1 in Appendix C that these quantities actually scale polynomially in d and other parameters like $L_{\mathsf{sc},*}$, e.g. when the data distribution is Gaussian. Altogether, this suggests that our non-asymptotic guarantees are of polynomial complexity in all relevant parameters from Assumption 4.1.

In practice, the process $(\widetilde{x}^{\leftarrow}_{kh})$ would be initialized at the stationary measure q^* of the forward process (after some suitable re-scaling), rather than at $q^{\leftarrow}_0$. As observed in (Lee et al., 2022a; Chen et al., 2022b; Lee et al., 2022b), the KL divergence between the final iterate of the process under the alternative initialization $\widetilde{x}^{\leftarrow}_0 \sim q^*$ and the final iterate under the initialization $\widetilde{x}^{\leftarrow}_0 \sim q^{\leftarrow}_0$ is at most the KL divergence between the inital iterates of these two processes. But by stationarity of q^*, the latter KL is equivalent to the KL between the stationary measure of the forward process and the the law of the forward process at time T. This KL is typically exponentially small in T, e.g. when the forward process is an Ornstein-Uhlenbeck process. By passing from KL to total variation via Pinsker's inequality and applying triangle inequality, we conclude that the total variation between $\widetilde{x}^{\leftarrow}_T$ under this alternative initialization and the data distribution q is at most the sum of the error bound in Theorem 4.3 plus the distance between q_T and the stationary distribution. Formally:

Corollary 4.5. *Let $\epsilon > 0$. Let $f_t(x) = -x$ and $g(t) = \sqrt{2}$, so that the forward process in (2) corresponds to the standard Ornstein-Uhlenbeck process. Define the process $(\overline{x}_{kh})$ to be the process given by the same updates as in (4.1) but with $\overline{x}_T$ sampled from $\mathcal{N}(0, \mathrm{Id})$ instead of $q^{\leftarrow}_0$. Let p denote the law of $\overline{x}_0$. Suppose $\nabla^2 \ln q^{\leftarrow}_t$ is L_{high}-Lipschitz in operator norm. Then there exist quantities $\mathfrak{C}_1$ and $\mathfrak{C}_2$ which are polynomially bounded in d, $L_{\mathsf{sc},*}$, L_{high}, Λ, Λ', and $1/\epsilon$ such that*

$$\mathsf{TV}(p, q) \le \epsilon + \sqrt{\mathsf{KL}\left(q\|\mathcal{N}(0, \mathrm{Id})\right)}\,\exp(-T)$$

provided $\ell \ge \mathfrak{C}_1$ and $\ell h \le \mathfrak{C}_2^{-1}$.

4.3. Proof overview

Our discretization analysis is an interpolation-style argument, similar to the kind used in the log-concave sampling literature (Vempala & Wibisono, 2019; Chewi et al., 2021; Wibisono & Yang, 2022) as well as some recent analyses of score-based generative modeling (Lee et al., 2022a;b; Chen et al., 2022a). Here we describe the setup for this argument and highlight the key technical differences that manifest when analyzing ODEs rather than SDEs.

We begin with a generic setting where we are given two stochastic processes $(y_t)_{t\in[0,T]}$ and $(y'_t)_{t\in[0,T]}$ as follows. The process (y_t) is given by an arbitrary ODE

$$\mathrm{d}y_t = \mu_t(y_t)\,\mathrm{d}t\,.$$

We will ultimately take μ_t to be $-f_{T-t} + \frac{1}{2}g(T-t)^2\nabla \ln q_{T-t}$ so that (4.3) is the probability flow ODE associated to the forward process in (2). The process (y'_t) is given by first taking a discrete-time approximation to (y_t), e.g. via the update rules

$$y'_{(k+1)h} = y'_{kh} + h \cdot \mu'_{kh}(y_{kh})$$

for all integers $k = 0, 1, \ldots, T/h$. We will ultimately take μ'_{kh} to be $-f_{T-kh} + \frac{1}{2}g(T-kh)^2\nabla \ln q_{T-kh}$ plus error terms coming from the approximations in (3.2) and from taking $\ell \to \infty$ to ensure (3.2) approximates the Euler discretization of the probability flow ODE.

Then to get y'_t for all real values $t \in [0, T]$, we consider a linear interpolation of these iterates: if $k = \lfloor t/h \rfloor$, then we define $y_t = y_{kh} + (t - kh)\mu'_{kh}(y_{kh})$. We write this as

$$\mathrm{d}y'_t = \mu'_{kh}(y'_{kh})\,\mathrm{d}t\,.$$

Provided these processes are both initialized at the same distribution, that is, $y_0, y'_0 \sim \pi$ for some probability measure π over $\mathbb{R}^d$, then we would like to control the statistical distance between the marginal distributions on y_t and on y'_t as a function of t. Denoting these distributions by π_t and π'_t respectively, we prove the following generic bound which is the technical core of our work. First, we make the following assumptions about the two processes; when we specialize these processes to $(x^{\leftarrow}_t)$ and $(\widetilde{x}^{\leftarrow}_t)$, these assumptions will follow from Assumption 4.1:

Assumption 4.6. For all $0 \le t \le T$, there are parameters $L_t, L'_t, M \ge 1$ and $\zeta_t > 0$ such that:

1. $\nabla \ln \pi_t$ and μ_t are L_t-Lipschitz.
2. $\nabla \mu_t$ is M-Lipschitz in operator norm.
3. μ'_t is L'_t-Lipschitz.
4. $\mathbb{E}[\|\mu_t(y'_t) - \mu'_{kh}(y'_{kh})\|^2] \le \zeta_t^2$.
5. $h \le 1/2L'_t$ for all $0 \le t \le T$.

We briefly interpret these assumptions in the context of our eventual application to bounding the error of our discretization procedure. There, Conditions 1 and 2 apply to the true continuous process. The former is an immediate consequence of our (standard) assumption on the second-order smoothness of the marginals of the true process. The latter is an immediate consequence of our assumption on the third-order smoothness, which is stronger than what is needed for analyses of the reverse *SDE* but is likely necessary for our analysis of the reverse ODE.

Conditions 3 and 4 are properties that we will eventually establish for our discretization procedure (see Section D). Roughly, they stipulate that the drift term in the discretized probability flow ODE is Lipschitz and close on average to the drift of the true ODE.

Lastly, Condition 5 simply corresponds to a constraint on the step size of our discretization procedure.

For convenience, we will also define the quantities

$$L \triangleq \max_t L_t, \quad L' \triangleq \max_t L'_t, \quad \zeta^2 \triangleq \int_0^T \zeta_t^2\,\mathrm{d}t$$
$$\Lambda \triangleq \exp\Big(\int_0^T L_t\,\mathrm{d}t\Big), \quad \Lambda' \triangleq \exp\Big(\int_0^T L'_t\,\mathrm{d}t\Big)\,.$$

The main result of this section is a bound on the KL divergence between π'_T and π_T:

Theorem 4.7.

$$\mathsf{KL}\,(\pi'_T\|\pi_T) \lesssim \Lambda^{O(1)} L'^{1/2}\zeta^2$$
$$+(\Lambda^{O(1)} + \Lambda'^{O(1)})(L_0'^{1/2} d^{1/2} + MdT^{1/2})\,\zeta T^{1/2}\,.$$

The main ingredient in proving this is to bound the time derivative of $\mathsf{KL}\,(\pi'_t\|\pi_t)$ uniformly across $t \in [0,T]$, from which a bound on $\mathsf{KL}\,(\pi'_t\|\pi_t)$ follows by integrating.

One can explicitly compute this time derivative by appealing to the time derivatives of the densities of π'_t, π_t, given by the Fokker-Planck equations for the two processes:

$$\partial_t \pi_t = -\mathrm{div}(\pi_t \cdot \mu_t), \quad \partial_t \pi'_t = -\mathrm{div}(\pi'_t \cdot \widehat{\mu}_{t,kh})\,,$$

for $\widehat{\mu}_{t,kh}(x) \triangleq \mathbb{E}[\mu'_{kh}(y'_{kh}) \mid y'_t = x]$. Here $\widehat{\mu}_{t,kh}$ is the expectation over the drift at time kh conditioned on the position at the *future* time t. A calculation (see Lemma C.5) then reveals that

$$\partial_t \mathsf{KL}\,(\pi'_t\|\pi_t) = \int \pi'_t \langle \nabla \ln \frac{\pi'_t}{\pi_t}, \widehat{\mu}_{t,kh} - \mu_t\rangle\,.$$

It is here where our analysis departs from typical applications of the interpolation method. Indeed, if the ODEs driving y_t and y'_t were SDEs equipped with an additional Brownian motion term, then (4.3) would come with an additional negative term given by a multiple of the *Fisher information* between π'_t and π_t. In equations, this means that in lieu of (4.3), we would have

$$\partial_t \mathsf{KL}\,(\pi'_t\|\pi_t) = \int \pi'_t \langle \nabla \ln \frac{\pi'_t}{\pi_t}, \widehat{\mu}_{t,kh} - \mu_t\rangle$$
$$-C \int \pi'_t \|\nabla \ln \frac{\pi'_t}{\pi_t}\|^2,$$

for some $C > 0$ depending on the amount of Brownian motion. The advantage of the Fisher information term in (4.3) is that we can apply Young's inequality to conveniently upper bound the above by a multiple of

$$\int \pi'_t\,\|\widehat{\mu}_{t,kh} - \mu_t\|^2,$$

and avoid having to deal with $\nabla \ln \frac{\pi'_t}{\pi_t}$ altogether. Roughly speaking, the quantity (4.3) corresponds to the expected squared difference between the drift of the discrete process at time kh versus the drift of the continuous process at time t. This is small provided the former process doesn't move around too much between times kh and t, and provided the drifts μ'_{kh} and μ_t are sufficiently close on average. We verify in Section D that both of these conditions are satisfied by the probability flow ODE.

The situation is trickier in the ODE setting. To handle (4.3), we instead apply Cauchy-Schwarz to get

$$\partial_t \mathsf{KL}\,(\pi'_t \| \pi_t) \leq \left(\int \pi'_t \| \nabla \ln \frac{\pi'_t}{\pi_t} \|^2 \right)^{1/2} \times \left(\int \pi'_t \| \widehat{\mu}_{t,kh} - \mu_t \| \right)^{1/2} ,$$

after which the main technical obstacle is to ensure the first term on the right-hand side, again corresponding to the Fisher information between π'_t and π_t, does not explode with t. In Lemmas C.6 and C.8, we show how to bound the time derivative of this quantity polynomially in various problem-specific parameters like dimension and smoothness of μ_t. Altogether, this leads to the following bounds. We defer the technical details to the supplement and provide a brief proof sketch of how to control the time derivatives of these quantities:

Lemma 4.8 (See Lemmas C.6 and C.8)**.** *For all* $0 \leq t \leq T$,

$$\mathbb{E}_{\pi'_t}[\|\nabla \ln \pi'_t\|^2] \lesssim \Lambda'^{O(1)}(L'_0 d + M^2 d^2 t)$$
$$\mathbb{E}_{\pi'_t}[\|\nabla \ln \pi_t\|^2] \lesssim \Lambda^{O(1)}(L'_0 d + M^2 d^2 t + L' \zeta^2)$$

Proof sketch. When computing $\partial_t \int \pi'_t \|\nabla \ln \pi'_t\|^2$, one term that shows up is $\partial_t \ln \pi'_t$. Using the Fokker-Planck equation for π'_t, we can derive an expression for $\partial_t \ln \pi'_t$ (see Proposition C.7). This and a calculation with integration by parts reveals that

$$\begin{aligned} \partial_t \int \pi'_t \|\nabla \ln \pi'_t\|^2 &= -2 \int \pi'_t \big(\langle \nabla \mathrm{div}\, \widehat{\mu}_{t,kh}, \nabla \ln \pi'_t \rangle \\ &\quad + (\nabla \ln \pi'_t)^\top (\nabla \widehat{\mu}_{t,kh})(\nabla \ln \pi'_t) \big) \\ &\lesssim \sup_x \|\nabla \widehat{\mu}_{t,kh}(x)\|_{\mathsf{op}} \int \pi'_t \|\nabla \ln \pi'_t\|^2 + \\ &\quad + \int \pi'_t \|\nabla \mathrm{div}\, \widehat{\mu}_{t,kh}\|^2, \end{aligned}$$

where in the last step we used Young's inequality. Lipschitzness of μ_t allows us to bound $\sup \|\nabla \widehat{\mu}_{t,kh}\|_{\mathsf{op}}$, and higher-order smoothness of μ_t allows us to bound $\|\nabla \mathrm{div} \widehat{\mu}_{t,kh}\|$. □

This is the only part of the analysis where third-order derivatives appear and where Part 6 of Assumption 4.1, which corresponds to Part 2 of Assumption 4.6, comes into play. One subtlety in the argument above is deducing smoothness of $\widehat{\mu}_{t,kh}$, a complicated-looking conditional expectation, to smoothness of the true drift μ_t. To connect the two, we exploit the fact that for step size h sufficiently small, the discrete-time process is *invertible* (Lemma C.3) so that $\widehat{\mu}_{t,kh}$ can be expressed as μ'_{kh} composed with a deterministic function.

Altogether, the bound on the Fisher information which is implied by Lemma 4.8 allows us, as in the SDE case, to reduce controlling $\partial_t \mathsf{KL}\,(\pi'_t \| \pi_t)$ to controlling the difference in drifts as captured by Eq. (4.3), which we then carry out in Appendix D.

5. Related Work

There has been great recent progress on diffusion models including recently outperforming other deep generative models such as Generative Adversarial Networks (GANs) (Dhariwal & Nichol, 2021; Song et al., 2020; Daras et al., 2022a; Kim et al., 2022). Applications range from protein generation (Anand & Achim, 2022; Trippe et al., 2022; Schneuing et al., 2022; Corso et al., 2022), medical imaging (Jalal et al., 2021; Arvinte et al., 2022), 3-D data (Poole et al., 2022) and many more, e.g. see Yang et al. (2022) for a comprehensive survey.

Non-asymptotic analysis of stochastic samplers, (Block et al., 2022; De Bortoli et al., 2021; De Bortoli, 2022; Liu et al., 2022; Lee et al., 2022a; Pidstrigach, 2022; Lee et al., 2022b; Chen et al., 2022a;b) has drawn upon tools from the rich literature on log-concave sampling (see (Chewi, 2022) for a recent survey) to yield convergence guarantees for diffusion models. These works focus on the setting where the forward process is an Ornstein-Uhlenbeck process, and the reverse process is given by a *stochastic* differential equation. Notably, the very recent works of Chen et al. (2022b); Lee et al. (2022b); Chen et al. (2022a) show under mild assumptions on the data distribution q (e.g. smooth and bounded second moment) that a suitable discretization of the reverse SDE run for polynomially many steps generates samples that are close in statistical distance to the data distribution.

Prior to our work, no previous non-asymptotic bounds in KL divergence were known for the probability flow ODE associated to any forward process. Prior analyses are insufficient because they either rely on Girsanov's theorem (Chen et al., 2022b) or a chain rule-based variant (Chen et al., 2022a) of the interpolation argument of Vempala & Wibisono (2019).

Informally, Girsanov's theorem allows one to bound not just the distance between the distributions over the final iterates of the algorithm but even the distance between the distributions over the *trajectories* of the two processes – note that the latter distance upper bounds the former by the data processing inequality. Stochasticity in every step of the reverse process ensures that even the latter distance is small. Without stochasticity however, there is no reason this distance should even be finite.

The chain rule-based argument of (Chen et al., 2022a) establishes a similar bound to Girsanov's; in particular, when the algorithm and the idealized process are initialized to the same distribution, the bounds these two arguments give are identical.

Lastly, we remark that (Lee et al., 2022a) used an interpolation argument without chain rule, but their analysis, similar to existing analyses of Langevin Monte Carlo in the log-sampling literature (Vempala & Wibisono, 2019; Chewi et al., 2021), exploits the appearance of a certain Fisher information term in the expression for the time derivative of the KL divergence between the algorithm and the idealized process (see Section 4.3 for further details). For ODEs however, this Fisher information term does not appear.

6. Conclusion

In this work we gave an operational interpretation for the probability flow ODE as iterating a two-step process of restoration via gradient ascent and degradation towards the current iterate. This perspective also extends to reverse processes with a Brownian motion component. Our operational interpretation closely aligns with the samplers introduced in (Bansal et al., 2022; Daras et al., 2022b) and generalizes the framework of denoising diffusion implicit models (Song et al., 2021) to general, non-linear forward processes.

The main technical contribution of our work was a non-asymptotic analysis of the deterministic sampler arising from our framework. While previous works (Chen et al., 2022b; Lee et al., 2022b; Chen et al., 2022a) gave non-asymptotic analyses for diffusion models when the underlying reverse process is an SDE, to our knowledge our analysis is the first of its kind in the ODE setting. Our proof is based on an interpolation argument, but the key difference with prior applications of this method is that the deterministic nature of the sampler necessitates controlling the time derivative of the Fisher information between the algorithm and the true reverse process.

Limitations and future directions. The most obvious area for improvement would be to sharpen the quantitative dependence on various parameters like dimension and Lipschitz-ness of the score functions. Intuitively, the absence of Brownian motion in the probability flow ODE should lead to better dimension dependence compared to using an SDE, but in this work we are only able to establish an iteration complexity for the deterministic sampler which is some polynomial in d. Additionally, for convenience in this work we ignore issues of score estimation error. While it should be possible to use change-of-measure-type arguments like in (Wibisono & Yang, 2022) to obtain guarantees when the score estimation error has sub-Gaussian tails, new ideas are needed to handle merely an L_2 bound on the score estimation error like in (Chen et al., 2022b; Lee et al., 2022b; Chen et al., 2022a).

We also leave as an open question whether our assumption of higher-order smoothness is really necessary to obtain non-asymptotic guarantees for the probability flow ODE.

Apart from these technical improvements, we mention some empirical directions to explore. First, our discretization procedure introduces a number of new hyperparameters that one can try tuning to get improved performance in practice. Even for linear diffusions, it would be interesting to explore the effect of tuning ℓ, which under DDIM is currently taken to be $(T-t)/h$. In addition, it seems interesting to explore how parameters of the restoration procedure like the learning rate and number of steps of gradient ascent, or the use of momentum or higher-order optimization methods can lead to better samplers. We expect that different restoration procedures can recover other discretization frameworks, e.g. second-order ones like Heun's method. Empirically, we expect that optimizing the learning rate and number of steps can lead to deterministic samplers with smaller computational overhead and higher sample quality.

Acknowledgments. SC would like to thank Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim for many enlightening discussions about deterministic score-based generative modeling and the interpolation method. The authors would also like to thank Sinho Chewi, Holden Lee, and Adil Salim for helpful feedback on an earlier version of this work.

This research has been supported by NSF Grants CCF 1763702, AF 1901292, CNS 2148141, Tripods CCF 1934932, IFML CCF 2019844, the Texas Advanced Computing Center (TACC) and research gifts by Western Digital, WNCG IAP, UT Austin Machine Learning Lab (MLL), Cisco and the Archie Straiton Endowed Faculty Fellowship. SC has been supported by NSF Award 2103300. GD has been supported by the Onassis Fellowship, the Bodossaki Fellowship and the Leventis Fellowship.

References

Anand, N. and Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Arvinte, M., Jalal, A., Daras, G., Price, E., Dimakis, A., and Tamir, J. I. Single-shot adaptation using score-based models for mri reconstruction. In *International Society for Magnetic Resonance in Medicine, Annual Meeting*, 2022.

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., and Goldstein, T. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv e-prints*, art. arXiv:2002.00107, 2022.

Cattiaux, P., Conforti, G., Gentil, I., and Léonard, C. Time reversal of diffusion processes under a finite entropy condition. September 2022.

Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022a.

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 preprint arXiv:2209.11215*, 2022b.

Chewi, S. *Log-concave sampling*. 2022. Book draft available at https://chewisinho.github.io/.

Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R., and Zhang, M. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *arXiv e-prints*, art. arXiv:2112.12662, 2021.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Daras, G., Dagan, Y., Dimakis, A. G., and Daskalakis, C. Score-guided intermediate layer optimization: Fast langevin mixing for inverse problem. *arXiv preprint arXiv:2206.09104*, 2022a.

Daras, G., Delbracio, M., Talebi, H., Dimakis, A. G., and Milanfar, P. Soft diffusion: Score matching for general corruptions. *arXiv preprint arXiv:2209.05442*, 2022b.

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

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion Schrödinger bridge with applications to score-based generative modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709. Curran Associates, Inc., 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Föllmer, H. An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lect. Notes Control Inf. Sci.*, pp. 156–163. Springer, Berlin, 1985.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A. G., and Tamir, J. Robust compressed sensing mri with deep generative priors. *Advances in Neural Information Processing Systems*, 34:14938–14954, 2021.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

Kim, D., Shin, S., Song, K., Kang, W., and Moon, I.-C. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, pp. 11201–11228. PMLR, 2022.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *arXiv e-prints*, art. arXiv:2206.06227, 2022a.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. *arXiv preprint arXiv:2209.12381*, 2022b.

Liu, X., Wu, L., Ye, M., and Liu, Q. Let us build bridges: understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Pidstrigach, J. Score-based generative models detect manifolds. *arXiv e-prints*, art. arXiv:2206.01018, 2022.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Schneuing, A., Du, Y., Harris, C., Jamasb, A., Igashov, I., Du, W., Blundell, T., Lió, P., Gomes, C., Welling, M., et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., and Canton, C. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Trippe, B. L., Yim, J., Tischer, D., Broderick, T., Baker, D., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, pp. 8094–8106. Curran Associates, Inc., 2019.

Wibisono, A. and Yang, K. Y. Convergence in kl divergence of the inexact langevin algorithm with application to score-based generative models. *arXiv preprint arXiv:2211.01512*, 2022.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Cui, B., and Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

Zhang, Q., Tao, M., and Chen, Y. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022.

Supplement Roadmap. In Appendix A we motivate the choice of certain learning rate parameter that arises in Section 3. In Appendix B we provide preliminary calculations for the proof of Theorem 4.3. In Appendix C, we give a generic bound on the distance between two processes driven by ODEs with similar drifts, one of which is an interpolation of a discrete-time process. Finally, in Appendix D we apply this generic bound to our setting, bound the difference in drifts between the probability flow ODE and our sampler, and prove Theorem 4.3.

A. Tuning the Learning Rate

In this section we justify the choice of learning rate (3.2) in our gradient ascent interpretation of the restoration operator by considering the special case where the data distribution q is isotropic Gaussian and the forward process is an Ornstein-Uhlenbeck process.

First, recall the definition of the loss function $\ell_{x_t^{\leftarrow}}$ from (3.2). In general, one step of gradient ascent with learning rate η starting from $x_t^{\leftarrow}$ gives the iterate

$$x_t^{\leftarrow} + \eta \nabla \ell_{x_t^{\leftarrow}}(x_t^{\leftarrow}) = x_t^{\leftarrow} + \eta\Big(-\frac{1}{g(T-t)^2}\big(\mathrm{Id} + (t-s)\nabla f_{T-s}(x_t^{\leftarrow})\big) f_{T-s}(x_t^{\leftarrow}) + \nabla \ln q_s^{\leftarrow}(x_t^{\leftarrow})\Big)\,.$$

Now suppose $q \sim \mathcal{N}(0, \sigma^2\,\mathrm{Id})$ and furthermore

$$f_t(x) = -\alpha x \qquad \text{and} \qquad g(t) = \beta\sqrt{2}\,.$$

Then $q_t^{\leftarrow}$ is given by $\mathcal{N}(0, (e^{-2\alpha t}\sigma^2 + \frac{\beta^2}{\alpha}(1-e^{-2\alpha t}))\,\mathrm{Id})$, and the conditional log-likelihood $\ln q_s^{\leftarrow}(x \mid x_t^{\leftarrow})$ is quadratic in x and is thus maximized at x for which $\nabla \ell_{x_t^{\leftarrow}}$ vanishes.

In this case, (3.2) simplifies to

$$\nabla \ell_{x_t^{\leftarrow}} \approx \frac{1}{2\beta^2(t-s)}(1+\alpha(s-t))(x_t^{\leftarrow} - x(1+\alpha(s-t))) - \frac{1}{\mathbb{V}[q_s^{\leftarrow}]}\,x\,,$$

where $\mathbb{V}[q_s^{\leftarrow}]$ denotes the variance of $q_s^{\leftarrow}$. Setting the right-hand side to zero and solving for x shows that

$$x = \frac{1+\alpha(s-t)}{\mathbb{V}[q_{T-s}] + (1+\alpha(s-t))^2}\,x_t^{\leftarrow} = \Big(1 + \big(\alpha - \frac{2\beta^2}{\mathbb{V}[q_{T-s}]}\big)\cdot(t-s) + O(|t-s|^2)\Big)\,x_t^{\leftarrow}$$

is an (approximate) stationary point of $\nabla \ell_{x_t^{\leftarrow}}$.

The next iterate (A) after one gradient step simplifies to

$$x_t^{\leftarrow} + \eta\nabla\ell_{x_t^{\leftarrow}}(x_t^{\leftarrow}) \approx x_t^{\leftarrow} - \frac{\eta}{2\beta^2(t-s)}(1+\alpha(s-t))\cdot\alpha(s-t)\cdot x_t^{\leftarrow} - \frac{\eta}{e^{-2\alpha s}\sigma^2 + \frac{\beta^2}{\alpha}(1-e^{-2s})}\,x_t^{\leftarrow}\,.$$

Finally, we observe that by taking

$$\eta \triangleq 2\beta^2(t-s)\,,$$

the above simplifies to

$$x_t^{\leftarrow} - (1+\alpha(s-t))\cdot\alpha(s-t)\cdot x_t^{\leftarrow} - \frac{2\beta^2(t-s)}{e^{-2\alpha s}\sigma^2 + \frac{\beta^2}{\alpha}(1-e^{-2s})}\,x_t^{\leftarrow} = \Big(1 + \big(\alpha - \frac{2\beta^2}{\mathbb{V}[q_{T-s}]}\big)\cdot(t-s) + O(|t-s|^2)\Big)\,x_t^{\leftarrow}\,,$$

which agrees up to second-order terms with (A). Therefore, when the data is Gaussian and the forward process is an Ornstein-Uhlenbeck process, as $t-s \to 0$ the right choice of η to ensure to first order approximation that a single gradient step takes us from $x_t^{\leftarrow}$ to the maximizer of the conditional log-likelihood $\ln q_s^{\leftarrow}(x \mid x_t^{\leftarrow})$ is given by (A), which corresponds to (3.2) in the main text as claimed.

B. Proof Preliminaries

Let $h > 0$ and $\ell \in \mathbb{N}$ be discretization parameters. Define

$$\delta_\ell \triangleq 1 - \sqrt{1-1/\ell} = \frac{1}{2\ell} + O(1/\ell^2)$$

and

$$\xi_\ell = \delta_\ell - \frac{1}{2\ell} = O(1/\ell^2)$$

Recall the definition of the process $(\widetilde{x}_{kh})_{k\in\{0,\dots,T/h\}}$ in Eq. (4.1). Here we rewrite the update rule in (4.1) to make clear its similarity to the Euler-Maruyama discretization:

$$\begin{aligned}
\widetilde{x}_{(k-1)h} &= z + (\ell-1)h\, f_{(k-\ell)h}(z) + g((k-\ell)h)\sqrt{(\ell-1)h}\cdot\gamma \\
&= z + (\ell-1)h\, f_{(k-\ell)h}(z) + g((k-\ell)h)\sqrt{(\ell-1)h}\cdot\frac{\widetilde{x}_{kh} - z - \ell h\, f_{(k-\ell)h}(z)}{g((k-\ell)h)\sqrt{\ell h}} \\
&= (1-\delta_\ell)\widetilde{x}_{kh} + \delta_\ell z + (\ell\delta_\ell - 1)h f_{(k-\ell)h}(z) \\
&= \widetilde{x}_{kh} + (\ell\xi_\ell - 1/2)\, h f_{(k-\ell)h}(z) \\
&\qquad - \delta_\ell\big(\ell h\, f_{kh}(\widetilde{x}_{kh}) - \ell h\, g(kh)^2 \nabla \ln q_{kh}(\widetilde{x}_{kh})\big).
\end{aligned}$$

Note that $\delta_\ell\eta_k = hg(kh)^2\,(1/2 + \ell\xi_\ell)$, so we can further rewrite this as

$$\begin{aligned}
&= \widetilde{x}_{kh} + (\ell\xi_\ell - 1/2)\, h f_{(k-\ell)h}(z) \\
&\qquad - h(1/2 + \ell\xi_\ell)\big(f_{kh}(\widetilde{x}_{kh}) - g(kh)^2\nabla\ln q_{kh}(\widetilde{x}_{kh})\big) \\
&= \widetilde{x}_{kh} - h\left\{f_{kh}(\widetilde{x}_{kh}) - \frac{1}{2}g(kh)^2\nabla\ln q_{kh}(\widetilde{x}_{kh})\right\} + v^{(1)}_{kh}(\widetilde{x}_{kh}) + \cdots + v^{(3)}_{kh}(\widetilde{x}_{kh}),
\end{aligned}$$

where the excess terms are given by

$$\begin{aligned}
v^{(1)}_{kh}(\widetilde{x}_{kh}) &\triangleq \ell\xi_\ell h f_{(k-\ell)h}(z)\cdot\mathbb{1}[k\geq\ell] \\
v^{(2)}_{kh}(\widetilde{x}_{kh}) &\triangleq \frac{h}{2}(f_{(k-\ell)h}(z) - f_{kh}(\widetilde{x}_{kh}))\cdot\mathbb{1}[k\geq\ell] \\
v^{(3)}_{kh}(\widetilde{x}_{kh}) &\triangleq h\ell\xi_\ell\big(-f_{kh}(\widetilde{x}_{kh}) + g(kh)^2\nabla\ln q_{kh}(\widetilde{x}_{kh})\big)\mathbb{1}[k\geq\ell]\cdot\,.
\end{aligned}$$

Note that as $\ell\to\infty$ and $h\ell\to 0$, the excess terms tend to zero and the process $(\widetilde{x}_{kh})$ converges to the one given by the Euler-Maruyama discretization.

In the subsequent sections, we make this quantitative via an interpolation argument. Let $(\widetilde{x}_t)_{0\leq t\leq T}$ denote the linear interpolation of the discrete process $(\widetilde{x}_{kh})_{h=0,\dots,T/h}$, and let $(\widetilde{x}^{\leftarrow}_t)$ denote the time-reversed process $\widetilde{x}^{\leftarrow}_t \triangleq \widetilde{x}_{T-t}$. Concretely, for any $kh \leq t < (k+1)h$,

$$\begin{aligned}
\mathrm{d}\widetilde{x}^{\leftarrow}_t = -\Big\{& f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh}) - \frac{1}{2}g(T-kh)^2\nabla\ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh}) \\
&- \frac{1}{h}\,(v^{(1)}_{T-kh}(\widetilde{x}^{\leftarrow}_{kh}) + \cdots + v^{(3)}_{T-kh}(\widetilde{x}^{\leftarrow}_{kh}))\Big\}\,\mathrm{d}t.
\end{aligned}$$

We note that even in the absence of the excess terms above, in which case the above process would just be the Euler-Maruyama discretization of the probability flow ODE, no existing works gave a non-asymptotic analysis showing that this discretization converges polynomially to the continuous-time probability flow ODE. Our analysis in the sequel allows us to both control the excess terms and establish such a non-asymptotic analysis.

C. Interpolation Argument

In this section we give general bounds for how the KL divergence between two distributions, one driven by a discretized ODE and the other by a continuous-time one, changes over time. Throughout this section, we work with two stochastic processes $(y_t)_{t\in[0,T]}$ and $(y'_t)_{t\in[0,T]}$ over $\mathbb{R}^d$ given by the ODEs

$$\begin{aligned}
\mathrm{d}y_t &= \mu_t(y_t)\,\mathrm{d}t \\
\mathrm{d}y'_t &= \mu'_{kh}(y'_{kh})\,\mathrm{d}t, \quad k = \lfloor t/h\rfloor,
\end{aligned}$$

where $y_0, y'_0 \sim \pi$ for some probability measure π over $\mathbb{R}^d$. The process (y'_t) is equivalent to a linear interpolation of a discrete-time process where one goes from the k-th iterate y'_{kh} to the $(k+1)$-st iterate $y'_{(k+1)h}$ via the update

$$y'_{(k+1)h} = y'_{kh} + h\,\mu'_{kh}(y'_{kh})\,.$$

We let π_t, π'_t denote the law of y_t, y'_t respectively. When we eventually apply the estimates obtained in this section, we will take (y'_t) to be given by our discretization of the probability flow ODE, and we will take (y_t) to be the true probability flow ODE in continuous time.

The bounds in this section hold under the conditions of Assumption 4.6, restated here for convenience:

Assumption C.1. For all $0 \le t \le T$, there are parameters $L_t, L'_t, M \ge 1$ and $\zeta_t > 0$ such that:

1. $\nabla \ln \pi_t$ and μ_t are L_t-Lipschitz.
2. $\nabla \mu_t$ is M-Lipschitz in operator norm.
3. μ'_t is L'_t-Lipschitz.
4. $\mathbb{E}[\|\mu_t(y'_t) - \mu'_{kh}(y'_{kh})\|^2] \le \zeta_t^2$.
5. $h \le 1/2L'_t$ for all $0 \le t \le T$.

For convenience, we also recall the quantities defined in (4.3):

$$L \triangleq \max_t L_t, \quad L' \triangleq \max_t L'_t, \quad \Lambda \triangleq \exp\Big(\int_0^T L_t\,\mathrm{d}t\Big), \quad \Lambda' \triangleq \exp\Big(\int_0^T L'_t\,\mathrm{d}t\Big), \quad \zeta^2 \triangleq \int_0^T \zeta_t^2\,\mathrm{d}t$$

and restate the main claimed bound on the KL divergence between π'_T and π_T:

Theorem 4.7.

$$\begin{aligned}\mathsf{KL}\,(\pi'_T \| \pi_T) &\lesssim \Lambda^{O(1)} L'^{1/2} \zeta^2 \\ &+ (\Lambda^{O(1)} + \Lambda'^{O(1)})(L_0'^{1/2} d^{1/2} + MdT^{1/2})\,\zeta T^{1/2}\,.\end{aligned}$$

Example 1. Here we work out a simple example showing that when (y_t) corresponds to the probability flow ODE that reverses the Ornstein-Uhlenbeck process starting from a Gaussian distribution, Λ' scales polynomially, rather than exponentially, in d and L'.

Define $\pi^{\rightarrow}_t$ for $0 \le t \le T$ as the marginal distribution of running the Ornstein-Uhlenbeck process for time t starting from $\mathcal{N}(0, \frac{1}{L}\mathrm{Id})$ for some large L, and consider the associated reverse ODE

$$\mathrm{d}y_t = (y_t + \nabla \ln \pi_t(y_t))\,\mathrm{d}t,$$

where $\pi_t \triangleq \pi^{\rightarrow}_{T-t}$ denotes the marginal laws of $(y_t)_{t\in[0,T]}$. Concretely, π_t is given by $\mathcal{N}(0, \frac{1}{L_t}\mathrm{Id})$ for $L_t = (e^{-2(T-t)}/L + 1 - e^{-2(T-t)})^{-1}$. Note that

$$\Lambda' = \exp\Big(\int_0^T L_t\,\mathrm{d}t\Big) = \exp\Big(\frac{1}{2}\ln(1 + (e^{2T} - 1)L)\Big).$$

Because $\mathsf{KL}\,\big(\mathcal{N}(0, \frac{1}{L}\mathrm{Id}) \| \mathcal{N}(0, \mathrm{Id})\big) = \frac{d}{2}(\ln L - 1 + \frac{1}{\ell}) \lesssim d \ln L$, we must run the forward process for time $T \approx \frac{1}{2}\ln(d \ln L)$ for $\pi^{\rightarrow}_T$ to be close to $\mathcal{N}(0, \mathrm{Id})$. In this case, $\Lambda' \lesssim \sqrt{dL \ln L}$.

We begin by working out the Fokker-Planck equations for (π'_t) and (π_t).

Proposition C.2. *The laws (π'_t) and (π_t) satisfy*

$$\begin{aligned}\partial_t \pi_t &= -\mathrm{div}(\pi_t \cdot \mu_t) \\ \partial_t \pi'_t &= -\mathrm{div}(\pi'_t \cdot \widehat{\mu}_{t,kh}),\end{aligned}$$

where

$$\widehat{\mu}_{t,kh}(x) \triangleq \mathbb{E}[\mu'_{kh}(y'_{kh}) \mid y'_t = x].$$

When k is clear from context, we will denote $\widehat{\mu}_{t,kh}$ by $\widehat{\mu}_t$ to ease notation.

Proof. The Fokker-Planck equation for (π_t) is given by

$$\partial_t \pi_t = -\operatorname{div}(\pi_t \cdot \mu_t).$$

For the interpolated process (π'_t), the Fokker-Planck for $(\pi'_t)_{kh \le t < (k+1)h}$ conditioned on time kh, which we will denote by $(\pi'_{t|kh})_{kh \le t < (k+1)h}$, is given by

$$\partial_t \pi'_{t|kh}(x) = -\operatorname{div}_x(\pi'_{t|kh}(x) \cdot \mu'_{kh}(y'_{kh})).$$

If Π'_{kh} denotes the probability measure over $\sigma(y'_t \mid 0 \le t \le kh)$, then if we integrate both sides of (C) with respect to Π'_{kh}, we get

$$\begin{aligned}
\partial_t \pi'_t(x) &= -\int \operatorname{div}_x(\pi'_t(x \mid \xi) \cdot \mu'_{kh}(y'_{kh}))\, \Pi'_{kh}(\mathrm{d}\xi) \\
&= -\operatorname{div}_x \int \pi'_t(x \mid \xi) \cdot \mu'_{kh}(y'_{kh})\, \Pi'_{kh}(\mathrm{d}\xi) \\
&= -\operatorname{div}_x\Big(\pi'_t(x) \int \mu'_{kh}(y'_{kh})\, \Pi'_{kh|t}(\mathrm{d}\xi \mid y'_t = x)\Big) \\
&= -\operatorname{div}_x(\pi'_t(x) \cdot \mathbb{E}[\mu'_{kh}(y'_{kh}) \mid y'_t = x]) \\
&= -\operatorname{div}_x(\pi'_t(x) \cdot \widehat{\mu}_{t,kh}(x)).
\end{aligned}$$

□

It turns out that because we are assuming the step size h is sufficiently small in Condition 5 of Assumption 4.6, the conditional expectation $\widehat{\mu}_{t,kh}$ has a simple form. For any k, the ODE $\mathrm{d}y'_t = \mu'_{kh}(y'_{kh})\,\mathrm{d}t$ defines a map $F_{kh\to t} : \mathbb{R}^d \to \mathbb{R}^d$ for any $kh \le t \le (k+1)h$ via

$$F_{kh\to t}(z) = z + (t - kh)\mu'_{kh}(z)$$

so that starting at z at time kh and running the ODE to time t, we end up at $F_{kh\to t}(z)$. When h is sufficiently small, $F_{kh\to t}$ is invertible:

Lemma C.3. *Let $h \le 1/2L'$. Then for any $z, z' \in \mathbb{R}^d$,*

$$\frac{1}{2}\|z - z'\| \le \|F_{kh\to t}(z) - F_{kh\to t}(z')\| \le \frac{3}{2}\|z - z'\|.$$

In particular, $F_{kh\to t}$ has a unique, 2-Lipschitz inverse $F^{-1}_{kh\to t} : \mathbb{R}^d \to \mathbb{R}^d$, so

$$\widehat{\mu}_{t,kh}(x) = \mu'_{kh}(F^{-1}_{kh\to t}(x)).$$

Furthermore, $\widehat{\mu}_{t,kh}$ is $O(L'_t)$-Lipschitz.

Henceforth, when k, h, t are clear from context, we will refer to the inverse $F^{-1}_{kh\to t}$ simply as F^{-1}.

Proof. For the first bound, note that

$$\|F_{kh\to t}(z) - F_{kh\to t}(z')\| \ge \|z - z'\| - (t - kh)\|\mu'_{kh}(z) - \mu'_{kh}(z')\| \ge (1 - h \cdot L'_{kh})\,\|z - z'\|,$$

so the lower bound in (C.3) follows by the fact that $h \le 1/2L'$. The upper bound follows analogously.

For the second part of the lemma, recall that bi-Lipschitz functions on $\mathbb{R}^d$ are bijective, so $F_{kh\to t}$ has a unique inverse $F^{-1}_{kh\to t}$. To see why the latter function is 2-Lipschitz, for any z_0, z'_0 we can take $z = F^{-1}_{kh\to t}(z_0)$ and $z' = F^{-1}_{kh\to t}(z'_0)$ in the lower bound of (C.3) to conclude that $\frac{1}{2}\|F^{-1}_{kh\to t}(z_0) - F^{-1}_{kh\to t}(z'_0)\| \le \|z_0 - z'_0\|$ as desired. Eq. (C.3) then follows from the fact that the distribution of y'_{kh} conditioned on $y'_t = x$ is the point mass at $F^{-1}_{kh\to t}(x)$.

The only part that remains to be verified is Lipschitzness of $\widehat{\mu}_{t,kh}$. This follows from the fact that $\widehat{\mu}_{t,kh}$ is the composition of a L'_t-Lipschitz function with a 2-Lipschitz function. □

We will also use the following simple consequence of the third-order smoothness of μ_t (Condition 2 of Assumption 4.6):

Lemma C.4. *For all $x, x' \in \mathbb{R}^d$, then*

$$\sup \|\nabla \mathrm{div}\, \mu_t\| \le Md \qquad \textit{and} \qquad \sup \|\nabla \mathrm{div}\, \widehat{\mu}_{t,kh}\| \le 2Md$$

Proof. The first bound is immediate from

$$|\mathrm{div}\, \mu_t(x) - \mathrm{div}\, \mu_t(x')| = |\operatorname{Tr} \nabla\, \mu_t(x) - \operatorname{Tr} \nabla\, \mu_t(x')| \le \|\nabla \mu_t(x) - \nabla \mu_t(x')\|_{\mathsf{tr}} \le Md\|x - x'\|.$$

For the second bound, note that

$$|\mathrm{div}\, \widehat{\mu}_t(x) - \mathrm{div}\, \widehat{\mu}_t(x')| \le \|\nabla \mu'_{kh}(F^{-1}(x)) - \nabla \mu'_{kh}(F^{-1}(x'))\|_{\mathsf{op}} \le dM\|F^{-1}(x) - F^{-1}(x')\| \le 2Md$$

as claimed. □

We are now ready to compute the time derivative of the KL divergence between π'_t and π_t.

Lemma C.5.

$$\partial_t \mathsf{KL}\,(\pi'_t \| \pi_t) \le \zeta_t \left(\int \pi'_t \|\nabla \ln \pi'_t - \nabla \ln \pi_t\|^2\right)^{1/2}$$

Proof. We can compute

$$\begin{aligned}
\partial_t \mathsf{KL}\,(\pi'_t \| \pi_t) &= \int (\partial_t \pi'_t) \ln \frac{\pi'_t}{\pi_t} + \int \pi'_t\, \partial_t \ln \frac{\pi'_t}{\pi_t} = \int (\partial_t \pi'_t) \ln \frac{\pi'_t}{\pi_t} + \int \pi'_t \, \frac{\partial_t (\pi'_t/\pi_t)}{\pi'_t/\pi_t} \\
&= \int (\partial_t \pi'_t) \ln \frac{\pi'_t}{\pi_t} + \int \pi_t \cdot \frac{\pi_t \partial_t \pi'_t - \pi'_t \partial_t \pi_t}{{\pi_t}^2} \\
&= \int (\partial_t \pi'_t) \ln \frac{\pi'_t}{\pi_t} - \int \frac{\pi'_t}{\pi_t}\, \partial_t \pi_t \\
&= -\int \mathrm{div}(\pi'_t \cdot \widehat{\mu}_{t,kh}) \, \ln \frac{\pi'_t}{\pi_t} + \int \frac{\pi'_t}{\pi_t}\, \mathrm{div}(\pi_t \cdot \mu_t) \\
&= \int \pi'_t \, \langle \widehat{\mu}_{t,kh}, \nabla \ln \frac{\pi'_t}{\pi_t} \rangle - \int \pi_t \, \langle \nabla \frac{\pi'_t}{\pi_t}, \mu_t \rangle \\
&= \int \pi'_t \, \langle \nabla \ln \frac{\pi'_t}{\pi_t}, \widehat{\mu}_{t,kh} - \mu_t \rangle.
\end{aligned}$$

The lemma then follows by Cauchy-Schwarz, as

$$\int \pi'_t \|\widehat{\mu}_{t,kh} - \mu_t\|^2 = \mathbb{E}_{\pi'_t}[\|\mu'_{kh}(F^{-1}(y'_t)) - \mu_t(y'_t)\|^2] = \mathbb{E}_{\pi'_t}[\|\mu'_{kh}(y'_{kh}) - \mu_t(y'_t)\|^2] \le \zeta_t^2.$$ □

We need to control the Fisher information $\int \pi'_t \|\nabla \ln \pi'_t - \nabla \ln \pi_t\|^2$ in Lemma C.5. To do this, we will bound the time derivatives of $\int \pi'_t \|\nabla \ln \pi'_t\|^2$ and $\int \pi'_t \|\nabla \ln \pi_t\|^2$ in Lemmas C.6 and C.8 below and apply triangle inequality.

Lemma C.6.

$$\partial_t \int \pi'_t \|\nabla \ln \pi'_t\|^2 \lesssim L'_t \int \pi'_t \|\nabla \ln \pi'_t\|^2 + M^2 d^2$$

In particular, by Grönwall's inequality, for any $0 \le t \le T$ *we have*

$$\int \pi'_t \|\nabla \ln \pi'_t\|^2 \lesssim \Lambda'^{O(1)}(L'_0 d + M^2 d^2 t)$$

Proof. We have

$$\begin{aligned}
\partial_t \int \pi'_t \|\nabla \ln \pi'_t\|^2 &= -\int \mathrm{div}(\pi'_t \cdot \widehat{\mu}_t) \|\nabla \ln \pi'_t\|^2 + \int \pi'_t \partial_t \|\nabla \ln \pi'_t\|^2 \\
&= 2 \int \pi'_t \big(\langle \widehat{\mu}_t, (\nabla^2 \ln \pi'_t) \nabla \ln \pi'_t \rangle + \langle \partial_t \nabla \ln \pi'_t, \nabla \ln \pi'_t \rangle \big) \\
&= 2 \int \pi'_t \big(\langle \widehat{\mu}_t, (\nabla^2 \ln \pi'_t) \nabla \ln \pi'_t \rangle + \langle \nabla(-\mathrm{div}\, \widehat{\mu}_t - \langle \nabla \ln \pi'_t, \widehat{\mu}_t \rangle), \nabla \ln \pi'_t \rangle \big),
\end{aligned}$$

where in the last step we used the first part of Proposition C.7 below. Note that we can write the latter term in the parentheses in (C) as

$$\langle -\nabla \mathrm{div}\, \widehat{\mu}_t - (\nabla^2 \ln \pi'_t)\widehat{\mu}_t - (\nabla\widehat{\mu}_t)\nabla \ln \pi'_t, \nabla \ln \pi'_t\rangle.$$

Of these three terms, the second one exactly cancels with the first term in (C). Putting everything together, we get

$$\begin{aligned}\partial_t \int \pi'_t \|\nabla \ln \pi'_t\|^2 &= -2\int \pi'_t\big(\langle \nabla \mathrm{div}\, \widehat{\mu}_t, \nabla \ln \pi'_t\rangle + (\nabla \ln \pi'_t)^\top (\nabla\widehat{\mu}_t)(\nabla \ln \pi'_t)\big)\\ &\lesssim \sup \|\nabla \widehat{\mu}_t\|_{\mathsf{op}} \int \pi'_t \|\nabla \ln \pi'_t\|^2 + \int \pi'_t \|\nabla \mathrm{div}\, \widehat{\mu}_t\|^2,\end{aligned}$$

where in the last step we used Young's inequality. The first part of the lemma follows by Lemmas C.3 and C.4. For the second part, Grönwall's inequality tells us that

$$\mathbb{E}_{\pi'_t}[\|\nabla \ln \pi'_t\|^2] \le \Lambda'^{O(1)}(\int \pi \|\nabla \ln \pi\|^2 + M^2 d^2 t).$$

We conclude by noting that

$$\int \pi \|\nabla \ln \pi\|^2 = -\int \pi \Delta \ln \pi \le L'_0 d$$

by integration by parts and Condition 1 of Assumption 4.6. □

We remark that Lemma C.6 is tight as $h \to 0$ when the marginals $\{\pi'_{T-t}\}_{t\in[0,T]}$ are given by running the Ornstein-Uhlenbeck process starting with a spherical Gaussian distribution.

In the above proof, we needed the following calculation:

Proposition C.7.

$$\begin{aligned}\partial_t \ln \pi'_t &= -\mathrm{div}\, \widehat{\mu}_{t,kh} - \langle \nabla \ln \pi'_t, \widehat{\mu}_{t,kh}\rangle\\ \partial_t \ln \pi_t &= -\mathrm{div}\, \mu_t - \langle \nabla \ln \pi_t, \mu_t\rangle.\end{aligned}$$

Next, we carry out a calculation analogous to Lemma C.6 to bound the time derivative of $\mathbb{E}_{\pi'_t}[\|\nabla \ln \pi_t\|^2]$:

Lemma C.8.

$$\partial_t \int \pi'_t \|\nabla \ln \pi_t\|^2 \lesssim L_t \int \pi'_t \|\nabla \ln \pi_t\|^2 + M^2 d^2 + L_t \zeta_t^2.$$

In particular, by Grönwall's inequality, for any $0 \le t \le T$ *we have*

$$\mathbb{E}_{\pi'_t}[\|\nabla \ln \pi_t\|^2] \lesssim \Lambda^{O(1)}(L'_0 d + M^2 d^2 t + L' \zeta^2)$$

Proof. We have

$$\begin{aligned}\partial_t \int \pi'_t \|\nabla \ln \pi_t\|^2 &= -\int \mathrm{div}(\pi'_t \cdot \widehat{\mu}_t)\|\nabla \ln \pi_t\|^2 + \int \pi'_t \partial_t \|\nabla \ln \pi_t\|^2\\ &= 2\int \pi'_t\big(\langle \widehat{\mu}_t, (\nabla^2 \ln \pi_t)\nabla \ln \pi_t\rangle + \langle \partial_t \nabla \ln \pi_t, \nabla \ln \pi_t\rangle\big)\\ &= 2\int \pi'_t\big(\langle \widehat{\mu}_t, (\nabla^2 \ln \pi_t)\nabla \ln \pi_t\rangle + \langle \nabla(-\mathrm{div}\, \mu_t - \langle \nabla \ln \pi_t, \mu_t\rangle), \nabla \ln \pi_t\rangle\big),\end{aligned}$$

where in the last step we used the second part of Proposition C.7. Note that we can write the latter term in the parentheses in (C) as

$$\langle -\nabla \mathrm{div}\, \mu_t - (\nabla^2 \ln \pi_t)\mu_t - (\nabla\mu_t)\nabla \ln \pi_t, \nabla \ln \pi_t\rangle.$$

Of these three terms, the second one nearly cancels with the first term in (C). Putting everything together, we get the inequality

$$\begin{aligned}\partial_t \int \pi'_t \|\nabla \ln \pi_t\|^2 &= -2 \int \pi'_t \big(\langle \nabla \mathrm{div}\, \mu_t, \nabla \ln \pi_t\rangle + (\nabla \ln \pi_t)^\top (\nabla \mu_t)(\nabla \ln \pi_t) \\ &\qquad + (\mu_t - \widehat{\mu}_t)^\top (\nabla^2 \ln \pi_t) \nabla \ln \pi_t\big) \\ &\lesssim \sup \|\nabla \mu_t\|_{\mathsf{op}} \int \pi'_t \|\nabla \ln \pi_t\|^2 + \int \pi'_t \|\nabla \mathrm{div}\, \mu_t\|^2 \\ &\qquad + 2 \sup \|\nabla^2 \ln \pi_t\|_{\mathsf{op}} \Big(\int \pi'_t \|\nabla \ln \pi_t\|^2\Big)^{1/2} \Big(\int \pi'_t \|\mu_t - \widehat{\mu}_t\|^2\Big)^{1/2} \\ &\lesssim L_t \int \pi'_t \|\nabla \ln \pi_t\|^2 + \int \pi'_t \|\nabla \mathrm{div}\, \mu_t\|^2 + L_t \int \pi'_t \|\mu_t - \widehat{\mu}_t\|^2\end{aligned}$$

where in the penultimate and final steps we used Young's inequality, and in the final step we used Condition 1 of Assumption 4.6. The first part of the lemma follows by Lemmas C.3 and Condition 4 of Assumption 4.6. The second part of the lemma follows by Grönwall's inequality and (C). □

We can now combine Lemmas C.5, C.6, and C.8 to prove Theorem 4.7:

Proof of Theorem 4.7. By triangle inequality and Eqs. (C.6) and (C.8),

$$\Big(\int \pi'_t \|\nabla \ln \pi'_t - \nabla \ln \pi_t\|^2\Big)^{1/2} \lesssim (\Lambda^{O(1)} + \Lambda'^{O(1)})(L_0'^{1/2} d^{1/2} + M d t^{1/2}) + \Lambda^{O(1)} L'^{1/2} \zeta_t,$$

so integrating the bound in Lemma C.5 over $t \in [0, T]$, we get

$$\mathsf{KL}\,(\pi'_T \| \pi_T) \lesssim (\Lambda^{O(1)} + \Lambda'^{O(1)})(L_0'^{1/2} d^{1/2} + M d T^{1/2}) \int_0^T \zeta_t \,\mathrm{d}t + \Lambda^{O(1)} L'^{1/2} \zeta^2 \,.$$

We conclude by bounding $\int_0^T \zeta_t \,\mathrm{d}t \le \zeta T^{1/2}$ by Cauchy-Schwarz. □

Finally, we record a norm bound which will be useful in the sequel:

Lemma C.9. *For any $0 \le t \le T$ and any $c > 0$,*

$$\partial_t \, \mathbb{E}\|y'_t\|^2 \le \mathbb{E}\|\mu'_{kh}\|^2 + \mathbb{E}\|y'_t\|^2 \,.$$

Proof. Recall that $y'_t = y'_{kh} + (t - kh)\, \mu'_{kh}(y'_{kh})$, so

$$\mathbb{E}\|y'_t\|^2 = \mathbb{E}\|y'_{kh}\|^2 + (t-kh)^2\, \mathbb{E}\|\mu'_{kh}(y'_{kh})\|^2 + 2(t-kh)\, \mathbb{E}\langle y'_{kh}, \mu'_{kh}(y'_{kh})\rangle \,.$$

Differentiating with respect to t, we get

$$\partial_t \mathbb{E}\|y'_t\|^2 = 2(t-kh)\, \mathbb{E}\|\mu'_{kh}(y'_{kh})\|^2 + 2\, \mathbb{E}\langle y'_{kh}, \mu'_{kh}(y'_{kh})\rangle = 2\, \mathbb{E}\langle y'_t, \mu'_{kh}(y'_{kh})\rangle \,,$$

so the lemma follows by Young's inequality. □

D. Bounding the Difference in Drifts

We wish to apply Theorem 4.7 with (y_t) and (y'_t) given by $(x^{\leftarrow}_t)$ and $(\widetilde{x}^{\leftarrow}_t)$ defined in Eqs. (2) and (B). For these processes, the drifts (μ_{kh}) and (μ'_t) in Eqs. (C) and (C) are given by

$$\begin{aligned}\mu_t(x) &\triangleq -f_{T-t}(x) + \frac{1}{2} g(T-t)^2 \nabla \ln q^{\leftarrow}_t(x) \\ \mu'_{kh}(x) &\triangleq -f_{T-kh}(x) + \frac{1}{2} g(T-kh)^2 \nabla \ln q^{\leftarrow}_{kh}(x) - \frac{1}{h}(v^{(1)}_{T-kh}(x) + \cdots + v^{(3)}_{T-kh}(x)) \,,\end{aligned}$$

and both processes are initialized at the distribution $\pi = q_T$. In general, the marginal laws (π_t) of the former process are given by $(q^{\leftarrow}_t)$. We will denote the marginal laws (π'_t) of the latter process by (p_t).

D.1. Smoothness of drift

We now verify the first three parts of Assumption 4.6.

Lemma D.1. *Part 1 of Assumption 4.6 holds with*

$$L_t \triangleq \Theta(L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},t}) \,.$$

Proof. By Part 5 of Assumption 4.1, $\nabla \ln q_t^{\leftarrow}$ is $L_{\mathsf{sc},t}$-Lipschitz. As μ_t is the sum of an $L_{f;\mathsf{x}}$-Lipschitz function and a $\frac{1}{2}g_{\max}^2 L_{\mathsf{sc},t}$-Lipschitz function, the claim follows. □

Lemma D.2. *Part 2 of Assumption 4.6 holds with*

$$M \triangleq (1 + g_{\max}^2/2)L_{\mathsf{high}} = \Theta(g_{\max}^2 L_{\mathsf{high}}) \,.$$

Proof. By Part 6 of Assumption 4.1, $\nabla\mu_t$ is the sum of a L_{high}-Lipschitz function and a $g_{\max}^2 L_{\mathsf{high}}/2$-Lipschitz function. □

Lemma D.3. *The restoration operator $R_{kh\to(k-\ell)h}$ is $O(1)$-Lipschitz for all integers $\ell \le k \le T/h$.*

Proof. For any x, x', we have

$$\begin{aligned}
\|R_{kh\to(k-\ell)h}(x) - R_{kh\to(k-\ell)h}(x')\| &\le \|x - x'\| + \ell h \, \|f_{kh}(x) - f_{kh}(x')\| \\
&\quad + \ell h \, g(kh)^2 \, \|\nabla \ln q_{kh}(x) - \nabla \ln q_{kh}(x')\| \\
&\le (1 + \ell h L_{f;\mathsf{x}} + \ell h g_{\max}^2 L_{\mathsf{sc},kh}) \, \|x - x'\| \lesssim \|x - x'\|.
\end{aligned}$$
□

Lemma D.4. $\frac{1}{h}(v_1 + \cdots + v_3)$ *is* $O(L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},kh})$*-Lipschitz.*

Proof. By Lemma D.3, $f_{(k-\ell)h}(z) = f_{(k-\ell)h}(R_{kh\to(k-\ell h)}(\widetilde{x}_{kh}^{\leftarrow})$ is a composition of an $L_{f;\mathsf{x}}$-Lipschitz function with an $O(1)$-Lipschitz function in $\widetilde{x}_{kh}^{\leftarrow}$, so $\frac{1}{h}v_1$ is $O(\ell\xi_\ell L_{f;\mathsf{x}})$-Lipschitz. Similarly, $\frac{1}{h}v_2$ is the difference between an $O(L_{f;\mathsf{x}})$-Lipschitz function and an $L_{f;\mathsf{x}}/2$-Lipschitz function in $\widetilde{x}_{kh}^{\leftarrow}$, so it is $O(L_{f;\mathsf{x}})$-Lipschitz. Finally, $\frac{1}{h}v_3$ is the sum of an $\ell\xi_\ell L_{f;\mathsf{x}} \ll L_{f;\mathsf{x}}$-Lipschitz function and a $g_{\max}^2 L_{\mathsf{sc},kh}$-Lipschitz function, so it is $(L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},kh}$-Lipschitz. □

Lemma D.5. μ'_{kh} *as defined in* (D) *is* $O(L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},kh})$*-Lipschitz. In particular, Part 3 of Assumption 4.6 holds with*

$$L'_t \triangleq \Theta(L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},kh})$$

for all $kh \le t < (k+1)h$.

Proof. Note that $f_{T-kh}(\cdot) - \frac{1}{2}g(T-kh)^2\nabla \ln q_{kh}^{\leftarrow}(\cdot)$ is $O(L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},kh})$-Lipschitz, so the claim follows by Lemma D.4. □

D.2. Distance between drifts

The bulk of our discretization analysis is devoted to verifying Part 4 of Assumption 4.6. For convenience, we will denote $v_{T-kh}^{(1)}(z), \ldots, v_{T-kh}^{(3)}(z)$ by $v_1, \ldots, v_3$. Henceforth, assume that

$$h \ll \min((RL_{f;\mathsf{x}}\ell)^{-1}, (g_{\max}^2 L_{\mathsf{sc},*})^{-1})$$

For any $kh \le t \le (k+1)h$, we have

$$\begin{aligned}
\mathbb{E}\|\mu_t(\widetilde{x}_t^{\leftarrow}) - \mu'_{kh}(\widetilde{x}_{kh}^{\leftarrow})\|^2 &\lesssim \mathbb{E}\|f_{T-t}(\widetilde{x}_t^{\leftarrow}) - f_{T-kh}(\widetilde{x}_{kh}^{\leftarrow})\|^2 \\
&\quad + \mathbb{E}\|g(T-t)^2\nabla \ln q_t^{\leftarrow}(\widetilde{x}_t^{\leftarrow}) - g(T-kh)^2\nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow})\|^2 \\
&\quad + \frac{1}{h^2}(\mathbb{E}\|v_1\|^2 + \cdots \mathbb{E}\|v_3\|^2) \,.
\end{aligned}$$

We first bound the excess terms $v_1, \ldots, v_3$. We focus on the case $T - kh \ge \ell h$, as otherwise $v_1 = v_2 = v_3 = 0$ by definition.

Lemma D.6.
$$\frac{1}{h^2}\mathbb{E}[\|v_1\|^2 + \cdots + \|v_3\|^2] \lesssim \epsilon_1 \max_{k' \in \{0,1,\ldots,T/h\}} \mathbb{E}\|\nabla \ln q^{\leftarrow}_{k'h}(\widetilde{x}^{\leftarrow}_{k'h})\|^2 + \epsilon_2$$

for

$$\begin{aligned}\epsilon_1 &\triangleq \exp(O(L_{f;\mathrm{x}}^2 T))(\ell^{-2} + \ell^2 h^2 L_{f;\mathrm{x}}^2)\, g_{\max}^4 \\ \epsilon_2 &\triangleq \exp(O(L_{f;\mathrm{x}}^2 T))(\ell^{-2} + \ell^2 h^2 L_{f;\mathrm{x}}^2)(\mathbb{E}\|\widetilde{x}^{\leftarrow}_0\|^2 + R^2 + \ell^2 h^2 L_{f;\mathrm{t}}^2)\,.\end{aligned}$$

Proof. Recall that
$$v_1 = \ell \xi_\ell h f_{T-(k-\ell)h}(z)\,,$$
so we have
$$\begin{aligned}\mathbb{E}\|v_1\|^2 &= \ell^2 \xi_\ell^2 h^2\, \mathbb{E}\|f_{T-(k-\ell)h}(z)\|^2 \\ &\lesssim \ell^{-2} h^2 (\mathbb{E}\|f_{T-(k-\ell)h}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + L_{f;\mathrm{x}}^2\, \mathbb{E}\|z - \widetilde{x}^{\leftarrow}_{kh}\|^2) \\ &\lesssim \ell^{-2} h^2 (L_{f;\mathrm{x}}^2\, \mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2 + R^2 + L_{f;\mathrm{x}}^2\, \mathbb{E}\|z - \widetilde{x}^{\leftarrow}_{kh}\|^2)\,.\end{aligned}$$

Recall that
$$v_2 = \frac{h}{2}(f_{T-(k-\ell)h}(z) - f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh}))\,,$$
so we have
$$\begin{aligned}\mathbb{E}\|v_2\|^2 &= \frac{h^2}{4}\, \mathbb{E}\|f_{T-(k-\ell)h}(z) - f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 \\ &\lesssim h^2(\ell^2 h^2 L_{f;\mathrm{t}}^2 + L_{f;\mathrm{x}}^2\, \|z - \widetilde{x}^{\leftarrow}_{kh}\|^2)\,.\end{aligned}$$

Recall that
$$v_3 = h\ell\xi_\ell\big(-f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh}) + g(T-kh)^2 \nabla \ln q^{\leftarrow}_{(k+\ell)h}(\widetilde{x}^{\leftarrow}_{kh})\big)\,,$$
so we have
$$\begin{aligned}\mathbb{E}\|v_3\|^2 &= h^2 \ell^2 \xi_\ell^2\, \mathbb{E}\|-f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh}) + g(T-kh)^2 \nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 \\ &\lesssim \ell^{-2} h^2 (R^2 + L_{f;\mathrm{x}}^2\, \mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2 + g_{\max}^4\, \mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2)\end{aligned}$$

Combining Eqs. (D.2), (D.2), and (D.2) we get
$$\begin{aligned}\frac{1}{h^2}\mathbb{E}[\|v_1\|^2 + \cdots + \|v_3\|^2] \lesssim (\ell^2 h^2 L_{f;\mathrm{t}}^2 + \ell^{-2} R^2) &+ \ell^{-2} g_{\max}^4\, \mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 \\ &+ \ell^{-2} L_{f;\mathrm{x}}^2\, \mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2 + L_{f;\mathrm{x}}^2\, \mathbb{E}\|z - \widetilde{x}^{\leftarrow}_{kh}\|^2\,.\end{aligned}$$

Recall from (4.1) that
$$z = \widetilde{x}_{kh} - \ell h\,(f_{kh}(\widetilde{x}_{kh}) - g(T-kh)^2 \nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}_{kh}))\,,$$
so
$$\begin{aligned}\|z - \widetilde{x}^{\leftarrow}_{kh}\|^2 &\lesssim \ell^2 h^2 (1 + \ell^2 h^2 L_{f;\mathrm{x}}^2)\, \|f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + \ell^2 h^2 g_{\max}^4\, \|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 \\ &\lesssim \ell^2 h^2\, (L_{f;\mathrm{x}}^2\, \|\widetilde{x}^{\leftarrow}_{kh}\|^2 + R^2) + \ell^2 h^2 g_{\max}^4\, \|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2,\end{aligned}$$

where in the second step we used (D.2). Substituting this into (D.2) and using Lemma C.9 below to bound $\mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2$, we obtain the desired bound. □

Lemma D.7. *For any integer* $0 \le k \le T/h$ *and any* $kh \le t < (k+1)h$*,*
$$\mathbb{E}\|\mu_t(\widetilde{x}^{\leftarrow}_t) - \mu'_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 \lesssim \epsilon'_1 \max_{k' \in \{0,1,\ldots,T/h\}} \mathbb{E}\|\nabla \ln q^{\leftarrow}_{k'h}(\widetilde{x}^{\leftarrow}_{k'h})\|^2 + \epsilon'_2$$

for

$$\begin{aligned}\epsilon_1' &\triangleq \epsilon_1 + h^2 L_g^2 + g_{\max}^4 (h^2 L_{f;\mathsf{x}}^2 + h^2 g_{\max}^4 L_{\mathsf{sc},*}^2 + g_{\max}^4 \beta^2 h^{2c}) \cdot \exp(O(L_{f;\mathsf{x}}^2 T))\\ \epsilon_2' &\triangleq \epsilon_2 + g_{\max}^4 \beta^2 h^{2c} + (\mathbb{E}\|\widetilde{x}_0^{\leftarrow}\|^2 + R^2 + \ell^2 h^2 L_{f;\mathsf{t}}^2)\\ &\qquad \times (h^2 L_{f;\mathsf{x}}^2 + h^2 g_{\max}^4 L_{\mathsf{sc},*} + g_{\max}^4 \beta^2 h^{2c}) \cdot \exp(O(L_{f;\mathsf{x}}^2 T))\,.\end{aligned}$$

In particular, for any $\delta > 0$, if

$$\begin{aligned}&\ell \gtrsim \delta^{-1/2} (g_{\max}^2 + R + \ell h L_{f;\mathsf{t}} + \mathbb{E}\|x_0^{\leftarrow}\|^2) \cdot \exp(O(L_{f;\mathsf{x}}^2 T))\\ &h \lesssim \min\bigl\{\operatorname{poly}(L_g, L_{f;\mathsf{t}}, R, g_{\max}, L_{\mathsf{sc},*}, L_{f;\mathsf{x}}, \mathbb{E}\|x_0^{\leftarrow}\|^2)^{-1}\, \ell^{-1} \delta^{1/2}, (\delta/(g_{\max}^4 \beta^2))^{1/2c}\bigr\} \cdot \exp(O(L_{f;\mathsf{x}}^2 T))\,,\end{aligned}$$

then $\epsilon_1', \epsilon_2' \leq \delta$.

Proof. We can bound the first term on the right-hand side of (D.2) using Lipschitzness of f in time and space:

$$\mathbb{E}\|f_{T-kh}(\widetilde{x}_{kh}^{\leftarrow}) - f_{T-t}(\widetilde{x}_t^{\leftarrow})\|^2 \lesssim L_{f;\mathsf{x}}^2\, \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow} - \widetilde{x}_t^{\leftarrow}\|^2 + h^2 L_{f;\mathsf{t}}^2\,.$$

For the second term on the right-hand side of (D.2), we can use Lipschitzness of g^2 and the score:

$$\begin{aligned}&\mathbb{E}\|g(T-t)^2\, \nabla \ln q_t^{\leftarrow}(\widetilde{x}_t^{\leftarrow}) - g(T-kh)^2\, \nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow})\|^2\\ &\quad\lesssim h^2 L_g^2\, \mathbb{E}\|\nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow})\|^2 + g(T-t)^4\, \mathbb{E}\Bigl\|\nabla \ln \frac{q_{kh}^{\leftarrow}}{q_t^{\leftarrow}}(\widetilde{x}_{kh}^{\leftarrow})\Bigr\|^2 + g(T-t)^4\, L_{\mathsf{sc},t}^2 \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow} - \widetilde{x}_t^{\leftarrow}\|^2\\ &\quad\lesssim (h^2 L_g^2 + g(T-t)^4 \beta^2 h^{2c})\, \mathbb{E}\|\nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow})\|^2\\ &\qquad\quad + g(T-t)^4 \beta^2 h^{2c}\, \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow}\|^2 + g(T-t)^4 \beta^2 h^{2c} + g(T-t)^4 L_{\mathsf{sc},t}^2\, \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow} - \widetilde{x}_t^{\leftarrow}\|^2\\ &\quad\lesssim (h^2 L_g^2 + g_{\max}^4 \beta^2 h^{2c})\, \mathbb{E}\|\nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow})\|^2\\ &\qquad\quad + g_{\max}^4 \beta^2 h^{2c}\, \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow}\|^2 + g_{\max}^4 \beta^2 h^{2c} + g_{\max}^4 L_{\mathsf{sc},t}^2\, \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow} - \widetilde{x}_t^{\leftarrow}\|^2\,.\end{aligned}$$

Substituting the above bounds into (D.2), we get that

$$\begin{aligned}&\mathbb{E}\|\mu_t(\widetilde{x}_t^{\leftarrow}) - \mu_{kh}'(\widetilde{x}_{kh}^{\leftarrow})\|^2\\ &\quad\lesssim (L_{f;\mathsf{x}}^2 + g_{\max}^4 L_{\mathsf{sc},t}^2)\, \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow} - \widetilde{x}_t^{\leftarrow}\|^2 + (h^2 L_g^2 + g_{\max}^4 \beta^2 h^{2c})\, \mathbb{E}\|\nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow})\|^2\\ &\qquad\quad + g_{\max}^4 \beta^2 h^{2c}\, \mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow}\|^2 + g_{\max}^4 \beta^2 h^{2c} + \frac{1}{h^2}\, \mathbb{E}[\|v_1\|^2 + \cdots + \|v_3\|^2].\end{aligned}$$

By applying the bounds for $\mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow} - \widetilde{x}_t^{\leftarrow}\|^2$ and $\mathbb{E}\|\widetilde{x}_{kh}^{\leftarrow}\|^2$ in Lemma D.8 and D.9 and noting that $L_{f;\mathsf{x}}^2 + g_{\max}^4 L_{\mathsf{sc},t}^2 \ll 1/h^2$ by (D.2), we see that the lemma follows from Lemma D.6 and the definition of ϵ_1', ϵ_2' in Eqs. (D.7), (D.7). Note that in the assumed bounds on ℓ, h in the lemma statement, we substituted $\mathbb{E}\|x_0^{\leftarrow}\|^2$ for $\mathbb{E}\|\widetilde{x}_0^{\leftarrow}\|^2$; this is because these two quantities are identical. □

D.3. Movement and norm bounds

Lemma D.8. *For any integer $0 < k \leq T/h$ and any $kh \leq t < (k+1)h$,*

$$\begin{aligned}\mathbb{E}\|\widetilde{x}_t^{\leftarrow} - \widetilde{x}_{kh}^{\leftarrow}\|^2 \lesssim h^2 \cdot \exp(O(L_{f;\mathsf{x}}^2 T))\Bigl(&\mathbb{E}\|\widetilde{x}_0\|^2 + R^2 + \ell^2 h^2 L_{f;\mathsf{t}}^2\\ &+ g_{\max}^4 \max_{k \in \{0,1,\ldots,T/h\}} \mathbb{E}\|\nabla \ln q_t^{\leftarrow}(\widetilde{x}_t^{\leftarrow})\|^2\Bigr) + \mathbb{E}[\|v_1\|^2 + \cdots + \|v_3\|^2]\,.\end{aligned}$$

Proof. By definition of the interpolated process,

$$\widetilde{x}_t^{\leftarrow} = \widetilde{x}_{kh}^{\leftarrow} - (t - kh)\, \{f_{T-kh}(\widetilde{x}_{kh}^{\leftarrow}) - \frac{1}{2} g(T-kh)^2 \nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow}) + \frac{1}{h}(v_1 + \cdots + v_3)\},$$

so

$$\mathbb{E}\|\widetilde{x}_t^{\leftarrow} - \widetilde{x}_{kh}^{\leftarrow}\|^2 \lesssim h^2\, \mathbb{E}\|f_{T-kh}(\widetilde{x}_{kh}^{\leftarrow})\|^2 + h^2 g_{\max}^4\, \mathbb{E}\|\nabla \ln q_{kh}^{\leftarrow}(\widetilde{x}_{kh}^{\leftarrow})\|^2 + \mathbb{E}[\|v_1\|^2 + \cdots + \|v_3\|^2]\,.$$

The proof is complete upon using Part 1 of Assumption 4.1 and Lemma D.9 to get

$$\mathbb{E}\|f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 \lesssim \exp(O(L^2_{f;\mathsf{x}}T))\Big(\mathbb{E}\|\widetilde{x}_0\|^2 + R^2 + \ell^2h^2L^2_{f;\mathsf{t}} + g^4_{\max} \max_{k\in\{0,1,\dots,T/h\}} \mathbb{E}\|\nabla \ln q^{\leftarrow}_t(\widetilde{x}^{\leftarrow}_t)\|^2\Big)\,,$$

where we have used that $\exp(O(L^2_{f;\mathsf{x}}T)) \cdot L^2_{f;\mathsf{x}} = \exp(O(L^2_{f;\mathsf{x}}T))$. □

Lemma D.9. *For all* $0 \le t \le T$*,*

$$\mathbb{E}\|\widetilde{x}^{\leftarrow}_t\|^2 \lesssim \exp(O(L^2_{f;\mathsf{x}}T))\left(\mathbb{E}\|\widetilde{x}^{\leftarrow}_0\|^2 + R^2 + \ell^2h^2L^2_{f;\mathsf{t}} + g^4_{\max} \max_{k\in\{0,1,\dots,T/h\}} \mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2\right).$$

Proof. By Lemma C.9,

$$\begin{aligned}\partial_t\,\mathbb{E}\|\widetilde{x}^{\leftarrow}_t\|^2 &\lesssim \mathbb{E}\|\widetilde{x}^{\leftarrow}_t\|^2 + \mathbb{E}\|f_{T-kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + g^4_{\max}\,\mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + \frac{1}{h^2}\,\mathbb{E}[\|v_1\|^2 + \cdots + \|v_3\|^2]\\ &\lesssim \mathbb{E}\|\widetilde{x}^{\leftarrow}_t\|^2 + L^2_{f;\mathsf{x}}\,\mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2 + g^4_{\max}\,\mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + \mathbb{E}\|z - \widetilde{x}^{\leftarrow}_{kh}\|^2 + R^2 + \ell^2h^2L^2_{f;\mathsf{t}}\\ &\lesssim \mathbb{E}\|\widetilde{x}^{\leftarrow}_t\|^2 + L^2_{f;\mathsf{x}}\,\mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2 + g^4_{\max}\,\mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + R^2 + \ell^2h^2L^2_{f;\mathsf{t}}\,,\end{aligned}$$

where in the second step we used (D.2) and z is defined in (4.1), and in the third step we used (D.2) and the fact that $\ell h \ll 1$ by (D.2). By Grönwall applied to the interval of times $t \in [kh, (k+1)h]$ along the reverse process, we find that

$$\begin{aligned}\mathbb{E}\|\widetilde{x}^{\leftarrow}_t\|^2 &\lesssim \exp(O(h)) \cdot \big((1 + hL^2_{f;\mathsf{x}})\,\mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2 + h(g^4_{\max}\,\mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + R^2 + \ell^2h^2L^2_{f;\mathsf{t}})\big)\\ &\lesssim \exp(cL^2_{f;\mathsf{x}}h)\,\mathbb{E}\|\widetilde{x}^{\leftarrow}_{kh}\|^2 + h\,\exp(O(h)) \cdot (g^4_{\max}\,\mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 + R^2 + \ell^2h^2L^2_{f;\mathsf{t}})\end{aligned}$$

for all $t \in [kh, (k+1)h]$ for some absolute constant $c > 0$. In particular, this bound holds for $t = (k+1)h$. Iterating this T/h times, we obtain the desired bound. □

Recall the definition of Λ, Λ' in (4.3).

Lemma D.10. *For all integers* $0 \le k \le T/h$*,*

$$\begin{aligned}\mathbb{E}\|\nabla \ln q^{\leftarrow}_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2 \lesssim \Lambda^{O(1)}\Big(&(L_{f;\mathsf{x}} + g^2_{\max}L_{\mathsf{sc},*})d + g^4_{\max}L^2_{\mathsf{high}}d^2T\\ &+ L_{\mathsf{sc},*}T \max_{t\in[0,T]} \mathbb{E}\|\mu_t(\widetilde{x}^{\leftarrow}_t) - \mu'_{\lfloor t/h\rfloor h}(\widetilde{x}^{\leftarrow}_{\lfloor t/h\rfloor h})\|^2\Big)\end{aligned}$$

Proof. The proof follows from Lemmas D.1, D.2, D.5, and the bound in Lemma C.8 with $\zeta_t \triangleq \mathbb{E}\|\mu_t(\widetilde{x}^{\leftarrow}_t) - \mu'_{kh}(\widetilde{x}^{\leftarrow}_{kh})\|^2$ and $\zeta^2 = \int_0^T \zeta_t^2\,\mathrm{d}t \le T\max_t \zeta_t^2$. Note that in the definition of Λ and Λ', we have a $L^2_{f;\mathsf{x}}$ term in the integrand even though there is only an $L_{f;\mathsf{x}}$ term in the definition of L_t in Lemma D.1. The reason for this looseness is to absorb the $\exp(O(L^2_{f;\mathsf{x}}T))$ terms that appear elsewhere in the above analysis. □

D.4. Putting everything together

Proof of Theorem 4.3. Let $\delta > 0$ be a small parameter to be tuned later, and suppose h, ℓ satisfy (D.7). Then by integrating the bound in Lemma D.7 over $0 \le t \le T$ and applying Lemma D.10, we conclude that

$$\begin{aligned}\zeta^2 &\triangleq \int_0^T \mathbb{E}\|\mu_t(\widetilde{x}^{\leftarrow}_t) - \mu'_{\lfloor t/h\rfloor h}(\widetilde{x}^{\leftarrow}_{\lfloor t/h\rfloor h})\|^2\,\mathrm{d}t\\ &\lesssim \delta T + \delta\Lambda^{O(1)}\Big((L_{f;\mathsf{x}} + g^2_{\max}L_{\mathsf{sc},*})dT + (1 + g^2_{\max})^2L^2_{\mathsf{high}}d^2T^2\\ &\qquad + L_{\mathsf{sc},*}T\int_0^T \mathbb{E}\|\mu_t(\widetilde{x}^{\leftarrow}_t) - \mu'_{\lfloor t/h\rfloor h}(\widetilde{x}^{\leftarrow}_{\lfloor t/h\rfloor h})\|^2\,\mathrm{d}t\Big)\,.\end{aligned}$$

Provided that

$$\delta \le \frac{1}{2}\Lambda^{-O(1)}L^{-1}_{\mathsf{sc},*}T^{-1}\,,$$

we can rearrange to conclude that

$$\zeta^2 \lesssim \delta \Lambda^{O(1)} \big((L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},*})dT + g_{\max}^4 L_{\mathsf{high}}^2 d^2 T^2\big).$$

By Theorem 4.7,

$$\mathsf{KL}\,(\pi'_T \| \pi_T) \lesssim (\Lambda^{O(1)} + \Lambda'^{O(1)})(L_0'^{1/2} d^{1/2} + MdT^{1/2})\,\zeta T^{1/2} + \Lambda^{O(1)} L'^{1/2} \zeta^2$$

We will take δ sufficiently small that $\zeta^2 \leq 1$, in which case by upper bounding L_0' by L', the above is at most

$$\begin{aligned}
&\lesssim (\Lambda^{O(1)} + \Lambda'^{O(1)})(L'^{1/2} d^{1/2} + MdT^{1/2})\,\zeta T^{1/2} \\
&\lesssim (\Lambda^{O(1)} + \Lambda'^{O(1)})\big((L_{f;\mathsf{x}}^{1/2} + g_{\max} L_{\mathsf{sc},*}^{1/2})d^{1/2} + g_{\max}^2 L_{\mathsf{high}} dT^{1/2}\big) \\
&\qquad \times \big((L_{f;\mathsf{x}}^{1/2} + g_{\max} L_{\mathsf{sc},*}^{1/2})d^{1/2} T^{1/2} + g_{\max}^2 L_{\mathsf{high}} dT\big)\delta^{1/2} T^{1/2} \\
&\lesssim (\Lambda^{O(1)} + \Lambda'^{O(1)})\big((L_{f;\mathsf{x}} + g_{\max}^2 L_{\mathsf{sc},*})dT + g_{\max}^4 L_{\mathsf{high}}^2 d^2 T^2\big)\delta^{1/2} T^{1/2}
\end{aligned}$$

We take δ so that the above is at most the target accuracy ϵ. By (D.7), this can be achieved by taking h, ℓ satisfying the bounds in the theorem statement. □