Fast Score-Based Sampling via Log-Concave Reductions

Martin J. Wainwright
MJWAIN@MIT.EDU
Laboratory for Information and Decision Systems
Statistics and Data Science Center
EECS & Mathematics
Massachusetts Institute of Technology

Editors: Steve Hanneke and Tor Lattimore

Abstract

Sampling based on score diffusions has led to striking empirical results, and has attracted considerable attention from various research communities. It depends on the availability of (approximate) Stein score functions for various levels of additive noise. We show how, in some generality, the availability of scores allows the general problem to be "reduced" to sampling from an adaptively constructed sequence of K strongly log-concave (SLC) sub-problems. The reduction is simple, constructive and algorithm-independent, so that any SLC sampler can be used as a subroutine. Various bounds on score-based sampling complexity follow directly: for instance, high-accuracy SLC samplers yield $\widetilde{O}(\sqrt{d}\,\mathrm{polylog}(1/\varepsilon))$ guarantees for accuracy ε in dimension d, whereas randomized midpoint SLC schemes yield $\widetilde{O}(d^{1/3}\,\mathrm{poly}(1/\varepsilon))$ guarantees. When the original distribution itself is SLC, we prove that $K \leq 1 + \log_2(\kappa)$, thereby obtaining the first efficient procedure with logarithmic dependence on the condition number κ; for general distributions, the quantity K depends on the geometry of the score Hessian across the trajectory. Our analysis is direct and simple, involving techniques and insights complementary to those in standard analyses of discretized diffusions.

Keywords: Log-concave sampling; diffusion sampling; score-based methods.

1. Introduction

The problem of drawing samples from a d-dimensional density is a core computational challenge. Efficient samplers are essential for Monte Carlo approximation (e.g., Robert (2004); Rubinstein and Kroese (2008)); exploration of posterior distributions in Bayesian statistics and inverse problems (e.g., Gelman et al. (2013); Brooks et al. (2011)); and generation of images, audio and other structured data in generative AI (e.g., Rombach et al. (2022); Croitoru et al. (2023); Chen et al. (2024)).

Score-based diffusions: In recent years, researchers have demonstrated dramatic advances in sampling through the use of score-based diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021). All these procedures are based on a forward noising process: beginning with a sample X from the target distribution p_{x}, it converts it to some form of "noise", most often a standard Gaussian vector. In continuous time, this forward process can be described by a stochastic differential equation (SDE), and the problem of drawing samples corresponds to simulating the evolution of the reverse-time SDE (Haussmann and Pardoux, 1986; Anderson, 1982) that tracks backward from the noise W to a fresh sample from p_{x}. Stochastic sampling schemes are based on careful discretizations of this reverse-time SDE (e.g., Ho et al. (2020);

Song et al. (2021); Lee et al. (2022); Li et al. (2023); Lee et al. (2023); Chen et al. (2023c,a); Benton et al. (2024)), whereas other sampling schemes make use of an ordinary differential equation (ODE) that describes the backwards evolution (e.g., Song et al. (2021); Benton et al. (2023); Albergo et al. (2023); Chen et al. (2023b); Cai and Li (2025)). In both cases, the forward process is useful because it provides the data needed to estimate the Stein score functions that describe the backwards evolution, using methods such as score matching or Tweedie-based denoising (e.g., Robbins (1956); Miyasawa (1961); Hyvärinen (2005); Vincent (2011)). There are now a wide variety of schemes within the general diffusion framework along with a relatively rich theoretical understanding; see Section 1.1 for further discussion.

Fast sampling for "nice" distributions: In parallel, the past ten years have witnessed tremendous advances in sampling from "nice" distributions, based only on information from the original distribution (and not invoking a diffusion path). For instance, there are highly efficient methods, along with associated theoretical guarantees, for sampling from strongly log-concave (SLC) distributions, as well as those that satisfy a geometric inequality (e.g., log-Sobolev, or Poincaré). Various methods have been studied, including the unadjusted Langevin algorithm (ULA), its Metropolis-corrected variant (MALA), higher-order extensions including Hamiltonian Monte Carlo, as well as proximal schemes (e.g., Dalalyan (2016); Cheng and Bartlett (2018); Durmus and Moulines (2017); Dwivedi et al. (2019); Mou et al. (2022); Chen et al. (2020); Vempala and Wibisono (2022); Chewi et al. (2022); Chen and Gatmiry (2023b)). The modular scheme of this paper allows *any strongly log-concave (SLC) sampler* to be applied, and we obtain a spectrum of rates depending on this choice. Notably, for SLC problems with order-one conditioning, there exist high-accuracy Metropolized samplers with $\sqrt{d}\ \mathrm{polylog}(1/\varepsilon)$ iteration complexity (Altschuler and Chewi, 2023; Chewi, 2025). While the standard Metropolis implementation requires density evaluations, concurrent work (Chen et al., 2026) provides a randomized scheme that achieves the Metropolized complexity using only first-order gradient information.

Our contributions: In this paper, we bring these two lines of research into close contact, in particular by describing and analyzing a simple and modular scheme for score-based sampling. We show how, given the availability of diffused Stein scores, it is possible to "reduce" the problem of sampling from a general target density p_{x} to a sequence of K calls to *any SLC sampler* applied to distributions with condition number at most 2. We give guarantees in which the number of calls K to the SLC sampler is independent of the target accuracy ε, but depends on the geometric structure of the problem. More specifically, we prove two main theorems:

- In Theorem 1, we study the problem of sampling from a strongly log-concave density on $\mathbb{R}^d$ with condition number κ. This is a very well-studied problem, and standard Langevin-type procedures exhibit *linear* scaling with κ. By adapting our modular scheme, we show that the availability of diffused Stein scores reduces the dependence to *logarithmic* in κ. In particular, we do so by constructing a sequence of K SLC problems with $K \leq 1+\log_2(\kappa)$. To the best of our knowledge, this is the first efficient scheme that exhibits such logarithmic dependence. When high-accuracy SLC samplers are applied to the sub-problems, we guarantee an overall iteration complexity of $\widetilde{O}\big(\sqrt{d}\ \log(\kappa)\,\mathrm{polylog}(1/\varepsilon)\big)$.
- In Theorem 2, we study the use of our modular scheme for sampling from a general multi-modal density. We specify a forward trajectory of length K, specified by an *adaptive stepsize sequence*, that ensures a SLC sequence of sub-problems. This leads to concrete bounds on sampling complexity in terms of trajectory length K, and the complexity of solving these sub-problems, and we

exhibit a sampler with iteration complexity $\widetilde{O}(K\sqrt{d}\,\mathrm{polylog}(1/\varepsilon))$. In Corollary 1, we provide a worst-case bound on K in terms of a geometric Lipschitz constant, but suspect that this guarantee can be improved.

We put these results in the context of past work in the next section, as well as in the discussion following the statement of our theorems.

1.1. Related work

There is a long line of work on fast algorithms for sampling from strongly log-concave distributions (SLC), as well as more general families, including those satisfying log-Sobolev and Poincaré inequalities (e.g., Dalalyan (2016); Cheng and Bartlett (2018); Durmus and Moulines (2017); Dwivedi et al. (2019); Mou et al. (2022); Chen et al. (2020); Vempala and Wibisono (2022); Chewi et al. (2022); Chen and Gatmiry (2023b)). The modular reduction in this paper allows any of these procedures to be called as a black box routine. Various algorithms have been analyzed, including the unadjusted Langevin (ULA) algorithm, its Metropolis-adjusted variant (known as MALA), higher-order schemes including randomized midpoint and Hamiltonian Monte Carlo; and samplers based on proximal updates. Suitable variants can achieve iteration complexity proportional to $\sqrt{d}$; of particular relevance to this paper are high-accuracy samplers for SLC distributions with iteration complexity scaling polynomially in $\log(1/\varepsilon)$ (e.g., Dwivedi et al. (2019); Chen et al. (2020); Chen and Gatmiry (2023b); Altschuler and Chewi (2023)). Other schemes, such as randomized midpoint discretizations (Shen and Lee, 2019), sacrifice the $\mathrm{polylog}(1/\varepsilon)$ scaling but reduce dimension dependence to $d^{1/3}$. Our general scheme allows any of these SLC samplers to be applied.

For diffusion-based samplers, there is now a wide range of theoretical results, applying to both stochastic (SDE-based) samplers (e.g., Lee et al. (2022); Li et al. (2023); Lee et al. (2023); Chen et al. (2023c,a); Benton et al. (2024)) as well as (ODE or flow-based) deterministic ones (e.g., Song et al. (2021); Albergo et al. (2023); Benton et al. (2023); Chen et al. (2023b); Cai and Li (2025)). Earlier analyses of the iteration complexity, meaning the number of iterations needed to obtain ε-accurate samples, exhibited polynomial scaling in the dimension. Focusing on the KL divergence, recent results have reduced this dependence to linear in dimension d for both stochastic samplers (Conforti et al., 2023; Benton et al., 2024) and ODE-based samplers (Li et al., 2024), and both classes of methods have polynomial scaling in $(1/\varepsilon)$. By comparison, our modular scheme yields a method with KL iteration complexity scaling as $\sqrt{d}$, and polynomially in $\log(1/\varepsilon)$; see Theorem 2 for details.

The idea of decomposing a "hard" sampling problem into strongly log-concave problems can be understood as an auxiliary variable method (Liu, 2001; Higdon, 1998), and has been exploited for sampling from diffusions (Huang et al., 2024), as well as posterior distributions in neural networks (McDonald and Barron, 2024), and sparse linear regression (Montanari and Wu, 2024). Most closely related to our work is the paper of Huang et al. (2024), who proposed two different algorithms (RTK-ULA and RTK-MALA) that exploit log-concave segments within a diffusion. Among other results, they showed that the RTK-MALA procedure has TV mixing time scaling as $d^2\,\mathrm{polylog}(1/\varepsilon)$, which is the first high-accuracy guarantee for a diffusion-based procedure. Our work extends this idea of extracting SLC sub-problems in a way that is *adaptive*, depending on the problem structure, and *algorithm-independent*, thereby allowing any SLC sampler to be applied to the sub-problems. Consequently, our guarantees inherit the best available bounds for the SLC subroutine for TV, KL and Wasserstein distances. For instance, by using high-accuracy SLC

samplers, we obtain TV and KL mixing time guarantees for multi-modal distributions that scale as $\sqrt{d}\,\mathrm{polylog}(1/\varepsilon)$, and for sampling from a κ-SLC distribution, the adaptivity of our scheme reduces the dependence to logarithmic in κ.

1.2. Tweedie-based structure of diffused scores

We now present the background that underlies our development. Score-based sampling procedures operate on a sequence of random variables that are transformed by a simple linear operation with Gaussian noise. So as to guide the reader, here we lay out the induced Tweedie structure. Given a random vector $U \in \mathbb{R}^d$ and a pair of positive scalars a and b, consider the update

$$V = aU + bW \qquad \text{where } W \sim \mathcal{N}(0, \mathbf{I}) \text{ is standard Gaussian.} \tag{1}$$

This transformation is a form of annealing: the density p_{v} of V will be smoother than the density p_{u} of U, since it is obtained by convolving p_{u} with the Gaussian density.

First-order Tweedie and sampling: The classical Robbins–Tweedie formula (Robbins, 1956; Miyasawa, 1961; Efron, 2011) guarantees that the *Stein score* function $\nabla \log p_{\mathsf{v}}$ can be expressed in terms of $\mathbb{E}[U \mid V = v]$, which allows for denoising-based score estimation. Knowledge of this score function allows us to draw samples from p_{v} using gradient-based sampling procedures. For our scheme, it is essential that the *conditional score* $\nabla_u \log p_{\mathsf{u}|\mathsf{v}}$ also has a simple representation. In particular, we have

$$\nabla_u \log p_{\mathsf{u}|\mathsf{v}}(u \mid v) = \nabla_u \log p_{\mathsf{u}}(u) + \nabla_u \log p_{\mathsf{v}|\mathsf{u}}(v \mid u) \;=\; \nabla_u \log p_{\mathsf{u}}(u) - \frac{a}{b^2}\big(au - v\big), \tag{2}$$

where the second equality follows from the fact that $(V \mid U = u) \sim \mathcal{N}(au, b^2\mathbf{I})$, and the form of the Gaussian density. Consequently, knowledge of the marginal score $\nabla_u \log p_{\mathsf{u}}(u)$ gives us knowledge of the conditional score. In particular, we can then use gradient-based algorithms to draw samples from the backwards conditional distribution $p_{\mathsf{u}|\mathsf{v}}$. In summary, for the 1-step model $V = aU + bW$, knowledge of the marginal score functions enables us to exploit fast algorithms for both (a) generating samples from the marginal distributions p_{u} and p_{v}, and (b) generating samples from the backward conditional $p_{\mathsf{u}|\mathsf{v}}$.

Second-order Tweedie and Hessian structure: Instead of focusing on the score function—that is, the first derivative of the log density—the bulk of our analysis is instead focused on second derivatives. More precisely, still focusing on the update $V = aU + bW$, we introduce the two Hessian matrices

$$\mathbf{H}_{\mathsf{u}}(u) \coloneqq -\nabla^2 \log p_{\mathsf{u}}(u) \quad \text{and} \quad \mathbf{H}_{\mathsf{v}}(v) \coloneqq -\nabla^2 \log p_{\mathsf{v}}(v), \tag{3}$$

associated with the marginal distributions p_{u} and p_{v} over U and V respectively, along with the Hessian $\mathbf{J}_{\mathsf{u}|\mathsf{v}}(u, v) \coloneqq -\nabla_u^2 \log p_{\mathsf{u}|\mathsf{v}}(u \mid v)$ associated with the conditional distribution of $U \mid V$. To be clear, our analysis makes central use of these second-order objects, but the standard sampling schemes that we use to solve sub-problems are still based on first-order information only. The following standard results, proved for completeness in Appendix C, play a central role in our analysis:

$$\textbf{Second-order Tweedie:} \qquad \underbrace{\mathbf{H}_{\mathsf{v}}(v)}_{-\nabla^2 \log p_{\mathsf{v}}(v)} = \frac{1}{b^2}\Big\{\mathbf{I} - \frac{a^2}{b^2}\operatorname{cov}(U \mid V = v)\Big\}, \qquad \text{and} \quad (4a)$$

$$\textbf{Backward conditional Hessian:} \qquad \underbrace{\mathbf{J}_{\mathsf{u}|\mathsf{v}}(u,v)}_{-\nabla_u^2 \log p_{\mathsf{u}|\mathsf{v}}(u|v)} = \mathbf{H}_{\mathsf{u}}(u) + \frac{a^2}{b^2}\mathbf{I}. \qquad (4b)$$

2. Statement of main results

We now turn to the statement of our main results, including exponentially accelerated log-concave sampling (Theorem 1 in Section 2.1), and guarantees for general multi-modal distributions (Theorem 2 in Section 2.2).

2.1. Exponential acceleration for log-concave sampling

We begin by studying the consequences of our modular scheme for the problem of sampling from a smooth and strongly log-concave (SLC) distribution. For a given pair of scalars $0 < m \leq M < \infty$, we say that a twice differentiable density p_{x} is (m, M)-SLC if its negative log Hessian satisfies the sandwich relation $m\mathbf{I} \preceq -\nabla^2 \log p_{\mathsf{x}}(x) \preceq M\mathbf{I}$ for all $x \in \mathbb{R}^d$. The ratio $\kappa := M/m \geq 1$ defines the *condition number* of the problem. While standard procedures exhibit iteration complexity scaling linearly in κ, the main result of this section gives a simple modular scheme that, by exploiting the availability of diffused score functions, provides an *exponential acceleration*, in particular reducing the dependence to $\log \kappa$.

Procedure and a general guarantee: Given a trajectory length K and initial value $Y_0 = X$, consider the Gaussian forward path given by

$$Y_{k+1} = a_k Y_k + \sqrt{1 - a_k^2}\, W_k \qquad \text{for } k = 0, 1, \ldots, K-1, \qquad (5)$$

where $W_k \sim \mathcal{N}(0, \mathbf{I})$, and $a_k \in (0,1)$ are stepsizes to be chosen. The *sub-problems* to be solved in traversing the backward path are (i) sampling the terminal distribution $Y_K \sim p_K$, and (ii) for each $k = K-1, \ldots, 0$, sampling from the backward conditionals $p_{k|k+1}$ of $Y_k \mid Y_{k+1}$. Letting $\mathcal{S}_0^K := \{p_K\} \cup \{p_{k|k+1}\}_{k=0}^{K-1}$ denote the collection of all such sampling sub-problems, the following result guarantees the existence of a "short" trajectory such that all sub-problems $\mathcal{S}_0^K$ are strongly log-concave with condition number at most 2:

Theorem 1 (Logarithmic reduction to SLC black box sampling) *Given any (m, M)-strongly log-concave target density p, there is a forward trajectory (5) of length at most*

$$K + 1 \leq 2 + \log_2\big(M/m\big) \;=\; 2 + \log_2(\kappa), \tag{6a}$$

such that all sampling sub-problems $\mathcal{S}_0^K$ are SLC with condition number at most 2. Consequently, for each distance $D \in \{D_{TV}, D_{KL}, \sqrt{m}\mathcal{W}_2\}$, given any *black-box SLC sampler with query complexity N_{SLC}, we can perform ε-accurate sampling (in distance D) from the target p in at most*

$$T(\varepsilon) = \sum_{k=0}^{K} N_{SLC}\big(\varepsilon/(K+1)\big) \;\leq\; \big\{2 + \log_2(\kappa)\big\}\, N_{SLC}\Big(\frac{\varepsilon}{2 + \log_2(\kappa)}\Big) \qquad \textit{queries.} \tag{6b}$$

In the proof, given in Section 3.1, we give an explicit but adaptively chosen sequence $\{a_k\}_{k=1}^K$ that certifies the theorem's claims. Analyzing the resulting sequence exploits the second-order Tweedie formula (4a) as well as the backward Hessian structure (4b). The key technical challenge is showing the logarithmic scaling (6a). We do so via a combination of the Cramér–Rao bound and the Brascamp–Lieb bound to sandwich the SLC-conditioning (m_k, M_k) of the intermediate problems along the sequence (5); we highlight Lemma 2 as a key result.

2.1.1. SOME SPECIFIC CONSEQUENCES

Theorem 1 is a general reduction that allows for any black-box SLC sampler to be used in solving the sub-problems. Specific consequences of the iteration complexity bound (6b) are easy to extract for a range of samplers; in all cases, we obtain a $\mathrm{polylog}(\kappa)$ dependence, whereas the (d, ε)-dependence changes with the chosen sampler.[1]

Low-accuracy samplers: We begin with the low-accuracy samplers, meaning that the iteration complexity scales with $1/\varepsilon$. Recall from Section 1.2 that access to annealed score functions gives us access to the required gradients for both marginal and backward conditional sampling. Using the ULA updates, it is known that it suffices to take $N_{\mathrm{ULA}}(\varepsilon) \asymp d/\varepsilon^2$ steps to achieve ε-accuracy in KL; combined with our general guarantee, this yields a total iteration complexity of $T(\varepsilon) \asymp \frac{d \log^3(\kappa)}{\varepsilon^2}$. The linear-in-$d$ dimension dependence can be reduced via the use of faster low-accuracy samplers; for instance, schemes based on underdamped Langevin would achieve $\sqrt{d}$-scaling, so an overall complexity $T(\varepsilon) \asymp \frac{\sqrt{d}\, \log^3(\kappa)}{\varepsilon^2}$ in KL distance. As another example, if we use randomized midpoint discretization (Shen and Lee, 2019), then we obtain a total iteration complexity $T(\varepsilon) \asymp \frac{d^{1/3}}{\varepsilon^{2/3}} \log^{5/3}(\kappa)$ in the $\mathcal{W}_2$-distance.

High-accuracy guarantees: Turning to high-accuracy samplers, as previously discussed, there are Metropolized samplers (Altschuler and Chewi, 2023; Chewi, 2025) with KL iteration complexity $N_{\mathtt{fast}}(\varepsilon) \asymp \sqrt{d}\,\mathrm{polylog}(1/\varepsilon)$. Coupled with the bound (6b), this yields an overall iteration

1. Since the sub-problems all have condition number at most 2, we need only focus on their dependence on the pair (d, ε).

complexity scaling as

$$T(\varepsilon) \asymp \sqrt{d}(1+\log(\kappa))\log(\frac{1+\log(\kappa)}{\varepsilon}). \tag{7}$$

To be clear, classical Metropolized samplers require access to both the score function (first derivative), and log density values; in certain cases, log densities can be estimated (e.g., Guth et al. (2025)), but this is not always the case. In their work on RTK-MALA, Huang et al. (2024) tackled this issue, and showed how to approximate Metropolis sampling using only gradient information, but did not achieve $N_{\texttt{fast}}$ complexity. In concurrent work, Chen et al. (2026) analyze a "Bernoulli-factory" randomized approximation, and show how it can be leveraged to implement a purely gradient-based sampler that achieves iteration complexity $N_{\texttt{fast}}$. Since the reduction in Theorem 1 allows *for any SLC sampler*, we can adopt their particular scheme; doing so leads to a purely score-based sampler that achieves the guarantee (7). This bootstrapping via our scheme yields a refinement of their initial guarantee, sharpening the linear in κ scaling from their guarantee to logarithmic in κ.

2.1.2. ROBUSTNESS TO SCORE ERRORS

Letting p_k denote the marginal distribution of Y_k at round k, we have stated Theorem 1 for samplers that operate using the exact score functions $s_k(x) := \nabla_x \log p_k(x)$. In practice, these exact score functions are not available, but are instead estimated using samples (e.g., via denoising using the first-order Tweedie representation). Here we describe how our guarantees remain robust when the procedure is implemented using estimates $\widehat{s}_k$ of the true score functions. In addition to the computational error tracked by Theorem 1, this guarantee involves additional error in terms of the score differences $\widehat{s}_k - s_k$.

For concreteness, we focus on the TV error, and state a general stability result for score errors, one that can be applied both to Theorem 1, as well as to Theorem 2, which is stated in Section 2.2. At time k, let $\mathcal{Q}_k(\cdot \mid y)$ denote the family of backward transition kernels defined by the score estimate $\widehat{s}_k$, and let $\widetilde{\mathcal{Q}}_k(\cdot \mid y)$ be the backwards transition defined by a sampler used to implement this backward transition. Suppose that: (a) the backward sampler is ε_k-accurate, meaning that $D_{\mathrm{TV}}(\widetilde{\mathcal{Q}}_k(\cdot \mid y), \mathcal{Q}_k(\cdot \mid y)) \le \varepsilon_k$ for each y; and (b) the true backward transition $\mathcal{P}_k(\cdot \mid y)$ has density that is α_k-strongly log-concave.

Lemma 1 (Robustness to score errors) *Under the stated conditions, for each backward step $k = 0, 1, \ldots, K-1$, the marginal $\widetilde{q}_k$ satisfies the TV norm recursion*

$$D_{TV}(\widetilde{q}_k, p_k) \le \varepsilon_k + \frac{\|\widehat{s}_k - s_k\|_{L^2(\bar{q}_k)}}{2\sqrt{\alpha_k}} + D_{TV}(\widetilde{q}_{k+1}, p_{k+1}) \quad \textit{where } \bar{q}_k = \widetilde{q}_{k+1}\mathcal{Q}_k. \tag{8}$$

See Appendix F.1 for the proof of this claim.

Proving the TV guarantee in Theorem 1 exploits the bound (8) with no score error ($\widehat{s}_\ell = s_\ell$, and the settings $\varepsilon_\ell = \varepsilon/(K+1)$, so that the total error scales as $\sum_{\ell=0}^{K} \varepsilon_\ell = \varepsilon$. Since Lemma 1 allows

for score error,[2] we can prove a generalized result that involves the total error

$$\sum_{\ell=0}^{K} \varepsilon_\ell + \sum_{\ell=0}^{K} \frac{\|\widehat{s}_\ell - s_\ell\|_{L^2(\bar{q}_\ell)}}{2\sqrt{\alpha_\ell}}, \tag{9}$$

where, in the setting of Theorem 1, the strong log-concavity parameter is given by $\alpha_\ell = 2 + 2^{-\ell}(\kappa - 1)$ in terms of the condition number κ.

2.2. Multi-modal setting

We now turn to sampling from a general multi-modal distribution. In this case, given some $\delta > 0$ that is user-specified, our goal is to draw samples from the random vector $Z := \widetilde{X} + \delta W_0$, where $W_0 \sim \mathcal{N}(0, \mathbf{I})$. As in diffusion analyses, we refer to the parameter δ as the early stopping error. With this parameter fixed, our goal is to develop algorithms that produce ε-accurate samples from the distribution p_Z of Z. In order to do so, it is convenient to work with the rescaled random vector $Y_1 = \frac{Z}{\sqrt{2}\,\delta}$. This rescaling has no effect on the distances D_{KL} and D_{TV}.

2.2.1. Adaptive scheme and its analysis

Having reduced our problem to sampling from $p_1 \equiv p_{Y_1}$, we note that Y_1 can be written as $Y_1 := \frac{1}{\sqrt{2}} X + \frac{1}{\sqrt{2}} W_0$, where $X := \widetilde{X}/\delta$ is a rescaled version of the original variable $\widetilde{X}$. In our analysis, we state conditions directly on X. Given this Y_1, we then generate the forward sequence

$$Y_{k+1} = a_k Y_k + \sqrt{1 - a_k^2} W_k, \qquad \text{for stepsizes } a_k \in (0,1) \text{ to be chosen adaptively.} \tag{10a}$$

Our adaptive choice of stepsizes depends on the sequence

$$B_k \;:=\; \sup_{y \in \mathbb{R}^d} \bigl\| \operatorname{cov}(X \mid Y_k = y) \bigr\|_{\mathrm{op}}, \tag{10b}$$

defined for each $k = 1, 2, \ldots$ in the forward trajectory. We construct the forward sequence with a stepsize sequence $\{a_k\}_{k \geq 0}$ based on the initialization $a_0 = 1/\sqrt{2}$, and for $k = 1, 2, \ldots$, the adaptive updates

$$\lambda_k \;:=\; 4\, B_k \prod_{\ell=0}^{k-1} a_\ell^2 \quad \text{and} \quad a_k^2 = \frac{2\lambda_k + 2}{2\lambda_k + 3}. \tag{10c}$$

As before, for a trajectory length K, we let $\mathcal{S}_0^K := \{p_K\} \cup \{p_{k|k+1}\}_{k=0}^{K-1}$ be the collection of all sampling sub-problems: the terminal distribution p_K, and each of the backward conditional distributions.

2. The lemma applies to the backward transitions, and we can also prove an analogous stability result for the terminal sub-problem p_K.

Theorem 2 (SLC reduction for multi-modal case) *Given the adaptive stepsize choice (10c), consider any trajectory of length K such that $\prod_{\ell=0}^{K-1} a_\ell^2 \leq \frac{1}{8B_K}$. Then we are guaranteed that all of the sampling sub-problems $\mathcal{S}_0^K$ are strongly log-concave (SLC) with condition number at most 4. Consequently, for any distance $D \in \{D_{TV}, D_{KL}\}$, by using* any *SLC sampler with query complexity N_{SLC}, we can obtain ε-accurate samples using*

$$T(\varepsilon) = \sum_{k=0}^{K} N_{SLC}\big(\varepsilon/(K+1)\big) \qquad \textit{queries.} \tag{11}$$

See Section 3.2 for the proof. Here the main technical challenge is showing how the adaptive stepsize choice (10c) suffices to ensure the SLC property of the full collection $\mathcal{S}_0^K$ of sub-problems. As with Theorem 1, this reduction allows for any SLC sampler, and so has consequences for both low- and high-accuracy schemes.

Low-accuracy guarantees: If we use ULA as our base sampler, then we achieve an overall guarantee of $T(\varepsilon) \asymp K^3 \frac{d}{\varepsilon^2}$; this is comparable to the RTK-ULA guarantee in the paper (Huang et al., 2024). On the other hand, we could also use underdamped Langevin to achieve the scaling $T(\varepsilon) \asymp K^3 \frac{\sqrt{d}}{\varepsilon^2}$. If the trajectory length K is independent of dimension, then the end-to-end procedure exhibits $\sqrt{d}$-dependence, as opposed to the linear-in-d scaling that arises from optimal KL bounds for ULA-based discretizations of diffusions (Conforti et al., 2023; Benton et al., 2024). We can also use a more sophisticated Langevin discretization to go below the $\sqrt{d}$ barrier; for instance, if we use a randomized midpoint scheme (RMC), then known TV guarantees for RMC-Langevin (Gupta et al., 2025) give us an overall complexity of $T(\varepsilon) \asymp K^{7/3} d^{5/12}/\varepsilon^{4/3}$.

High-accuracy results: As discussed following Theorem 1, we can achieve ε-accurate sampling in KL for each sub-problem with complexity $N_{\texttt{fast}}(\varepsilon) \asymp \sqrt{d}\,\mathrm{polylog}(1/\varepsilon)$, either by using Metropolized schemes (Altschuler and Chewi, 2023) that use gradients and density values, or by applying the gradient-only SLC-Metropolis sampler (Chen et al., 2026), as discussed following Theorem 1. In either case, when combined with the guarantee (11), our reduction yields an overall iteration complexity of

$$T(\varepsilon) = \widetilde{O}\Big(K\sqrt{d}\ \mathrm{polylog}\,\big(\frac{K}{\varepsilon}\big)\Big). \tag{12}$$

In general, the trajectory length K is a function of the full sequence B_k from equation (10b); in the next section, we bound K under a worst-case assumption.

2.2.2. TRAJECTORY LENGTH UNDER WORST-CASE ASSUMPTIONS

We now turn to an analysis of the trajectory length K under a particular worst-case assumption, namely that the only knowledge about the sequence $\{B_k\}_{k\geq 1}$ is the existence of some $B_{\max} < \infty$ such that

$$B_k \coloneqq \sup_{y\in\mathbb{R}^d} \|\operatorname{cov}(X \mid Y_k = y)\|_{\mathrm{op}} \leq B_{\max} \qquad \text{for all } k = 1, 2, \ldots, \tag{13}$$

but no further structure is given. By comparison to the second-order Tweedie formula (4a), we see that this type of uniform bound corresponds to assuming the score function has a bounded Lipschitz constant. By careful analysis of the adaptive sequence (10c), we obtain the following guarantee:

Corollary 1 *Under the uniform bound (13), the guarantees of Theorem 2 apply with trajectory length at most* $K \leq 7(1 + B_{\max})$.

See Appendix B.2 for the proof.

Substituting the upper bound on K from Corollary 1 into our earlier bound (12), we find a worst-case iteration complexity scaling as

$$T_{worst}(\varepsilon) \asymp B_{\max}\sqrt{d}\,\mathrm{polylog}(1/\varepsilon), \tag{14}$$

where $B_{\max}$ is the Lipschitz constant. Comparing to past results (Chen et al., 2023a,c) on discretized diffusions, these bounds exhibit a quadratic dependence on a Lipschitz condition, and d/ε^2 dependence on (d, ε), as opposed to $B_{\max}\sqrt{d}\,\mathrm{polylog}(1/\varepsilon)$ in the guarantee (14). The $\mathrm{polylog}(1/\varepsilon)$-scaling for diffusion models first appeared in the RTK-MALA guarantee of Huang et al. (2024), albeit with d^2-scaling, and applicable only to TV distance. In concurrent work, Chen et al. (2026) impose a related Lipschitz assumption, and obtain guarantees with the same $\sqrt{d}\ \mathrm{polylog}(1/\varepsilon)$ scaling using a diffusion approach.

While Corollary 1 leads to concrete results, we note that imposing the global bound via $B_{\max}$ is a crude and worst-case assumption. For future work, it would be interesting to use our framework to obtain bounds on K depending on a milder geometric quantity; we strongly suspect this should be possible, since the covariance sequence $\{\mathrm{cov}(X \mid Y_k = y)\}_{k\geq 1}$ and associated operator norms $\{B_k\}_{k\geq 1}$ evolve in a very structured way.

2.2.3. SOME OPEN QUESTIONS AND EXTENSIONS

Our scheme and results suggest a variety of other open questions and extensions. Let us comment on a few of them here.

Best of all worlds? It is interesting to make further explicit comparisons in a specific setting, which we refer to as the *bounded* (R, σ)*-model*. Suppose that we wish to sample $Y_1 = \widetilde{X} + \sigma W$, where $\|\widetilde{X}\|_2 \leq R$. In this case, the worst-case bound (13) holds with $B_{\max} = R^2/\sigma^2$, and our approach yields the overall iteration complexity $T_{R,\sigma}(\varepsilon) \asymp (R/\sigma)^2\sqrt{d}\,\mathrm{polylog}(1/\varepsilon)$. State-of-the-art results on discretized diffusions (Conforti et al., 2023; Benton et al., 2024; Li et al., 2024) provide iteration complexities with d/ε^2 scaling, inferior to the $\sqrt{d}\log(1/\varepsilon)$ scaling in $T_{R,\sigma}(\varepsilon)$, but with a milder poly-logarithmic dependence on R/σ. We are thus led to a natural open question: is it possible to obtain a "best-of-both-worlds" guarantee: i.e., with poly-logarithmic dependence on both (R/σ) and $(1/\varepsilon)$?

Reductions beyond the SLC class: Theorem 1 and Theorem 2 both exploit trajectories of strongly log-concave (SLC) problems, which allows us to exploit fast SLC sampling algorithms. However, there are many other classes of distributions for which fast samplers are available, including those satisfying geometric relations such as log-Sobolev (LSI) or Poincaré inequalities. For instance, the class of LSI distributions is a strict superset of the SLC class, allowing for considerable multimodality depending on the LSI constant. There are various fast samplers available for LSI distributions (e.g., Vempala and Wibisono (2022); Chen and Gatmiry (2023a); Altschuler and Chewi

(2023)), so that the scope of our scheme could be enlarged substantially by instead reducing to a sequence of well-conditioned LSI problems. One natural open question is whether (for 1-Lipschitz problems) one can obtain sampling guarantees with logarithmic dependence on the LSI constant, thereby generalizing the logarithmic dependence on condition number κ in Theorem 1.

3. Proof sketches

We now provide sketches of our two main results: Theorem 1 in Section 3.1 and Theorem 2 in Section 3.2.

3.1. Proof sketch of Theorem 1

At a high level, the proof consists of three main steps:

(1) Rescaling: We apply a rescaling argument to reduce the problem to a sampling problem that is $(1, M/m)$-strongly log-concave; see Appendix A.1 for this argument.

(2) Adaptive stepsizes: We prescribe an adaptive choice of stepsizes $\{a_k\}_{k=0}^{K-1}$, and analyze the evolution of marginal distributions $p_k \equiv p_{\mathsf{y}_k}$ of the forward process, and the conditional distributions $p_{k|k+1} \equiv p_{\mathsf{y}_k|\mathsf{y}_{k+1}}$ of the backward process; and

(3) Stability analysis: We control error propagation throughout the entire backward process; see Section 3.1.3 for details of this step.

In the main body, we provide more detail on the adaptive stepsize and trajectory analysis in Step 2, and the stability analysis in Step 3.

3.1.1. FORWARD RECURSION WITH ADAPTIVE STEPSIZE CHOICE

Introduce the shorthand $M_0 = M/m$ for the smoothness constant of the target distribution after the rescaling step. We run the forward recursion (5) with the following choice of stepsizes $\{a_k\}_{k\geq 1}$. Given the initial value M_0, we initialize with $a_0^2 = M_0/(1+M_0)$, and then evolve the pair (M_k, a_k) according to the recursion

$$M_{k+1} \overset{(i)}{=} \frac{M_k}{M_k(1-a_k^2)+a_k^2}, \quad \text{and} \quad a_k^2 \overset{(ii)}{=} \frac{M_k}{1+M_k} \qquad \text{at each round.} \tag{15}$$

Given these stepsize choices, the key technical steps in our proof are to establish that after $K \leq 1 + \log_2 M_0$ rounds, the Hessian $\mathbf{H}_{\mathsf{K}}$ satisfies

$$\mathbf{I} \preceq \mathbf{H}_{\mathsf{K}}(y) \preceq 2\mathbf{I} \qquad \text{uniformly in } y \in \mathbb{R}^d, \tag{16a}$$

and at each round $k \in [K-1]$, the Hessian $\mathbf{J}_k$ of the backward conditional $p_{k|k+1}$ satisfies the sandwich

$$1 + M_k \preceq \mathbf{J}_k(y) \preceq 2M_k \qquad \text{uniformly in } y \in \mathbb{R}^d, \tag{16b}$$

Note that the sandwich relation (16a) ensures that the terminal distribution p_K is SLC with condition number at most 2, whereas the sandwich relations (16b) guarantee that each of the backward conditionals $p_{k|k+1}$ is SLC with condition number at most 2. Together, these results certify that all the sampling sub-problems $\mathcal{S}_0^K$ satisfy the claims of Theorem 1.

3.1.2. PROPAGATION OF SPECTRAL CONTROL

Proving the claims (16a) and (16b) hinges on the following auxiliary result. Each step in the sequence (5) is of the generic form $V = aU + bW$, where $W \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian. Recalling the Hessian matrices from equation (3), the following result shows how control of the spectrum of $\mathbf{H}_{\mathsf{u}}(u)$ yields spectral control of $\mathbf{H}_{\mathsf{v}}(v)$.

Lemma 2 (Propagation of spectral control)
(a) Suppose there is some $m \geq 0$ such that $\mathbf{H}_{\mathsf{u}}(u) \succeq m\mathbf{I}$ uniformly in $u \in \mathbb{R}^d$. Then we have

$$\mathbf{H}_{\mathsf{v}}(v) \succeq \frac{m}{mb^2 + a^2}\mathbf{I} \qquad \text{uniformly in } v \in \mathbb{R}^d. \tag{17a}$$

(b) Suppose there is some $M < \infty$ such that $\mathbf{H}_{\mathsf{u}}(u) \preceq M\mathbf{I}$ uniformly in $u \in \mathbb{R}^d$. Then we have

$$\mathbf{H}_{\mathsf{v}}(v) \preceq \frac{M}{Mb^2 + a^2}\mathbf{I} \qquad \text{uniformly in } v \in \mathbb{R}^d. \tag{17b}$$

See Appendix A.3 for the proof. Part (a) requires strong log-concavity (SLC), and exploits the Brascamp–Lieb inequality applicable in this case. Part (b) applies in general, and makes use of the Cramér-Rao lower bound on location estimation.

3.1.3. BACKWARDS ERROR PROPAGATION

Finally, we provide a lemma that allows us to control propagation of errors in moving along the backward path. On one hand, we have the Markov kernel

$$\mathcal{P}_k(r)(\cdot) := \int_{\mathbb{R}^d} p_{k|k+1}(\cdot \mid y) r(y) dy \tag{18a}$$

that defines the backwards evolution of the true marginals $p_{k+1} \to p_k$. Our backwards sampler defines a second kernel $\mathcal{Q}_k$ that underlies the backwards evolution $q_{k+1} \to q_k$ of the algorithm's marginals. By our set-up, with a sufficient number of iterations, we can assume that our backwards sampler is δ_k-accurate uniformly in its inputs, meaning that

$$D(\mathcal{Q}_k(e_y), \mathcal{P}_k(e_y)) \leq \delta_k \qquad \text{for all } y \in \mathbb{R}^d, \tag{18b}$$

Lemma 3 (Error propagation in backwards kernel $\mathcal{P}_k$) *Under the condition (18b), for any of the distances $D \in \{D_{TV}, D_{KL}, \mathcal{W}_2\}$, we have*

$$\underbrace{D(q_k, p_k)}_{\text{Round-}k\text{ error}} \leq \delta_k + \underbrace{D(q_{k+1}, p_{k+1})}_{\text{Round-}(k+1)\text{ error}}. \tag{19}$$

See Appendix A.4 for the proof. For the TV distance D_{TV}, the bound (19) is a special case of Lemma 1 without score error. Proof of the bound (19) for the KL divergence D_{KL} is an easy consequence of the data processing inequality; the corresponding proof for Wasserstein is more involved (cf. Lemma 8 for details).

3.1.4. COMBINING THE PIECES

Let us now combine the pieces so as to complete the proof of Theorem 1. Recall that we say that a problem is 2-SLC if it is strongly log-concave (SLC) with condition number at most 2. Let $(p_K, p_{K-1}, \ldots, p_1, p_0)$ denote the true sequence of marginal distributions, and let $(q_K, q_{K-1}, \ldots, q_1, q_0)$ denote the sequence of distributions generated by our approximate sampling algorithm. We assume that:

- At the terminal stage K, we generate samples from a distribution q_K that is $\varepsilon/(K+1)$-close to p_K. From the guarantee (16a), this terminal stage problem is 2-SLC-controlled, so that doing so using our black-box SLC procedure requires $N_{\text{SLC}}(\varepsilon/(K+1))$ calls.
- In moving backward from $(k+1)$ to k, we run enough iterations of the sampler so that the bound (18b) holds with $\delta_k = \varepsilon/(K+1)$. From the guarantee (16b), each backward conditional is 2-SLC-controlled, so that doing so requires $N_{\text{SLC}}(\varepsilon/(K+1))$ calls.

By construction, the total number of calls required is $\sum_{k=0}^{K} N_{\text{SLC}}(\varepsilon/(K+1))$, matching the claim (6b). For a given distance $D \in \{D_{\text{TV}}, D_{\text{KL}}, \mathcal{W}_2\}$, let us compute the error $D(q_0, p_0)$ at the initial stage. By iterating the bound (19) at each round with $\delta_k = \varepsilon/(K+1)$ and using $D(q_K, p_K) \leq \varepsilon/(K+1)$, we have $D(q_0, p_0) \;\leq\; \varepsilon/(K+1) + D(q_1, p_1) \leq \frac{1}{K+1} \sum_{k=0}^{K} \varepsilon = \varepsilon$, as claimed.

3.2. Proofs of Theorem 2 and Corollary 1

We now turn to the proof sketch for Theorem 2, along with the related Corollary 1. Here we sketch the result for the KL divergence; it can be used to upper bound D_{TV}^2 via Pinsker's inequality, but a sharper TV guarantee can be obtained by direct analysis. This proof does not require the rescaling step, and we can re-use the backward error propagation for D_{KL} given in Lemma 3. (However, the Wasserstein error propagation there does not apply to the multi-modal case, as the Wasserstein proof in Lemma 3 exploits the SLC structure.) Consequently, the new technical effort required is analyzing the evolution of the forward and backward processes induced by the adaptive stepsize sequence (10c).

Recall the Hessian matrices from equation (3). The following result shows how control of the spectrum of $\mathbf{H}_{\mathsf{u}}(u)$ yields spectral control of $\mathbf{H}_{\mathsf{v}}(v)$. The following lemma controls the Hessian structure of the trajectory, both in terms of the marginal Hessians of the forward process, and the conditional Hessians of the backward trajectory. It allows us to show that all relevant sampling sub-problems satisfy the requisite SLC conditions.

Lemma 4 (Trajectory control for multi-modal case) *Given the adaptive stepsize choice (10c), the following properties hold:*

(a) At each round $k \in [K]$, we have

$$-\lambda_k \mathbf{I} \;\preceq\; (1-\lambda_k)\mathbf{I} \;\preceq\; \mathbf{H}_k(y) \;\preceq\; 2\mathbf{I}. \tag{20a}$$

(b) For each $k \in [K-1]$, the backwards conditional distribution $p_{k|k+1}$ has a Hessian $\mathbf{J}_k(y_k)$ that satisfies the sandwich

$$(\lambda_k + 2)\,\mathbf{I} \;\preceq\; \mathbf{J}_k(y) \;\preceq\; 2\,(\lambda_k + 2)\,\mathbf{I} \qquad \text{for all } y \in \mathbb{R}^d. \tag{20b}$$

See Section B.1 for the proof of this claim.

Completing the proof: We now sketch how to complete the proof of Theorem 2. From the theorem set-up, recall that the trajectory length K is chosen to ensure that $\prod_{k=0}^{K-1} a_k^2 \leq \frac{1}{8B_K}$. From the definition of the sequence λ_k, this bound implies that $\lambda_K \leq 1/2$. Thus, from part (a) of Lemma 4, we see that the terminal Hessian $\mathbf{H}_{\mathsf{K}}$ satisfies the sandwich bound $\frac{1}{2}\mathbf{I} \preceq \mathbf{H}_{\mathsf{K}}(y) \preceq 2\mathbf{I}$ uniformly in $y \in \mathbb{R}^d$, so that it is SLC with condition number at most 4, as claimed. From part (b) of Lemma 4, we see that the backward Hessians $\mathbf{J}_k$ have condition number at most 2. Thus, we have established that the adaptive stepsize sequence constructs a sequence of K sub-problems that are each SLC with condition number at most 4. Finally, the sampling guarantee (11) given in Theorem 2 follows from the same proof template used in Section 3.1.4.

4. Discussion

In this paper, we have described a modular approach to sampling from a given target distribution p_{x} based on the availability of annealed score functions. It can be understood as a form of "divide-and-conquer", showing how the original sampling problem is reducible to a K-length sequence of sub-problems, each defined by a well-conditioned and strongly log-concave (SLC) distribution. Using this reduction, we proved novel results both for uni-modal and multi-modal distributions. For sampling from a SLC distribution (Theorem 1), we exhibited efficient methods scaling as $\log \kappa$ with the condition number κ, and dimension dependence ranging in $\{d^{1/3}, \sqrt{d}, d\}$, depending on the SLC sampler used to solve the sub-problems. Using high-accuracy SLC samplers leads to $\sqrt{d}\,\mathrm{polylog}(1/\varepsilon)$ scaling. We established similar guarantees for general multi-modal distributions in Theorem 2; here the $\sqrt{d}\,\mathrm{polylog}(1/\varepsilon)$ scaling obtained by high-accuracy samplers improves upon the best known results for diffusion samplers prior to our work.

Our work leaves open various questions and potential extensions. In Corollary 1, we provided a worst-case bound on the trajectory length K for multi-modal problems; this crude analysis does not exploit any structure of the process, and we suspect that it could be sharpened. At a broader level, we focused on adaptive schemes that generate sequences of SLC sub-problems with constant conditioning, but our general reduction scheme is not limited to this choice. It could be enriched by instead reducing to richer classes of distributions for which fast schemes are available (e.g., those satisfying a log-Sobolev inequality (LSI)), since being forced to ensure SLC on the backward path can be restrictive. Finally, it is likely that our current bounds on the trajectory length K for the multi-modal case can be improved by exploiting structure of the covariance process.

Acknowledgments

This work was partially supported by a Guggenheim Fellowship, grant NSF DMS-2311072, and the Ford Professorship at MIT. We thank Curtis McDonald and Sasha Rakhlin for bringing related work to our attention, as well as the anonymous reviewers for helpful comments.

References

M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

J. M. Altschuler and S. Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. Technical Report arXiv:2302.10249, arXiv, February 2023. URL https://arxiv.org/abs/2302.10249. Preprint.

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

J. Benton, G. Deligiannidis, and A. Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations (ICLR) 2024*, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

S. Brooks, A. Gelman, G. L. Jones, and X. L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

C. Cai and G. Li. Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*, 2025.

F. Chen, S. Chewi, C. Daskalakis, and A. Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions. *arXiv preprint arXiv:2602.01338*, 2026.

H. Chen, H. Lee, and J. Lu. Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions. In *International Conference on Machine Learning*, 2023a.

M. Chen, S. Mei, J. Fan, and M. Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024. URL https://arxiv.org/abs/2404.07771.

S. Chen, S. Chewi, H. Lee, Y. Li, J. Lu, and A. Salim. The probability flow ODE is provably fast. *arXiv preprint arXiv:2305.11798*, 2023b.

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023c.

Y. Chen and K. Gatmiry. A simple proof of the mixing of Metropolis-adjusted Langevin algorithm under smoothness and isoperimetry. Technical report, June 2023a. arxiv:2304.04095v2.

Y. Chen and K. Gatmiry. A simple proof of the mixing of metropolis-adjusted langevin algorithm under smoothness and isoperimetry. Technical Report arXiv:2304.04095v2, arXiv, jun 2023b.

Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21(92):1–71, 2020.

X. Cheng and P. L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83, pages 186–211, may 2018.

S. Chewi. Log-concave sampling. Technical report, 2025. Book in preparation.

S. Chewi, M. A. Erdogdu, M. B. Li, R. Shen, and M. Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. In *Proceedings of Thirty Fifth Conference on Learning Theory*, 2022.

G. Conforti, A. Durmus, and M. G. Silveri. Score diffusion models without early stopping: finite Fisher information is all you need. *arXiv preprint arXiv:2308.12240*, aug 2023.

F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. doi: 10.1109/TPAMI.2023.3261988.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Jour. Royal. Stat. Soc. B*, 2016.

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

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017. doi: 10.1214/16-AAP1238.

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

A. Gelman, J. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. K. Salomon. Bayesian data analysis. *CRC Press*, 2013.

N. Gozlan and C. Léonard. Transport inequalities. a survey. *Markov Processes and Related Fields*, 16(4):635–736, 2010.

S. Gupta, L. Cai, and S. Chen. Faster diffusion sampling with randomized midpoints: Sequential and parallel. In *International Conference on Learning Representations (ICLR) 2025 Poster*, 2025. URL https://openreview.net/forum?id=MT3aOfXIbY.

F. Guth, Z. Kadkhodaie, and E. P. Simoncelli. Learning normalized image densities via dual score matching. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=wtYcS4kxpF.

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986.

D. M. Higdon. Auxiliary variable methods for markov chain monte carlo with applications. *Journal of the American Statistical Association*, 93(442):585–595, 1998.

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

X. Huang, D. Zou, H. Dong, Y. Zhang, Y.-A. Ma, and T. Zhang. Reverse transition kernel: A flexible framework to accelerate diffusion inference. *arXiv preprint arXiv:2405.16387*, 2024.

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

H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022.

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985, 2023.

G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models. *arXiv preprint arXiv:2306.09251*, 2023.

G. Li, Y. Wei, Y. Chi, and Y. Chen. A sharp convergence theory for the probability flow ODEs of diffusion models. Technical Report arXiv:2304.04095, arXiv, August 2024.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2001. ISBN 978-0-387-95230-7.

C. McDonald and A. R. Barron. Rapid Bayesian computation and estimation for neural networks via log-concave coupling. *arXiv preprint arXiv:2411.17667v3*, 2024. doi: 10.48550/arXiv.2411.17667.

K. Miyasawa. An empirical Bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38:181–188, 1961. pp. 1–2.

A. Montanari and Y. Wu. Provably efficient posterior sampling for sparse linear regression via measure decomposition. *arXiv preprint arXiv:2406.19550*, 2024. doi: 10.48550/arXiv.2406.19550.

W. Mou, N. Flammarion, M. J. Wainwright, and P. L. Bartlett. An efficient sampling algorithm for non-smooth composite potentials. *Journal of Machine Learning Research*, 23:1–50, 2022.

H. E. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163. University of California Press, 1956.

C. P. Robert. *Monte Carlo methods*. Wiley Online Library, 2004.

R. Rombach, E. Blattmann, S. L. Dhariwal, A. M. D. M. L., and P. E. S. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10694, 2022.

R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo Method*. John Wiley and Sons, Hoboken, NJ, 2nd edition, 2008.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. *arXiv preprint arXiv:1909.05503*, 2019.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11895–11907, 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

S. S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. Technical Report arXiv:1903.08568v4, arXiv, March 2022.

P. Vincent. Connection between score matching and denoising autoencoders. *Neural Networks*, 24(8):971–978, 2011. doi: 10.1016/j.neunet.2011.03.003.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1:1—305, December 2008.

Appendix A. Auxiliary results for Theorem 1

In this appendix, we collect together the proofs of various auxiliary results related to Theorem 1.

A.1. Rescaling argument

It is convenient, as a first step, to reduce the problem of sampling from an (m, M)-conditioned distribution to an equivalent one that is $(1, M/m)$-conditioned. In order to do so, letting $X \sim p$ be the original distribution, we define the rescaled vector $Y = \sqrt{m}X$. If the original distribution p is (m, M)-conditioned, then the rescaled vector Y has a distribution that is $(1, M/m)$-conditioned.

Now suppose that we can generate samples of the random vector $\widetilde{Y}$ whose distribution is ε-close to that of Y in a given distance D. We then define $\widetilde{X} = \widetilde{Y}/\sqrt{m}$, and consider the quality of the $\widetilde{X}$ samples as approximations to the original X. The argument is slightly different, depending on whether D is the TV or KL distance, as opposed to the Wasserstein distance.

When D is the KL distance, it is invariant to this linear transformation, so that we have

$$D_{\mathrm{KL}}(p_{\widetilde{X}}, p_X) = D_{\mathrm{KL}}(p_{\widetilde{Y}}, p_Y) \; = \; \varepsilon.$$

Consequently, the samples $\widetilde{X}$ are ε-close to X in KL distance. A similar argument applies for the TV distance.

When D is the Wasserstein distance, we use $\mathcal{W}_2(X, Y)$ to denote the Wasserstein distance between p_X and p_Y. In this case, we have

$$\sqrt{m}\mathcal{W}_2(\widetilde{X}, X) = \mathcal{W}_2\big(\sqrt{m}\widetilde{X}, \sqrt{m}X\big) \; = \; \mathcal{W}_2(\widetilde{Y}, Y) \; = \; \varepsilon$$

so that the samples are ε-close in the rescaled Wasserstein norm $\sqrt{m}\mathcal{W}_2$. (Note that the theorem statement involves this rescaled Wasserstein distance.)

Thus, for the remainder of the proof, we study the rescaled problem with initial values $m_0 = 1$ and $M_0 = \frac{M}{m}$. We prove bounds in terms of M_0, and then recall this transformation.

A.2. Analysis of forward-backward trajectory

The following lemma summarizes some key properties of the forward/backward trajectory:

Lemma 5 (Properties of forward/backward trajectory)

(a) We have the Hessian sandwich

$$\mathbf{I} \preceq \mathbf{H}_{\mathsf{k}}(y) \preceq M_k \mathbf{I} \qquad \text{at each round } k \in [K]. \tag{21a}$$

(b) After $K \leq 1 + \log_2 M_0$ rounds, the Hessian $\mathbf{H}_{\mathsf{K}}$ satisfies

$$\mathbf{I} \preceq \mathbf{H}_{\mathsf{K}}(y) \preceq 2\mathbf{I} \qquad \text{uniformly in } y \in \mathbb{R}^d, \tag{21b}$$

so that $\sup_{y \in \mathbb{R}^d} \operatorname{cond}(\mathbf{H}_{\mathsf{K}}(y)) \leq 2$.

(c) At each round $k \in [K-1]$, the Hessian $\mathbf{J}_k$ of the backwards conditional $p_{k|k+1}$ satisfies the sandwich

$$(1 + M_k)\mathbf{I} \preceq \mathbf{J}_k(y) \preceq 2M_k \mathbf{I} \qquad \text{uniformly in } y \in \mathbb{R}^d, \tag{21c}$$

so that $\sup_{y \in \mathbb{R}^d} \operatorname{cond}(\mathbf{J}_k(y)) \leq 2$.

Proof We first prove the Hessian sandwich (21a), in particular via induction on the iteration number k. Beginning with the base case $k = 0$, the claim holds because the original problem is $(1, M_0)$-conditioned by construction. Suppose that the sandwich (21a) holds at step k, and let us prove that it holds at step $k + 1$.

Beginning with the lower bound, we apply the bound (17a) from Lemma 2 with $m = 1, U = Y_k$ and $V = Y_{k+1}$, and $a^2 = a_k^2$ and $b^2 = 1 - a_k^2$. Doing so yields

$$\mathbf{H}_{\mathsf{k}+1}(y_{k+1}) \succeq \frac{1}{a_k^2 + (1 - a_k^2)} \mathbf{I} \;=\; \mathbf{I}$$

as required. Similarly, we apply the upper bound (17b) with $M = M_k$ and the same choices as above. Doing so yields

$$\mathbf{H}_{\mathsf{k}+1}(y_{k+1}) \preceq \frac{M_k}{M_k(1 - a_k^2) + a_k^2} \mathbf{I} \;=\; M_{k+1}\mathbf{I},$$

using the definition (15) of the (M_k, a_k) recursion.

Now we establish the claim (21b). Based on the Hessian sandwich (21a) from part (a), the proof of this claim amounts to showing that $M_K \leq 2$ after at most $K \leq 1 + \log_2 M_0$ rounds. We claim that the sequence M_k evolves in a very simple way—namely as

$$M_{k+1} = 1 + \tfrac{1}{2}\big(M_k - 1\big). \tag{22}$$

To prove this claim, note that since $1 - a_k^2 = \frac{1}{1+M_k}$, the denominator in the recursion (15) is given by

$$M_k(1 - a_k^2) + a_k^2 = \frac{M_k}{1 + M_k} + \frac{M_k}{1 + M_k} \;=\; \frac{2M_k}{1 + M_k}$$

Consequently, we have $M_{k+1} = \frac{M_k}{2M_k/(1+M_k)} = \frac{1+M_k}{2} = 1 + \frac{1}{2}(M_k - 1)$, as claimed.

From the decay rate (22), we see that taking $K = \lceil \log_2 M_0 \rceil \leq 1 + \log_2 M_0$ steps suffices to ensure that $M_K \leq 2$. Since we also have $\mathbf{H}_{\mathsf{k}}(y) \succeq \mathbf{I}$ uniformly in y, it follows that $\sup_y \operatorname{cond}(\mathbf{H}_{\mathsf{k}}(y)) \leq M_k$. Consequently, this choice of K ensures that $\sup_y \operatorname{cond}(\mathbf{H}_{\mathsf{K}}(y)) \leq 2$, as claimed.

Proof of the claim (21c)**:** By the representation (29b) from Lemma 6 applied with $V = Y_{k+1}$ and $U = Y_k$, we have $\mathbf{J}_k(y) = \mathbf{H}_{\mathsf{k}}(y) + \frac{a_k^2}{1-a_k^2}$. With our choice $a_k^2 = \frac{M_k}{1+M_k}$, we have $1 - a_k^2 = \frac{1}{1+M_k}$, and hence $\frac{a_k^2}{1-a_k^2} = M_k$. Since the eigenvalues of $\mathbf{H}_{\mathsf{k}}(y)$ all lie in the interval $[1, M_k]$ by construction, the eigenvalues of $\mathbf{J}_k(y)$ all lie in the interval $[1+M_k, 2M_k]$, so that we are guaranteed to have

$$\operatorname{cond}(\mathbf{J}_k(y)) \leq \frac{2M_k}{1+M_k} \leq 2,$$

as claimed. ■

A.3. Proof of Lemma 2

We prove each of these two claims in turn.

Proof of the lower bound (17a)**:** Combining the representation of $\mathbf{J}_{\mathsf{u}|\mathsf{v}}$ in equation (29b) with our assumed lower bound on $\mathbf{H}_{\mathsf{u}}$, we have the uniform lower bound $\mathbf{J}_{\mathsf{u}|\mathsf{v}}(u) \succeq \big(m + \frac{a^2}{b^2}\big)\mathbf{I}$. Thus, the conditional distribution $p_{\mathsf{u}|\mathsf{v}}$ is strongly log-concave, so that the Brascamp–Lieb inequality (cf. Appendix D) is in force. It allows us to argue that

$$\operatorname{cov}(U \mid V = v) \overset{(i)}{\preceq} \mathbb{E}_{p_{\mathsf{u}|\mathsf{v}}}\big[\big(\mathbf{J}_{\mathsf{u}|\mathsf{v}}(U)\big)^{-1}\big] \overset{(ii)}{\preceq} \big(m + \tfrac{a^2}{b^2}\big)^{-1}\mathbf{I},$$

where step (i) follows from the bound (34b) in Appendix D; and step (ii) follows from the uniform lower bound on $\mathbf{J}_{\mathsf{u}|\mathsf{v}}(u)$.

Consequently, we have the lower bound

$$\mathbf{I} - \frac{a^2}{b^2}\operatorname{cov}(U \mid V = v) \succeq \Big(1 - \frac{a^2}{b^2}\frac{1}{m + \frac{a^2}{b^2}}\Big)\mathbf{I} = \Big(\frac{mb^2}{mb^2 + a^2}\Big)\mathbf{I}.$$

Combining with the second-order Tweedie formula (29a) from Lemma 6, we have shown that $\mathbf{H}_{\mathsf{v}}(v) \succeq \frac{m}{mb^2+a^2}\mathbf{I}$, as claimed.

Proof of the upper bound (17b)**:** Returning to the second-order Tweedie representation (29a), we see that upper bounds on $\mathbf{H}_{\mathsf{v}}$ require lower bounds on the covariance matrix $\operatorname{cov}(U \mid V = v)$. By a suitable embedding of our model into a parametric family, we can obtain such a lower bound via the Cramer–Rao approach, as we now describe.

Fix a realization $v \in \mathbb{R}^d$, and for each vector $\theta \in \mathbb{R}^d$, define the shifted density $q_\theta(u) := p_{\mathsf{u}|\mathsf{v}}(u - \theta)$. (To keep the notation clean, we are suppressing the dependence on v, since it remains fixed throughout the argument.) We can now apply the Cramer–Rao bound to the parametric family $\{q_\theta \mid \theta \in \mathbb{R}^d\}$. By construction, we have

$$-\nabla_\theta^2 \log q_\theta(u)\Big|_{\theta=0} = -\nabla_u^2 \log p_{\mathsf{u}|\mathsf{v}}(u) = \mathbf{J}_{\mathsf{u}|\mathsf{v}}(u).$$

Consequently, the Fisher information for estimating $\theta = 0$ is given by $\mathbf{F} := \mathbb{E}_{U \sim p_{\mathsf{u}|\mathsf{v}}}[\mathbf{J}_{\mathsf{u}|\mathsf{v}}(U)]$. From the representation (29b) of $\mathbf{J}_{\mathsf{u}|\mathsf{v}}$ and our assumed upper bound on $\mathbf{H}_{\mathsf{u}}$, we have

$$\mathbf{F} = \mathbb{E}_{U \sim p_{\mathsf{u}|\mathsf{v}}}\Big[\mathbf{H}_{\mathsf{u}}(U) + \frac{a^2}{b^2}\mathbf{I}\Big] \preceq \Big(M + \tfrac{a^2}{b^2}\Big)\mathbf{I},$$

Inverting and negating the relation, we have $-\mathbf{F}^{-1} \preceq -\Big(M + \frac{a^2}{b^2}\Big)^{-1}\mathbf{I}$.

Now observe that $\psi(u) = u - \mathbb{E}[U \mid V = v]$ is an unbiased estimate of $\theta = 0$ in this model, so that the Cramer–Rao bound implies that $\operatorname{cov}(U \mid V = v) \succeq \mathbf{F}^{-1}$. Putting together the pieces, we have

$$\mathbf{I} - \tfrac{a^2}{b^2}\operatorname{cov}(U \mid V = v) \preceq \mathbf{I} - \frac{a^2}{b^2}\mathbf{F}^{-1} \preceq \Big(1 - \frac{a^2}{b^2}\frac{1}{M + \frac{a^2}{b^2}}\Big)\mathbf{I} = \Big(\frac{Mb^2}{Mb^2 + a^2}\Big)\mathbf{I}.$$

Combining with the Tweedie form (29a) of $\mathbf{H}_{\mathsf{v}}$, we conclude that $\mathbf{H}_{\mathsf{v}}(v) \preceq \frac{M}{Mb^2+a^2}\mathbf{I}$, as claimed.

A.4. Proof of Lemma 3

We divide our proof into two parts, corresponding to each of the two distances.

Proof for KL divergence: For the KL-divergence, let Q be the joint distribution over (Y_k, Y_{k+1}) defined by the marginal q_{k+1} and the backward kernel $\mathcal{Q}_k$, so that Y_k has marginal q_k. Similarly, let P be the joint (Y_k, Y_{k+1}) defined by the marginal p_{k+1} and the backward kernel $\mathcal{P}_k$. By the data-processing inequality, we have $D_{\mathrm{KL}}(q_k \mid p_k) \leq D_{\mathrm{KL}}(Q \mid P)$. Combined with the chain rule for KL divergence, we find that

$$\begin{aligned} D_{\mathrm{KL}}(q_k \mid p_k) &\leq \mathbb{E}_{q_{k+1}}\Big[D_{\mathrm{KL}}\big(\mathcal{Q}_k(e_Y) \| \mathcal{P}_k(e_Y)\big)\Big] + D_{\mathrm{KL}}(q_{k+1} \mid p_{k+1}) \\ &\leq \delta_k + D_{\mathrm{KL}}(q_{k+1} \mid p_{k+1}), \end{aligned}$$

as claimed.

Proof for Wasserstein $\mathcal{W}_2$: The Wasserstein distance satisfies the triangle inequality, so that we can write

$$\mathcal{W}_2(q_k, p_k) = \mathcal{W}_2\big(\mathcal{Q}_k(q_{k+1}), \mathcal{P}_k(p_{k+1})\big) \leq \mathcal{W}_2\big(\mathcal{Q}_k(q_{k+1}), \mathcal{P}_k(q_{k+1})\big) + \mathcal{W}_2\big(\mathcal{P}_k(q_{k+1}), \mathcal{P}_k(p_{k+1})\big).$$

We have

$$\mathcal{W}_2^2(\mathcal{Q}_k(q_{k+1}), \mathcal{P}_k(q_{k+1})) \overset{(i)}{\leq} \mathbb{E}_{q_{k+1}}\mathcal{W}_2^2\big(\mathcal{Q}_k(\cdot \mid Y), \mathcal{P}_k(\cdot \mid Y)\big) \overset{(ii)}{\leq} \delta_k^2$$

where step (i) follows from joint convexity of $\mathcal{W}_2^2$ in its arguments, and step (ii) follows from the assumed sampler accuracy (18b). Taking square roots of this upper bound, and combining the two bounds yields

$$\mathcal{W}_2(q_k, p_k) \leq \delta_k + \mathcal{W}_2\big(\mathcal{P}_k(q_{k+1}), \mathcal{P}_k(p_{k+1})\big) \tag{23}$$

To complete the proof, it remains to show that $\mathcal{W}_2\big(\mathcal{P}_k(q_{k+1}), \mathcal{P}_k(p_{k+1})\big) \leq \mathcal{W}_2(q_{k+1}, p_{k+1})$. We make use of Lemma 8, proved in Appendix E, which controls the Wasserstein stability of the kernels that arise in our backward analysis. It allows us to prove that

$$\mathcal{W}_2(\mathcal{P}_k(q_{k+1}), \mathcal{P}_k(p_{k+1})) \leq a_k \mathcal{W}_2(q_{k+1}, p_{k+1}) \;\leq\; \mathcal{W}_2(q_{k+1}, p_{k+1}), \tag{24}$$

and the remainder of the proof goes through as in the KL case.

In order to prove inequality (24), we first show how our backward kernel can be converted to the form assumed in Lemma 8. Let $\propto$ denote the proportionality relation, keeping only terms dependent on y_k. Since $Y_{k+1} \mid y_k \sim \mathcal{N}(a_k y_k, \mathbf{I}/(1-a_k^2))$ by construction, we can write

$$\begin{aligned} p_{k|k+1}(y_k \mid y_{k+1}) &\propto \exp\Big\{ \log p_k(y_k) - \tfrac{1}{2(1-a_k^2)} \|a_k y_k - y_{k+1}\|_2^2 \Big\} \\ &\propto \exp\Big\{ -\psi(y_k) + \tfrac{a_k}{(1-a_k^2)} \langle y_k, \, y_{k+1} \rangle \Big\}, \end{aligned}$$

where $\psi(y_k) = -\log p_k(y_k) + \frac{a_k^2}{2(1-a_k^2)} \|y_k\|_2^2$. Since $-\nabla^2 \log p_k(y_k) = \mathbf{H}_k(y_k) \succeq \mathbf{I}$, we have the lower bound $\nabla^2 \psi(y_k) \succeq 1 + \frac{a_k^2}{1-a_k^2}$, so that we can apply Lemma 8 with $\alpha = 1 + \frac{a_k^2}{1-a_k^2}$ and $\beta = \frac{a_k}{1-a_k^2} \geq 0$. Finally, we observe that $\frac{\beta}{\alpha} = \frac{a_k/(1-a_k^2)}{1+\frac{a_k^2}{1-a_k^2}} = a_k$, so that the claim (24) follows by application of Lemma 8.

Appendix B. Auxiliary results related to Theorem 2

In this section, we collect the proofs of the auxiliary results related to Theorem 2.

B.1. Proof of Lemma 4

Define the sequence $\theta_0^2 = a_0^2 = 1/2$ and $\theta_k^2 = a_k^2 \theta_{k-1}^2$ for $k = 1, 2, \ldots$. We split our proof into two parts. By induction on k, it is easy to show that

$$Y_k = \theta_{k-1} X + \sqrt{1 - \theta_{k-1}^2} W'_{k-1} \qquad \text{where } W'_{k-1} \sim \mathcal{N}(0, \mathbf{I}) \text{ is independent of } X. \tag{25}$$

We make use of this representation repeatedly.

Proof of the forward sandwich (20a): Applying the second-order Tweedie formula (29a) from Lemma 6 to the representation (25), we find that

$$\mathbf{H}_k(y) \;=\; \frac{1}{1-\theta_{k-1}^2} \Big\{ \mathbf{I} \;-\; \frac{\theta_{k-1}^2}{1-\theta_{k-1}^2} \operatorname{Cov}\big(X \mid Y_k = y\big) \Big\}.$$

Since the covariance is positive semidefinite, it follows immediately that $\mathbf{H}_k(y) \preceq \frac{1}{1-\theta_{k-1}^2} \mathbf{I} \preceq 2\mathbf{I}$, where the final inequality follows since $\theta_{k-1}^2 \leq 1/2$ for all $k = 1, 2, \ldots$.

As for the lower bound, we have $1 \leq \frac{1}{(1-\theta_{k-1}^2)^2} \leq 4$, and hence

$$\mathbf{H}_k(y) \succeq \Big\{ 1 - 4\theta_{k-1}^2 \sup_{y \in \mathbb{R}^d} \|\!| \operatorname{cov}(X \mid Y_k = y) |\!\|_{\mathrm{op}} \Big\} \mathbf{I} = \big\{ 1 - 4\theta_{k-1}^2 B_k \big\} \mathbf{I} \;\succeq\; \underbrace{-4\theta_{k-1}^2 B_k}_{\equiv -\lambda_k} \mathbf{I},$$

as claimed.

Proof of the backward sandwich (20b): In this case, we use the equation $Y_{k+1} = a_k Y_k + \sqrt{1-a_k^2} W_k$. By the backward Hessian representation (29b) from Lemma 6 applied with $U = Y_k$, $V = Y_{k+1}$, $a^2 = a_k^2$ and $b^2 = 1 - a_k^2$, we find that

$$\mathbf{J}_k(y_k) \;=\; \mathbf{H}_k(y_k) \;+\; \frac{a_k^2}{1-a_k^2}\mathbf{I} \qquad \text{where } s_k \;:=\; \tfrac{a_k^2}{1-a_k^2}. \tag{26}$$

The sandwich condition (20a) ensures that $(-\lambda_k + s_k)\,\mathbf{I} \;\preceq\; \mathbf{J}_k(y_k) \;\preceq\; (2+s_k)\,\mathbf{I}$. With the adaptively chosen stepsizes (10c), we have

$$s_k \;=\; \tfrac{2\lambda_k+2}{(2\lambda_k+2)+1}\Big(1 - \tfrac{2\lambda_k+2}{(2\lambda_k+2)+1}\Big)^{-1} \;=\; 2\lambda_k + 2.$$

Combining the pieces yields

$$\big(-\lambda_k + 2\lambda_k + 2\big)\mathbf{I} \;=\; \big(\lambda_k + 2\big)\mathbf{I} \;\preceq\; \mathbf{J}_k(y_k) \;\preceq\; \big(2 + (2\lambda_k + 2)\big)\mathbf{I} \;=\; 2\,\big(\lambda_k + 2\big)\,\mathbf{I},$$

as claimed.

B.2. Proof of Corollary1

Define the auxiliary sequence $\{\theta_k\}_{k=0}^{\infty}$ via $\theta_0^2 = a_0^2 = 1/2$ and $\theta_k^2 = a_k^2\theta_{k-1}^2$ for $k = 1, 2, \ldots$. Introducing the shorthand $L = 8B_{\max}$, it suffices to specify a choice of K that ensures $\theta_{K-1}^2 \le 1/L$, so that Theorem 2 can be applied. The remainder of our argument is devoted to showing that

$$K \le 1 + 4B_{\max} + 3\Big\lceil \log_2\Big(\frac{L}{2}\Big)\Big\rceil \;\le\; 7\,\big(1 + B_{\max}\big) \tag{27}$$

rounds are sufficient.

For our analysis, it is convenient to make use of the auxiliary sequence defined by $u_0 = 2B_{\max}$, and $u_k \;:=\; 4B_{\max}\,\theta_k^2$. Suppose that we can establish that

$$u_{K-1} \le u^\star := 4B_{\max}/L, \qquad \text{with } K \text{ from equation (27).} \tag{28a}$$

It then follows that $\theta_{K-1}^2 = u_{K-1}/(4B_{\max}) \le 1/L$, as desired.

Our proof of the bound (28a) is based on the following descent guarantee: for $k = 1, 2, \ldots$, we have

$$u_k - u_{k-1} \le -g(u_{k-1}) \quad \text{where} \quad g(s) := \frac{s}{2(s+3/2)}. \tag{28b}$$

We return to prove it momentarily; taking it as given for the moment, let us prove the bound (28a).

Proof of the bound (28a): Introduce the shorthand $T = K - 1$. Our goal is to establish that $u_T \le u^\star = 4B_{\max}/L$. Since $u_0 = 2B_{\max}$, we have the ratio $u_0/u^\star = L/2$, and we analyze the evolution of the iterates $\{u_k\}_{k\ge 0}$ as they move through a sequence of $J := \lceil \log_2(L/2) \rceil$ epochs. Each epoch is constructed so that the value u_k drops by a factor of $1/2$ as it transitions from the start to the end of the interval. Concretely, for $j = 1, \ldots, J+1$, we define the intervals

$$\mathcal{I}^{(j)} := \Big(\underbrace{\frac{u_0}{2^j}}_{\equiv v^{(j)}}, \quad \underbrace{\frac{u_0}{2^{j-1}}}_{\equiv v^{(j-1)}} \Big],$$

so that $v^{(0)} = u_0$ and $v^{(J)} = u_0/2^J \leq u^\star$. At the terminal round K, we will ensure that $u_T \in \mathcal{I}^{(J+1)}$, so that $u_T \leq v^{(J)} \leq u^\star$.

The total number of steps T is defined by the sum $T = \sum_{j=1}^J N^{(j)}$ where $N^{(j)}$ is the number of steps k for which $u_k \in \mathcal{I}^{(j)}$. In order to bound T, we need to bound $N^{(j)}$. During epoch j, we need to reduce the value of u from $v^{(j-1)}$ down to $v^{(j)}$, so that the total decrease is given by $\Delta^{(j)} := v^{(j-1)} - v^{(j)} = v^{(j)}$. Observe that the function g from the descent condition (28b) is strictly increasing. Consequently, for values of $u \in \mathcal{I}^{(j)}$, we have $g(u) \geq g(v^{(j)}) = \frac{v^{(j)}}{2(v^{(j)}+3/2)}$. Using this fact, we have the bound

$$N^{(j)} \leq \frac{\Delta^{(j)}}{g(v^{(j)})} = \frac{v^{(j)}}{v^{(j)}/(2(v^{(j)}+3/2))} = 2\big(v^{(j)} + 3/2\big).$$

Using this upper bound and summing over the epochs $j = 1, \ldots, J$ yields

$$T = \sum_{j=1}^J N^{(j)} \leq \sum_{j=1}^J 2\big(v^{(j)} + 3/2\big) = 2\sum_{j=1}^J v^{(j)} + 3J.$$

Since $\sum_{j=1}^J v^{(j)} = u_0(1 - 2^{-J})$ and $u_0 = 2B_{\max}$, we obtain

$$T \leq 4B_{\max}\big(1 - 2^{-J}\big) + 3J \leq 4B_{\max} + 3\lceil \log_2(L/2) \rceil.$$

Since $K = T + 1$, we have proved the claim (27).

Proof of the descent bound (28b)**:** It remains to establish the descent condition. From the definition of λ_k and the assumption that $B_k \leq B_{\max}$, we have

$$\lambda_k = 4\theta_{k-1}^2 B_k \leq 4\theta_{k-1}^2 B_{\max} = u_{k-1}.$$

For $s > 0$, define the function $f(s) = \frac{2s+2}{2s+3}$, and note that $a_k^2 = f(\lambda_k)$ by construction. Since f is an increasing function and $\lambda_k \leq u_{k-1}$, we have

$$a_k^2 = f(\lambda_k) \leq f(u_{k-1}) = \frac{2u_{k-1}+2}{2u_{k-1}+3} = 1 - \frac{1}{2(u_{k-1}+3/2)}.$$

Multiplying both sides by u_{k-1} yields the one-step recursion

$$u_k = a_k^2\, u_{k-1} \leq u_{k-1}\Big(1 - \frac{1}{2(u_{k-1}+3/2)}\Big),$$

and re-arranging yields the claim (28b).

Appendix C. Second-order Tweedie formula

In this section, we provide a proof of the following lemma:

Lemma 6 (Forward and conditional Hessians) *We have the second-order Tweedie formula*

$$\textbf{Second-order Tweedie:} \qquad \underbrace{\mathbf{H}_{\mathsf{v}}(v)}_{-\nabla^2 \log p_{\mathsf{v}}(v)} = \frac{1}{b^2}\Big\{\mathbf{I} - \frac{a^2}{b^2}\operatorname{cov}(U \mid V = v)\Big\}. \tag{29a}$$

Moreover, we have

$$\textbf{Backward conditional Hessian:} \qquad \underbrace{\mathbf{J}_{\mathsf{u}|\mathsf{v}}(u,v)}_{-\nabla_u^2 \log p_{\mathsf{u}|\mathsf{v}}(u|v)} = \mathbf{H}_{\mathsf{u}}(u) + \frac{a^2}{b^2}\mathbf{I}. \tag{29b}$$

The second-order Tweedie formula (29a) is a known result (see the papers Efron (2011); Chen et al. (2023a); De Bortoli (2022); Benton et al. (2024) for variants), whereas the backward conditional Hessian formula (29b) follows directly from the structure of the joint distribution. For completeness, we provide a proof here.

We prove this lemma using a slightly more general result, which we begin by stating. Given a pair of random vectors (U, V), suppose that the integral representation

$$p_{\mathsf{v}}(v) = \int_{\mathbb{R}^d} p_{\mathsf{v}|\mathsf{u}}(v \mid u)\, p_{\mathsf{u}}(u)\, du \tag{30}$$

of the marginal density admits sufficient regularity so that differentiation under the integral is permitted.

Lemma 7 (Hessian identity for marginal log densities) *Under the above conditions, for every v in the support of V, we have the identity*

$$\nabla_v^2 \log p_{\mathsf{v}}(v) = \mathbb{E}\big[G(U,v) \mid V = v\big] \;+\; \operatorname{cov}\big(s(U,v) \mid V = v\big), \tag{31}$$

where $s(u,v) \coloneqq \nabla_v \log p_{V|U}(v \mid u)$ is the conditional score, and $G(u,v) \coloneqq \nabla_v^2 \log p_{V|U}(v \mid u)$ is the conditional Hessian.

We first use this lemma to prove the two claims (29a) and (29b) from Lemma 6. In Appendix C.3, we return to prove Lemma 7.

C.1. Proof of equation (29a):

We apply Lemma 7 to our generative model $V = aU + bW$, where W is standard Gaussian. By definition, we have $\big(V \mid U = u\big) \sim \mathcal{N}(au, b^2\mathbf{I})$, so that

$$s(u,v) = \nabla_v \log p_{\mathsf{v}|\mathsf{u}}(v \mid u) = \nabla_v\Big\{-\frac{1}{2b^2}\|v - au\|_2^2\Big\} = \frac{au - v}{b^2}, \quad \text{and}$$
$$G(u,v) = \nabla^2 \log p_{\mathsf{v}|\mathsf{u}}(v \mid u) = -\frac{1}{b^2}\mathbf{I}.$$

Thus, we have $\operatorname{cov}(s(U,v) \mid V = v) = \operatorname{cov}\left(\frac{aU}{b^2} \mid V = v\right) = \frac{a^2}{b^4}\operatorname{cov}(U \mid V = v)$. Substituting into (31) yields $\nabla^2 \log p_{\mathsf{v}}(v) = -\frac{1}{b^2}\mathbf{I} + \frac{a^2}{b^4}\operatorname{cov}(U \mid V = v)$, and re-arranging yields the claim (29a).

C.2. Proof of equation (29b):

Since $p_{\mathsf{u}|\mathsf{v}}(u \mid v) = p_{\mathsf{v}|\mathsf{u}}(v \mid u) p_{\mathsf{u}}(u)/p_{\mathsf{v}}(v)$, we have

$$\begin{aligned}\mathbf{J}(u,v) \equiv -\nabla_u^2 \log p_{\mathsf{u}|\mathsf{v}}(u \mid v) &= -\nabla_u^2 \log p_{\mathsf{u}}(u) - \nabla_u^2 \log p_{\mathsf{v}|\mathsf{u}}(v \mid u) \\ &= \mathbf{H}_{\mathsf{u}}(u) + \nabla_u^2 \Big\{ \frac{1}{2b^2} \|au - v\|_2^2 \Big\} \;=\; \mathbf{H}_{\mathsf{u}}(u) + \frac{a^2}{b^2}\mathbf{I},\end{aligned}$$

as claimed.

C.3. Proof of Lemma 7

By chain rule, we can compute $\nabla_v \log p_{\mathsf{v}}(v) = \nabla p_{\mathsf{v}}(v)/p_{\mathsf{v}}(v)$, and hence

$$\begin{aligned}\nabla_v^2 \log p_{\mathsf{v}}(v) &= \frac{1}{p_{\mathsf{v}}(v)} \nabla_v^2 p_{\mathsf{v}}(v) \;-\; \frac{1}{p_{\mathsf{v}}(v)^2} \big(\nabla_v p_{\mathsf{v}}(v)\big)\big(\nabla_v p_{\mathsf{v}}(v)\big)^\top \\ &= \frac{1}{p_{\mathsf{v}}(v)} \nabla_v^2 p_{\mathsf{v}}(v) \;-\; \big(\nabla_v \log p_{\mathsf{v}}(v)\big)\big(\nabla_v \log p_{\mathsf{v}}(v)\big)^T. \qquad (32a)\end{aligned}$$

Suppose that we can show that

$$\nabla_v \log p_{\mathsf{v}}(v) = \mathbb{E}\big[s(U,v) \mid V = v\big], \quad \text{and} \qquad (32b)$$

$$\nabla_v^2 p_{\mathsf{v}}(v) = p_{\mathsf{v}}(v)\, \mathbb{E}\big[G(U,v) + s(U,v)s(U,v)^\top \mid V = v\big]. \qquad (32c)$$

Substituting these expressions into our decomposition (32a) then yields the claim (31).

Proof of the relation (32b): Beginning with the representation (30), differentiating under the integral yields the relation $\nabla_v p_{\mathsf{v}}(v) \;=\; \int \nabla_v p_{V|U}(v \mid u)\, p_U(u)\, du$. Next we observe that

$$\nabla_v p_{\mathsf{v}|\mathsf{u}}(v \mid u) = p_{V|U}(v \mid u)\, \nabla_v \log p_{\mathsf{v}|\mathsf{u}}(v \mid u) = p_{\mathsf{v}|\mathsf{u}}(v \mid u)\, s(u,v), \qquad \text{and hence} \qquad (33a)$$

$$\nabla_v p_{\mathsf{v}}(v) = \int p_{\mathsf{v}|\mathsf{u}}(v \mid u) s(u,v)\, p_{\mathsf{u}}(u)\, du. \qquad (33b)$$

Using Bayes' rule, we can rewrite this as $\nabla_v p_{\mathsf{v}}(v) = p_{\mathsf{v}}(v)\, \mathbb{E}\big[s(U,v) \mid V = v\big]$, and dividing by $p_{\mathsf{v}}(v)$ yields the claim (32b).

Proof of the relation (32c): Differentiating p_{v} twice under the integral yields the expression $\nabla_v^2 p_{\mathsf{v}}(v) \;=\; \int \nabla_v^2 p_{\mathsf{v}|\mathsf{u}}(v \mid u)\, p_{\mathsf{u}}(u)\, du$. Now introduce the shorthand $s(u,v) \coloneqq \nabla_v \log p_{\mathsf{v}|\mathsf{u}}(v \mid u)$ and $G(u,v) \coloneqq \nabla_v s(u,v) \;=\; \nabla_v^2 \log p_{\mathsf{v}|\mathsf{u}}(v \mid u)$, so that $\nabla_v p_{\mathsf{v}|\mathsf{u}}(v \mid u) \;=\; p_{\mathsf{v}|\mathsf{u}}(v \mid u) s(u,v)$. Differentiating once more gives

$$\nabla_v^2 p_{\mathsf{v}|\mathsf{u}}(v \mid u) = \nabla_v \big\{ p_{\mathsf{v}|\mathsf{u}}(v \mid u) s(u,v) \big\} \;=\; p_{\mathsf{v}|\mathsf{u}}(v \mid u) \big\{ s(u,v)s(u,v)^\top + G(u,v) \big\}.$$

Putting together the pieces, we have

$$\begin{aligned}\nabla_v^2 p_{\mathsf{v}}(v) &= \int p_{V|U}(v \mid u) \Big(G(u,v) + s(u,v)s(u,v)^\top \Big) p_U(u)\, du \\ &= p_{\mathsf{v}}(v)\, \mathbb{E}\big[G(U,v) + s(U,v)s(U,v)^\top \mid V = v\big],\end{aligned}$$

again using Bayes' rule. This completes the proof of the claim (32c).

Appendix D. Elementary consequence of Brascamp–Lieb inequality

Consider a density on $\mathbb{R}^d$ of the form $p(x) \propto \exp(-\psi(x))$ where ψ is twice differentiable and strictly convex. The Brascamp–Lieb inequality asserts that for any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we have

$$\operatorname{var}_p(f(X)) \leq \mathbb{E}_p\Big[\big\langle \nabla f(X),\, (\nabla^2\psi(X))^{-1}\nabla f(X)\big\rangle\Big]. \tag{34a}$$

Let us derive an elementary consequence that leads to an upper bound on the matrix $\operatorname{cov}_p(X)$. For any given $v \in \mathbb{R}^d$, define the linear function $f_v(x) = \langle v,\, x - \mu\rangle$, where $\mu \coloneqq \mathbb{E}[X]$. Computing the derivative $\nabla f_v(x) = v$ and then applying inequality (34a) yields the upper bound

$$v^T \operatorname{cov}_p(X) v \;=\; \operatorname{var}_p(f_v(X)) \;\leq\; v^T \mathbb{E}_p\big[(\nabla^2\psi(X))^{-1}\big] v.$$

Since the choice of $v \in \mathbb{R}^d$ was arbitrary, it follows that $\operatorname{cov}_p(X) \preceq \mathbb{E}_p\Big[(\nabla^2\psi(X))^{-1}\Big]$. In particular, when $\nabla^2\psi(x) \succeq m\mathbf{I}$ uniformly in x, it follows that

$$\operatorname{cov}_p(X) \preceq \frac{1}{m}\mathbf{I}. \tag{34b}$$

Appendix E. Wasserstein stability of the backward kernel

Consider the Markov kernel $q \mapsto \mathcal{P}(q)(\cdot) \coloneqq \int_{\mathbb{R}^d} p_{\mathsf{u}|\mathsf{v}}(\cdot \mid v) q(v) dv$ where, for some $\beta \in \mathbb{R}$, the conditional density takes the form

$$p_{\mathsf{u}|\mathsf{v}}(u \mid v) \propto \exp\Big\{ -\psi(u) + \beta\,\langle u,\, v\rangle \Big\}, \tag{35a}$$

and the Hessian $\nabla^2\psi$ satisfies

$$\nabla^2\psi(u) \succeq \alpha\mathbf{I} \qquad \text{uniformly over } u \in \mathbb{R}^d. \tag{35b}$$

Lemma 8 (Wasserstein stability of the backward kernel) *Under the above conditions, we have*

$$\mathcal{W}_2\big(\mathcal{P}(q), \mathcal{P}(\widetilde{q})\big) \leq \frac{|\beta|}{\alpha}\,\mathcal{W}_2(q, \widetilde{q}) \qquad \textit{valid for any pair of distributions } q, \widetilde{q}. \tag{36}$$

Proof The Wasserstein distance is defined via an infimum over couplings; thus, by a standard "gluing" argument, it is sufficient to show that

$$\mathcal{W}_2\big(p_{\mathsf{u}|\mathsf{v}}(\cdot \mid v),\; p_{\mathsf{u}|\mathsf{v}}(\cdot \mid \widetilde{v})\big) \leq \frac{|\beta|}{\alpha}\,\|v - \widetilde{v}\|_2 \qquad \text{for all } v, \widetilde{v} \in \mathbb{R}^d. \tag{37a}$$

In order to do so, it is convenient to introduce the d-dimensional exponential family

$$r_\theta(u) \coloneqq \exp\big\{\, \langle \theta,\, u\rangle - \psi(u) - A(\theta)\big\}$$

where $A(\theta) = \log \int e^{\langle \theta, u \rangle - \psi(u)} du$ is the log normalization constant. By construction, for each $v \in \mathbb{R}^d$, we have $p_{\mathsf{u}|\mathsf{v}}(u \mid v) \equiv r_{\theta(v)}(u)$ where $\theta(v) := \beta v$. Consequently, in order to prove the bound (37a), it suffices to show that

$$\mathcal{W}_2\big(r_\theta, r_{\widetilde{\theta}}\big) \leq \frac{1}{\alpha} \|\theta - \widetilde{\theta}\|_2 \qquad \text{for each } \theta, \widetilde{\theta}. \tag{37b}$$

Since $\nabla^2 \psi(u) \succeq \alpha \mathbf{I}$ uniformly in $u \in \mathbb{R}^d$, the density $r_{\widetilde{\theta}}$ is α-strongly-log-concave. Hence, Talagrand's T_2-bound holds. (In particular, the Bakry–Emery criterion implies that $r_{\widetilde{\theta}}$ satisfies a log-Sobolev inequality (LSI) with parameter $2/\alpha$, and the Otto–Villani translation from LSI to T_2 then implies that $r_{\widetilde{\theta}}$ satisfies Talagrand's T_2-inequality with parameter $2/\alpha$. See Corollary 7.3 and Theorem 8.12 in the survey paper by Gozlan and Léonard (2010) for details.) Consequently, we have

$$\mathcal{W}_2^2(r_\theta, r_{\widetilde{\theta}}) \leq \frac{2}{\alpha} D_{\mathrm{KL}}(r_\theta \| r_{\widetilde{\theta}}), \tag{38a}$$

where D_{KL} denotes the Kullback-Leibler divergence between r_θ and $r_{\widetilde{\theta}}$. From the exponential family structure (Wainwright and Jordan, 2008), we have $D_{\mathrm{KL}}(r_\theta \| r_{\widetilde{\theta}}) = A(\widetilde{\theta}) - A(\theta) - \Big\langle \nabla A(\theta), \, \widetilde{\theta} - \theta \Big\rangle$. Taking one more derivative, we can write

$$D_{\mathrm{KL}}(r_\theta \| r_{\widetilde{\theta}}) = \frac{1}{2} (\widetilde{\theta} - \theta)^\top \nabla^2 A(\bar{\theta}) (\widetilde{\theta} - \theta), \tag{38b}$$

where $\bar{\theta}$ is some vector on the line joining θ and $\widetilde{\theta}$. Thus, it remains to bound the Hessian $\nabla^2 A$. Again using standard properties of exponential families (Wainwright and Jordan, 2008), we have $\nabla^2 A(\bar{\theta}) = \mathrm{cov}_{\bar{\theta}}(U)$, where $\mathrm{cov}_{\bar{\theta}}$ denotes the covariance computed under the exponential family density $r_{\bar{\theta}}$. Since $r_{\bar{\theta}}$ is defined by a potential function that is α-strongly-convex, the Brascamp–Lieb inequality (see equation (34b) in Appendix D) can be applied to assert that

$$\nabla^2 A(\bar{\theta}) = \mathrm{cov}_{\bar{\theta}}(U) \preceq \frac{1}{\alpha} \mathbf{I} \qquad \text{uniformly for all } \bar{\theta}. \tag{38c}$$

Applying the Brascamp–Lieb bound (38c) to equation (38b), we find that the KL can be upper bounded as $\mathrm{KL}(r_\theta \mid r_{\widetilde{\theta}}) \leq \frac{1}{2\alpha} \|\theta - \widetilde{\theta}\|_2^2$. Combining with our T_2-bound (38a), we find that

$$\mathcal{W}_2^2(r_\theta, r_{\widetilde{\theta}}) \leq \frac{2}{\alpha} D_{\mathrm{KL}}(r_\theta \| r_{\widetilde{\theta}}) \leq \frac{2}{\alpha} \left\{ \frac{1}{2\alpha} \|\theta - \widetilde{\theta}\|_2^2 \right\} = \frac{1}{\alpha^2} \|\theta - \widetilde{\theta}\|_2^2.$$

Taking square roots, we have proved the desired bound (37b). ■

Appendix F. Results on score perturbations

This section is devoted to results on score perturbations, beginning in Appendix F.1 with the proof of the robustness guarantee stated in Lemma 1. This proof makes use of an auxiliary result of independent interest: Lemma 9 on perturbed drift functions in Appendix F.2.

F.1. Proof of Lemma 1

By definition of the approximate and exact updates, we have $\widetilde{q}_k = \widetilde{q}_{k+1}\widetilde{\mathcal{Q}}_k$ and $p_k = p_{k+1}\mathcal{P}_k$. Thus, we can write

$$D_{\mathrm{TV}}(\widetilde{q}_k, p_k) = D_{\mathrm{TV}}\left(\widetilde{q}_{k+1}\widetilde{\mathcal{Q}}_k,\, p_{k+1}\mathcal{P}_k\right) \overset{(i)}{\leq} D_{\mathrm{TV}}(\widetilde{q}_{k+1}\widetilde{\mathcal{Q}}_k, \widetilde{q}_{k+1}\mathcal{P}_k) + D_{\mathrm{TV}}(\widetilde{q}_{k+1}\mathcal{P}_k,\, p_{k+1}\mathcal{P}_k)$$
$$\overset{(ii)}{\leq} D_{\mathrm{TV}}(\widetilde{q}_{k+1}\widetilde{\mathcal{Q}}_k, \widetilde{q}_{k+1}\mathcal{P}_k) + D_{\mathrm{TV}}(\widetilde{q}_{k+1},\, p_{k+1}) \tag{39}$$

where step (i) follows from the triangle inequality; and step (ii) follows since the TV distance is non-expansive under application of the Markov kernel $\mathcal{P}_k$. As for the first term in the inequality, since both measures are mixtures over the same input law $\widetilde{q}_{k+1}$, we have

$$D_{\mathrm{TV}}(\widetilde{q}_{k+1}\widetilde{\mathcal{Q}}_k,\, \widetilde{q}_{k+1}\mathcal{P}_k) \leq \mathbb{E}_{Y\sim\widetilde{q}_{k+1}} D_{\mathrm{TV}}\big(\widetilde{\mathcal{Q}}_k(\cdot \mid Y), \mathcal{P}_k(\cdot \mid Y)\big) \tag{40a}$$

By triangle inequality, we have

$$D_{\mathrm{TV}}\big(\widetilde{\mathcal{Q}}_k(\cdot \mid y), \mathcal{P}_k(\cdot \mid y)\big) \leq D_{\mathrm{TV}}(\widetilde{\mathcal{Q}}_k(\cdot \mid y), \mathcal{Q}_k(\cdot \mid y)) + D_{\mathrm{TV}}\big(\mathcal{Q}_k(\cdot \mid y), \mathcal{P}_k(\cdot \mid y)\big)$$
$$\leq \varepsilon_k + D_{\mathrm{TV}}\big(\mathcal{Q}_k(\cdot \mid y), \mathcal{P}_k(\cdot \mid y)\big), \tag{40b}$$

where the second step uses our ε_k-accuracy assumption on the sampler $\widetilde{\mathcal{Q}}_k$. Next, we have

$$D^2_{\mathrm{TV}}\big(\mathcal{Q}_k(\cdot \mid y), \mathcal{P}_k(\cdot \mid y)\big) \overset{(i)}{\leq} \frac{1}{2} D_{\mathrm{KL}}\big(\mathcal{Q}_k(\cdot \mid y) \| \mathcal{P}_k(\cdot \mid y)\big)$$
$$\overset{(ii)}{\leq} \frac{1}{4\alpha_k} I(\mathcal{Q}_k(\cdot \mid y) \| \mathcal{P}_k(\cdot \mid y)), \tag{40c}$$

where step (i) follows from Pinsker's inequality; and step (ii) follows since any α_k-log-concave density satisfies a log-Sobolev inequality with parameter α_k. By definition, the distribution $\mathcal{P}_k(\cdot \mid y)$ has score function $f(x) = s_k(x) - \frac{a_k}{1-a_k^2}(a_k x - y)$, whereas $\mathcal{Q}_k(\cdot \mid y)$ is stationary for the SDE with drift function $\widehat{f}(x) = \widehat{s}_k(x) - \frac{a_k}{1-a_k^2}(a_k x - y)$. Thus, the drift $\widehat{f}$ is a perturbation of the score function f, so that we can apply Lemma 9 to assert that $I(\mathcal{Q}_k(\cdot \mid y) \| \mathcal{P}_k(\cdot \mid y)) \leq \|\widehat{s}_k - s_k\|^2_{L^2(\mathcal{Q}_k(\cdot|y))}$.

Combining with the bounds (40b) and (40c), we have shown that

$$D_{\mathrm{TV}}\big(\widetilde{\mathcal{Q}}_k(\cdot \mid y), \mathcal{P}_k(\cdot \mid y)\big) \leq \varepsilon_k + \frac{1}{2\sqrt{\alpha_k}} \|\widehat{s}_k - s_k\|_{L^2(\mathcal{Q}_k(\cdot|y))}.$$

Taking expectations with respect to $\widetilde{q}_{k+1}$ and using the bound (40a), we find that

$$D_{\mathrm{TV}}(\widetilde{q}_{k+1}\widetilde{\mathcal{Q}}_k,\, \widetilde{q}_{k+1}\mathcal{P}_k) \leq \varepsilon_k + \frac{1}{2\sqrt{\alpha_k}} \mathbb{E}_{Y\sim\widetilde{q}_{k+1}} \|\widehat{s}_k - s_k\|_{L^2(\mathcal{Q}_k(\cdot|y))}$$
$$\leq \varepsilon_k + \frac{1}{2\sqrt{\alpha_k}} \|\widehat{s}_k - s_k\|_{L^2(\bar{q}_k)},$$

where the final step follows by Jensen's inequality, and the definition of $\bar{q}_k$. Combining with inequality (39) completes the proof.

F.2. Stationary drift perturbations

Let π be a smooth density with score function f, and let $\widehat{\pi}$ be a smooth stationary distribution for the perturbed diffusion $dX_t = \widehat{f}(X_t)\,dt + \sqrt{2}\,dB_t$.

Lemma 9 (Stationary drift perturbations) *Under standard regularity conditions permitting integration by parts, we have*

$$I(\widehat{\pi}\|\pi) := \int \big\|\nabla \log \tfrac{\widehat{\pi}(x)}{\pi(x)}\big\|_2^2\, \widehat{\pi}(x)dx \;\leq\; \|\widehat{f} - f\|^2_{L^2(\widehat{\pi})}. \tag{41}$$

Proof Introduce the shorthand $e(x) := \widehat{f}(x) - f(x)$ for the perturbation, and $r(x) := \log \frac{\widehat{\pi}(x)}{\pi(x)}$ for the log density ratio. Observe that $I(\widehat{\pi}\|\pi) = \|\nabla r\|^2_{L^2(\widehat{\pi})}$ by definition. Moreover, since $\widehat{\pi}$ is stationary for the perturbed diffusion, it satisfies the stationary Fokker–Planck equation

$$0 = -\nabla \cdot \{\widehat{f}\widehat{\pi}\} + \Delta\widehat{\pi} \;=\; -\nabla \cdot \{(f+e)\widehat{\pi}\} + \Delta\widehat{\pi}.$$

Now we have

$$\nabla\widehat{\pi} = \widehat{\pi}\nabla \log \widehat{\pi} \;=\; \widehat{\pi}\nabla\big(\log \pi + r\big) \;=\; \widehat{\pi} f + \widehat{\pi}\nabla r,$$

using the definition of r, and the fact that f is the score function of π. Consequently, the FP equation can be rewritten as $0 = \nabla \cdot (\widehat{\pi}\nabla r) - \nabla \cdot (e\widehat{\pi})$. Multiplying by r and integrating by parts[3] gives

$$0 = -\int \|\nabla r\|_2^2\, d\widehat{\pi} + \int \langle e, \nabla r\rangle\, d\widehat{\pi}.$$

Re-arranging, we find that

$$I(\widehat{\pi}\|\pi) = \int \|\nabla r\|_2^2\, d\widehat{\pi} \;\leq |\int \langle e, \nabla r\rangle\, d\widehat{\pi}| \;\leq\; I(\widehat{\pi}\|\pi)^{1/2} \left(\int \|e\|_2^2\, d\widehat{\pi}\right)^{1/2},$$

where the last step follows by applying the Cauchy–Schwarz inequality in $L^2(\widehat{\pi})$. ■

3. Indeed, for any smooth vector field A, integration by parts gives $\int_{\mathbb{R}^d} r\,\nabla\cdot A\,dx = -\int_{\mathbb{R}^d}\langle \nabla r, A\rangle\,dx$ provided the boundary term $\lim_{R\to\infty}\int_{\partial B_R} r\,\langle A, n\rangle\,dS$ vanishes. Applying this identity first with $A = \widehat{\pi}\nabla r$, and then with $A = e\widehat{\pi}$ yields $0 = \int r\,\nabla\cdot(\widehat{\pi}\nabla r)\,dx - \int r\,\nabla\cdot(e\widehat{\pi})\,dx = -\int\|\nabla r\|_2^2\,d\widehat{\pi} + \int\langle e,\nabla r\rangle\,d\widehat{\pi}$.