Non-Asymptotic Convergence of Diffusion Models

Alan Chung, Ben Schiffer, Kevin Luo

December 24, 2023

Contents

1 Introduction

Diffusion models, collectively introduced in the past 5 or so years through [SDWMG15, SE20, SSDK+21, HJA20], have revolutionized the field of generative modeling, now well known in the form of DALL-E for the task of image synthesis. Supplanting generative adversarial networks as the state of the art in this field, the intuition behind their inner workings is quite different from their predecessors. Diffusion models are trained by corruptting samples from a target distribution with increasing levels of noise, and then learning the denoising process. This then enables one to sample from the target distribution by running the denoising process on samples of pure noise.

While practical usage of diffusion models has taken off, yielding visibly impressive results, our theoretical understanding of them remains rather nascent. Here, we overview recent results surrounding the non-asymptotic convergence of diffusion models. The main result is that of [BBDD23], which shows that $\tilde{O}\left(\frac{d\log^2(1/\delta)}{\epsilon^2}\right)$ steps are necessary to arbitrarily approximate a δ-noised version of any data distribution on $\mathbb{R}^d$ within ϵ^2, measured by KL divergence, assuming one has access to a sufficiently accurate score estimator. This work builds on a recent body of literature including [CLL23, CCL+23, LLT23] which offer looser results on the convergence rates of diffusion models. We overview in depth both the theoretical tools used in these works to analyze diffusion models, the proof of the major result itself, as well as the practical implications of the results.

We begin with an overview of the theoretical model of diffusion in Section 2. We then describe the main proof technique of [BBDD23] and comment that previous works such as [CLL23] use similar methods in

Section 3. Here, we highlight the role of the stochastic localization in yielding the improved bounds, and in Section 4, we give a self-contained overview of the relevant theory needed to obtain those results. In particular, we will explore the proofs of Propositions 1 and 2 of [BBDD23], which are the main tools through which they managed to achieve tighter bounds; .

2 Diffusion Models

In this section, we first introduce the basic principles of diffusion models.

The main idea behind diffusion models is to corrupt samples with noise, and then learn the denoising process. Running the denoising process on pure noise should then yield samples from the target distribution. Formally, suppose that we have some data distribution p_{data} supported on $\mathbb{R}^d$ from which want to generate more samples. Consider the following $\mathbb{R}^d$-valued stochastic process: we initialize $X_0 \sim p_{\text{data}}$ and then run an SDE of the form

$$\mathrm{d}X_t = F(t, X_t)\mathrm{d}t + \sqrt{g(t)}\mathrm{d}B_t \tag{1}$$

on $t \in [0, T]$. In our setting, we choose the coefficients of the OU process, i.e.,

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{2}$$

The reason this choice is canonical is because the formal solution to this SDE is given by

$$X_t = X_0 e^{-t} + B_{1-e^{-2t}}, \tag{3}$$

so that this process converges exponentially quickly (in some norm) to $B_1 \sim \mathsf{N}(0, \mathrm{Id}_d)$. This, along with the fact that the transition densities of the OU process have closed analytic form, make it a convenient choice.

The stochastic prescribed in (2) known as the *forward process*. Our goal is to learn the dynamics of the reverse process $Y_t := X_{T-t}$, which satisfies the following SDE

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log q_{T-t}(Y_t))\,\mathrm{d}t + \sqrt{2}\mathrm{d}B'_t, \quad Y_0 \sim q_T, \tag{4}$$

with q_t being the the marginal distribution of X_t and B'_t is another Brownian motion on the same space. That the reverse process satisfies this SDE is proven in Section 4. If the quantities in (4) were known, then to sample from p_{data}, we need only sample $Y_0 \sim q_T$ and run the reverse dynamics, outputting $Y_T \sim p_{\text{data}}$.

We also introduce the notation $\mathbf{m}_t(\mathbf{x}_t) = \mathbb{E}_{q_{0|t}(\mathbf{x}_0|\mathbf{x}_t)}[\mathbf{x}_0]$ and $\mathbf{\Sigma}_t(\mathbf{x}_t) := \mathrm{Cov}_{q_{0|t}(\mathbf{x}_0|\mathbf{x}_t)}(\mathbf{x}_0)$ to be the expectation and variance of $\mathbf{x}_0$ under the conditional distribution given $\mathbf{x}_t$.

3 Main results of [BBDD23]

3.1 Informal Main Theorem

The core idea underlying all of [BBDD23, CLL23, CCL+23] is to start with a sample from $\mathsf{N}(0, \mathrm{Id}_d)$ and simulate the reverse process in Equation (4) as best possible. There are three main sources of error when using this as an approximation for the distribution p_{data}.

1. The first source of error comes from the fact that Equation (4) uses $\nabla \log q_{T-t}(Y_t)$, the score. We do not have access to these quantities because the true distribution p_{data} is unknown, and thus must use some approximation to the score function. Formally, we let $s_\theta(\mathbf{x}, t)$ be an approximation of $\nabla \log q_t(\mathbf{x}_t)$ that is a sufficiently close in the sense that the quantity

$$\mathcal{L}(s_\theta) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)}\left[\|s_\theta(\mathbf{x}_t, t) - \nabla \log q_t(\mathbf{x}_t)\|^2\right]\,\mathrm{d}t \tag{5}$$

is bounded by some ϵ^2_{score}. We will refer $s_\theta(\mathbf{x}, t)$ as the "score estimator". In practice, we gain access to such an estimator by training one on data samples.

2. Because we do not have access to q_T to initialize the reverse process, we instead start with $p_0 \sim \pi_d := \mathsf{N}(0, \mathrm{Id}_d)$. As noted above, because we chose the OU process we know that as $T \longrightarrow \infty$, q_T approaches $\mathsf{N}(0, \mathrm{Id}_d)$ exactly. However, we are working with finite T, and therefore X_T will not exactly be distributed according to π_d. Fortunately, the convergence rate is fast enough that we can easily bound this source of error.

3. Simulating a continuous time process exactly is computationally impossible, and therefore we must approximate the continuous process with a discretization. This discretization will include N points $0 = t_0 < t_1 < \cdots < t_N \leq T$. Note that the step size will be denoted $\gamma_k = t_{k+1} - t_k$. The *approximate reverse process* is then given by

$$\mathrm{d}\hat{Y}_t = \{\hat{Y}_t + 2s_\theta(\hat{Y}_{t_k}, T - t_k)\}\mathrm{d}t + \mathrm{d}\hat{B}_t. \tag{6}$$

   The key contribution of [BBDD23] relative to [CLL23] is that they prove a tighter bound on the error caused by discretization using results from stochastic localization. This allows [BBDD23] to achieve stronger bounds relative to other recent papers.

In addition to these three approximations, instead of approximating $p_{\text{data}} = q_0$ which is the goal of this problem, we will instead approximate q_δ. In other words, we will perform "early stopping" and end the process p_n at time $t_N = T - \delta$. Such an approximation is allowed and standard in many of these works because for small δ, q_δ and p_{data} are close.

Our goal in this project is then to describe the proof of the following main result of [BBDD23], stated informally below:

Theorem 3.1 ([BBDD23, Theorem 1], informal)**.** *Assuming a "sufficiently accurate" score estimator, there exists a sequence of discretization times $t_1, \ldots, t_N$ such that*

$$\mathrm{KL}(q_\delta \| p_{t_N}) = O(\epsilon^2), \tag{7}$$

with N at most $\tilde{O}\left(\frac{d \log^2(1/\delta)}{\epsilon^2}\right)$.

Recall that p_{t_N} is the output of the algorithm described above, which we want to be a good approximation of p_{data} (but we will settle for being a good approximation of q_δ. The interpretation of this theorem is that the distance between q_δ and p_{T_n} can get as close as the accuracy of the score estimator with only $\tilde{O}\left(\frac{d \log^2(1/\delta)}{\epsilon^2}\right)$ steps. Therefore, the constraining factor in designing better diffusion models does reduce to simply designing better score estimators.

3.2 Formal Results of [BBDD23]

Theorem 3.2 ([BBDD23, Theorem 1])**.** *Suppose $s_\theta(\mathbf{x}, t)$ satisfies*

$$\sum_{k=0}^{N-1} \gamma_k \, \mathbb{E}_{q_{t_k}(\mathbf{x})} \left[\|\nabla \log q_{T-t_k}(\mathbf{x}) - s_\theta(\mathbf{x}, T - t_k)\|^2 \right] \leq \epsilon^2_{score} \tag{8}$$

Further assume that p_{data} is normalized such that $\mathrm{Cov}(p_{data}) = \mathrm{Id}_d$, $T \geq 1$, and there exists $\kappa > 0$ such that for $k = 0, ..., N-1$ we have $\gamma_k \leq \kappa \min\{1, T - t_{k+1}\}$. Then

$$\mathrm{KL}(q_\delta \parallel p_{t_N}) \leq \underbrace{\epsilon^2_{score}}_{\textit{Error 1}} + \underbrace{\kappa^2 dN + \kappa dT}_{\textit{Error 3}} + \underbrace{de^{-2T}}_{\textit{Error 2}} \tag{9}$$

where d is the dimension of p_{data}.

Note that in Theorem 3.2, we explicitly can separate the error terms in the KL divergence as being caused by the three sources of error we listed above. The additional variable κ introduced in the theorem forces an exponential decay in the step size near the end of the reverse process and can be viewed as simply a control on the discretization that allows for tighter bounds.

In order to transition from Theorem 3.2 to the informal theorem statement, we need a choice of κ, N, T that satisfy the necessary step inequality. To do this, they take $T = \frac{1}{2}\log(d/\epsilon_{score}^2)$ and $N = \Theta\left(\frac{d(T+\log(1/\delta))^2}{\epsilon_{score}^2}\right)$ and $\kappa = \Theta\left(\frac{T+\log(1/\delta)}{N}\right)$. The choice of discretization steps γ_k that satisfy $\gamma_k \leq \kappa \min(1, T - t_{k+1})$ will be taking the first half of the times to be equally spaced between $[0, T-1]$. The other half of the times will then be exponentially spaced in the remaining distance between $[T-1, T-\delta]$ with spacing starting at $\frac{\kappa}{1+\kappa}$ and decaying by $\frac{1}{1+\kappa}$ with each time step.

3.3 Proof Sketch of Theorem 3.2

As somewhat discussed in class, the general method for proofs of this nature is to instead reduce this problem to comparing the path measures of the true and approximate reverse processes, rather than just the measures of the termination point. With that in mind, let Q be the path measure of the true reverse process (4), initialized at q_T, let P^{π_d} be the path measure of the approximate reverse process (6) which is initialized at π_d (recall $\pi_d = \mathsf{N}(0, \mathrm{Id}_d)$, and let P^{q_T} be the path measure of the approximate reverse process initialized via q_T. We only consider paths up to time $T - \delta$ for each of these processes, as we are only interested in approximating q_δ.

First, note that the data processing inequality yields

$$\mathrm{KL}(q_\delta \| p_{t_N}) \leq \mathrm{KL}(Q \| P^{\pi_d}). \tag{10}$$

The data processing inequality simply states that given a conditional distribution $P_{Y|X}$ and two distributions P_X and Q_X over X, and P_Y (resp Q_Y) is a distribution over Y formed by first sampling X according to P_X (resp Q_X), then one has $\mathrm{KL}(P_X \| Q_X) \geq \mathrm{KL}(P_Y \| Q_Y)$. The intuition for this is that one can simply regard $P_{Y|X}$ as a stochastic function taking in inputs x; then essentially the DPI states that one cannot increase the KL divergence of two distributions by applying the same stochastic function to each of their outputs. Here, $P_{Y|X}$ is given simply as reading off the value of the path at time $t_N = T - \delta$.

In the following sections, we sacrifice rigor for intuition, especially when explaining topics from stochastic calculus.

3.3.1 Moving between approximate and true reverse processes - Girsanov's theorem

The first important step is controlling the difference between the true and approximate reverse paths using Girsanov's theorem. We first review Girsanov's theorem itself:

Proposition 3.1 (Girsanov's Theorem)**.** *Suppose we have a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, Q)$ and that $(b_t)_{t\in[0,T]}$ is a previsible process such that $\mathbb{E}_Q[\int_0^T \|b_s\|^2 \,\mathrm{d}s] < \infty$. Let $\mathcal{L}_t = \int_0^t b_s \,\mathrm{d}B_s$. Then $\mathcal{L}_t$ is a square integrable Q-martingale. Moreover, let*

$$\mathcal{E}(\mathcal{L})_t = \exp\left\{\mathcal{L}_t - \frac{1}{2}\langle\mathcal{L}\rangle_t\right\} = \exp\left\{\int_0^t b_s \,\mathrm{d}B_s - \frac{1}{2}\int_0^t \|b_s\|^2 \,\mathrm{d}s\right\}, \tag{11}$$

be the stochastic exponential *of $\mathcal{L}_t$, where $\langle\mathcal{L}\rangle_t$ denotes the* quadratic variation *of the process $\mathcal{L}$, detailed in Section 4. Assuming that $\mathbb{E}_Q[\mathcal{E}(\mathcal{L})_T] = 1$, then $\mathcal{E}(\mathcal{L})$ is also a Q-martingale. Furthermore, a probability measure P can then be defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0})$ such that the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is given by $\mathcal{E}(\mathcal{L})_T$ (i.e. $P = \mathcal{E}(\mathcal{L})_T Q$). Then, the process*

$$\beta_t = B_t - \int_0^t b_s \,\mathrm{d}s \tag{12}$$

is a P-Brownian motion.

Essentially, Girsanov's theorem enables us to define a tilted measure under which a different process is a Brownian motion; importantly, this process has a nonzero drift under the original measure (the drift being $\int_0^t b_s \,\mathrm{d}s$). Thus, in some sense, Girsanov's theorem enables us to change the "mean" of the process to some prescribed form.

We now move to proving the bound between the true and approximate reverse paths:

Proposition 3.2 ([BBDD23, Proposition 3])**.** *Let Q and P^{q_T} be the path measures of the solutions to* (4) *and* (6) *respectively, both started in $Y_0 \sim q_T$ and run from $t = 0$ to $t = t_N$. Assume that*

$$\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q[\|\nabla \log_{T-t}(Y_t) - s_\theta(Y_{t_k}, T - t_k)\|^2] \ \mathrm{d}t < \infty. \tag{13}$$

Then, we have

$$\mathrm{KL}(Q\|P^{q_T}) \leq \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T - t_k)\|^2] \ \mathrm{d}t. \tag{14}$$

Proof. We give an informal proof of this proposition. We take $(B'_t)_{t\geq 0}$ to be our Q-Brownian motion and define the process

$$b_t = \sqrt{2}\,(s_\theta(Y_{t_k}, T - t_k) - \nabla \log q_{T-t}(Y_t))$$

for $t \in [t_k, t_{k+1}]$ and each $k = 0, \ldots, N-1$. b_t is previsible (this is not immediately clear to us, since it involves Y_t, but it is probably because the natural filtration is left-continuous). Moreover, b_t satisfies

$$\mathbb{E}_Q\left[\int_0^{t_N} \|b_s\|^2 \ \mathrm{d}s\right] = 2\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T - t_k)\|^2] \ \mathrm{d}t < \infty$$

by assumption, and thus first assumption of Girsanov's theorem are satisfied. Then if we define $\mathcal{L}_t = \int_0^t b_s \ \mathrm{d}B'_s$, then $(\mathcal{E}(\mathcal{L}))_{t\in[0,T]}$ is a continuous local martingale.

We will assume that it is in fact a continuous martingale. There are some ways to get around this (discussed at the end), but this simplifies the argument.

Following Girsanov's theorem, we define the new probability measure $P = \mathcal{E}(\mathcal{L})_{t_N} Q$ and the new process

$$\beta_t = B'_t - \int_0^t b_s \ \mathrm{d}s$$

such that $(\beta_t)_{t\in[0,t_N]}$ is a P-Brownian motion. Furthermore, note that by Itô's lemma (see Section 4 for explanation of Itô's Lemma), we have

$$\mathrm{d}\beta_t = \mathrm{d}B'_t - b_t \ \mathrm{d}t.$$

Since the SDE of the true reverse process (4) holds almost surely under Q, we have that, using the above and the definition of b_t,

$$\begin{aligned} \mathrm{d}Y_t &= (Y_t + 2\nabla \log q_{T-t}(Y_t)) \ \mathrm{d}t + \sqrt{2} \ \mathrm{d}B'_t \\ &= (Y_t + 2s_\theta(Y_{t_k}, T - t_k)) \ \mathrm{d}t + \sqrt{2} \ \mathrm{d}\beta_t. \end{aligned}$$

Note that under the measure P, this is exactly the approximate reverse process (6), as β_t is a P-Brownian motion. Moreover, by construction of P, we now have

$$\begin{aligned} \mathrm{KL}(Q\|P) &= \mathbb{E}_Q\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}\right] = -\,\mathbb{E}_Q[\log \mathcal{E}(\mathcal{L})_{t_N}] \\ &= \mathbb{E}_Q[-\mathcal{L}_{t_N} + \frac{1}{2}\int_0^{t_N} \|b_s\|^2 \ \mathrm{d}s] = \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T - t_k)\|^2] \ \mathrm{d}t \end{aligned}$$

where the last line is by the fact that $\mathcal{L}_t$ is a martingale.

It turns out we can remove the assumption of $\mathcal{E}(\mathcal{L})_t$ being a continuous martingale and just work with it being a continuous local martingale by considering a sequence of times $T_n \to t_N$ and truncating the process β_t. The argument is technical and does not improve the intuition and is the reason for the inequality in 14 rather than the equality attained above.

Essentially, what has occurred here is that the difference between the approximate and true reverse processes is given up to scale by b_s, a previsible process. Using Girsanov's theorem, we are able to produce another measure under which this shifted process is a Brownian motion, and in fact it allows us to directly characterize the KL divergence, since it explicitly gives us the form of the Radon-Nikodym derivative. □

This result then not only gives us this explicit bound, but also shows that Q is absolutely continuous with respect to P^{π_d}. Note that P^{q_T} and P^{π_d} have the same dynamics and instead only differ in their choice of starting distribution. Hence, conditioned on the value of the process at time 0, their densities are equal, meaning

$$\frac{\mathrm{d}P^{q_T}}{\mathrm{d}P^{\pi_d}}(\mathbf{y}) = \frac{\mathrm{d}q_T}{\mathrm{d}\pi_d}(\mathbf{y}_0).$$

It then follows that

$$\frac{\mathrm{d}Q}{\mathrm{d}P^{\pi_d}}(\mathbf{y}) = \frac{\mathrm{d}Q}{\mathrm{d}P^{q_T}}(\mathbf{y})\frac{\mathrm{d}P^{q_T}}{\mathrm{d}P^{\pi_d}}(\mathbf{y}) = \frac{\mathrm{d}Q}{\mathrm{d}P^{q_T}}(\mathbf{y})\frac{\mathrm{d}q_T}{\mathrm{d}\pi_d}(\mathbf{y}_0)$$

and we hence have

$$\mathrm{KL}(Q\|P^{\pi_d}) = \mathrm{KL}(Q\|P^{q_T}) + \mathrm{KL}(q_T\|\pi_d). \tag{15}$$

Combining 10, 15, and Prop 3.2 now yields

$$\mathrm{KL}(q_\delta\|p_{t_N}) \leq \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}[\|\nabla\log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T-t_k)\|^2]\ \mathrm{d}t + \mathrm{KL}(q_T\|\pi_d), \tag{16}$$

and passing through the triangle inequality yields

$$\underbrace{\leq 2\sum_{k=0}^{N-1}\gamma_k\mathbb{E}_Q\left[\|\nabla\log q_{T-t_k}(Y_{t_k}) - s_\theta(Y_{t_k}, T-t_k)\|^2\right]}_{\text{Error 1}}$$

$$\underbrace{+\,2\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_Q\left[\|\nabla\log q_{T-t}(Y_t) - \nabla\log q_{T-t_k}(Y_{t_k})\|^2\right]\mathrm{d}t}_{\text{Error 3}} + \underbrace{\mathrm{KL}(q_T\|\pi_d)}_{\text{Error 2}} \tag{17}$$

Bounding each of these terms is then the remainder of the proof, and they each contribute to each of the errors.

3.3.2 Error from score estimator and from q_T (Errors 1 and 2)

Note that the first term is exactly $\epsilon^2_{\text{score}}$, the error of our score estimator.

Next, we bound the approximation error from starting with normal distribution as opposed to q_T. Under the regularity assumptions, we have the following Lemma.

Lemma 3.3. *For $T \geq 1$, under the assumed regularity conditions we have*

$$\mathrm{KL}(q_T\|\pi_d) \lesssim de^{-2T} \tag{18}$$

Proof. Sketch: Note that $q_T(x)$ can be written as $\int_{\mathbb{R}^d} q_{T|0}(x \mid x_0)p_{\text{data}}(\mathrm{d}x_0)$, and that since the forward process is the OU process, we know the form of $q_{T|0}$ to be $\mathsf{N}(e^{-t}x_0; \sigma_t^2\mathrm{Id}_d)$. Then, because the KL divergence is convex in the first argument, $\mathrm{KL}(q_T\|\pi_d) \leq \mathbb{E}_{x_0\sim p_{\text{data}}}[\mathrm{KL}(q_{T|0}(\cdot \mid x_0)\|\pi_d)]$. The KL divergence between two Gaussians also has a closed form, and ultimately this evaluates to $-d\log(1-e^{-2T}) \lesssim de^{-2T}$, where we utilize the normalization condition $\mathrm{Cov}(p_{\text{data}}) = \mathrm{Id}_p$. □

This lemma directly gives the desired bounds on Error 2.

3.3.3 Bounding discretization error (Error 3)

The final and novel step of [BBDD23] a refinement in the bound of the discretization error (Error 3), which we recall is the error caused by taking discrete time steps instead of continuously reconstructing the reverse process. As written above, this error is bounded as

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_Q\left[\|\nabla\log q_{T-t}(Y_t) - \nabla\log q_{T-t_k}(Y_{t_k})\|^2\right]\mathrm{d}t \leq \kappa^2 dN + \kappa dT \tag{19}$$

This is where the key stochastic localization lemma below is used.

Lemma 3.4 (Application of Stochastic Localization)**.** *For all $t > 0$, if $\sigma_t^2 = 1 - e^{-2t}$, then*

$$\frac{\sigma_t^3}{2\dot{\sigma}_t} \frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[\mathbf{\Sigma}_t] = \mathbb{E}[\mathbf{\Sigma}_t^2] \tag{20}$$

Recall as defined above that $\mathbf{\Sigma}_t$ is the covariance of the conditional distribution of X_0 given X_t. This algebraic result is key in obtaining the strong bound on discretization error in Equation (19). While the exact application is buried within the mathematical calculations, the key idea is that it is easier to work with $\mathbf{\Sigma}_t$ rather than $\mathbf{\Sigma}_t^2$, and this Lemma allows us to substitute one for the other.

We proceed with a sketch of the proof. Instead of thinking of two separate processes $(Y_t)_{t\in[0,T]}$ and $(\hat{Y}_t)_{t\in[0,T]}$ which are solutions to (4) and (6) respectively, under the *same* probability measure, we instead consider a single process $(Y_t)_{t\in[0,T]}$ where under a probability measure Q it is a solution to (4) but under a measure P^{π_d}, it is a solution to (6). One way to think about how this is possible is that in the first way of thinking about the two processes, each of $(Y_t)_{t\in[0,T]}$ and $(\hat{Y}_t)_{t\in[0,T]}$ induce measures on paths, and thus we can simply take Q and P^{π_d} to be the induced path measures and the probability space to be over these paths. Additionally, we define P^{q_T} to be the measure under which $(Y_t)_{t\in[0,T]}$ is a solution to (6), where Y_0 is instead initialized according to q_T rather than π_d.

Note that in the sum we want to control, we are essentially dealing with expectations of differences of the form

$$E_{s,t} = \mathbb{E}_Q[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)\|^2] \qquad s < t \tag{21}$$

Since we expect the score to evolve slowly, we should be able to bound this quantity pointwise (for fixed t, s). One way we might try to do this is to control the time derivative of $E_{s,t}$. With this in mind, regarding t as evolving forward and s fixed, we apply Itô's lemma (see Section 4) to $\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)$ to obtain the following result:

Lemma 3.5 ([BBDD23, Lemma 2])**.** *If $(Y_t)_{t\in[0,T]}$ is the solution to the SDE (4) and $s \in [0,T)$ is fixed, then we have*

$$\begin{aligned} &\mathrm{d}(\|\nabla \log q_{T-t}(Y_t) - \nabla q_{T-s}(Y_s)\|^2) \\ &\quad = -2\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)\|^2 \ \mathrm{d}t - 2(\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)) \cdot \nabla \log q_{T-s}(Y_s) \ \mathrm{d}t \\ &\qquad + 2 \left\|\nabla^2 \log q_{T-t}\left(Y_t\right)\right\|_F^2 \ \mathrm{d}t + 2\sqrt{2}\left\{\nabla \log q_{T-t}\left(Y_t\right) - \nabla \log q_{T-s}\left(Y_s\right)\right\} \cdot \nabla^2 \log q_{T-t}\left(Y_t\right) \cdot \mathrm{d}B_t' \end{aligned} \tag{22}$$

for all $s \le t < T$.

Proof. Sketch: One applies Itô's lemma to $\nabla \log q_{T-t}(Y_t)$ and $\|\nabla \log q_{T-t}(Y_t)) - \nabla \log q_{T-s}(Y_t)\|^2$, as well as using the Fokker-Plank equation for the forward process to compute $\mathrm{d}(\nabla \log q_t)(\mathbf{x})/\mathrm{d}t$. Combining terms and simplifying yields the above result. □

We then take expectations and integrate (22), followed by differentiating with respect to t to obtain

$$\begin{aligned} \frac{\mathrm{d}E_{s,t}}{\mathrm{d}t} &= -2\,\mathbb{E}_Q[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)\|^2] \\ &\qquad + 2\,\mathbb{E}_Q[(\nabla \log q_{T-s}(Y_s) - \nabla \log q_{T-t}(Y_t)) \cdot \nabla \log q_{T-s}(Y_s)] + 2\,\mathbb{E}_q[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2] \end{aligned}$$

Applying AM-GM to the middle term:

$$\le \mathbb{E}_Q[\|\nabla \log q_{T-s}(Y_s)\|^2] + 2\,\mathbb{E}_Q[\|\nabla^2 \log q_{T-t}\|_F^2] \tag{23}$$

Thus it suffices to bound these two terms separately, and then integrate back out over t. To this, we rely on the following lemma:

Lemma 3.6 ([BBDD23, Lemma 3])**.** *For all $t > 0$, $\nabla \log q_t(\mathbf{x}_t) = -\sigma_t^{-2}\mathbf{x}_t + e^{-t}\sigma_t^{-2}\mathbf{m}_t$ and $\nabla^2 \log q_t(\mathbf{x}_t) = -\sigma_t^{-2}\mathrm{Id}_d + e^{-2t}\sigma_t^{-4}\mathbf{\Sigma}_t$, where $\sigma_t^2 = 1 - e^{-2t}$.*

Proof. The first portion is actually just Tweedie's formula (up to rescaling and rearrangement), since, using the transition densities for the OU process, we see that $\mathbf{x}_t \mid \mathbf{x}_0 \sim \mathsf{N}(e^{-t}\mathbf{x}_0, \sigma_t^2\mathrm{Id}_d)$, which we discussed in class. The proof of the second portion just involves swapping integrals and derivatives. □

Now, the first term in (23) can be controlled using the first part of Lemma 3.6 through a Nishimori-style trick, yielding $\mathbb{E}_Q[\|\nabla \log q_{T-s}(Y_s)\|^2] \leq d\sigma_{T-s}^{-2}$. The important portion is the bounding of the second term, which involves the use of Lemma 3.4.

Let $\dot{\sigma}_t$ denote the time derivative of σ_t. The, using the second part of Lemma 3.6 and expanding, we have

$$\mathbb{E}_{q_t(\mathbf{x}_t)}[\|\nabla^2 \log q_t(\mathbf{x}_t)\|_F^2)] = d\sigma_t^{-4} - 2\dot{\sigma}_t\sigma_t^{-5}\,\mathbb{E}[\mathrm{Tr}(\mathbf{\Sigma}_t)] + \dot{\sigma}_t^2\sigma_t^{-6}\,\mathbb{E}[\mathrm{Tr}(\mathbf{\Sigma}_t^2)].$$

Crucially, Lemma 3.4 then yields

$$\frac{\sigma_t^4}{2e^{-2t}}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[\mathrm{Tr}(\mathbf{\Sigma}_t)] = \mathbb{E}[\mathrm{Tr}(\mathbf{\Sigma}_t^2)],$$

and substituting this above obtains

$$\mathbb{E}_{q_t(\mathbf{x}_t)}\left[\left\|\nabla^2 \log q_t\left(\mathbf{x}_t\right)\right\|_F^2\right] = d\sigma_t^{-4} - 2\dot{\sigma}_t\sigma_t^{-5}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{\Sigma}_t\right)\right] + \frac{1}{2}\dot{\sigma}_t\sigma_t^{-3}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{\Sigma}_t\right)\right]$$

Using $0 \leq \sigma_t\dot{\sigma}_t \leq 1$:

$$\leq d\sigma_t^{-4} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\sigma_t^{-4}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{\Sigma}_t\right)\right]\right)$$

This immediately yields

$$\mathbb{E}_Q\left[\left\|\nabla^2 \log q_{T-t}\left(Y_t\right)\right\|_F^2\right] \leq d\sigma_{T-t}^{-4} - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r}\left(\sigma_{T-r}^{-4}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{\Sigma}_{T-r}\right)\right]\right)\Big|_{r=t}$$

giving the complete bound

$$\mathbb{E}_Q\left[\left\|\nabla \log q_{T-s}\left(Y_s\right)\right\|^2\right] + 2\mathbb{E}_Q\left[\left\|\nabla^2 \log q_{T-t}\left(Y_t\right)\right\|_F^2\right] \leq \underbrace{d\sigma_{T-s}^{-2} + 2d\sigma_{T-t}^{-4}}_{E_{s,t}^{(1)}} \underbrace{- \frac{\mathrm{d}}{\mathrm{d}r}\left(\sigma_{T-r}^{-4}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{\Sigma}_{T-r}\right)\right]\right)\Big|_{r=t}}_{E_{s,t}^{(2)}}.$$

We now integrate back out over t. Recall this expression is an upper bound on $\frac{\mathrm{d}E_{s,t}}{\mathrm{d}t}$, and thus

$$E_{t_k,t} \leq \int_{t_k}^{t} E_{t_k,s}^{(1)} + E_{t_k,s}^{(2)}\ \mathrm{d}s. \tag{24}$$

The remainder of the proof is mostly computational. We can write

$$\begin{aligned}&\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_Q\left[\left\|\nabla \log q_{T-t}\left(Y_t\right) - \nabla \log q_{T-t_k}\left(Y_{t_k}\right)\right\|^2\right]\mathrm{d}t\\ &\quad\leq \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} E_{t_k,t}\ \mathrm{d}t = \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t} E_{t_k,s}^{(1)}\ \mathrm{d}s\ \mathrm{d}t + \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t} E_{t_k,s}^{(2)}\ \mathrm{d}s\ \mathrm{d}t\end{aligned}$$

We bound the two sums separately. Recall γ_k satisfies $\gamma_k \leq \kappa \min\{1, T - t_{k+1}\}$ for some κ. As a result of this, assume there exists some index M such that $t_M = T - 1$, so that when analyzing each sum, we consider the first M terms and the last $N - M$ terms separately due to the form of γ_k. Ultimately, some crude bounds suffice, and one eventually obtains that the first sum is $\lesssim \kappa^2 dN$ and the second is $\lesssim \kappa d + \kappa^2 dN$. This then yields the final bound

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_Q\left[\left\|\nabla \log q_{T-t}\left(Y_t\right) - \nabla \log q_{T-t_k}\left(Y_{t_k}\right)\right\|^2\right]\mathrm{d}t \leq \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} E_{t_k,t}\ \mathrm{d}t \lesssim \kappa^2 dN + \kappa dT. \tag{25}$$

This is then the bound on Error 3, and plugging in everything to (17) yields the desired result.

3.4 Previous Works

The results in [BBDD23] built upon a strong foundation of works that also analyze the convergence rates of diffusion models, following [CCL$^+$23] and [CLL23]. The key differences between these three works is the assumptions needed in order to guarantee the results, with the methodology in each work building with slightly weaker assumptions and slightly looser bounds on convergence rates. Starting with [CCL$^+$23], the main difference with [BBDD23] is that they make the (stronger) assumption that $\nabla \log(q_t)$ is L-Lipschitz $\forall t$ (recall that $\nabla \log(q_t)$ is the score that was instrumental in the proof above. Even in [CCL$^+$23] they note that this L-Lipschitz condition could be relaxed, as was done in the later works. Note that while this work also presents bounds in terms of Total Variation distance rather than KL divergence, these two quantities can be interpreted side-by-side with a variety of inequalities such as Pinsker's inequality as is done in [CCL$^+$23]. The proof structure in [CCL$^+$23] is very similar to what we described above, with the $TV(p_T, q_0)$ being decomposed into three components, 1. convergence of forward process 2. discretization error, 3. score estimation error. The first result in [CCL$^+$23] is that $TV(p_{data}, p_T)^2 \leq \tilde{O}(\epsilon^2)$ with $\tilde{O}(dL^2/\epsilon^2)$ samples. Note that, unlike above, this result is with respect to the true distribution we are trying to estimate $p_{data} = q_0$ rather than q_δ, and uses the full T step reverse process p_T rather than $p_{T_n} = p_{T-\delta}$. The second main result from this paper is that they do an initial result of relaxing the L-Lipschitz condition, and instead assume only that the distribution q_0 is only supported on the ball of radius $R \geq 1$. In this result, they also use early stopping to bound $TV(q_\delta, p_{T_n})^2$ with $\tilde{O}(dR^4/(\epsilon^2\delta^4)$ samples.

[CLL23] builds on the work in [CCL$^+$23] by relaxing the assumption of L-Lipschitz to only hold for $\nabla \log p_0$ rather than holding for $\nabla \log p_t$ for all t. In this work they also switch to bounding KL divergence, and bound $\mathrm{KL}(q_0, p_T)$ with $\tilde{O}\left(\frac{d^2 \log^2(L)}{\epsilon}^2\right)$ samples. The other main new result from this paper is that they use early stopping (as described above) to bound $\mathrm{KL}(q_\delta \parallel p_{T-\delta})$ with $\tilde{O}\left(\frac{d^2 \log^2(1/\delta)}{\epsilon_0^2}\right)$ samples without any Lipschitz assumptions. Again, recall that this result must be relative to the early stopping time as with the weak assumptions, it is impossible to obtain KL or TV closeness to the true p_{data}. Note that this result is exactly a factor of d worse than the results described above by [BBDD23]. This paper can obtain the correct upper bound (in terms of d) when the score is assumed Lipschitz; when they drop this assumption, they instead attempt to prove a high probability bound on the norm of the Hessian in order to control the discretization error. It is this bound that incurs the factor of d^2 that ultimately causes the entire bound to be of order d^2. In the main paper we survey, the entire approach of bounding the discretization error is quite different, and the results from stochastic localization enable them to prove this tighter bound.

4 Diffusion, Stochastic Localization, and the Key Lemma

In this section, we provide a proof of the key Lemma 3.4, which, as stated above, leads to the improvement in [BBDD23]. This lemma is based on a fundamental connection between diffusion models and stochastic localization [Mon23], a self-contained treatment of which we provide here before proceeding to the proof.

Before we begin, we give a brief overview of some basic notions in stochastic calculus which were used a handful times in the sections above, and will be used move heavily below.

4.1 Itô Calculus

Theorem 4.1 ((1-dimensional) Itô's Lemma, from Oskendal)**.** *Let $\{X_t\}$ be an Itô process given by $\mathrm{d}X_t = v_t \mathrm{d}B_t + u_t \mathrm{d}t$. Let $g(t,x) \in C^2([0,\infty) \times \mathbb{R})$. Then $Y_t := g(t, X_t)$ is again an Itô process given by*

$$dY_t = \frac{\partial g}{\partial t}(t, X_t)\ \mathrm{d}t + \frac{\partial g}{\partial x}(t, X_t)\ \mathrm{d}X_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t)\ \mathrm{d}\langle X \rangle_t \tag{26}$$

The intuition behind this is that the quadratic variation of Brownian motion does not vanish, while it does for deterministic functions. The result is that when one Taylor expands, they must consider up to second order terms in x as a result. This is why there is an additional third term. We will use the multidimensional version of this, but it does not change much.

Theorem 4.2 (The Itô Isometry). *For any 1-dimensional stochastic process X_t which is adapted to the natural filtration of the 1-dimensional Brownian motion B_t, one has*

$$\mathbb{E}\left[\left(\int_0^T X_t \,\mathrm{d}B_t\right)^2\right] = \mathbb{E}\left[\int_0^T X_t^2 \,\mathrm{d}t\right] \tag{27}$$

Similar results again hold for higher dimensions.

4.2 Stochastic Localization

The main idea of stochastic localization is to construct a sequence of measures μ_t that "localizes" around some point mass. In other words, we would like that $\mu_t \to \delta_{x^*}$ as $t \to \infty$, where x^* is distributed according to μ. More concretely, consider the following canonical construction. Our ultimate goal is to sample from μ. Suppose at first that x^* is such a sample from μ. Now, consider the process

$$U_t = tx^* + B_t, \tag{28}$$

where W_t is a Brownian motion. Intuitively, taking μ_t to be the conditional distribution of x^* given U_t, then μ_t converges to δ_{x^*}, since this process becomes more and more informative about x^* as $t \to \infty$ (in the sense that the signal-to-noise ratio grows arbitrarily large). One can more formally compute (from [Mon23])

$$\mu_t(dx) = \frac{1}{Z'}\mu(dx)\exp\left(-\frac{1}{2t}||U_t - tx||_2^2\right) = \frac{1}{Z}\mu(dx)\exp\left(\langle U_t, x\rangle - \frac{t}{2}||x||^2\right) \tag{29}$$

Hence, the construction of the process $(\mu_t)_{t\geq 0}$ allows us to sample from μ. Even though we have this convenient localization result, this appears not to be useful at first glance. Firstly, this definition of μ_t depends on U_t, which in itself depends on x^*. Hence, constructing this stochastic process appears to rely on us being able to sample in the first place. However, this problem is resolved through the following proposition in Section 7.4, [LS01].

Proposition 4.1. *Suppose that μ has finite second moment. Then, $(U_t)_{t\geq 0}$ is the unique solution of the following SDE with initial condition $U_0 = 0$:*

$$\mathrm{d}U_t = \mathbf{a}_t(U_t)\mathrm{d}t + \mathrm{d}B_t, \tag{30}$$

where

$$\mathbf{a}_t(u) := \mathbb{E}[x|tx + \sqrt{t}G = u], \quad (x, G) \sim \mu \otimes N(0, I_n). \tag{31}$$

Additionally define $\mathbf{A}_t(U_t) = \mathrm{Cov}(\mu_t)$.

4.3 Relationship Between Diffusion and Stochastic Localization

The key observation that relates diffusion and stochastic localization is that the processes in (2) and (28) are equivalent up to a change of time. In particular, we have the following theorem.

Theorem 4.3. *If $(X_t)_{t\geq 0}$ is defined by (2) and $(U_s)_{s\geq 0}$ is defined by (28), then, for $t(s) = \frac{1}{2}\log(1 + s^{-1})$, $(U_s)_{s\geq 0}$ and $(se^{t(s)}X_{t(s)})_{s\geq 0}$ have the same law. Similarly, we will have that $\mathbf{a}_s(U_s)$ and $\mathbf{m}_t(X_t)$ have the same law when $t = t(s)$.*

Note that this will be *reversing* the flow of time, and one can already see the resemblance in the SDEs defining the SL process in (30) and the true reverse process in (4), since we know the form of $\mathbf{a}_t$ will be something similar, as a result of Tweedie's formula.

Proof. Suppose that $(X_t)_{t\geq 0}$ follows the OU SDE in (2), that is, $\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$. Then, integration by parts for continuous semimartingales yields that

$$\mathrm{d}(e^tX_t) = e^tX_t\mathrm{d}t + e^t\left(-X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\right) = \sqrt{2}e^t\mathrm{d}B_t. \tag{32}$$

Hence, the DDS theorem states that there exists some standard Brownian motion $(W_s)_{s\geq 0}$ so that

$$W_{e^{2s}-1} = \int_0^s \sqrt{2}e^r \mathrm{d}B_r, \tag{33}$$

since we have the equality $\langle \int_0^s \sqrt{2}e^r \mathrm{d}B_r \rangle = \int_0^s 2e^{2r}dr = e^{2s} - 1$. Hence, by (32), we have that $e^{\tau(s)}X_{\tau(s)} = X_0 + W_s$, where $\tau(s) = \frac{1}{2}\log(1+s)$. Substituting $s := 1/s$ and multiplying both sides by s yields $se^{\tau(1/s)}X_{\tau(1/s)} = sX_0 + sW_{1/s}$. However, we note that this exactly satisfies Equation (28), as $sW_{1/s}$ has the law of a standard Brownian motion. This shows one of the desired results. To argue that $\mathbf{a}_s(U_s)$ and $\mathbf{m}_t(X_t)$ have the same laws when $t = \tau(s)$, it suffices to note that conditioning on U_s is equivalent to conditioning on $X_{t(s)}$. □

Hence, the main technique in this paper is that because there is this relationship/equivalence between diffusion and stochastic localization, the tools developed in the stochastic localization literature can be used to obtain tighter bounds.

4.4 Flow Reversal

This is maybe not a rigorous proof, but does what was suggested in class in that it checks that the Fokker-Plank equations for the forward and reverse processes are exactly negated. Recall from class the Fokker-Plank equation:

Theorem 4.4. *For any smoothly varying family of smooth vector fields $v_t : \mathbb{R}^d \to \mathbb{R}^d$, the iterates x_t of the SDE*

$$\mathrm{d}X_t = v_t(X_t)\ \mathrm{d}t + \sqrt{2}\ \mathrm{d}B_t \tag{34}$$

are distributed according to q_t satisfying the PDE

$$\frac{\partial q_t}{\partial t} = -\mathrm{div}(q_t(x)v_t(x)) + \Delta q_t. \tag{35}$$

For the forward process, a straightforward calculations yields

$$\frac{\partial q_t}{\partial t} = dq_t + \nabla q_t \cdot x + \Delta q_t.$$

For the reverse process, we should have the marginal distribution (denote them q') at time $T-t$ is equal to the marginal distribution distribution of the forward process at time t. We check this:

$$\frac{\partial q'_{T-t}}{\partial t} = -\mathrm{div}(q_t \cdot (Y + 2\nabla \log q_t(Y))) + \Delta q_t$$

Chain ruling and doing basic calculations yields

$$= -dq_t - \nabla q_t \cdot y - \Delta q_t$$

which is exactly what is expected.

4.5 Proof of the Key Lemma

The relevant results in the stochastic localization literature as stated and proven as follows:

Proposition 4.2 ([BBDD23, Proposition 1]). *If we define $L_s(\mathbf{x}) = \frac{\mathrm{d}\mu_s}{\mathrm{d}\mu}(\mathbf{x})$, then $\mathrm{d}L_s(\mathbf{x}) = L_s(\mathbf{x})(\mathbf{x}-\mathbf{a}_s)\cdot \mathrm{d}B_s$ for all $s \geq 0$.*

Proof. (29) yields

$$L_s(\mathbf{x}) = \frac{1}{Z_s}\exp\left(U_t^\top \mathbf{x} - \frac{s}{2}\|\mathbf{x}\|^2\right)\ \mu(d\mathbf{x}) \qquad Z_s = \int \exp\left(U_s\cdot \mathbf{x} - \frac{s}{2}\|\mathbf{x}\|^2\right)\ \mu(d\mathbf{x}) \tag{36}$$

and thus

$$\mathrm{d}\log L_s(\mathbf{x}) = \mathbf{x}\cdot \mathrm{d}U_s - \frac{1}{2}\|\mathbf{x}\|^2\mathrm{d}s - \mathrm{d}\log Z_s. \tag{37}$$

Define $h_s(\mathbf{x}) = U_s \cdot \mathbf{x} - \frac{s}{2}\|\mathbf{x}\|^2$ and note $\mathrm{d}h_s(\mathbf{x}) = \mathbf{x}\cdot \mathrm{d}U_s - \frac{1}{2}\|\mathbf{x}\|^2\mathrm{d}s$, so $h_s(\mathbf{x})$ is also an Itô process. Then $Z_s = \int \exp h_s(\mathbf{x})\ \mu(d\mathbf{x})$ and by Itô's Lemma (the function here is $\int e^x\ \mathrm{d}x$),

$$\begin{aligned}\mathrm{d}Z_s &= \int (\mathrm{d}h_s(\mathbf{x}) + \frac{1}{2}\mathrm{d}\langle h(\mathbf{x})\rangle_s)e^{h_s(\mathbf{x})}\ \mu(\mathrm{d}\mathbf{x})\\ &= \int (\mathbf{x}\cdot \mathrm{d}U_s - \frac{1}{2}\|\mathbf{x}\|^2\ \mathrm{d}s + \frac{1}{2}\|\mathbf{x}\|^2)e^{h_s(\mathbf{x})}\ \mu(\mathrm{d}\mathbf{x})\\ &= Z_s\left(\frac{1}{Z_s}\int \mathbf{x}e^{h_s(\mathbf{x})}\ \mu(\mathrm{d}\mathbf{x})\right)\cdot \mathrm{d}U_s\end{aligned}$$

Recall that $\frac{1}{Z_s}e^{h_s(\mathbf{x})}\mu(\mathrm{d}\mathbf{x})$ is exactly the conditional density of $\mathbf{x}^*$, conditional on observing $\mathbf{x}_s$, and hence this integral is just the conditional mean $\mathbf{a}_s$:

$$= Z_s(\mathbf{a}_s\cdot \mathrm{d}U_s).$$

Hence Z_s is also an Itô process, and thus applying the Itô Lemma to $\log Z_s$ yields

$$\mathrm{d}\log Z_s = \frac{\mathrm{d}Z_s}{Z_s} - \frac{1}{2}\frac{\mathrm{d}\langle Z\rangle_s}{Z_s^2} = (\mathbf{a}_s\cdot \mathrm{d}U_s) - \frac{\|\mathbf{a}_s\|^2}{2}\ \mathrm{d}s$$

upon which substituting into (37) yields

$$\begin{aligned}\mathrm{d}\log L_s(\mathbf{x}) &= (\mathbf{x}-\mathbf{a}_s)\cdot \mathrm{d}U_s - \frac{1}{2}(\|\mathbf{x}\|^2 - \|\mathbf{a}_s\|^2)\ \mathrm{d}s\\ &= (\mathbf{x}-\mathbf{a}_s)\cdot(\mathbf{x}\ \mathrm{d}s + \mathrm{d}B_s) - \frac{1}{2}(\|\mathbf{x}\|^2 - \|\mathbf{a}_s\|^2)\ \mathrm{d}s\\ &= (\mathbf{x}-\mathbf{a}_s)\cdot \mathrm{d}B_s - \frac{1}{2}\|\mathbf{x}-\mathbf{a}_s\|^2\ \mathrm{d}s.\end{aligned}$$

To finish, we apply Itô's lemma one last time, to find $\mathrm{d}L_s(\mathbf{x}) = \mathrm{d}\exp(\log L_s(\mathbf{x}))$:

$$\mathrm{d}L_s(\mathbf{x}) = L_s(\mathbf{x})\cdot \mathrm{d}\log L_s(\mathbf{x}) + \frac{1}{2}L_s(\mathbf{x})\ \mathrm{d}\langle \log L(\mathbf{x})\rangle_s = L_s(\mathbf{x})(\mathbf{x}-\mathbf{a}_s)\ \mathrm{d}B_s$$

as desired. □

Proposition 4.3 ([BBDD23, Proposition 2])**.** *For all $s \geq 0$, $\frac{\mathrm{d}}{\mathrm{d}s}\,\mathbb{E}[\mathbf{A}_s] = -\,\mathbb{E}[\mathbf{A}_s^2]$.*

Proof. From above, we obtain

$$\begin{aligned}\mathrm{d}\mathbf{a}_s &= \mathrm{d}\left(\int_{\mathbb{R}^d}\mathbf{x}\frac{\mathrm{d}\mu_s}{\mathrm{d}\mu}(\mathbf{x})\ \mu(\mathrm{d}\mathbf{x})\right)\\ &= \mathrm{d}\left(\int_{\mathbb{R}^d}\mathbf{x}L_s(\mathbf{x})\ \mu(\mathrm{d}\mathbf{x})\right)\end{aligned}$$

Again by Itô,

$$\begin{aligned}&= \int_{\mathbb{R}^d}\mathbf{x}\mathrm{d}L_s(\mathbf{x})\ \mu(\mathrm{d}\mathbf{x})\\ &= \int_{\mathbb{R}^d}\mathbf{x}\otimes(\mathbf{x}-\mathbf{a}_s)\,L_s(\mathbf{x})\cdot \mathrm{d}B_s\ \mu(\mathrm{d}\mathbf{x}).\end{aligned}$$

Because, again, $L_s(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) = \mu_s(\mathrm{d}\mathbf{x})$, we have

$$\mathrm{d}\mathbf{a}_s = \mathbb{E}_{\mu_s(\mathbf{x})}[\mathbf{x}\otimes(\mathbf{x}-\mathbf{a}_s)]\cdot \mathrm{d}B_s = (\mathbb{E}_{\mu_s}[(\mathbf{x}-\mathbf{a}_s)\otimes(\mathbf{x}-\mathbf{a}_s)] + \underbrace{\mathbb{E}_{\mu_s}[\mathbf{a}_s\otimes(\mathbf{x}-\mathbf{a}_s)]}_{=0})\cdot \mathrm{d}B_s = \mathbf{A}_s\cdot \mathrm{d}B_s$$

Thus

$$\mathbf{a}_t = \int_0^t \mathbf{A}_s \cdot \mathrm{d}B_s$$

By the Itô Isometry:

$$\mathbb{E}[\mathbf{a}_t^{\otimes 2}] = \mathbb{E}\left[\left(\int_0^t \mathbf{A}_s \cdot \mathrm{d}B_s\right)\right] = \mathbb{E}\left[\int_0^t (\mathbf{A}_s)^2 \, \mathrm{d}s\right]$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[\mathbf{a}_t^{\otimes 2}] = \mathbb{E}[\mathbf{A}_t^2]$$

Finally, recall that $\mathbb{E}[\mathbf{A}_t] = \mathbb{E}[\mathbf{x}^{\otimes 2}] - \mathbb{E}[\mathbf{a}_t^{\otimes 2}]$, yielding the desired result. □

We now need to pass back through the relation between stochastic localization and the OU process to obtain the relevant result, Lemma 3.4

Recall $\mathbf{A}_s \sim \mathbf{\Sigma}_t$ when $t = t(s) = \frac{1}{2}\log(1 + s^{-1})$ (so $s = 1/(e^{2t} - 1)$). Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[\mathbf{\Sigma}_t] = \frac{\mathrm{d}\,\mathbb{E}[\mathbf{A}_{s(t)}]}{\mathrm{d}s(t)}\frac{\mathrm{d}s(t)}{\mathrm{d}t} = -\,\mathbb{E}[\mathbf{A}_s^2] \cdot -\frac{2e^{2t}}{(e^{2t}-1)^2}$$
$$\frac{(e^{2t}-1)^2}{2e^{2t}}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[\mathbf{\Sigma}_t] = \mathbb{E}[\mathbf{\Sigma}_t^2]$$

One can then check that the expression on the right is indeed equal to $\sigma_t^3/\dot{\sigma}_t$.

References

[BBDD23] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization, 2023.

[CCL+23] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. 2023.

[CLL23] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. 2023.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity, 2023.

[LS01] Robert Shevilevich Liptser and Al'bert Nikolaevich Shiryaev. *Statistics of Random Processes II: II. Applications*, volume 2. Springer Science & Business Media, 2001.

[Mon23] Andrea Montanari. Sampling, diffusions, and stochastic localization, 2023.

[SDWMG15] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[SE20] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020.

[SSDK+21] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.