# Posterior Sampling by Combining Diffusion Models with Annealed Langevin Dynamics

**Zhiyang Xun**
UT Austin
zxun@cs.utexas.edu

**Shivam Gupta**
UT Austin*
shivamgupta@utexas.edu

**Eric Price**
UT Austin & Microsoft Research
ecprice@cs.utexas.edu

## Abstract

Given a noisy linear measurement $y = Ax + \xi$ of a distribution $p(x)$, and a good approximation to the prior $p(x)$, when can we sample from the posterior $p(x \mid y)$? Posterior sampling provides an accurate and fair framework for tasks such as inpainting, deblurring, and MRI reconstruction, and several heuristics attempt to approximate it. Unfortunately, approximate posterior sampling is computationally intractable in general.

To sidestep this hardness, we focus on (local or global) log-concave distributions $p(x)$. In this regime, Langevin dynamics yields posterior samples when the exact scores of $p(x)$ are available, but it is brittle to score–estimation error, requiring an MGF bound (sub-exponential error). By contrast, in the unconditional setting, diffusion models succeed with only an $L^2$ bound on the score error. We prove that combining diffusion models with an *annealed* variant of Langevin dynamics achieves conditional sampling in polynomial time using merely an $L^4$ bound on the score error.

## 1 Introduction

Diffusion models are currently the leading approach to generative modeling of images. Diffusion models are based on learning the "smoothed scores" $s_{\sigma^2}(x)$ of the modeled distribution $p(x)$. Such scores can be approximated from samples of $p(x)$ by optimizing the score matching objective [HD05]; and given good $L^2$-approximations to the scores, $p(x)$ can be efficiently sampled using an SDE [SE19, HJA20, SSDK$^+$21] or an ODE [SME20].

Much of the promise of generative modeling lies in the prospect of applying the modeled $p(x)$ as a *prior*: combining it with some other information $y$ to perform a search over the manifold of plausible images. Many applications, including MRI reconstruction, deblurring, and inpainting, can be formulated as *linear measurements*

$$y = Ax + \xi \qquad \text{for} \qquad \xi \sim \mathcal{N}(0, \eta^2 I_m) \tag{1}$$

for some (known) matrix $A \in \mathbb{R}^{m \times d}$. *Posterior sampling*, or sampling from $p(x \mid y)$, is a natural and useful goal. When aiming to reconstruct $x$ accurately, it is 2-competitive with the optimal in any metric [JAD$^+$21] and satisfies fairness guarantees with respect to protected classes [JKH$^+$21].

Researchers have developed a number of heuristics to approximate posterior sampling using the smoothed scores, including DPS [CKM$^+$23], particle filtering methods [WTN$^+$23, DS24], DiffPIR [ZZL$^+$23], and second-order approximations [RCK$^+$24]. Unfortunately, unlike for unconditional sampling, these methods do not converge efficiently and robustly to the posterior distribution. In fact, a lower bound shows that *no* algorithm exists for efficient and robust posterior sampling in

*Now at Google DeepMind.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

general [GJP+24]. But the lower bound uses an adversarial, bizarre distribution $p(x)$ based on one-way functions; actual image manifolds are likely much better behaved. Can we find an algorithm for provably efficient, robust posterior sampling for relatively nice distributions $p$? That is the goal of this paper: we describe conditions on $p$ under which efficient, robust posterior sampling is possible.

A close relative to diffusion model sampling is *Langevin dynamics*, which is a different method for sampling that uses an SDE involving the *unsmoothed* score $s_0$. Unlike diffusion, Langevin dynamics is in general *slow* and *not robust* to errors in approximating the score. To be efficient, Langevin dynamics needs stronger conditions, like that $p(x)$ is log-concave and that the score estimation error satisfies an MGF bound (meaning that large errors are exponentially unlikely).

However, Langevin dynamics adapts very well to posterior sampling: it works for posterior sampling under exactly the same conditions as it does for unconditional sampling. The difference from diffusion models is that the *unsmoothed* conditional score $s_0(x \mid y)$ can be computed from the unconditional score $s_0(x)$ and the explicit measurement model $p(y \mid x)$, while the *smoothed* conditional score (which diffusion needs) cannot be easily computed.

So the current state is: diffusion models are efficient and robust for unconditional sampling, but essentially always inaccurate or inefficient for posterior sampling. No algorithm for posterior sampling is efficient and robust in general. Langevin dynamics is efficient for log-concave distributions, but still not robust. Can we make a robust algorithm for this case?

*Can we do posterior sampling with* log-concave $p(x)$ *and* $L^p$*-accurate scores?*

### 1.1 Our Results

Our first result answers this in the affirmative. Algorithm 1 uses a diffusion model for initialization, followed by an *annealed* version of Langevin dynamics, to do posterior sampling for log-concave $p(x)$ with just $L^4$-accurate scores. Annealing is necessary here; see Section F for why standard Langevin dynamics would not suffice in this setting.

**Assumption 1** ($L^4$ score accuracy)**.** *The score estimates* $\widehat{s}_{\sigma^2}(x)$ *of the smoothed distributions* $p_{\sigma^2}(x) = p(x) * \mathcal{N}(0, \sigma^2 I_d)$ *have finite* $L^4$ *error, i.e.,*

$$\mathop{\mathbb{E}}_{p_{\sigma^2}(x)}[\|\widehat{s}_{\sigma^2}(x) - s_{\sigma^2}(x)\|^4] \le \varepsilon_{score}^4 < \infty.$$

**Theorem 1.1** (Posterior sampling with global log-concavity)**.** *Let* $p(x)$ *be an* $\alpha$*-strongly log-concave distribution over* $\mathbb{R}^d$ *with* $L$*-Lipschitz score. For any* $0 < \varepsilon < 1$*, there exist* $K_1 = \mathrm{poly}(d, m, \frac{\|A\|}{\eta\sqrt{\alpha}}, \frac{1}{\varepsilon})$ *and* $K_2 = \mathrm{poly}(d, m, \frac{\|A\|}{\eta\sqrt{\alpha}}, \frac{1}{\varepsilon}, \frac{L}{\alpha})$ *such that: if* $\varepsilon_{score} \le \frac{\sqrt{\alpha}}{K_1}$*, then there exists an algorithm that takes* $K_2$ *iterations to sample from a distribution* $\widehat{p}(x \mid y)$ *with*

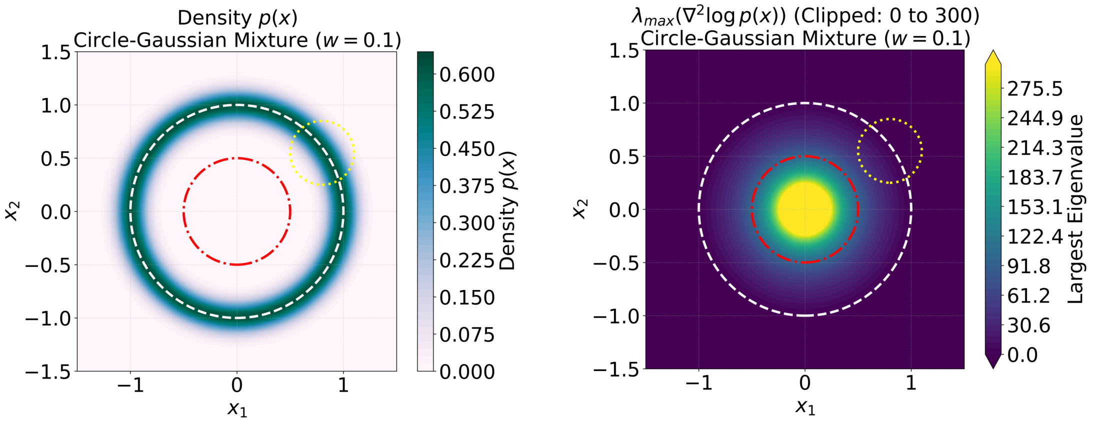
$$\mathbb{E}\left[\mathrm{TV}(\widehat{p}(x \mid y), p(x \mid y))\right] \le \varepsilon.$$

For precise bounds on the polynomials, see Theorem E.6. To understand the parameters, $\frac{\|A\|}{\eta\sqrt{\alpha}}$ should be viewed as the signal-to-noise ratio of the measurement.

**Local log-concavity.** Global log-concavity, as required by Theorem 1.1, is simple to state but a fairly strong condition. In fact, Algorithm 1 only needs a *local* log-concavity condition.

As motivation, consider MRI reconstruction. Given the MRI measurement $y$ of $x$, we would like to get as accurate an estimate $\widehat{x}$ of $x$ as possible. We expect the image distribution $p(x)$ to concentrate around a low-dimensional manifold. We also know that existing compressed sensing methods (e.g., the LASSO [Tib96, CRT06]) can give a fairly accurate reconstruction $x_0$; not *as* accurate as we are hoping to achieve with the full power of our diffusion model for $p(x)$, but still pretty good. Then *conditioned on* $x_0$, we know basically where $x$ lies on the manifold; if the manifold is well behaved, we only really need to do posterior sampling on a single branch of the manifold. The posterior distribution on this branch can be log-concave even when the overall $p(x)$ is not.

In the theorem below, we suppose we are given a Gaussian measurement $x_0 = x + \mathcal{N}(0, \sigma^2 I_d)$ for some $\sigma$, and that the distribution $p$ is nearly log-concave in a ball polynomially larger than $\sigma$. We can then converge to $p(x \mid x_0, y)$.



(a) Density of $p$, the uniform distribution over the unit circle (white), convolved with $\mathcal{N}(0, w^2 I_2)$.

(b) $\lambda_{\max}(\nabla^2 \log p(x))$ reaches $\Omega(1/w^4)$ near the center, demonstrating strong non-log-concavity.

Figure 1: A "locally nearly log-concave" distribution suitable for Theorem 1.2: uniform on the unit circle plus $\mathcal{N}(0, w^2 I_2)$. The Hessian's largest eigenvalue is much smaller near the bulk of the density than it is globally. Specifically, for $\|A\| w/\eta = O(1)$, a Gaussian measurement $\tilde{x}$ with $\sigma \le cw$ and $\varepsilon_{\text{score}} \le cw^{-1}$ for small enough $c > 0$ enables sampling from $p(x \mid y, \tilde{x})$.

**Theorem 1.2** (Posterior sampling with local log-concavity)**.** *For any $\varepsilon, \tau, R, L > 0$, suppose $p(x)$ is a distribution over $\mathbb{R}^d$ such that*

$$\Pr_{x' \sim p}\left[\forall x \in B(x', R) : -L I_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2) I_d\right] \ge 1 - \varepsilon.$$

*Then, there exist $K_1, K_2 = \operatorname{poly}(d, m, \frac{\|A\|\sigma}{\eta}, \frac{1}{\varepsilon})$ and $K_3 = \operatorname{poly}(d, m, \frac{\|A\|\sigma}{\eta}, \frac{1}{\varepsilon}, L\sigma^2)$ such that: Given a Gaussian measurement $x_0 = x + \mathcal{N}(0, \sigma^2 I_d)$ of $x \sim p$ with $\sigma \le R/(K_1 + 2\tau)$. If $\varepsilon_{score} \le \frac{1}{K_2 \sigma}$, then there exists an algorithm that takes $K_3$ iterations to sample from a distribution $\widehat{p}(x \mid x_0, y)$ such that*

$$\mathop{\mathbb{E}}_{y, x_0}\left[\operatorname{TV}(\widehat{p}(x \mid x_0, y), p(x \mid x_0, y))\right] \lesssim \varepsilon.$$

If $p$ is globally log-concave, we can set $\sigma = \infty$ so $x_0$ is independent of $x$ and recover Theorem 1.1; but if we have local information then this just needs local log-concavity. For precise bounds and a detailed discussion of the algorithm, see Section E.2.

The largest eigenvalue of $\nabla^2 \log p(x)$ quantifies the extent to which the distribution departs from log-concavity at a given point. In Figure 1, we show an instance of a locally nearly log-concave distribution: $x$ is uniformly on the unit circle plus $\mathcal{N}(0, w^2 I_2)$. This distribution is very far from globally log-concave, but it is nearly log-concave within a $w$-width band of the unit circle. See Section E.4 for details.

**Compressed Sensing.** In compressed sensing, one would like to estimate $x$ as accurately as possible from $y$. There are many algorithms under many different structural assumptions on $x$, most

Table 1: Summary of theorems and corresponding algorithms.

| Theorem | Setting | Method | Target |
|---|---|---|---|
| Theorem 1.1 | Global log-concavity | Algorithm 1 | $p(x \mid y)$ |
| Theorem 1.2 | Local log-concavity with a Gaussian measurement $x_0$ | Run Algorithm 1 using $p(x \mid x_0)$ as the prior (Algorithm 2) | $p(x \mid x_0, y)$ |
| Corollary 1.3 | Local log-concavity with an arbitrary noisy measurement $x_0$ | Run Algorithm 2 but replace $x_0$ with $x_0' = x_0 + \mathcal{N}(0, \sigma^2 I_d)$ (Algorithm 3) | small $\|x - x_0\|$ |

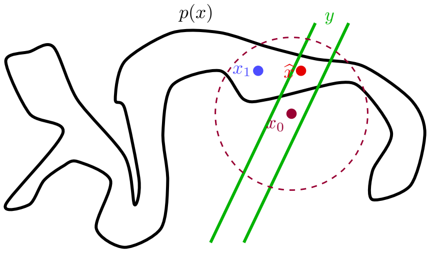


Figure 2: Corollary 1.3 sampling process. Given the distribution $p(x)$ and measurement $y$, we (1) start with a warm start estimate $x_0$, which may not lie on the effective manifold containing $p(x)$; (2) use the diffusion process to sample from $p(x)$ in a ball around $x_0$, getting $x_1$ on the manifold but not matching $y$; and finally (3) use annealed Langevin dynamics to converge to $p(x \mid y)$. This works if $p(x)$ is *locally* close to log-concave, even if it is *globally* complicated. See Section E.3 for a more detailed discussion.

notably the LASSO if $x$ is known to be approximately sparse [Tib96, CRT06]. The LASSO does not use much information about the structure of $p(x)$, and one can hope for significant improvements when $p(x)$ is known. Posterior sampling is known to be near-optimal for compressed sensing: if any algorithm achieves $r$ error with probability $1-\delta$, then posterior sampling achieves at most $2r$ error with probability $1-2\delta$. But, as we discuss above, posterior sampling cannot be efficiently computed in general.

We can use Theorem 1.2 to construct a competitive compressed sensing algorithm under a "local" log-concavity condition on $p$. Suppose we have a naive compressed sensing algorithm (e.g., the LASSO) that recovers the true $x$ to within $R$ error; and $p$ is usually log-concave within an $R \cdot \operatorname{poly}$ ball; then if *any exponential time* algorithm can get $r$ error from $y$, our algorithm gets $2r$ error in polynomial time.

**Corollary 1.3** (Competitive compressed sensing)**.** *Consider attempting to accurately reconstruct $x$ from $y = Ax + \xi$. Suppose that:*

- *Information theoretically (but possibly requiring exponential time or using exact knowledge of $p(x)$), it is possible to recover $\widehat{x}$ from $y$ satisfying $\|\widehat{x} - x\| \leq r$ with probability $1-\delta$ over $x \sim p$ and $y$.*
- *We have access to a "naive" algorithm that recovers $x_0$ from $y$ satisfying $\|x_0 - x\| \leq R$ with probability $1-\delta$ over $x \sim p$ and $y$.*
- *For $R' = R \cdot \operatorname{poly}(d, m, \frac{\|A\|R}{\eta}, \frac{1}{\delta})$,*
$$\Pr_{x' \sim p}\left[\forall x \in B(x', R') : -LI_d \preceq \nabla^2 \log p(x) \preceq 0\right] \geq 1 - \delta.$$

*Then we give an algorithm that recovers $\widehat{x}$ satisfying $\|\widehat{x} - x\| \leq 2r$ with probability $1 - O(\delta)$, in $\operatorname{poly}(d, m, \frac{\|A\|R}{\eta}, \frac{1}{\delta})$ time, under Assumption 1 with $\varepsilon_{score} < \frac{1}{\operatorname{poly}(d, m, \frac{\|A\|R}{\eta}, \frac{1}{\delta}, LR^2)R}$.*

That is, we can go from a decent warm start to a near-optimal reconstruction, so long as the distribution is locally log-concave, with radius of locality depending on how accurate our warm start is. To our knowledge this is the first known guarantee of this kind. Per the lower bound [GJP$^+$24], such a guarantee would be impossible without any warm start or other assumption.

Figure 2 illustrates the sampling process of Corollary 1.3. The initial estimate $x_0$ may lie well outside the bulk of $p(x)$; with just an $L^4$ error bound, the unsmoothed score at $x_0$ could be extremely bad. We add a bit of spherical Gaussian noise to $x_0$, then treat this as a *spherical* Gaussian measurement of $x$, i.e., $x + \mathcal{N}(0, RI)$; for spherical Gaussian measurements, the posterior $p(x \mid x_0)$ *can* be sampled robustly and efficiently using the diffusion SDE. We take such a sample $x_1$, which now won't be too far outside the distribution of $p(x)$, then use $x_1$ as initialization for annealed Langevin dynamics to sample from $p(x \mid y)$. The key part of our paper is that this process will never evaluate a score with respect to a distribution far from the distribution it was trained on, so the process is robust to error in the score estimates.

**Algorithm 1** Sampling from $p(x \mid Ax + \mathcal{N}(0, \eta^2 I_m) = y)$

```
function PosteriorSampler($p : \mathbb{R}^d \to \mathbb{R}$, $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times d}$, $\eta \in \mathbb{R}$)
    Let $\eta_1 > \eta_2 > \cdots > \eta_N = \eta$ and $T_1, \dots, T_{N-1}$ be an admissible schedule.
    Initialize $y_N = y$
    for $i = N - 1$ down to $1$ do
        $y_i = y_{i+1} + \mathcal{N}(0, (\eta_i^2 - \eta_{i+1}^2) I_m)$
    end for
    Sample $X_1 \sim p(x)$                    ▷ Approximately, using the diffusion SDE (5)
    for $i = 1$ to $N - 1$ do
        Let $\widehat{s}_{i+1}$ be the estimated score function for $s_{i+1}(x) = \nabla \log p(x \mid y_{i+1})$.
        Initialize $x_0 = X_i$.
        Simulate the SDE for time $T_i$:
            $\mathrm{d}x_t = \widehat{s}_{i+1}(x_t^{(h)})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$        (2)
        Here, $x_t^{(h)} = x_{h \cdot \lfloor t/h \rfloor}$ is the discretized $x_t$, where $h$ is a small enough step size.
        Set $X_{i+1} \leftarrow x_{T_i}$
    end for
    Return: $X_N$ as an approximation of $p(x \mid Ax + \mathcal{N}(0, \eta^2 I_m) = y)$.
end function
```

We summarize our results in Table 1.

## 2 Notation and Background

We consider $x \sim p(x)$ over $\mathbb{R}^d$. The "score function" $s(x)$ of $p$ is $\nabla \log p(x)$. The "smoothed score function" $s_{\sigma^2}(x)$ is the score of $p_{\sigma^2}(x) = p(x) * \mathcal{N}(0, \sigma^2 I_d)$.

**Unconditional sampling.** There are several ways to sample from $p$ using the scores. Langevin dynamics is a classical MCMC method that considers the following overdamped Langevin Stochastic Differential Equation (SDE):

$$dX_t = s(X_t)dt + \sqrt{2}dB_t, \tag{3}$$

where $B_t$ is standard Brownian motion. The stationary distribution of this SDE is $p$, and discretized versions of it, such as the Unadjusted Langevin Algorithm (ULA), are known to converge rapidly to $p(x)$ when $p(x)$ is strongly log-concave [Dal17]. One can replace the true score $s(x)$ with an approximation $\widehat{s}$, as long as it satisfies a (fairly strong) MGF condition

$$\mathop{\mathbb{E}}_{x \sim p(x)} \left[ \exp\left( \|s(x) - \widehat{s}(x)\|^2 / \varepsilon_{mgf}^2 \right) \right] < \infty, \quad \text{for some } \varepsilon_{mgf} > 0. \tag{4}$$

In particular, [YW22] showed that Langevin dynamics needs an MGF bound for convergence, and an $L^p$-accurate score estimator for any $1 \le p < \infty$ is insufficient.

An alternative approach, used by diffusion models, is to involve the smoothed scores. Starting from $x_0 \sim \mathcal{N}(0, I_d)$, one can follow a different SDE [And82]:

$$dX_t = (X_t + 2s_{\sigma_t^2}(X_t))dt + \sqrt{2}dB_t \tag{5}$$

for a particular smoothing schedule $\sigma_t$; the result $x_T$ is exponentially close (in $T$) to being drawn from $p(x)$. This also has efficient discretizations [CCL+22, CCSW22, BBDD24], does not require log-concavity, and only requires an $L^2$ guarantee such as [CCL+22]

$$\mathop{\mathbb{E}}_{x \sim p_{\sigma^2}(x)} \left[ \|s_{\sigma^2}(x) - \widehat{s}_{\sigma^2}(x)\|^2 \right] < \varepsilon^2$$

to accurately sample from $p(x)$. One can also run a similar ODE with similar guarantees but faster [CCL+23].

**Posterior sampling.** Now, in this paper we are concerned with *posterior sampling*: we observe a noisy linear measurement $y \in \mathbb{R}^m$ of $x$, given by

$$y = Ax + \xi \qquad \text{for} \qquad \xi \sim \mathcal{N}(0, \eta^2 I_m),$$

and want to sample from $p(x \mid y)$. The unsmoothed score $s_y(x) := \nabla_x \log p(x \mid y)$ is easily computed by Bayes' rule:

$$\nabla_x \log p(x \mid y) = \nabla_x \log p(x) + \nabla_x \log p(y \mid x) = s(x) + \frac{A^\top (y - Ax)}{\eta^2}.$$

Thus we can run the Langevin SDE (3) with the same properties: if $p(x \mid y)$ is strongly log-concave and the score estimate satisfies the MGF error bound (4), it will converge quickly and accurately.

Naturally, researchers have looked to diffusion processes for more general and robust posterior sampling methods. The main difficulty is that the smoothed score of the posterior involves $\nabla_x \log p(y \mid x_{\sigma_t^2})$ rather than the tractable unsmoothed term $\nabla_x \log p(y \mid x)$. Because the smoothed score is hard to evaluate exactly, a range of approximation techniques has been proposed [BGP+24, CKM+23, MK25, RCK+24, SVMK23, WYZ23]. One prominent example is the DPS algorithm [CKM+23]. Other methods include Monte Carlo/MCMC-inspired approximations [CJeILCM24, DS24, WSC+24, EKZL25], singular value decomposition and transport tilting [KVE21, KSEE22, WYZ23, BH24], and schemes that combine corrector steps with standard diffusion updates [CL23, CY22, CSRY22, KBBW23, LKA+24, SKZ+24, SSXE22, ZZL+23, AVTT21, XC24, RLdB+24, RRD+23]. These approaches have shown strong empirical performance, and several provide guarantees under additional structure of the linear measurement; however, general guarantees for fast and robust posterior sampling remain limited beyond these restricted regimes.

Several recent studies [JAD+21, ZCB+25, KVE21] use various annealed versions of the Langevin SDE as a key component in their diffusion-based posterior sampling method and achieve strong empirical results. Still, these methods provide no theoretical guidance on two key aspects: how to design the annealing schedule and why annealing improves robustness. None of these approaches come with correctness guarantees for the overall sampling procedure.

**Comparison with Computational Lower Bounds.** Recent work of [GJP+24] shows that it is actually *impossible* to achieve a general algorithm that is guaranteed fast and robust: there is an exponential computational gap between unconditional diffusion and posterior sampling. Under standard cryptographic assumptions, they construct a distribution $p$ over $\mathbb{R}^d$ such that

1. One can efficiently obtain an $L^p$-accurate estimate of the smoothed score of $p$, so diffusion models can sample from $p$.
2. Any sub-exponential time algorithm that takes $y = Ax + \mathcal{N}(0, \eta^2 I_m)$ as input and outputs a sample from the posterior $p(x \mid y)$ fails on most $y$ with high probability.

Our algorithm shows that, once an additional noisy observation $\tilde{x}$ that is close to $x$ is provided, then we can efficiently sample from $p(x \mid y, \tilde{x})$, circumventing the impossibility result.

To illustrate why the extra observation helps, consider the following simplified version of the hardness instance:

$$p := q * \mathcal{N}(0, \sigma^2 I_d), \quad q(x) := \frac{1}{2^{d/2}} \sum_{s \in \{0,1\}^{d/2}} \delta((s, f(s)) - x).$$

Here, $f : \{0,1\}^{d/2} \to \{0,1\}^{d/2}$ is a one-way permutation — it takes exponential time to compute $f^{-1}(x)$ for most $x \in \{0,1\}^{d/2}$. $\delta(\cdot)$ is the Dirac delta function, and we choose $\sigma \ll d^{-1/2}$. Thus, $p(x)$ is a mixture of $2^{d/2}$ well-separated Gaussians centered at the points $(s, f(s))$.

Assume we observe

$$y = Ax + \mathcal{N}(0, \eta^2 I_d), \quad A = \begin{pmatrix} 0 & I_{d/2} \end{pmatrix}, \quad \sigma \ll \eta \ll d^{-1/2},$$

and let $\mathrm{rnd}(y)$ denote the vertex of $\{0,1\}^d$ closest to $y$. Then the posterior $p(x \mid y)$ is approximately a Gaussian centered at $(f^{-1}(\mathrm{rnd}(y)), \mathrm{rnd}(y))$ with covariance $\sigma^2 I_d$. Generating a single sample would therefore reveal $f^{-1}(\mathrm{rnd}(y))$, which requires $\exp(\Omega(d))$ time.

However, suppose we have a coarse estimate $x_0$ satisfying $\|x_0 - x\| < 1/3$ (e.g., obtained by compressed sensing). Then, $x_0$ uniquely identifies the correct $(s, f(s))$ with $f(s) = \mathrm{rnd}(y)$, and the remaining task is just sampling from a Gaussian. Therefore, this hard instance becomes easy once we have localized the task and does not contradict our Theorem 1.2.

We are able to handle the hard instance above well because it is exactly the type of distribution our approach is designed for: despite its complex global structure, it exhibits well-behaved local properties. This gives an important conceptual takeaway from our work: the hardness of posterior sampling may only lie in localizing $x$ within the exponentially large high-dimensional space.

Therefore, although posterior sampling is an intractable task in general, it is still possible to design a robust, provably correct posterior sampling algorithm — once we have localized the distribution. We view our work as a first step towards this goal.

# 3 Techniques

The algorithm we propose is clean and simple, but the proof is quite involved. Before we dive into the details, we provide a high-level overview of the intuitions behind the algorithm, concentrating on the illustrative case where the *prior* density $p(x)$ is $\alpha$-strongly log-concave. Under this assumption, every posterior density $p(x \mid y)$ is *also* $\alpha$-strongly log-concave. Therefore, posterior sampling could, in principle, be performed using classical Langevin dynamics.

The challenge arises because we lack access to the exact posterior score $s_y(x)$. We only possess an estimator derived from an estimate $\widehat{s}(x)$ of the *prior* score $s(x)$:

$$\widehat{s}_y(x) \;:=\; \widehat{s}(x) \;+\; \frac{A^\top(y - Ax)}{\eta^2}.$$

Assumption 1 implies an $L^4$ accuracy of $\widehat{s}_y$ on average, but how do we use this to support Langevin dynamics, which demands exponentially decaying error tails?

## 3.1 Score Accuracy: Langevin Dynamics vs. Diffusion Models

> *Why can diffusion models succeed with merely $L^2$-accurate scores, whereas Langevin dynamics require MGF accuracy?*

Both diffusion models and Langevin dynamics utilize SDEs. The $L^2$ error in the score-dependent drift term relates directly to the KL divergence between the true process (using $s(x)$) and the estimated process (using $\widehat{s}(x)$). Consequently, bounding the $L^2$ score error with respect to the current distribution $\widehat{p}_t$ controls the KL divergence.

Diffusion models leverage this property effectively. The forward process transforms data into a Gaussian, and the reverse generative process starts exactly from this Gaussian. At any time $t$, suppose $\widehat{p}_t$ is close to $p_{\sigma_t^2}$, then

$$\mathop{\mathbb{E}}_{x_t \sim \widehat{p}_t}[\|s_{\sigma_t^2}(x_t) - \widehat{s}_{\sigma_t^2}(x_t)\|^2] \approx \mathop{\mathbb{E}}_{x_t \sim p_{\sigma_t^2}}[\|s_{\sigma_t^2}(x_t) - \widehat{s}_{\sigma_t^2}(x_t)\|^2] \le \varepsilon_{\text{score}}^2$$

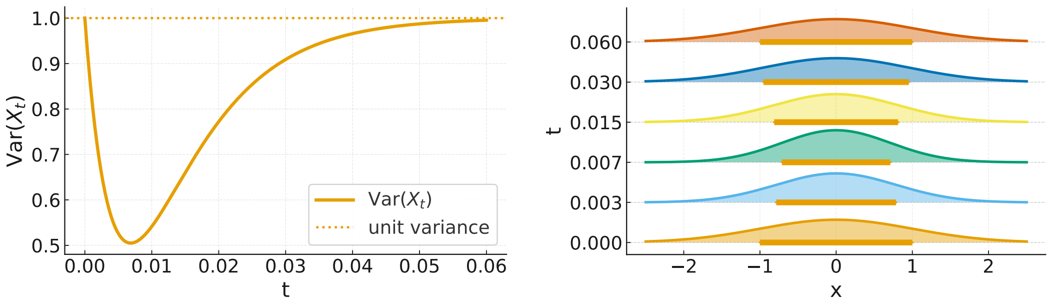
by the $L^2$ accuracy assumption. This keeps the process *close* to the ideal process, ensuring overall small error.

Langevin dynamics, by contrast, often starts from an arbitrary, not predefined initial distribution $p_{\text{initial}}$. An $L^p$ score accuracy guarantee with respect to $p_{\text{target}}$ alone does not ensure accuracy for points $x_t$ that are not on the distributional manifold of $p_{\text{target}}$ (consider running Langevin starting from $x_0$ in Figure 2). Therefore, a stronger MGF error bound is needed to prevent this from happening.

## 3.2 Adapting Langevin Dynamics for Posterior Sampling

While we can only use Langevin-type dynamics for posterior sampling, we possess a source of effective starting points: we can sample $x_0 \sim p(x)$ efficiently using the unconditional diffusion model. Intuitively, $x_0$ already lies on the data manifold. The score estimator $\widehat{s}_y(x)$ initially satisfies:

$$\mathop{\mathbb{E}}_{x_0 \sim p(x)}[\|s_y(x_0) - \widehat{s}_y(x_0)\|^2] = \mathop{\mathbb{E}}_{x_0 \sim p(x)}[\|s(x_0) - \widehat{s}(x_0)\|^2] \le \varepsilon_{\text{score}}^2.$$



(a) $\mathrm{Var}(X_t)$ as a function of $t$.

(b) $X_t \sim \mathcal{N}(0, \mathrm{Var}(X_t))$ at different times $t$.

Figure 3: Let $p = \mathcal{N}(0,1)$ and $y = x + \mathcal{N}(0, 0.01)$. Starting from $X_0 \sim p$, run the Langevin SDE $\mathrm{d}X_t = s_y(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$. Averaging over $y$, the marginal of $X_t$ remains Gaussian; its variance first contracts and then returns toward the prior. There is an intermediate time $t^*$ where $X_{t^*}$ has a constant factor lower variance; in high dimensions, this means $X_{t^*}$ is concentrated on an exponentially small region of $p$, so an $L^p$ bound on score error under $p$ does not effectively control the error under $X_{t^*}$. See Section F for details.

As the dynamics evolves, the distribution $p(x_t)$ transitions from $p(x)$ towards $p(x \mid y)$. If $x_t$ converges to $p(x \mid y)$, we again expect reasonable accuracy on average:

$$\mathop{\mathbb{E}}_{y}[\mathop{\mathbb{E}}_{x_t \sim p(x|y)}[\|s_y(x_t) - \widehat{s}_y(x_t)\|^2]] = \mathop{\mathbb{E}}_{y}[\mathop{\mathbb{E}}_{x_t \sim p(x|y)}[\|s(x_t) - \widehat{s}(x_t)\|^2]] \le \varepsilon_{\text{score}}^2.$$

Hence the estimator is accurate at the start and at convergence. The open question concerns the *intermediate* segment of the trajectory: does $x_t$ wander into regions where the prior score $\widehat{s}(x)$ is unreliable? Ideally, the time-marginal of $x_t$, averaged over $y$, remains close to $p(x)$ throughout.

### 3.3 Annealing via Mixing Steps

In fact, even though $x_0$ and $x_\infty$ both have marginal $p(x)$, so the score estimate $\widehat{s}(x)$ is accurate on average at those times, this is *not* true at intermediate times. In Figure 3, we illustrate this with a simple Gaussian example: $x_0$ and $x_\infty$ have distribution $\mathcal{N}(0, I)$ while $x_t$ has marginal $\mathcal{N}(0, cI)$ for a constant $c < 1$. An $L^p$ error bound under $x \sim \mathcal{N}(0, I)$ does not give an $L^2$ error bound under $x \sim \mathcal{N}(0, cI)$, which means Langevin dynamics may not converge to the right distribution. A very strong accuracy guarantee like the MGF bound is needed here.

However, consider the case where the target posterior $p(x \mid y)$ is very close to the initial prior $p(x)$, such as when the measurement noise $\eta$ is very large (low signal-to-noise ratio). Langevin dynamics between close distributions typically converges rapidly. This suggests a key insight: if the required convergence time $T$ is short, the process $x_t$ might not deviate substantially from its initial distribution $p(x_0)$. In such short-time regimes, an $L^2$ score error bound relative to $p(x_0)$ could potentially suffice to control the dynamics. While $p(x)$ itself is already a good approximation for $p(x \mid y)$ when $\eta$ is very large, this motivates a general strategy.

Instead of a single, potentially long Langevin run from $p(x)$ to $p(x \mid y)$, we introduce an annealing scheme using multiple *mixing steps*. Given the measurement parameters $(A, \eta, y)$, we construct a decreasing noise schedule $\eta_1 > \eta_2 > \cdots > \eta_N = \eta$. Correspondingly, we generate a sequence of auxiliary measurements $y_1, y_2, \ldots, y_N = y$ such that each $y_i$ is distributed as $Ax + \mathcal{N}(0, \eta_i^2 I_m)$ and $y_i$ is appropriately coupled to $y_{i+1}$ (specifically, $y_i \sim \mathcal{N}(y_{i+1}, (\eta_i^2 - \eta_{i+1}^2) I_m)$ conditional on $y_{i+1}$). This creates a sequence of intermediate posterior distributions $p(x \mid y_i)$.

An *admissible schedule* (formally defined in Definition D.1) ensures that:

- $\eta_1$ is sufficiently large, making $p(x \mid y_1)$ close to the prior $p(x)$.
- Consecutive $\eta_i$ and $\eta_{i+1}$ are sufficiently close, making $p(x \mid y_i)$ close to $p(x \mid y_{i+1})$.

Our algorithm proceeds as follows:

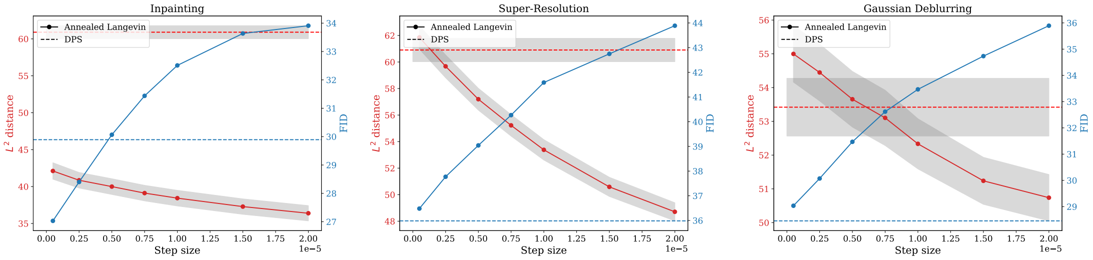


Figure 4: For each of the three settings (inpainting, super-resolution, and Gaussian deblurring), we plot the $L^2$ distance between samples obtained by our annealed Langevin method and the ground truth samples in red. We plot the FID of the distribution obtained by running annealed Langevin in blue. We plot the baseline $L^2$ distance and FID for samples obtained by the DPS algorithm using red and blue dashed lines.

1. Start with a sample $X_0 \sim p(x)$. Since $\eta_1$ is large, $p(x)$ is close to $p(x \mid y_1)$, so $X_0$ serves as an approximate sample $X_1 \sim \widehat{p}(x \mid y_1)$.
2. For $i = 1$ to $N - 1$: Run Langevin dynamics for a short time $T_i$, starting from the previous sample $X_i \sim \widehat{p}(x \mid y_i)$, targeting the next posterior $p(x \mid y_{i+1})$ using the score $\widehat{s}_{y_{i+1}}(x)$. Let the result be $X_{i+1} \sim \widehat{p}(x \mid y_{i+1})$.
3. The final sample $X_N \sim \widehat{p}(x \mid y_N)$ approximates a draw from the target posterior $p(x \mid y)$.

The core idea behind this annealing scheme is to actively control the process distribution $p(x_t)$, ensuring it remains on the manifold of the prior $p(x)$. By design, each mixing step $i \to i + 1$ connects two statistically close intermediate posteriors, $p(x \mid y_i)$ and $p(x \mid y_{i+1})$. This closeness guarantees that a short Langevin run $T_i$ can mix them, and this short duration prevents $p(x_t)$ from drifting significantly away from the step's starting distribution $\widehat{p}(x \mid y_i)$, and we can then argue that

$$\mathop{\mathbb{E}}_{y_i}\Big[\mathop{\mathbb{E}}_{x_t \sim \widehat{p}(x|y_i)}[\|s_{y_i}(x_t) - \widehat{s}_{y_i}(x_t)\|^2]\Big] \approx \mathop{\mathbb{E}}_{y_i}\Big[\mathop{\mathbb{E}}_{x_t \sim p(x|y_i)}[\|s(x_t) - \widehat{s}(x_t)\|^2]\Big] \le \varepsilon_{\text{score}}^2.$$

This contrasts fundamentally with a single long Langevin run, where $x_t$ could venture far "off-manifold" into regions of poor score accuracy. By inserting frequent checkpoints that re-anchor the process, our annealing method substitutes such strong assumptions with structural control: the frequent "checkpoints" $p(x \mid y_i)$ ensure the process is repeatedly localized to regions where the $L^4$ accuracy suffices. While error is incurred in each step, maintaining proximity to the manifold keeps this error small. The overall approach hinges on demonstrating that these small, per-step errors accumulate controllably across all $N$ steps.

This strategy, however, requires rigorous analysis of three key technical challenges:

1. How to bound the required convergence time $T_i$ for the transition from $p(x \mid y_i)$ to $p(x \mid y_{i+1})$? In particular, what happens when $p$ only has local strong log-concavity?
2. How to bound the error incurred during a single mixing step of duration $T_i$, given the $L^4$ score error assumption on the prior score estimate?
3. How to ensure the total error accumulated across all $N$ mixing steps remains small?

Addressing these questions forms the core of our proof.

**Proof Organization.** In Section A, we show that for globally strongly log-concave distributions $p$, Langevin dynamics converges rapidly from $p(x \mid y_i)$ to $p(x \mid y_{i+1})$. We extend this convergence analysis to locally strongly log-concave distributions in Section B. In Section C, we provide bounds on the errors incurred by score errors and discretization in Langevin dynamics. In Section D, we show how to design the noise schedule to control the accumulated error of the full process. In Section E, we conclude the analysis for Algorithm 1, and apply it to establish the main theorems.

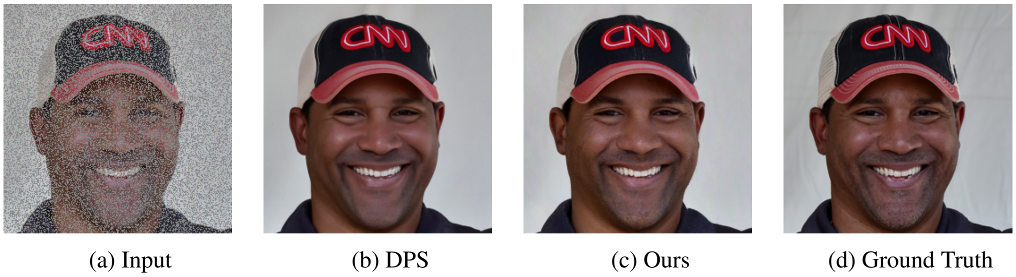

(a) Input (b) DPS (c) Ours (d) Ground Truth

Figure 5: A set of samples for the inpainting task.

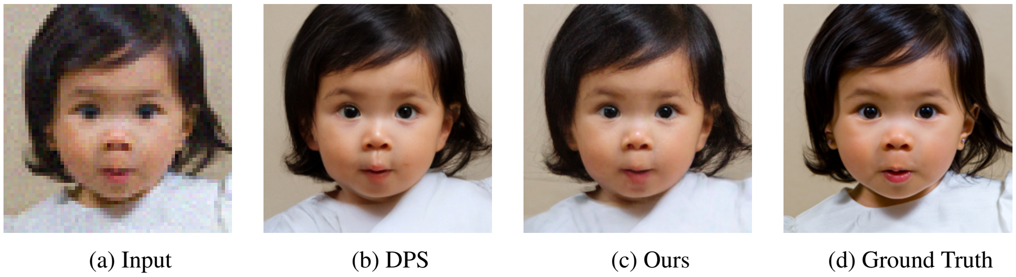

(a) Input (b) DPS (c) Ours (d) Ground Truth

Figure 6: A set of samples for the super-resolution task.

## 4 Experiments

To validate our theoretical analysis and assess real-world performance, we study three inverse problems on FFHQ–256 [KLA21]: inpainting, $4\times$ super-resolution, and Gaussian deblurring. Experiments use 1k validation images and the pre-trained diffusion model from [CKM+23]. Forward operators are specified as in [CKM+23]: inpainting masks $30\%$–$70\%$ of pixels uniformly at random; super-resolution downsamples by a factor of $4$; deblurring convolves the ground-truth with a Gaussian kernel of size $61 \times 61$ (std. 3.0). We first obtain initial reconstructions $x_0$ via Diffusion Posterior Sampling (DPS) [DS24], then refine them with our annealed Langevin sampler to draw samples close to $p(x \mid x_0, y)$. To control runtime, we sweep the step size while keeping the annealing schedule fixed.

For each step size, we report the per-image $L^2$ distance to the ground truth and the FID of the resulting sample distribution (Figure 4). Across all three tasks, increasing the time devoted to annealed Langevin decreases $L^2$ but increases FID; in the inpainting setting, when the step size is sufficiently small, our method surpasses DPS on both metrics. Qualitatively, our reconstructions better preserve ground-truth attributes compared to DPS (Figures 5 and 6). All experiments were run on a cluster with four NVIDIA A100 GPUs and required roughly two hours per task.

## Acknowledgments

This work is supported by the NSF AI Institute for Foundations of Machine Learning (IFML). ZX is supported by NSF Grant CCF-2312573 and a Simons Investigator Award (#409864, David Zuckerman).

## Bibliography

[And82] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[AVTT21] Marius Arvinte, Sriram Vishwanath, Ahmed H. Tewfik, and Jonathan I. Tamir. Deep j-sense: Accelerated mri reconstruction via unrolled alternating optimization. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021),* *Part VI*, volume 12905 of *Lecture Notes in Computer Science*, pages 350–360. Springer, 2021.

[BBDD24] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly $d$-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

[BGP$^+$24] Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and Omer Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*, 2024. TMLR (ICLR 2025 Journal Track).

[BH24] Joan Bruna and Jiequn Han. Provable posterior sampling with denoising oracles via tilted transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[CCL$^+$22] 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. *arXiv preprint arXiv:2209.11215*, 2022.

[CCL$^+$23] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68552–68575. Curran Associates, Inc., 2023.

[CCSW22] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

[CJeILCM24] Gabriel Cardoso, Yazid Janati el Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In *International Conference on Learning Representations (ICLR)*, 2024. Oral.

[CKM$^+$23] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

[CL23] Junqing Chen and Haibo Liu. An alternating direction method of multipliers for inverse lithography problem. *Numerical Mathematics: Theory, Methods and Applications*, 16(3):820–846, 2023.

[CRT06] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[CSRY22] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[CY22] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical Image Analysis*, 80:102479, 2022.

[Dal17] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

[DS24] Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

[EKZL25] Filip Ekström Kelvinius, Zheng Zhao, and Fredrik Lindsten. Solving linear-gaussian bayesian inverse problems with decoupled diffusion sequential monte carlo. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 15148–15181, 2025.

[GJP$^+$24] Shivam Gupta, Ajil Jalal, Aditya Parulekar, Eric Price, and Zhiyang Xun. Diffusion posterior sampling is computationally intractable. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17020–17059. PMLR, 21–27 Jul 2024.

[HD05] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[JAD$^+$21] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. *Advances in Neural Information Processing Systems*, 34:14938–14954, 2021.

[JCP24] Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2720–2721. PMLR, 7 2024.

[JKH$^+$21] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pages 4721–4732. PMLR, 2021.

[KBBW23] Ulugbek S. Kamilov, Charles A. Bouman, Gregery T. Buzzard, and Brendt Wohlberg. Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023.

[KLA21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, December 2021.

[KSEE22] Bahjat Kawar, Jiaming Song, Stefano Ermon, and Michael Elad. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[KVE21] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 21757–21769, 2021.

[LKA$^+$24] Xiang Li, Soo Min Kwon, Ismail R. Alkhouri, Saiprasad Ravishankar, and Qing Qu. Decoupled data consistency with diffusion purification for image restoration. *arXiv preprint arXiv:2403.06054*, 2024.

[LM00] B. Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28, 10 2000.

[MK25] Xiangming Meng and Yoshiyuki Kabashima. Diffusion model based posterior sampling for noisy linear inverse problems. In *Proceedings of the 16th Asian Conference on Machine Learning (ACML)*, volume 260 of *Proceedings of Machine Learning Research*, pages 623–638. PMLR, 2025.

[RCK$^+$24] Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order tweedie: Solving inverse problems using latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9472–9481, 2024.

[RLdB$^+$24] Marien Renaud, Jiaming Liu, Valentin de Bortoli, Andrés Almansa, and Ulugbek S. Kamilov. Plug-and-play posterior sampling under mismatched measurement and prior models. In *International Conference on Learning Representations (ICLR)*, 2024.

[RRD$^+$23] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G. Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. *arXiv preprint arXiv:2307.00619*, 2023.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[SKZ$^+$24] Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *International Conference on Learning Representations (ICLR)*, 2024.

[SME20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 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. In *International Conference on Learning Representations (ICLR)*, 2021.

[SSXE22] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations (ICLR)*, 2022.

[SVMK23] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[WSC$^+$24] Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[WTN$^+$23] Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

[WYZ23] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations (ICLR)*, 2023.

[XC24] Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[YW22] Kaylee Yingxi Yang and Andre Wibisono. Convergence in kl and rényi divergence of the unadjusted langevin algorithm using estimated score. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[ZCB$^+$25] Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[ZZL$^+$23] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023.

# A Langevin Convergence Between Strongly Log-concave Distributions

In this section, we study the following problem. Let $p$ be a probability distribution on $\mathbb{R}^d$, and let $A \in \mathbb{R}^{m \times d}$ be a matrix. For a sequence of parameters $\eta_i > \eta_{i+1}$ satisfying

$$\eta_i^2 = (1+\gamma_i)\eta_{i+1}^2,$$

consider two random variables $y_i$ and $y_{i+1}$ defined as follows. First, draw $x \sim p$. Then, generate

$$y_{i+1} = Ax + N(0, \eta_{i+1}^2 I_m),$$

and further perturb it by

$$y_i = y_{i+1} + N(0, (\eta_i^2 - \eta_{i+1}^2) I_m).$$

Define the score function

$$s_{i+1}(x) = \nabla_x \log p(x \mid y_{i+1}).$$

We analyze the following SDE:

$$\mathrm{d}x_t = s_{i+1}(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad x_0 \sim p(x \mid y_i). \tag{6}$$

This is the ideal (no discretization, no score estimation error) version of the process (2) that we actually run. Our goal is to establish the following lemma.

**Lemma A.1.** *Suppose the prior distribution $p(x)$ is $\alpha$-strongly log-concave. Then, running the process (6) for time*

$$T = O\left(\frac{m\gamma_i + \log(\lambda/\varepsilon)}{\alpha}\right)$$

*ensures that*

$$\Pr_{y_i, y_{i+1}}\left[\mathrm{TV}(x_T, p(x \mid y_{i+1})) \le \varepsilon\right] \ge 1 - \frac{1}{\lambda}.$$

## A.1 $\chi^2$-divergence Between Distributions

In this section, our goal is to bound $\chi^2\left(p(x \mid y_i) \,\|\, p(x \mid y_{i+1})\right)$. Since the posterior distributions can be expressed as

$$p(x \mid y_i) = \frac{p(y_i \mid x)p(x)}{p(y_i)}, \quad p(x \mid y_{i+1}) = \frac{p(y_{i+1} \mid x)p(x)}{p(y_{i+1})}.$$

The $\chi^2$ divergence is

$$\begin{aligned}
\chi^2\left(p(x \mid y_i) \,\|\, p(x \mid y_{i+1})\right) &= \mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(x \mid y_i)}{p(x \mid y_{i+1})}\right] - 1 \\
&= \mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)} \cdot \frac{p(y_{i+1})}{p(y_i)}\right] - 1 \\
&= \mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)}\right] \cdot \frac{p(y_{i+1})}{p(y_i)} - 1.
\end{aligned}$$

We bound the term $\mathbb{E}_{x \sim p(x|y_i)}\left[\frac{p(y_i|x)}{p(y_{i+1}|x)}\right]$ first.

**Lemma A.2.** *We have*

$$\mathop{\mathbb{E}}_{x, y_i, y_{i+1}}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)}\right] = 1$$

*Proof.* Let $Z_1 = y_{i+1} - Ax$, and let $Z_2 = y_i - Ax$. Then we have

$$\mathop{\mathbb{E}}_{x, y_i, y_{i+1}}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)}\right] = \mathop{\mathbb{E}}_{Z_1, Z_2}\left[\frac{p(Z_2)}{p(Z_1)}\right] = \iint \frac{p_{Z_2}(z_2)}{p_{Z_1}(z_1)} \cdot p_{Z_1, Z_2}(z_1, z_2)\,\mathrm{d}\,z_1\,\mathrm{d}\,z_2.$$

Note that

$$p_{Z_1, Z_2}(z_1, z_2) = p_{Z_1}(z_1) \cdot f(z_2 - z_1),$$

where $f$ is the density function for $N(0, (\eta_i^2 - \eta_{i+1}^2)I_m)$. Therefore,

$$\iint \frac{p_{Z_2}(z_2)}{p_{Z_1}(z_1)} \cdot p_{Z_1,Z_2}(z_1, z_2)\,\mathrm{d}\,z_1\,\mathrm{d}\,z_2 = \iint p_{Z_2}(z_2) \cdot f(z_2 - z_1)\,\mathrm{d}\,z_1\,\mathrm{d}\,z_2$$
$$= \int p_{Z_2}(z_2) \left( \int f(z_2 - z_1)\,\mathrm{d}\,z_1 \right) \mathrm{d}\,z_2.$$

Since $f$ is a density function, its integral over $\mathbb{R}^m$ is 1. This gives that

$$\int p_{Z_2}(z_2) \left( \int f(z_2 - z_1)\,\mathrm{d}\,z_1 \right) \mathrm{d}\,z_2 = \int p_{Z_2}(z_2)\,\mathrm{d}\,z_2 = 1.$$

Hence,

$$\mathop{\mathbb{E}}_{x,y_i,y_{i+1}} \left[ \frac{p(y_i \mid x)}{p(y_{i+1} \mid x)} \right] = 1.$$

□

**Corollary A.3.** *For any* $\lambda > 1$*, we have*

$$\Pr_{y_i,y_{i+1}} \left[ \mathop{\mathbb{E}}_{x \sim p(x|y_i)} \left[ \frac{p(y_i \mid x)}{p(y_{i+1} \mid x)} \right] \le \lambda \right] \ge 1 - \frac{1}{\lambda}.$$

*Proof.* By Lemma A.2, we have

$$\mathop{\mathbb{E}}_{y_i,y_{i+1}} \left[ \mathop{\mathbb{E}}_{x \sim p(x|y_i)} \left[ \frac{p(y_i \mid x)}{p(y_{i+1} \mid x)} \right] \right] = \mathop{\mathbb{E}}_{x,y_i,y_{i+1}} \left[ \frac{p(y_i \mid x)}{p(y_{i+1} \mid x)} \right] = 1.$$

Applying Markov's inequality gives the result. □

Now we bound $\frac{p(y_{i+1})}{p(y_i)}$. To make the lemma more self-contained, we abstract this a little bit.

**Lemma A.4.** *Let* $\eta_1 > \eta_2$ *be two positive numbers, and let* $X \in \mathbb{R}^d$ *be an arbitrary random variable. Define* $Y_1 = X + Z_1$ *and* $Y_2 = Y_1 + Z_2$*, where* $Z_1 \sim N(0, \eta_1^2 I_d)$ *and* $Z_2 \sim N(0, \eta_2^2 I_d)$*. Then,*

$$\mathop{\mathbb{E}}_{Y_1,Y_2 | \|Z_1\| \le t} \left[ \frac{p(Y_1)}{p(Y_2)} \right] \le \frac{1}{\Pr[\|Z_1\| \le t]} \cdot \exp\left( O\left( \frac{d\eta_2^2}{\eta_1^2} + \frac{t^2\eta_2^2}{\eta_1^4} \right) \right).$$

*where* $p(Y_1)$ *and* $p(Y_2)$ *are the densities of* $Y_1$ *and* $Y_2$*, respectively.*

*Proof.* First, we turn to bound

$$F_t(Y_1, Y_2) := \frac{p(Y_1)}{p(Y_2)} \cdot \Pr[\|Z_1\| \le t \mid Y_1].$$

Note that

$$F_t(Y_1, Y_2) = \frac{\int_{\|s-Y_1\| \le t} p(X = s) p(Y_1 \mid X = s)\,\mathrm{d}\,s}{\int_{\mathbb{R}^d} p(X = s) p(Y_2 \mid X = s)\,\mathrm{d}\,s} = \frac{\int_{\mathbb{R}^d} p_X(s) \phi_{\eta_1^2 I_d}(Y_1 - s) \cdot \mathbf{1}_{\{\|Y_1 - s\| \le t\}}\,\mathrm{d}\,s}{\int_{\mathbb{R}^d} p_X(s) \phi_{(\eta_1^2+\eta_2^2)I_d}(Y_2 - s)\,\mathrm{d}\,s}.$$

We have

$$F_t(Y_1, Y_2) \le \sup_{s \in \mathbb{R}^d} \frac{\phi_{\eta_1^2 I_d}(Y_1 - s) \cdot \mathbf{1}_{\{\|Y_1 - s\| \le t\}}}{\phi_{(\eta_1^2+\eta_2^2)I_d}(Y_2 - s)} \le \sup_{\|s - Y_1\| \le t} \frac{\phi_{\eta_1^2 I_d}(Y_1 - s)}{\phi_{(\eta_1^2+\eta_2^2)I_d}(Y_2 - s)}.$$

Write $Y_1 - s = e_1$, and note that $Y_2 - s = e_1 + Z_2$. Then define

$$G(e_1) = \frac{\phi_{\eta_1^2 I_d}(e_1)}{\phi_{(\eta_1^2+\eta_2^2)I_d}(e_1 + Z_2)}, \qquad \|e_1\| \le t.$$

This gives that for any $Y_1$, $Y_2$, and $t$,

$$F_t(Y_1, Y_2) \le \sup_{\|e_1\| \le t} G(e_1).$$

**Bounding $G(e_1)$** To bound $\sup_{\|e_1\|\le t} G(e_1)$, we expand $\phi$ as the $d$-dimensional Gaussian probability density function:

$$G(e_1) = \left(\frac{\eta_1^2+\eta_2^2}{\eta_1^2}\right)^{d/2} \exp\left(-\frac{\|e_1\|^2}{2\eta_1^2} + \frac{\|e_1+Z_2\|^2}{2(\eta_1^2+\eta_2^2)}\right).$$

Using the quadratic expansion $\|e_1+Z_2\|^2 = \|e_1\|^2 + 2\langle e_1, Z_2\rangle + \|Z_2\|^2$, we rewrite:

$$G(e_1) = \left(\frac{\eta_1^2+\eta_2^2}{\eta_1^2}\right)^{d/2} \exp\left(-\frac{\|e_1\|^2}{2\eta_1^2} + \frac{\|e_1\|^2 + 2\langle e_1, Z_2\rangle + \|Z_2\|^2}{2(\eta_1^2+\eta_2^2)}\right).$$

Since $\|e_1\|\le t$ and $\langle e_1, Z_2\rangle \le \|e_1\|\|Z_2\|$, we bound

$$\frac{2\langle e_1, Z_2\rangle}{2(\eta_1^2+\eta_2^2)} \le \frac{t\|Z_2\|}{\eta_1^2+\eta_2^2}.$$

Thus,

$$G(e_1) \le \left(\frac{\eta_1^2+\eta_2^2}{\eta_1^2}\right)^{d/2} \exp\left(\frac{\|Z_2\|^2}{2(\eta_1^2+\eta_2^2)} + \frac{t\|Z_2\|}{\eta_1^2+\eta_2^2}\right).$$

Therefore, for any $Y_1, Y_2$, and $t$, we have

$$F_t(Y_1, Y_2) \le \left(\frac{\eta_1^2+\eta_2^2}{\eta_1^2}\right)^{d/2} \exp\left(\frac{\|Z_2\|^2}{2(\eta_1^2+\eta_2^2)} + \frac{t\|Z_2\|}{\eta_1^2+\eta_2^2}\right).$$

This gives that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Y_1,Y_2 \mid \|Z_1\|\le t}\left[\frac{p(Y_1)}{p(Y_2)}\right] \\
&= \mathop{\mathbb{E}}_{Y_1,Y_2 \mid \|Z_1\|\le t}\left[\frac{F_t(Y_1,Y_2)}{\Pr\left[\|Z_1\|\le t \mid Y_1\right]}\right] \\
&\le \mathop{\mathbb{E}}_{Y_1,Y_2 \mid \|Z_1\|\le t}\left[\frac{1}{\Pr\left[\|Z_1\|\le t \mid Y_1\right]} \cdot \left(\frac{\eta_1^2+\eta_2^2}{\eta_1^2}\right)^{d/2} \exp\left(\frac{\|Z_2\|^2}{2(\eta_1^2+\eta_2^2)} + \frac{t\|Z_2\|}{\eta_1^2+\eta_2^2}\right)\right] \\
&= \left(\frac{\eta_1^2+\eta_2^2}{\eta_1^2}\right)^{d/2} \mathop{\mathbb{E}}_{Y_1 \mid \|Z_1\|\le t}\left[\frac{1}{\Pr\left[\|Z_1\|\le t \mid Y_1\right]}\right] \cdot \mathop{\mathbb{E}}_{Z_2}\left[\exp\left(\frac{\|Z_2\|^2}{2(\eta_1^2+\eta_2^2)} + \frac{t\|Z_2\|}{\eta_1^2+\eta_2^2}\right)\right].
\end{aligned}$$

**Bounding expectation over $Z_2$.** We have

$$\mathop{\mathbb{E}}_{Z_2}\left[\exp\left(\frac{\|Z_2\|^2}{2(\eta_1^2+\eta_2^2)} + \frac{t\|Z_2\|}{\eta_1^2+\eta_2^2}\right)\right] = \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,I_d)}\left[\exp\left(\frac{\eta_2^2\|Z\|^2}{2(\eta_1^2+\eta_2^2)} + \frac{t\eta_2\|Z\|}{\eta_1^2+\eta_2^2}\right)\right].$$

We can apply results on the Gaussian moment generating functions to bound this. Using Lemma A.10 by setting $\alpha = \frac{\eta_2^2}{2(\eta_1^2+\eta_2^2)}$, $\beta = \frac{t\eta_2}{\eta_1^2+\eta_2^2}$, and $\gamma = \frac{\eta_1^2}{4(\eta_1^2+\eta_2^2)}$, we have

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,I_d)}\left[\exp\left(\frac{\eta_2^2\|Z\|^2}{2(\eta_1^2+\eta_2^2)} + \frac{t\eta_2\|Z\|}{\eta_1^2+\eta_2^2}\right)\right] \le \exp\left(\frac{t^2\eta_2^2}{\eta_1^2(\eta_1^2+\eta_2^2)}\right) \cdot \left(\frac{2(\eta_1^2+\eta_2^2)}{\eta_1^2}\right)^{d/2}.$$

Finally, this gives

$$\mathop{\mathbb{E}}_{Y_1,Y_2 \mid \|Z_1\|\le t}\left[\frac{p(Y_1)}{p(Y_2)}\right] \le \frac{1}{\Pr\left[\|Z_1\|\le t\right]} \cdot \exp\left(O\left(\frac{d\eta_2^2}{\eta_1^2} + \frac{t^2\eta_2^2}{\eta_1^4}\right)\right).$$

One need to verify that

$$\mathop{\mathbb{E}}_{Y_1 \mid \|Z_1\|\le t}\left[\frac{1}{\Pr\left[\|Z_1\|\le t \mid Y_1\right]}\right] \le \frac{1}{\Pr\left[\|Z_1\|\le t\right]}.$$

Also,

$$\mathop{\mathbb{E}}_{Z_2}\left[F_t(Y_1,Y_2)\right] \le \exp\left(O\left(\frac{d\eta_2^2}{\eta_1^2} + \frac{t\eta_2\sqrt{d}}{\eta_1^2}\right)\right).$$

This gives the result. □

**Lemma A.5.** *Let $\eta_1 > \eta_2$ be two positive numbers, and let $X \in \mathbb{R}^d$ be an arbitrary random variable. Define $Y_1 = X + Z_1$ and $Y_2 = Y_1 + Z_2$, where $Z_1 \sim N(0, \eta_1^2 I_d)$ and $Z_2 \sim N(0, \eta_2^2 I_d)$. There exists a constant $C > 0$ such that for any $\lambda > 1$,*

$$\Pr_{Y_1,Y_2}\left[\frac{p(Y_1)}{p(Y_2)} \le \exp\left(C \cdot \left(\frac{d\eta_2^2}{\eta_1^2} + \ln\lambda\right)\right)\right] \ge 1 - \frac{1}{\lambda}.$$

*where $p(Y_1)$ and $p(Y_2)$ are the densities of $Y_1$ and $Y_2$, respectively.*

*Proof.* Let $t = (\sqrt{d} + \sqrt{2\ln(2\lambda)})\eta_1$. By applying Laurent-Massart bounds (Lemma A.11), we have

$$\Pr\left[\|Z_1\| \le t\right] \ge 1 - \frac{1}{2\lambda}.$$

Taking these into Lemma A.4, we have

$$\mathop{\mathbb{E}}_{Y_1,Y_2 \mid \|Z_1\| \le t}\left[\frac{p(Y_1)}{p(Y_2)}\right] \le \exp\left(O\left(\frac{d\eta_2^2}{\eta_1^2} + \frac{t^2\eta_2^2}{\eta_1^4}\right)\right) \le \exp\left(O\left(\frac{(d+\ln\lambda)\eta_2^2}{\eta_1^2}\right)\right).$$

By applying Markov's inequality, for a large enough constant $C > 0$, we have

$$\Pr_{Y_1,Y_2 \mid \|Z_1\| \le t}\left[\frac{p(Y_1)}{p(Y_2)} \le \lambda \exp\left(C \cdot \left(\frac{(d+\ln\lambda)\eta_2^2}{\eta_1^2}\right)\right)\right] \ge 1 - \frac{1}{2\lambda}.$$

Cleaning up the bound a little bit, this implies that for a large enough constant $C > 0$,

$$\Pr_{Y_1,Y_2 \mid \|Z_1\| \le t}\left[\frac{p(Y_1)}{p(Y_2)} \le \exp\left(C \cdot \left(\frac{d\eta_2^2}{\eta_1^2} + \ln\lambda\right)\right)\right] \ge 1 - \frac{1}{2\lambda}.$$

Combining this with the probability that $\|Z\| \le t$, a union bound gives that

$$\Pr_{Y_1,Y_2}\left[\frac{p(Y_1)}{p(Y_2)} \le \exp\left(C \cdot \left(\frac{d\eta_2^2}{\eta_1^2} + \ln\lambda\right)\right)\right] \ge 1 - \frac{1}{\lambda}.$$

□

The $\chi^2$ divergence is

$$\begin{aligned}\chi^2\left(p(x \mid y_i) \,\|\, p(x \mid y_{i+1})\right) &= \mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(x \mid y_i)}{p(x \mid y_{i+1})}\right] - 1 \\ &= \mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)} \cdot \frac{p(y_{i+1})}{p(y_i)}\right] - 1 \\ &= \mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)}\right] \cdot \frac{p(y_{i+1})}{p(y_i)} - 1.\end{aligned}$$

Now we can bound the $\chi^2$-diversity.

**Lemma A.6.** *There exists a constant $C > 0$ such that for any $\lambda > 1$,*

$$\Pr_{y_i,y_{i+1}}\left[\chi^2\left(p(x \mid y_i) \,\|\, p(x \mid y_{i+1})\right) \le \exp\left(C\left(\frac{m(\eta_i^2 - \eta_{i+1}^2)}{\eta_{i+1}^2} + \ln\lambda\right)\right)\right] \ge 1 - \frac{1}{\lambda}.$$

*Proof.* Note that

$$\chi^2\left(p(x \mid y_i) \,\|\, p(x \mid y_{i+1})\right) = \mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)}\right] \cdot \frac{p(y_{i+1})}{p(y_i)} - 1.$$

By Corollary A.3, we have

$$\Pr_{y_i,y_{i+1}}\left[\mathop{\mathbb{E}}_{x \sim p(x|y_i)}\left[\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)}\right] \le 2\lambda\right] \ge 1 - \frac{1}{2\lambda}.$$

By Lemma A.5, there exists a constant $C > 0$ such that

$$\Pr_{Y_1,Y_2}\left[\frac{p(Y_1)}{p(Y_2)} \le \exp\left(C\left(\frac{m(\eta_i^2 - \eta_{i+1}^2)}{\eta_{i+1}^2} + \ln\lambda\right)\right)\right] \ge 1 - \frac{1}{2\lambda}.$$

A union bound over these two implies that with probability of $1 - 1/\lambda$,

$$\frac{p(y_i \mid x)}{p(y_{i+1} \mid x)} \cdot \frac{p(y_{i+1})}{p(y_i)} - 1 \le 2\lambda \cdot \exp\left(C\left(\frac{m(\eta_i^2 - \eta_{i+1}^2)}{\eta_{i+1}^2} + \ln\lambda\right)\right) \le \exp\left(C'\left(\frac{m(\eta_i^2 - \eta_{i+1}^2)}{\eta_{i+1}^2} + \ln\lambda\right)\right),$$

where $C'$ is a positive constant. This concludes the lemma. □

### A.2 Convergence time of Langevin dynamics

We present the following result on the convergence of Langevin dynamics:

**Lemma A.7** ([Dal17])**.** *Let $p$ and $q$ be probability distributions such that $q$ is an $\alpha$-strong log-concave distribution. Consider the Langevin dynamics initialized with $p$ as the starting distribution. Then, for any $t \geq 0$, we have*

$$\mathrm{TV}(p_t, q) \leq \frac{1}{2}\chi^2(p \,\|\, q)^{1/2} e^{-t\alpha/2}.$$

This implies that

**Lemma A.8.** *Let $p$ and $q$ be probability distributions such that $q$ is an $\alpha$-strong log-concave distribution. Consider the Langevin dynamics initialized with $p$ as the starting distribution. By running the diffusion for time*

$$T = O\left(\frac{\log(1/\varepsilon) + \log \chi^2(p\|q)}{\alpha}\right),$$

*we have $\mathrm{TV}(p_T, q) \leq \varepsilon$.*

Now we show that the posterior distribution is even more strongly log-concave than prior distribution.

**Lemma A.9.** *Suppose that $p(x)$ is $\alpha$-strongly log-concave. Then, the posterior density*

$$p(x \mid Ax + N(\eta_i^2 I_m) = y_i)$$

*is $\alpha$-strongly log-concave.*

*Proof.* By Bayes' rule, the posterior density can be written (up to normalization) as

$$p(x \mid Ax + N(\eta_i^2 I_m) = y_i) \;\propto\; p(x)\, \exp\Bigl(-\frac{1}{2\eta_i^2}\|Ax - y_i\|_2^2\Bigr).$$

Define the negative log–posterior

$$\varphi(x) := -\log p(x) + \frac{1}{2\eta_i^2}\|Ax - y_i\|_2^2.$$

Since $p$ is $\alpha$-strongly log-concave, its negative log–density satisfies

$$\nabla^2(-\log p(x)) \;\succeq\; \alpha I.$$

Moreover, the Gaussian likelihood term has

$$\nabla^2\Bigl(\frac{1}{2\eta_i^2}\|Ax - y_i\|_2^2\Bigr) = \frac{1}{\eta_i^2}\,A^T A \;\succeq\; 0.$$

By the sum rule for Hessians,

$$\nabla^2\varphi(x) = \nabla^2(-\log p(x)) \;+\; \frac{1}{\eta_i^2}A^T A \;\succeq\; \alpha I.$$

Hence $\varphi$ is $\alpha$-strongly convex, and the posterior density $p(x \mid Ax + N(\eta_i^2 I_m) = y_i) \propto e^{-\varphi(x)}$ is $\alpha$-strongly log-concave. □

Now we are ready to prove Lemma A.1:

*Proof of Lemma A.1.* By Lemma A.9, $p(x \mid y_{i+1})$ is $alpha$-strongly log-concave. This allows us to apply Lemma A.8. Therefore, to achieve $\varepsilon$ TV error in convergence, we only need to run the process for

$$T = O\left(\frac{\log(1/\varepsilon) + \log \chi^2(p(x \mid y_i) \,\|\, p(x \mid y_{i+1}))}{\alpha}\right).$$

Taking in the result in Lemma A.6, we have with $1 - \frac{1}{\lambda}$ probability over $y_i$ and $y_{i+1}$, we only need

$$T = O\left(\frac{m\gamma_i + \log(\lambda/\varepsilon)}{\alpha}\right).$$

□

### A.3 Utility Lemmas.

**Lemma A.10.** *Let $Z \sim \mathcal{N}(0, I_d)$ be a $d$-dimensional standard Gaussian random vector, and let $\alpha, \beta \in \mathbb{R}$. For any $\gamma > 0$ satisfying $\alpha + \gamma < \frac{1}{2}$, we have*

$$\mathbb{E}\Big[\exp\Big(\alpha\|Z\|^2{+}\beta\|Z\|\Big)\Big] \leq \exp\Big(\frac{\beta^2}{4\gamma}\Big)\,(1-2(\alpha+\gamma))^{-d/2}.$$

*Proof.* For all $r \geq 0$ and any $\gamma > 0$, it is easy to check that by AM-GM inequality,

$$\beta\, r \leq \gamma\, r^2 + \frac{\beta^2}{4\gamma}.$$

Taking $r = \|Z\|$ and exponentiating both sides, we obtain

$$\exp\Big(\beta\|Z\|\Big) \leq \exp\Big(\gamma\,\|Z\|^2{+}\frac{\beta^2}{4\gamma}\Big).$$

Multiplying both sides by $\exp\Big(\alpha\|Z\|^2\Big)$ yields

$$\exp\Big(\alpha\|Z\|^2{+}\beta\|Z\|\Big) \leq \exp\Big(\frac{\beta^2}{4\gamma}\Big)\exp\Big((\alpha+\gamma)\|Z\|^2\Big).$$

This gives that

$$\mathbb{E}\Big[\exp\Big(\alpha\|Z\|^2{+}\beta\|Z\|\Big)\Big] \leq \exp\Big(\frac{\beta^2}{4\gamma}\Big)\,\mathbb{E}\Big[\exp\Big((\alpha+\gamma)\|Z\|^2\Big)\Big].$$

For $Z \sim \mathcal{N}(0, I_d)$ , when $\alpha + \gamma < \dfrac{1}{2}$ we have

$$\mathbb{E}\Big[\exp\Big((\alpha+\gamma)\|Z\|^2\Big)\Big] = (1-2(\alpha+\gamma))^{-d/2},$$

Hence,

$$\mathbb{E}\Big[\exp\Big(\alpha\|Z\|^2{+}\beta\|Z\|\Big)\Big] \leq \exp\Big(\frac{\beta^2}{4\gamma}\Big)\,(1-2(\alpha+\gamma))^{-d/2}.$$

□

**Lemma A.11** (Laurent-Massart Bounds[LM00])**.** *Let $v \sim \mathcal{N}(0, I_m)$. For any $t > 0$,*

$$\Pr[\|v\|^2 - m \geq 2\sqrt{mt} + 2t] \leq e^{-t},$$
$$\Pr[\|v\|^2 - m \leq -2\sqrt{mt}] \leq e^{-t}.$$

## B Convergence Between Locally Well-Conditioned Distributions

In the last section, we considered the convergence time between two posterior distributions of a globally strongly log-concave distribution. In this section, we will relax the assumption of global strong log-concavity and consider the convergence time between two distributions that are locally "well-behaved". We give the following formal definition:

**Definition B.1.** *For $\delta \in [0, 1)$ and $R, \widetilde{L}, \alpha \in (0, +\infty]$, we say that a distribution $p$ is $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned if there exists $\theta$ such that*

- $\nabla \log p(\theta) = 0$.
- $\Pr_{x\sim p}[x \in B(\theta, r)] \geq 1 - \delta$.
- *For $x, y \in B(\theta, R)$, we have that $\|s(x) - s(y)\| \leq \widetilde{L}\alpha\, \|x - y\|$.*
- *For $x, y \in B(\theta, R)$, we have that $\langle s(y) - s(x), x - y\rangle \geq \alpha\, \|x - y\|^2$.*

Again, we consider the following process $P$, which is identical to process (6) we considered in the last section:

$$dx_t = \left(s(x_t) + \frac{A^T y_{i+1} - A^T A x_t}{\eta_{i+1}^2}\right) dt + \sqrt{2} dB_t, \quad x_0 \sim p(x \mid y_i)$$

Our goal is to prove the following lemma:

**Lemma B.2.** *Suppose $p$ is a $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned distribution. Let $C > 0$ be a large enough constant. We consider the process $P$ running for time*

$$T \geq C \left(\frac{m\gamma_i + \log(\lambda/\varepsilon)}{\alpha}\right).$$

*Suppose that*

$$R \geq r + \frac{T\|A\|}{\eta_{i+1}^2}\Big(\|A\|r + \eta_{i+1}(\sqrt{m} + \sqrt{2\ln(1/\delta)})\Big) + 2\sqrt{dT\ln(2d/\delta)}.$$

*Then $x_T \sim P_T$ satisfies that*

$$\Pr_{y_i, y_{i+1}}[\mathrm{TV}(x_T, p(x \mid y_{i+1})) \leq \varepsilon + \lambda\delta] \geq 1 - O(\lambda^{-1}).$$

In this section, we will assume that $p$ is $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned. Without loss of generality, we assume that the mode of $p$ is at 0, i.e., $\theta = 0$.

### B.1 High Probability Boundness of Langevin Dynamics

We consider the process $P'$ defined as the process $P$ conditioned on $x_t \in B(0, R)$ for $t \in [0, T]$.

Our goal is to prove the following lemma:

**Lemma B.3.** *Suppose the following holds:*

$$R \geq r + \frac{T\|A\|}{\eta_{i+1}^2}\Big(\|A\|r + \eta_{i+1}(\sqrt{m} + \sqrt{2\ln(1/\delta)})\Big) + 2\sqrt{dT\ln(2d/\delta)}.$$

*We have that*

$$\mathbb{E}[\mathrm{TV}(P, P')] \lesssim \delta.$$

We start by decomposing the total variation distance between $P$ and $P'$ as follows:

**Lemma B.4.** *We have that*

$$\mathbb{E}[\mathrm{TV}(P, P')] \leq \mathop{\mathbb{E}}_{y_i, y_{i+1}}\left[\Pr_P\left[\exists\, t \in [0, T] : \|x_t\| \geq R \,\middle|\, x_0 \in B(0, r)\right]\right] + \delta.$$

*Proof.* Recall that the process $P'$ is defined as the law of $P$ conditioned on the event

$$\mathcal{F} := \{x_t \in B(0, R) \text{ for all } t \in [0, T]\}.$$

Thus, for any fixed $y_i$ we have

$$\mathrm{TV}(P, P') = \mathrm{TV}\Big(P,\, P(\cdot \mid \mathcal{F})\Big) = 1 - P(\mathcal{F}) = P(\mathcal{F}^c),$$

where $\mathcal{F}^c = \{\exists\, t \in [0, T] :\, \|x_t\| \geq R\}$.

Let $\mathcal{E} := \{x_0 \in B(0, r)\}$ denote the event that the initial condition is "good." Then, by the law of total probability,

$$P(\mathcal{F}^c) = P(\mathcal{F}^c \cap \mathcal{E}) + P(\mathcal{F}^c \cap \mathcal{E}^c) \leq P(\mathcal{F}^c \mid \mathcal{E}) + P(\mathcal{E}^c).$$

Taking the expectation with respect to $y_i$ and $y_{i+1}$, we obtain

$$\mathbb{E}\Big[\mathrm{TV}(P, P')\Big] \leq \mathbb{E}\Big[P(\mathcal{F}^c \mid \mathcal{E})\Big] + \mathbb{E}\Big[P(\mathcal{E}^c)\Big].$$

Since

$$P(\mathcal{F}^c \mid \mathcal{E}) = \Pr_P\Big[\exists\, t \in [0,T] : \|x_t\| \geq R \,\Big|\, x_0 \in B(0,r)\Big],$$

and by the law of total probability, we have

$$\mathbb{E}\Big[\, P(\mathcal{E}^c)\Big] = \Pr_{x \sim p}(\|x\| \geq r) \leq \delta,$$

it follows that

$$\mathbb{E}\Big[\mathrm{TV}(P, P')\Big] \leq \mathbb{E}\Big[\Pr_P\Big[\exists\, t \in [0,T] : \|x_t\| \geq R \,\Big|\, x_0 \in B(0,r)\Big]\Big] + \delta.$$

This completes the proof. □

Now we focus on bounding $\mathbb{E}_{y_i, y_{i+1}}\left[\Pr_P\left[\exists\, t \in [0,T] : \|x_t\| \geq R \,\Big|\, x_0 \in B(0,r)\right]\right]$. We start by observing the following lemma for log-concave distributions.

**Lemma B.5.** *Let p be a log-concave distribution such that p is continuously differentiable. Suppose the mode of p is at 0. Then, for all $x \in \mathbb{R}^d$,*

$$\langle \nabla \log p(x), x \rangle \leq 0.$$

*Proof.* Since $\log p$ is concave, for any $x, \theta \in \mathbb{R}^d$ the first-order condition for concavity yields

$$\log p(\theta) \leq \log p(x) + \langle \nabla \log p(x), -x \rangle.$$

Rearrange this inequality to obtain

$$\langle \nabla \log p(x), -x \rangle \geq \log p(\theta) - \log p(x).$$

Because $\theta$ is a mode, $\log p(\theta) \geq \log p(x)$ for every $x \in \mathbb{R}^d$; hence,

$$\langle \nabla \log p(x), x \rangle \leq 0.$$

□

**Lemma B.6.** *Let $x_t$ be the stochastic process*

$$dx_t = (f(x_t) + g(x_t))\, dt + \sqrt{2}\, dB_t, \quad x_0 \in \mathbb{R}^d,$$

*where $B_t$ is a standard $\mathbb{R}^d$-valued Brownian motion and the functions $f,\ g : \mathbb{R}^d \to \mathbb{R}^d$ satisfy*

$$\|f(x)\| \leq a \quad \text{and} \quad \langle g(x), x \rangle \leq 0 \quad \text{for all } x \in \mathbb{R}^d,$$

*with $a \geq 0$. Then, for any time horizon $T > 0$ and $\delta \in (0,1)$,*

$$\Pr\left[\sup_{t \in [0,T]} \|x_t\| \leq \|x_0\| + aT + 2\sqrt{T\, d\, \ln\!\Big(\frac{2d}{\delta}\Big)}\right] \geq 1 - \delta.$$

*Proof.* Define $r(t) = \|x_t\|$. Although the Euclidean norm is not smooth at the origin, an application of Itô's formula yields that, for $x_t \neq 0$, one has

$$dr(t) = \frac{\langle x_t,\ f(x_t) + g(x_t) \rangle}{\|x_t\|}\, dt + \sqrt{2}\, \langle u(t), dB_t \rangle + \frac{d-1}{\|x_t\|}\, dt,$$

where $u(t) = x_t / \|x_t\|$. Using the bound $\|f(x_t)\| \leq a$ and the hypothesis $\langle g(x_t), x_t \rangle \leq 0$, it follows by the Cauchy–Schwarz inequality that

$$\frac{\langle x_t, f(x_t) \rangle}{\|x_t\|} \leq a \quad \text{and} \quad \frac{\langle x_t, g(x_t) \rangle}{\|x_t\|} \leq 0.$$

Discarding the nonnegative Itô correction term $\frac{d-1}{\|x_t\|}\, dt$ (which can only increase the process), we deduce that

$$dr(t) \leq a\, dt + \sqrt{2}\, \langle u(t), dB_t \rangle.$$

Introduce the one-dimensional process

$$y(t) = \|x_0\| + at + \sqrt{2}\,\beta(t), \quad \text{with} \quad \beta(t) = \int_0^t \langle u(s), dB_s\rangle.$$

Since $\|u(s)\| = 1$ for all $s$, the process $\beta(t)$ is a standard one-dimensional Brownian motion with quadratic variation $\langle \beta\rangle_t = t$. By a standard comparison theorem for one-dimensional stochastic differential equations, it follows that $r(t) \le y(t)$ almost surely for all $t \ge 0$; hence,

$$\sup_{t\in[0,T]} \|x_t\| \le \|x_0\| + aT + \sqrt{2} \sup_{t\in[0,T]} \beta(t).$$

A classical application of the reflection principle for one-dimensional Brownian motion shows that, for any $\rho > 0$,

$$\Pr\Big[\sup_{t\in[0,T]} \beta(t) \ge \rho\Big] = 2\ \Pr(\beta(T) \ge \rho) \le 2\exp\Big(-\frac{\rho^2}{2T}\Big).$$

To incorporate the $d$-dimensional nature of the noise, one may use a union bound over the $d$ coordinate processes of $B_t$, which yields that

$$\Pr\left[\sqrt{2} \sup_{t\in[0,T]} \beta(t) \le 2\sqrt{T\,d\,\ln\Big(\frac{2d}{\delta}\Big)}\right] \ge 1-\delta.$$

Combining the foregoing estimates, we deduce that

$$\Pr\left[\sup_{t\in[0,T]} \|x_t\| \le \|x_0\| + aT + 2\sqrt{T\,d\,\ln\Big(\frac{2d}{\delta}\Big)}\right] \ge 1-\delta,$$

which is the desired result. □

**Lemma B.7.** *For any $\delta \in (0,1)$ and $T > 0$, it holds that*

$$\Pr_{x_t\sim P_t}\Big[\sup_{t\in[0,T]} \|x_t\| \ge r + T\cdot\frac{\|A^T y_{i+1}\|}{\eta_{i+1}^2} + 2\sqrt{T\,d\,\ln\Big(\frac{2d}{\delta}\Big)}\ \Big|\ x_0 \in B(0,r)\Big] < \delta.$$

*Proof.* We first note that by Lemma B.5, for any $x \in \mathbb{R}^d$, we have

$$\Big\langle s(x) - \frac{A^T A x}{\eta_{i+1}^2}, x\Big\rangle \le \langle s(x), x\rangle - \frac{1}{\eta_{i+1}^2}\|Ax\|^2 \le 0.$$

By Lemma B.6, we have that

$$\Pr_{x_t\sim P}\left[\sup_{t\in[0,T]} \|x_t\| \ge \|x_0\| + T\cdot\frac{\|A^T y_{i+1}\|}{\eta_{i+1}^2} + 2\sqrt{T\,d\,\ln\Big(\frac{2d}{\delta}\Big)}\right] < \delta,$$

This gives that

$$\Pr_{x_t\sim P_t}\Big[\sup_{t\in[0,T]} \|x_t\| \ge r + T\cdot\frac{\|A^T y_{i+1}\|}{\eta_{i+1}^2} + 2\sqrt{T\,d\,\ln\Big(\frac{2d}{\delta}\Big)}\ \Big|\ x_0 \in B(0,r)\Big] < \delta.$$

□

**Lemma B.8.** *For any $\delta \in (0,1)$, suppose*

$$R \ge r + \frac{T\|A\|}{\eta_{i+1}^2}\Big(\|A\|r + \eta_{i+1}(\sqrt{m} + \sqrt{2\ln(1/\delta)})\Big) + 2\sqrt{dT\ln(2d/\delta)}.$$

*It holds that*

$$\mathop{\mathbb{E}}_{y_i, y_{i+1}}\left[\Pr_{x_t\sim P}\Big[\sup_{t\in[0,T]} \|x_t\| \ge R\ \Big|\ x_0 \in B(0,r)\Big]\right] \lesssim \delta.$$

*Proof.* Recall that

$$y_{i+1} = Ax + \eta_{i+1} z, \quad z \sim \mathcal{N}(0, I_m).$$

With probability at least $1 - \delta$

$$\|z\| \leq \sqrt{m} + \sqrt{2 \ln(1/\delta)}.$$

Since $\|x\| \leq r$ with probability $1 - \delta$. Thus, with probability at least $1 - 2\delta$, it follows that

$$\|y_{i+1}\| \leq \|Ax\| + \eta_{i+1} \|z\| \leq \|A\| r + \eta_{i+1} \Big( \sqrt{m} + \sqrt{2 \ln(1/\delta)} \Big).$$

Hence, with the $1 - 2\delta$ probability,

$$T \cdot \frac{\|A^T y_{i+1}\|}{\eta_{i+1}^2} \leq \frac{T \|A\| \|y_{i+1}\|}{\eta_{i+1}^2} \leq \frac{T \|A\|}{\eta_{i+1}^2} \Big( \|A\| r + \eta_{i+1} (\sqrt{m} + \sqrt{2 \ln(1/\delta)}) \Big).$$

Therefore, ensuring that

$$R \geq r + T \cdot \frac{\|A^T y_{i+1}\|}{\eta_{i+1}^2} + 2 \sqrt{T \, d \, \ln\Big( \frac{2d}{\delta} \Big)}.$$

In this case, Lemma B.7 guarantees that

$$\Pr_{x_t \sim P} \Big[ \sup_{t \in [0,T]} \|x_t\| \geq R \,\Big|\, x_0 \in B(0, r) \Big] \lesssim \delta.$$

Since the probability satisfying the condition is at least $1 - 2\delta$, we have

$$\mathop{\mathbb{E}}_{y_i, y_{i+1}} \left[ \Pr_{x_t \sim P} \Big[ \sup_{t \in [0,T]} \|x_t\| \geq R \,\Big|\, x_0 \in B(0, r) \Big] \right] \lesssim \delta.$$

□

Putting Lemma B.4 and Lemma B.8 together, we directly obtain Lemma B.3.

### B.2 Concentration of Strongly Log-Concave Distributions

Before moving futher, we first prove that a strongly log-concave distribution is highly concentrated.

**Lemma B.9** (Norm Bound for $\alpha$-Strongly Logconcave Distributions). *Let $X$ be a random vector in $\mathbb{R}^d$ with density*

$$\pi(x) \propto \exp(-V(x)),$$

*where the potential $V : \mathbb{R}^d \to \mathbb{R}$ is $\alpha$-strongly convex; that is,*

$$\nabla^2 V(x) \succeq \alpha I \quad \textit{for all } x \in \mathbb{R}^d.$$

*Denote by $\mu = \mathbb{E}[X]$ the mean of $X$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have*

$$\|X - \mu\| \leq \sqrt{\frac{d}{\alpha}} + \sqrt{\frac{2 \ln(1/\delta)}{\alpha}}.$$

*Proof.* Since $V$ is $\alpha$-strongly convex, the density $\pi$ satisfies a logarithmic Sobolev inequality with constant $1/\alpha$. Consequently, for any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and any $t > 0$, one has the concentration inequality (via Herbst's argument)

$$\mathbb{P}\Big( f(X) - \mathbb{E}[f(X)] \geq t \Big) \leq \exp\Big( -\frac{\alpha t^2}{2} \Big).$$

Noting that the function

$$f(x) = \|x - \mu\|$$

is 1-Lipschitz (by the triangle inequality), it follows that

$$\mathbb{P}\Big( \|X - \mu\| - \mathbb{E}\|X - \mu\| \geq t \Big) \leq \exp\Big( -\frac{\alpha t^2}{2} \Big).$$

A standard calculation using the fact that the covariance matrix of $X$ satisfies $\operatorname{Cov}(X) \preceq \frac{1}{\alpha} I$ gives

$$\mathbb{E}\|X - \mu\| \leq \sqrt{\frac{d}{\alpha}}.$$

Thus, setting

$$t = \sqrt{\frac{2\ln(1/\delta)}{\alpha}},$$

we obtain

$$\mathbb{P}\Big(\|X - \mu\| \geq \sqrt{\frac{d}{\alpha}} + \sqrt{\frac{2\ln(1/\delta)}{\alpha}}\Big) \leq \delta.$$

This completes the proof. □

**Lemma B.10** ([JCP24])**.** *Let $\mu$ and $\theta$ denote the mean and the mode of distribution $p$, respectively, where $p$ is $\alpha$-strongly log-concave and univariate. Then, $|\mu - \theta| \leq \frac{1}{\sqrt{\alpha}}$.*

This immediately gives us the following corollary.

**Corollary B.11.** *Let $p$ be a $\alpha$–strongly log-concave distribution on $\mathbb{R}^d$. Let $\theta$ be the mode of $p$. For every $0 < \delta < 1$, we have*

$$\Pr_{X \sim p}\left[\|X - \theta\| \leq 2\sqrt{\frac{d}{\alpha}} + \sqrt{\frac{2\log(1/\delta)}{\alpha}}\right] \geq 1 - \delta.$$

This also implies that every $\alpha$-strongly log-concave distribution is mode-centered locally well-conditioned.

**Lemma B.12.** *Let $p$ be an $\alpha$-strongly log-concave distribution. Suppose the score function of $p$ is $L$-Lipschitz. Then, for any $0 < \delta < 1$, we have that $p$ is $(\delta, 2\sqrt{\frac{d}{\alpha}} + \sqrt{\frac{2\log(1/\delta)}{\alpha}}, \infty, L/\alpha, \alpha)$ mode-centered locally well-conditioned.*

### B.3 Convergence to Target Distribution

Since $p$ is not globally strongly log-concave, we need to extend the distribution $p$ to a globally strongly log-concave distribution. We will use the following lemma to extend the distribution.

**Lemma B.13.** *Suppose $g : B(0, R) \to \mathbb{R}$ is continuously differentiable with gradient $s := \nabla g \in C(B(0, R); \mathbb{R}^d)$ and satisfies*

$$\langle s(y) - s(x),\, x - y\rangle \geq \alpha \|x - y\|^2, \qquad \forall\, x, y \in B(0, R). \tag{7}$$

*For every $z \in B(0, R)$ define*

$$\varphi_z(x) = g(z) + \langle s(z),\, x - z\rangle - \frac{\alpha}{2}\|x - z\|^2, \qquad x \in \mathbb{R}^d,$$

*and set*

$$\tilde{g}(x) = \begin{cases} g(x), & \|x\| \leq R, \\ \inf\limits_{z \in B(0,R)} \varphi_z(x), & \|x\| > R. \end{cases} \tag{8}$$

*Then the density $\widetilde{p}(x) \propto e^{\tilde{g}(x)}$ is globally $\alpha$–strongly log–concave.*

*Proof.* For each fixed $z \in B(0, R)$ the mapping $\varphi_z$ has Hessian $-\alpha I_d$, hence is $\alpha$–strongly concave on the whole space. Because of (7) we have

$$g(x) \leq g(z) + \langle s(z),\, x - z\rangle - \frac{\alpha}{2}\|x - z\|^2 = \varphi_z(x), \qquad \forall\, x, z \in B(0, R),$$

with equality when $x = z$. Consequently $\tilde{g}$ defined in (8) agrees with $g$ on $B(0, R)$.

Fix $x \in \mathbb{R}^d$ and choose $z_x \in B(0, R)$ attaining the infimum in (8). Because $\varphi_{z_x}$ touches $\tilde{g}$ from above at $x$, the vector

$$\xi = \nabla\varphi_{z_x}(x) = s(z_x) - \alpha(x - z_x)$$

belongs to $\partial\tilde{g}(x)$. By $\alpha$–strong concavity of $\varphi_{z_x}$,

$$\varphi_{z_x}(y) \le \varphi_{z_x}(x) + \langle \xi,\, y - x\rangle - \frac{\alpha}{2}\,\|y-x\|^2\,, \qquad \forall\, y \in \mathbb{R}^d.$$

Taking the infimum over $z$ on the left and using $\tilde{g}(x) = \varphi_{z_x}(x)$ gives that

$$\tilde{g}(y) \le \tilde{g}(x) + \langle \xi,\, y - x\rangle - \frac{\alpha}{2}\,\|y-x\|^2\,, \qquad \forall\, x, y \in \mathbb{R}^d;$$

hence $\tilde{g}$ is globally $\alpha$–strongly concave, and therefore $\widetilde{p}$ is $\alpha$–strongly log-concave. □

**Lemma B.14.** *Let $p$ be a $d$-dimensional $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned probability distribution with $0 < \delta \le 1/2$ and $\alpha > 0$. Assume*

$$R \ge 2\sqrt{\frac{d}{\alpha}} + \sqrt{\frac{2\log(1/\delta)}{\alpha}}.$$

*Then there exists an $\alpha$-strongly log-concave distribution $\widetilde{p}$ on $\mathbb{R}^d$ such that*

$$\mathrm{TV}(p, \widetilde{p}) \le 3\delta.$$

*Proof.* Let $\theta$ be the point in Definition B.1 and without loss of generality, we assume $\theta = 0$. Write $B := B(0, R)$ and $B^{\mathrm{c}} := \mathbb{R}^d \setminus B$. By definition $p(B^{\mathrm{c}}) \le \delta$.

Set $g := \log p$, and let $\widetilde{g}$ be the function in Lemma B.13. Then, $\rho(x) := e^{\widetilde{g}(x)}$ is $\alpha$-*strongly* log-concave and $\rho = p$ on $B$. Let $Z := \int_{\mathbb{R}^d} \rho$ and define $\widetilde{p} := \rho/Z$.

Now we bound

$$\mathrm{TV}(p, \widetilde{p}) = \frac{1}{2}\int_B |p - \widetilde{p}| + \frac{1}{2}\int_{B^{\mathrm{c}}} |p - \widetilde{p}| =: I_B + I_{B^{\mathrm{c}}}.$$

Corollary B.11 implies that $\widetilde{p}(B^{\mathrm{c}}) \le \delta$. Therefore,

$$I_{B^{\mathrm{c}}} \le \frac{1}{2}[p(B^{\mathrm{c}}) + \widetilde{p}(B^{\mathrm{c}})] \le \delta.$$

Note that $\int_B \rho = p(B) \ge 1 - \delta$ and $\int_{B^{\mathrm{c}}} \rho \le \delta Z$ (since $\widetilde{p}(B^{\mathrm{c}}) \le \delta$). Thus,

$$1 - \delta \le Z = p(B) + \int_{B^{\mathrm{c}}} \rho \le 1 + 2\delta.$$

Since $\widetilde{p} = p/Z$ on $B$, we have

$$\left|1 - \frac{1}{Z}\right| \le \left|\frac{Z-1}{1-\delta}\right| \le \frac{2\delta}{1-\delta} \le 4\delta.$$

Therefore, $I_B \le \frac{1}{2} \cdot 4\delta = 2\delta$.

Combining,

$$\mathrm{TV}(p, \widetilde{p}) \le 2\delta + \delta = 3\delta.$$

□

Now, we can consider process $\widetilde{P}$ defined as

$$dx_t = \left(\nabla \log \widetilde{p}(x_t) + \frac{A^T y_{i+1} - A^T A x_t}{\eta_{i+1}^2}\right) dt + \sqrt{2} dB_t, \quad x_0 \sim p(x \mid y_i).$$

Then, we have the following lemma.

**Lemma B.15.** *Suppose the following holds:*

$$R \ge r + \frac{T\|A\|}{\eta_{i+1}^2}\Big(\|A\| r + \eta_{i+1}(\sqrt{m} + \sqrt{2\ln(1/\delta)})\Big) + 2\sqrt{dT\ln(2d/\delta)}.$$

*We have that*

$$\mathbb{E}\left[\mathrm{TV}(P, \widetilde{P})\right] \lesssim \delta.$$

*Proof.* Let

$$\mathcal{E} = \left\{ \sup_{t\in[0,T]} \|x_t\| \leq R \right\} \quad \text{and} \quad P' = P(\cdot \mid \mathcal{E}), \widetilde{P}' = \widetilde{P}(\cdot \mid \mathcal{E}).$$

Because $s(x) = \nabla \log \widetilde{p}(x)$ for every $x \in B(0,R)$, the drift coefficients of $P$ and $\widetilde{P}$ coincide on the event $\mathcal{E}$, and hence conditioning on $\mathcal{E}$ gives $P' = \widetilde{P}'$.

Then, we have

$$\mathrm{TV}(P, \widetilde{P}) \leq \mathrm{TV}(P, P') + \mathrm{TV}(\widetilde{P}, \widetilde{P}') = P(\mathcal{E}^{\mathrm{c}}) + \widetilde{P}(\mathcal{E}^{\mathrm{c}}).$$

Taking expectation over $(y_i, y_{i+1})$ gives

$$\mathbb{E}[\mathrm{TV}(P, \widetilde{P})] \leq \mathbb{E}[P(\mathcal{E}^{\mathrm{c}})] + \mathbb{E}[\widetilde{P}(\mathcal{E}^{\mathrm{c}})]. \tag{9}$$

Lemma B.3 implies that $\mathbb{E}[P(\mathcal{E}^{\mathrm{c}})] \lesssim \delta$. Furthermore, the same argument also implies that $\mathbb{E}[\widetilde{P}(\mathcal{E}^{\mathrm{c}})] \lesssim \delta$. Therefore, we have

$$\mathbb{E}[\mathrm{TV}(P, \widetilde{P})] \lesssim \delta.$$

□

*Proof of Lemma B.2.* We start by considering another process $\widetilde{P}^s$ defined as

$$dx_t = \left( \nabla \log \widetilde{p}(x_t) + \frac{A^T y_{i+1} - A^T A x_t}{\eta_{i+1}^2} \right) dt + \sqrt{2} dB_t, \quad x_0 \sim \widetilde{p}(x \mid y_i).$$

We can see that

$$\mathbb{E}\left[\mathrm{TV}(\widetilde{P}, \widetilde{P}^s)\right] \leq \mathbb{E}\left[\mathrm{TV}(p(x \mid y_i), \widetilde{p}(x \mid y_i))\right] \lesssim \delta.$$

Combining this with Lemma B.15, we have that

$$\mathbb{E}\left[\mathrm{TV}(P, \widetilde{P}^s)\right] \lesssim \delta.$$

By Markov's inequality, we have that

$$\Pr_{y_i, y_{i+1}}\left[\mathrm{TV}(P, \widetilde{P}^s) \geq \lambda\delta\right] \leq O(\lambda^{-1}).$$

Furthermore, by Lemma A.1 and our constraint on $T$, we have that

$$\Pr_{y_i, y_{i+1}}\left[\mathrm{TV}(\widetilde{P}_T^s, \widetilde{p}(x \mid y_{i+1})) \leq \varepsilon\right] \geq 1 - O(\lambda^{-1}).$$

Therefore, we have that

$$\Pr_{y_i, y_{i+1}}\left[\mathrm{TV}(P_T, \widetilde{p}(x \mid y_{i+1})) \leq \varepsilon + \lambda\delta\right] \geq 1 - O(\lambda^{-1}).$$

Combining this with $\Pr\left[\mathrm{TV}(\widetilde{p}(x \mid y_{i+1}), p(x \mid y_{i+1})) \leq \lambda\delta\right] \geq 1 - O(\lambda^{-1})$, we conclude that for $x_T \sim P_T$,

$$\Pr_{y_i, y_{i+1}}\left[\mathrm{TV}(x_T, p(x \mid y_{i+1})) \leq \varepsilon + \lambda\delta\right] \geq 1 - O(\lambda^{-1}).$$

□

## C Control of Score Approximation and Discretization Errors

In this section, we consider these processes running for time $T$:

- Process $P$:

$$dx_t = \left( s(x_t) + \frac{A^T y_{i+1} - A^T A x_t}{\eta_{i+1}^2} \right) dt + \sqrt{2} dB_t, \quad x_0 \sim p(x \mid y_i)$$

- Process $\widehat{P}$: Let $0 = t_1 < \cdots < t_M = T$ be the $M$ discretization steps with step size $t_{j+1} - t_j = h$. For $t \in [t_j, t_{j+1}]$,

$$dx_t = \left( \widehat{s}(x_{t_j}) + \frac{A^T y_{i+1} - A^T A x_{t_j}}{\eta_{i+1}^2} \right) dt + \sqrt{2} dB_t, \quad x_0 \sim p(x \mid y_i)$$

Note that $\widehat{P}$ is exactly the process (2) we run in Algorithm 1, except that we start from $x_0 \sim p(x \mid y_i)$.

We have shown that the process $P$ will converge to the target distribution $p(x \mid y_{i+1})$. We will show that the process $\widehat{P}$ will also converge to $p(x \mid y_{i+1})$ with a small error

**Lemma C.1.** *Let $p$ be a $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned. Suppose the followings hold for a large enough constant $C > 0$:*

- $T > C\left(\frac{m\gamma_i + \log(\lambda/\varepsilon)}{\alpha}\right)$.
- $\|A\|^4(T^2 m + TR^2) \le \frac{\eta_i^4}{C\gamma_i^2}$.
- $R \ge r + \frac{T\|A\|}{\eta_{i+1}^2}\left(\|A\|r + \eta_{i+1}(\sqrt{m} + \sqrt{2\ln(1/\delta)})\right) + 2\sqrt{dT\ln(2d/\delta)}$.

*Then running $\widehat{P}$ for time $T$ guarantees that with probability at least $1 - 1/\lambda$ over $y_i$ and $y_{i+1}$, we have:*

$$\mathrm{TV}(\widehat{P}_T, p(x \mid y_{i+1})) \lesssim \varepsilon + \lambda\delta + \lambda\sqrt{T} \cdot \left(\left(\widetilde{L}\alpha + \frac{\|A\|^2}{\eta_i^2}\right)\left(h\widetilde{L}\alpha R + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta_i}{\eta_i^2} + \sqrt{dh}\right) + \varepsilon_{score}\right).$$

In this section, we assume $p$ is $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned. Without loss of generality, we assume that the mode of $p$ is at 0, i.e., $\theta = 0$. Let $L := \widetilde{L}\alpha$, i.e., the Lipschitz constant inside the ball $B(0, R)$.

We will also consider the following stochastic processes:

- Process $Q$:
$$dx_t = \left(s(x_t) + \frac{A^T y_i - A^T A x_t}{\eta_i^2}\right) dt + \sqrt{2} dB_t, \quad x_0 \sim p(x \mid y_i)$$
- Process $Q'$ is the process $Q$ conditioned on $x_t \in B(0, R)$ for $t \in [0, T]$.
- Process $P'$ is the process $P$ conditioned on $x_t \in B(0, R)$ for $t \in [0, T]$.

We first note that following the same proof in Lemma B.3 that bounds $\mathrm{TV}(P, P')$, we can also bound $\mathrm{TV}(Q, Q')$.

**Lemma C.2.** *Suppose the following holds:*

$$R \ge r + \frac{T\|A\|}{\eta_i^2}\left(\|A\|r + \eta_i(\sqrt{m} + \sqrt{2\ln(1/\delta)})\right) + 2\sqrt{dT\ln(2d/\delta)}.$$

*We have that*

$$\mathbb{E}\left[\mathrm{TV}(Q, Q')\right] \lesssim \delta.$$

**Lemma C.3.** *We have*

$$\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|x_t - x_{t_j}\|^4\right] \lesssim \left(hLR + \frac{h\left\|A\right\|\left\|y_i\right\| + h\left\|A\right\|^2 R}{\eta_i^2}\right)^4 + d^2h^2$$

*Proof.*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|x_t - x_{t_j}\|^4\right] \\
=&\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\left\|\int_{t_j}^{t}\left(s(x_s) + \frac{A^T y_i - A^T A x_s}{\eta_i^2}\right)\mathrm{d}\,s + \sqrt{2}\,\mathrm{d}\,B_s\right\|^4\right] \\
\lesssim&\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\left(\int_{t_j}^{t}\|s(x_s)\|\,\mathrm{d}\,s\right)^4\right] + \mathop{\mathbb{E}}_{x_t \sim Q'}\left[\left(\int_{t_j}^{t}\left\|\frac{A^T y_i - A^T A x_s}{\eta_i^2}\right\|\mathrm{d}\,s\right)^4\right] + \mathop{\mathbb{E}}_{x_t \sim Q'}\left[\left\|\int_{t_j}^{t}\sqrt{2}\,\mathrm{d}\,B_s\right\|^4\right] \\
\lesssim&(hLR)^4 + \left(\frac{h\left\|A\right\|\left\|y_i\right\|}{\eta_i^2}\right)^4 + \left(\frac{h\left\|A\right\|^2 R}{\eta_i^2}\right)^4 + \mathbb{E}\left[\left\|\int_{t_j}^{t}\sqrt{2}\,\mathrm{d}\,B_s\right\|^4\right].
\end{aligned}$$

Since $\int_{t_j}^{t} \sqrt{2} dB_s \sim \mathcal{N}(0, (t - t_j) I_d)$, we have that $\mathbb{E}\|\int_{t_j}^{t} \sqrt{2} dB_s\|^4 \lesssim d^2(t - t_j)^2 \lesssim d^2 h^2$. This gives that

$$\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|x_t - x_{t_j}\|^4\right] \lesssim \left(hLR + \frac{h\|A\|\|y_i\| + h\|A\|^2 R}{\eta_i^2}\right)^4 + d^2 h^2$$

□

**Lemma C.4.** *Suppose* $\|A\|^4(T^2 m + TR^2) \leq \frac{\eta_i^4 \eta_{i+1}^4}{C(\eta_i^2 - \eta_{i+1}^2)^2}$ *for a large enough constant* $C$*.*

$$\mathop{\mathbb{E}}_{y_i, y_{i+1}, x_t \sim Q'}\left[\left(\frac{\mathrm{d}\, P'}{\mathrm{d}\, Q'}(x_t)\right)^2\right] = O(1).$$

*Proof.* By Girsanov's theorem, for any *trajectory* $x_{0,\dots,t}$,

$$\frac{dP'}{dQ'}(x_{0,\dots,t}) = \exp(M_t)$$

where the Girsanov exponent $M_t$ is given by

$$M_t = \frac{1}{\sqrt{2}} \int_0^t \Delta b_y(x_u) \cdot dB_u - \frac{1}{4}\int_0^t \|\Delta b_y(x_u)\|^2 du$$

for

$$\begin{aligned}\Delta b_y(x_u) &= \frac{A^T y_{i+1} - A^T A x_u}{\eta_{i+1}^2} - \frac{A^T y_i - A^T A x_u}{\eta_i^2} \\ &= \frac{\eta_i^2 A^T y_{i+1} - \eta_{i+1}^2 A^T y_i - A^T A x_u(\eta_i^2 - \eta_{i+1}^2)}{\eta_{i+1}^2 \eta_i^2}.\end{aligned}$$

Since $Q'$ is supported in $B(0, R)$,

$$\|\Delta b_y(x_u)\| \leq O\left(\frac{\|A\|\|\eta_i^2 y_{i+1} - \eta_{i+1}^2 y_i\| + \|A\|^2(\eta_i^2 - \eta_{i+1}^2)R}{\eta_{i+1}^2 \eta_i^2}\right) := \kappa_y$$

Now, for $\zeta_y := \int_0^t \|\Delta b_y(x_u)\|^2 du$, we have that $M_t \sim \mathcal{N}\left(-\frac{1}{4}\zeta_y, \frac{1}{2}\zeta_y\right)$
So,

$$\mathbb{E}\left[\exp(2M_t)\right] \leq \exp(\zeta_y/2) \leq \exp(\kappa_y^2 t/2)$$

Note that $\|\eta_i^2 y_{i+1} - \eta_{i+1}^2 y_i\|^2$ has mean $\|(\eta_i^2 - \eta_{i+1}^2)Ax\|^2$ and is subgamma with variance $m\left(\eta_{i+1}^2 \eta_i^4 - \eta_{i+1}^4 \eta_i^2\right)^2$ and scale $\eta_{i+1}^2 \eta_i^4 - \eta_{i+1}^4 \eta_i^2$. Thus, for $t\|A\|^2 \leq \frac{\eta_{i+1}^2 \eta_i^2}{C\left(\eta_i^2 - \eta_{i+1}^2\right)}$ we have

$$\begin{aligned}\mathop{\mathbb{E}}_{x, y_{i+1}, y_i}[\exp(2M_t)] &\leq \mathbb{E}\left[\exp\left(t\frac{\|A\|^2\|\eta_i^2 y_{i+1} - \eta_{i+1}^2 y_i\|^2 + (\eta_i^2 - \eta_{i+1}^2)^2\|A\|^4 R^2}{\eta_{i+1}^4 \eta_i^4}\right)\right] \\ &\lesssim \exp\left(2\left(\frac{t^2\|A\|^4(\eta_{i+1}^2\eta_i^4 - \eta_{i+1}^4\eta_i^2)^2 m}{\eta_{i+1}^8 \eta_i^8} + \frac{(\eta_i^2 - \eta_{i+1}^2)^2\|A\|^4 t R^2}{\eta_{i+1}^4 \eta_i^4}\right)\right) \\ &= \exp\left(2\left(\frac{t^2\|A\|^4(\eta_i^2 - \eta_{i+1}^2)^2 m + (\eta_i^2 - \eta_{i+1}^2)^2\|A\|^4 t R^2}{\eta_{i+1}^4 \eta_i^4}\right)\right) \\ &= \exp\left(\frac{\|A\|^4(\eta_i^2 - \eta_{i+1}^2)^2 \cdot (t^2 m + tR^2)}{\eta_{i+1}^4 \eta_i^4}\right) \\ &\lesssim 1.\end{aligned}$$

□

**Lemma C.5.** *Let $E$ be the event on $y_i$ such that $\mathrm{TV}(Q, Q') \leq \frac{1}{2}$. Suppose*

$$\|A\|^4(T^2 m + TR^2) \leq \frac{\eta_i^4 \eta_{i+1}^4}{C(\eta_i^2 - \eta_{i+1}^2)^2}.$$

*Then,*

$$\mathop{\mathbb{E}}_{y_i, y_{i+1}}\left[\mathrm{TV}(P', \widehat{P})\right] \lesssim 1 - \Pr\left[E\right] + \sqrt{T} \cdot \left(\left(L + \frac{A^T A}{\eta_i^2}\right)\left(hLR + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta_i}{\eta_i^2} + \sqrt{dh}\right) + \varepsilon_{score}\right).$$

*Proof.* Note that the bound is trivial when $\Pr\left[E\right] < 1/2$. Therefore, we can use the fact that $\mathbb{E}\left[\cdot \mid E\right] \lesssim \mathbb{E}\left[\cdot\right]$ throughout the proof. We have, for any $t \in [t_j, t_{j+1}]$, .

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y_i, y_{i+1} \mid E} \mathop{\mathbb{E}}_{x_t \sim P'}\left[\|s(x_t) - \widehat{s}(x_{t_j})\|^2 + \left\|\frac{A^T A}{\eta_i^2}(x_t - x_{t_j})\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{y_i, y_{i+1} \mid E} \mathop{\mathbb{E}}_{x_t \sim Q'}\left[\frac{\mathrm{d}\,P'}{\mathrm{d}\,Q'} \cdot \left(\|s(x_t) - \widehat{s}(x_{t_j})\|^2 + \left\|\frac{A^T A}{\eta_i^2}(x_t - x_{t_j})\right\|^2\right)\right] \\
&\lesssim \sqrt{\mathop{\mathbb{E}}_{y_i, y_{i+1}, x_t \sim Q'}\left[\left(\frac{\mathrm{d}\,P'}{\mathrm{d}\,Q'}(x_t)\right)^2\right] \cdot \mathop{\mathbb{E}}_{y_i \mid E} \mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|s(x_t) - \widehat{s}(x_{t_j})\|^4 + \left\|\frac{A^T A}{\eta_i^2}(x_t - x_{t_j})\right\|^4\right]}
\end{aligned}$$

The first term can be bounded using Lemma C.4. Now we focus on the second term. Note that

$$\mathop{\mathbb{E}}_{y_i \mid E}\left[\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|s(x_t) - \widehat{s}(x_{t_j})\|^4\right]\right] \leq \mathop{\mathbb{E}}_{y_i \mid E}\left[\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|s(x_t) - s(x_{t_j})\|^4\right]\right] + \mathop{\mathbb{E}}_{y_i \mid E}\left[\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|s(x_{t_j}) - \widehat{s}(x_{t_j})\|^4\right]\right].$$

Since $s$ is $L$-Lipschitz in $B(0, R)$, and using Lemma C.3, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y_i \mid E}\left[\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|s(x_t) - s(x_{t_j})\|^4 + \left\|\frac{A^T A}{\eta_i^2}(x_t - x_{t_j})\right\|^4\right]\right] \\
&\lesssim \mathop{\mathbb{E}}_{y_i}\left[\left(L + \frac{A^T A}{\eta_i^2}\right)^4 \mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|x_t - x_{t_j}\|^4\right]\right] \\
&\lesssim \left(L + \frac{A^T A}{\eta_i^2}\right)^4 \mathop{\mathbb{E}}_{y_i}\left[\left(hLR + \frac{h\left\|A\right\| y_i + h\left\|A\right\|^2 R}{\eta_i^2}\right)^4 + d^2 h^2\right] \\
&\lesssim \left(L + \frac{A^T A}{\eta_i^2}\right)^4 \left(hLR + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta_i}{\eta_i^2} + \sqrt{dh}\right)^4.
\end{aligned}$$

Since $Q'$ is a conditional measure of $Q$, conditioned on $E$, we have $\frac{\mathrm{d}\,Q'}{\mathrm{d}\,Q} \leq \frac{1}{1 - \mathrm{TV}(Q', Q)} \leq 2$. Therefore,

$$\begin{aligned}
\mathop{\mathbb{E}}_{y_i \mid E}\left[\mathop{\mathbb{E}}_{x_t \sim Q'}\left[\|s(x_{t_j}) - \widehat{s}(x_{t_j})\|^4\right]\right] &\leq \mathop{\mathbb{E}}_{y_i \mid E}\left[2 \cdot \mathop{\mathbb{E}}_{x_t \sim Q}\left[\|s(x_{t_j}) - \widehat{s}(x_{t_j})\|^4\right]\right] \\
&\lesssim \mathop{\mathbb{E}}_{y_i}\left[\mathop{\mathbb{E}}_{x_t \sim Q}\left[\|s(x_{t_j}) - \widehat{s}(x_{t_j})\|^4\right]\right] \\
&\leq \varepsilon_{score}^4
\end{aligned}$$

This gives that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y_i, y_{i+1} \mid E} \mathop{\mathbb{E}}_{x_t \sim P'}\left[\|s(x_t) - \widehat{s}(x_{t_j})\|^2 + \left\|\frac{A^T A}{\eta_i^2}(x_t - x_{t_j})\right\|^2\right] \\
&\lesssim \left(L + \frac{A^T A}{\eta_i^2}\right)^2 \left(hLR + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta_i}{\eta_i^2} + \sqrt{dh}\right)^2 + \varepsilon_{score}^2.
\end{aligned}$$

Thus, by Girsanov's theorem,

$$\begin{aligned}\mathop{\mathbb{E}}_{y_i,y_{i+1}|E}\left[\mathrm{KL}\left(P' \,\|\, \widehat{P}\right)\right] &\lesssim \sum_{j=0}^{M-1}\int_{t_j}^{t_{j+1}} \mathop{\mathbb{E}}_{y_i,y_{i+1},x_t\sim P'}\left[\|s(x_t)-\widehat{s}(x_{t_j})\|^2 + \left\|\frac{A^TA}{\eta_i^2}(x_t - x_{t_j})\right\|^2\right] \\ &\lesssim T\cdot\left(\left(L+\frac{\|A\|^2}{\eta_i^2}\right)^2\left(hLR+\frac{h\left\|A\right\|^2R+h\|A\|\sqrt{m}\eta_i}{\eta_i^2}+\sqrt{dh}\right)^2+\varepsilon_{score}^2\right).\end{aligned}$$

By Pinsker's inequality,

$$\mathop{\mathbb{E}}_{y_i,y_{i+1}|E}\left[\mathrm{TV}(P',\widehat{P})\right] \lesssim \sqrt{T}\cdot\left(\left(L+\frac{\|A\|^2}{\eta_i^2}\right)\left(hLR+\frac{h\left\|A\right\|^2R+h\|A\|\sqrt{m}\eta_i}{\eta_i^2}+\sqrt{dh}\right)+\varepsilon_{score}\right)$$

Hence,

$$\begin{aligned}\mathop{\mathbb{E}}_{y_i,y_{i+1}}\left[\mathrm{TV}(P',\widehat{P})\right] &\le 1-\Pr\left[E\right] + \mathop{\mathbb{E}}_{y_i,y_{i+1}|E}\left[\mathrm{TV}(P',\widehat{P})\right] \\ &\lesssim 1-\Pr\left[E\right]+\sqrt{T}\cdot\left(\left(L+\frac{\|A\|^2}{\eta_i^2}\right)\left(hLR+\frac{h\left\|A\right\|^2R+h\|A\|\sqrt{m}\eta_i}{\eta_i^2}+\sqrt{dh}\right)+\varepsilon_{score}\right).\end{aligned}$$

□

Then have the following as a corollary:

**Corollary C.6.** *Suppose*

$$\|A\|^4(T^2m+TR^2) \le \frac{\eta_i^4\eta_{i+1}^4}{C(\eta_i^2-\eta_{i+1}^2)^2}.$$

*Then,*

$$\begin{aligned}\mathop{\mathbb{E}}_{y_i,y_{i+1}}\left[\mathrm{TV}(P,\widehat{P})\right] \lesssim& \,\mathbb{E}\left[\mathrm{TV}(P,P')\right]+\mathbb{E}\left[\mathrm{TV}(Q,Q')\right] \\ &+\sqrt{T}\cdot\left(\left(L+\frac{\|A\|^2}{\eta_i^2}\right)\left(hLR+\frac{h\left\|A\right\|^2R+h\|A\|\sqrt{m}\eta_i}{\eta_i^2}+\sqrt{dh}\right)+\varepsilon_{score}\right).\end{aligned}$$

*Proof.* We have that

$$\mathop{\mathbb{E}}_{y_i,y_{i+1}}\left[\mathrm{TV}(P,\widehat{P})\right] \le \mathop{\mathbb{E}}_{y_i,y_{i+1}}[\mathrm{TV}(P,P')] + \mathop{\mathbb{E}}_{y_i,y_{i+1}}[\mathrm{TV}(P',\widehat{P})].$$

Furthermore,

$$\begin{aligned}&\mathop{\mathbb{E}}_{y_i,y_{i+1}}[\mathrm{TV}(P',\widehat{P})] \\ &\lesssim \Pr\left[\mathrm{TV}(Q,Q')>\frac{1}{2}\right]+\sqrt{T}\cdot\left(\left(L+\frac{\|A\|^2}{\eta_i^2}\right)\left(hLR+\frac{h\left\|A\right\|^2R+h\|A\|\sqrt{m}\eta_i}{\eta_i^2}+\sqrt{dh}\right)+\varepsilon_{score}\right) \\ &\lesssim \mathbb{E}\left[\mathrm{TV}(Q,Q')\right]+\sqrt{T}\cdot\left(\left(L+\frac{\|A\|^2}{\eta_i^2}\right)\left(hLR+\frac{h\left\|A\right\|^2R+h\|A\|\sqrt{m}\eta_i}{\eta_i^2}+\sqrt{dh}\right)+\varepsilon_{score}\right),\end{aligned}$$

where the last line follows from Markov's inequality. The gives the result. □

*Proof of Lemma C.1.* We note that by our definition of $\gamma_i$,

$$\|A\|^4(T^2m+TR^2) \le \frac{\eta_i^4\eta_{i+1}^4}{C(\eta_i^2-\eta_{i+1}^2)^2} \iff \|A\|^4(T^2m+TR^2) \le \frac{\eta_i^4}{C\gamma_i^2}$$

Then, combining Corollary C.6 with Lemmas B.3 and C.2, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{y_i, y_{i+1}}\left[\mathrm{TV}(P, \widehat{P})\right] &\lesssim \mathbb{E}\left[\mathrm{TV}(P, P')\right] + \mathbb{E}\left[\mathrm{TV}(Q, Q')\right] \\
&\quad + \sqrt{T} \cdot \left(\left(L + \frac{\|A\|^2}{\eta_i^2}\right)\left(hLR + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta_i}{\eta_i^2} + \sqrt{dh}\right) + \varepsilon_{score}\right) \\
&\lesssim \delta + \sqrt{T} \cdot \left(\left(L + \frac{\|A\|^2}{\eta_i^2}\right)\left(hLR + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta_i}{\eta_i^2} + \sqrt{dh}\right) + \varepsilon_{score}\right)
\end{aligned}$$

The conditions in Lemmas B.3 and C.2 are satisfied by our assumptions, noting that $\eta_{i+1} < \eta_i$ implies the bound on $R$ holds for both processes.

Applying Markov's inequality and combining Lemma B.2 with the above, we conclude the proof. □

## D Admissible Noise Schedule

Recall that we can define process $\widehat{P}_i$ that converges from $p(x \mid y_i)$ to $p(x \mid y_{i+1})$: Let $0 = t_1 < \cdots < t_M = T$ be the $M$ discretization steps with step size $t_{j+1} - t_j = h$. For $t \in [t_j, t_{j+1}]$,

$$dx_t = \left(\widehat{s}(x_{t_j}) + \frac{A^T y_{i+1} - A^T A x_{t_j}}{\eta_{i+1}^2}\right) dt + \sqrt{2} dB_t, \quad x_0 \sim p(x \mid y_i) \tag{10}$$

We have already proven that we can converge the process from $p(x \mid y_i)$ to $p(x \mid y_{i+1})$ with good probability, as long as some conditions are satisfied. Those conditions actually depend on the choice of the schedule of $\eta_i$ and $T_i$. In this section, we will specify the schedule of $\eta_i$ and $T_i$.

Now we specify the schedule of $\eta_i$ and $T_i$.

**Definition D.1.** *We say a noise schedule $\eta_1 > \cdots > \eta_N$ together with running times $T_1, \cdots, T_{N-1}$ is* admissible *(for a set of parameters $C, \alpha, \lambda, A, d, \varepsilon, \eta, R$) if:*

- $\eta_N = \eta$;
- $\eta_1 \geq \frac{\lambda\|A\|}{\varepsilon}\sqrt{\frac{d}{\alpha}}$;
- *For all $\gamma_i = (\eta_i/\eta_{i+1})^2 - 1$, we have $\gamma_i \leq 1$ and*

$$T_i \geq C\left(\frac{m\gamma_i + \log(\lambda/\varepsilon)}{\alpha}\right).$$

*Furthermore,*

$$\|A\|^4(T_i^2 m + T_i R^2) \leq \frac{\eta_i^4}{C\gamma_i^2}.$$

The reason we need to satisfy the last inequality is to satisfy the conditions in Lemma C.1. We formalize this in the following lemma.

**Lemma D.2.** *Let $C > 0$ be a sufficiently large constant and $p$ be a $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned distribution. For any $\delta, \varepsilon \in (0, 1)$ and $\lambda > 1$, suppose*

$$R \geq r + C\left(\frac{(m + \log\frac{\lambda}{\varepsilon})\|A\|}{\alpha\eta^2}\left(\|A\|\, r + \eta\sqrt{m + \log(1/\delta)}\right) + \sqrt{\frac{d\log(d/\delta)(m + \log(\lambda/\varepsilon))}{\alpha}}\right).$$

*For any admissible schedule $(\eta_i)_{i\in[N]}$ and $(T_i)_{i\in[N-1]}$, running the process $\widehat{P}_i$ for time $T_i$ guarantees that with probability at least $1 - 1/\lambda$ over $y_i$ and $y_{i+1}$:*

$$\mathrm{TV}(x_{T_i}, p(x \mid y_{i+1})) \lesssim \varepsilon + \lambda\delta + \lambda\sqrt{\frac{m + \log(\lambda/\varepsilon)}{\alpha}} \cdot (\varepsilon_{dis} + \varepsilon_{score}),$$

*where*

$$\varepsilon_{dis} := \left(\widetilde{L}\alpha + \frac{\|A\|^2}{\eta^2}\right)\left(h\widetilde{L}\alpha R + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta}{\eta^2} + \sqrt{dh}\right).$$

*Proof.* It is straightforward to verify that an admissible schedule satisfies the first two conditions of Lemma C.1.

For the third condition regarding $R$, our assumption states:

$$R \geq r + C\left(\frac{(m+\log\frac{\lambda}{\varepsilon})\|A\|}{\alpha\eta^2}\left(\|A\|\, r + \eta\sqrt{m+\log(1/\delta)}\right) + \sqrt{\frac{d\log(d/\delta)(m+\log(\lambda/\varepsilon))}{\alpha}}\right)$$

Given that $T_i \lesssim \frac{m+\log(\lambda/\varepsilon)}{\alpha}$, this choice of $R$ is sufficient to satisfy the third condition in Lemma C.1.

Therefore, applying Lemma C.1 at each step $i$, we obtain that with probability at least $1-1/\lambda$ over $y_i$ and $y_{i+1}$:

$$\begin{aligned}
&\mathrm{TV}(x_{T_i}, p(x \mid y_{i+1}))\\
&\lesssim \varepsilon + \lambda\delta + \lambda\sqrt{T_i}\cdot\left(\left(\widetilde{L}\alpha + \frac{\|A\|^2}{\eta_i^2}\right)\left(h\widetilde{L}\alpha R + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta_i}{\eta_i^2} + \sqrt{dh}\right) + \varepsilon_{score}\right)\\
&\lesssim \varepsilon + \lambda\delta + \lambda\sqrt{\frac{m+\log(\lambda/\varepsilon)}{\alpha}}\cdot(\varepsilon_{dis} + \varepsilon_{score}).
\end{aligned}$$

□

We also want to prove the following two lemmas:

**Lemma D.3.** *Let $p$ be a $d$-dimensional $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned distribution. For any $\delta \in (0,1)$, suppose*

$$R \;\geq\; 2\sqrt{\frac{d}{\alpha}} + \sqrt{\frac{2\log(1/\delta)}{\alpha}}.$$

*Then, suppose $\eta_1 \geq \frac{\lambda\|A\|}{\varepsilon}\sqrt{\frac{d}{\alpha}}$, with probability at least $1-\frac{1}{\lambda}$ over $y_1$,*

$$\mathrm{TV}(p(x \mid y_1),\, p(x)) \lesssim \varepsilon + \lambda\delta.$$

**Lemma D.4.** *There exists an admissible noise such that*

$$N \lesssim \rho^2\sqrt{m}\log(\lambda/\varepsilon) + \frac{\rho^2\alpha R^2}{\sqrt{m}} + \frac{m^2}{m\log(\lambda/\varepsilon)+\alpha R^2} + \log\left(2 + \frac{\lambda\sqrt{d}\rho}{\varepsilon}\right),$$

*where $\rho = \frac{\|A\|}{\eta\sqrt{\alpha}}$.*

### D.1 The Closeness Between $p(x \mid y_1)$ and $p(x)$

In this part, we prove Lemma D.3, showing that any admissible schedule has a large enough $\eta_1$, enabling us to use $p(x)$ to approximate $p(x \mid y_1)$.

We have the following standard information-theoretic result.

**Lemma D.5.** *Let $X \in \mathbb{R}^m$ be a random variable, and $Y = X + \mathcal{N}(0, \eta^2 I_m)$. Then,*

$$I(X;Y) \leq \frac{1}{2}\log\det\left(I_m + \frac{\mathrm{Cov}(X)}{\eta^2}\right).$$

**Lemma D.6.** *For any distribution $p$ with $\mathbb{E}_{x\sim p}\left[\|x - \mathbb{E}\, x\|^2\right] = m_2^2$, we have*

$$\mathbb{E}\left[\mathrm{TV}(p(x \mid y_1), p(x))\right] \leq \frac{\|A\|m_2}{2\eta_1}.$$

*Proof.* Note that $\mathbb{E}\left[\mathrm{KL}(p(x \mid y_i)\,\|\,p(x))\right]$ is exactly the mutual information between $x$ and $y_i$. In addition, we have

$$\mathbb{E}\left[\mathrm{KL}(p(x \mid y_i)\,\|\,p(x))\right] = I(x; y_i) \leq I(Ax; y_i) \leq \frac{1}{2}\log\det\left(I_m + \frac{\mathrm{Cov}(Ax)}{\eta_i^2}\right) \leq \frac{\|A\|^2 m_2^2}{2\eta_i^2}.$$

By Pinsker's inequality, we have

$$\mathbb{E}\left[\mathrm{TV}(p(x \mid y_1), p(x))\right] \leq \frac{\|A\| m_2}{2\eta_1}.$$

□

**Lemma D.7.** *Let $p$ be a $d$-dimensional $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well–conditioned probability distribution. Assume*

$$R \;\geq\; 2\sqrt{\frac{d}{\alpha}} + \sqrt{\frac{2\log(1/\delta)}{\alpha}}.$$

*Then*

$$\mathop{\mathbb{E}}_{y_1}[\mathrm{TV}(p(x \mid y_1),\, p(x))] \lesssim \frac{\|A\|}{\eta_1}\sqrt{\frac{d}{\alpha}} + \delta.$$

*Proof.* Lemma B.14 provides an $\alpha$-strongly log–concave density $\widetilde{p}$ satisfying

$$\mathrm{TV}(p, \widetilde{p}) \leq 3\delta.$$

For an $\alpha$-strongly log–concave law the Brascamp–Lieb inequality yields $\mathrm{Cov}_{\widetilde{p}} \preceq \alpha^{-1} I_d$; hence

$$m_2(\widetilde{p}) := (\mathop{\mathbb{E}}_{\widetilde{p}}\|x - \mathop{\mathbb{E}}_{\widetilde{p}} x\|^2)^{1/2} \;\leq\; \sqrt{\frac{d}{\alpha}}.$$

Applying Lemma D.6 to $\widetilde{p}$ gives

$$\mathop{\mathbb{E}}_{y_1}[\mathrm{TV}(\widetilde{p}(x \mid y_1), \widetilde{p}(x))] \;\leq\; \frac{\|A\|}{2\,\eta_1}\sqrt{\frac{d}{\alpha}}.$$

Note that

$$\mathrm{TV}(p(x \mid y_1), p(x)) \leq \mathrm{TV}(p(x \mid y_1), \widetilde{p}(x \mid y_1)) + \mathrm{TV}(\widetilde{p}(x \mid y_1), \widetilde{p}(x)) + \mathrm{TV}(\widetilde{p}(x), p(x)).$$

Integrating in $y_1$ and using the elementary fact

$$\mathop{\mathbb{E}}_{y_1}[\mathrm{TV}(p(x \mid y_1), \widetilde{p}(x \mid y_1))] \leq \mathrm{TV}(p, \widetilde{p}),$$

together with the above calculaion, yields

$$\mathop{\mathbb{E}}_{y_1}[\mathrm{TV}(p(x \mid y_1), p(x))] \leq 3\delta + \frac{\|A\|}{2\,\eta_1}\sqrt{\frac{d}{\alpha}} + 3\delta.$$

This proves the stated bound. □

Now we prove Lemma D.3.

*Proof of Lemma D.3.* By Lemma D.7, we have

$$\mathop{\mathbb{E}}_{y_1}[\mathrm{TV}(p(x \mid y_1),\, p(x))] \lesssim \frac{\|A\|}{\eta_1}\sqrt{\frac{d}{\alpha}} + \delta.$$

Since all admissible noise schedules satisfy $\eta_1 \geq \frac{\lambda\|A\|}{\varepsilon}\sqrt{\frac{d}{\alpha}}$. This implies

$$\frac{\|A\|}{\eta_1}\sqrt{\frac{d}{\alpha}} \leq \frac{\varepsilon}{\lambda}.$$

Consequently,

$$\mathop{\mathbb{E}}_{y_1}[\mathrm{TV}(p(x \mid y_1),\, p(x))] \lesssim \frac{\varepsilon}{\lambda} + \delta.$$

By Markov's inequality, with probability at least $1 - \frac{1}{\lambda}$ over $y_1$,

$$\mathrm{TV}(p(x \mid y_1),\, p(x)) \lesssim \varepsilon + \lambda\delta,$$

which proves the lemma. □

### D.2 Bound for $N$ Mixing Steps

In this part, we prove Lemma D.4.

**Lemma D.8.** *Let $a, x_0 > 0$, and let $c > 0$. Consider the number sequence*

$$x_{i+1} = (1 + \min((ax_i)^c, 1))x_i.$$

*For every $B > 0$, let $k(B)$ be the minimum integer $i$ such that $x_i \geq B$. Then*

$$k(B) = O\left((ax_0)^{-c} + \log\left(1 + \frac{B}{x_0}\right)\right).$$

*Proof.* We show in two steps that the time to go from $x_0$ to $1/a$, then to $B$. Define

$$k_1 = \min\{i \in \mathbb{N} : x_i \geq 1/a\},$$

**Bound for $k_1$.** We first show that $k_1 \lesssim (ax_0)^{-c}$. Consider the quantities

$$N_j = \min\{i \in \mathbb{N} : x_i \geq 2^j x_0\},$$

and let $j^*$ be the smallest $j$ such that $x_{N_j} \geq 1/a$. If instead $x_0 \geq 1/a$ already, then $k_1 = 0$ and there is nothing to prove.

Assume $x_0 < 1/a$. For each $j < j^*$ define

$$t_j = (2^j ax_0)^{-c}.$$

We claim that

$$N_{j+1} - N_j \leq t_j.$$

Indeed, for each $j < j^*$,

$$\begin{aligned} x_{N_j+t_j} &\geq x_{N_j} \prod_{i=N_j}^{N_j+t_j-1} \Big(1 + (ax_i)^c\Big) \\ &\geq x_{N_j} \prod_{i=N_j}^{N_j+t_j-1} \Big(1 + (ax_{N_j})^c\Big) = x_{N_j} \Big(1 + (ax_{N_j})^c\Big)^{t_j}. \end{aligned}$$

Since

$$(ax_{N_j})^c \geq (a \cdot 2^j x_0)^c = \frac{1}{t_j},$$

we get

$$x_{N_j+t_j} \geq \left(1 + \frac{1}{t_j}\right)^{t_j} x_{N_j} \geq 2\, x_{N_j} \geq 2^{\,j+1}\, x_0.$$

By monotonicity of the sequence $(x_i)$, it follows that $N_{j+1} \leq N_j + t_j$. Summing over $j$ up to $j^* - 1$ gives

$$N_{j^*} = \sum_{j=0}^{j^*-1} (N_{j+1} - N_j) \leq \sum_{j=0}^{j^*-1} (2^j ax_0)^{-c} \lesssim (ax_0)^{-c}.$$

By definition, $N_{j^*}$ is the first index $i$ such that $x_i \geq 1/a$, so $k_1 = N_{j^*} \lesssim (ax_0)^{-c}$.

**Bound to achieve $B$.** If $B \leq 1/a$, the bound already holds. Now we analyze how many steps Note that for every $i \geq k_1$,

$$x_{i+1} = (1 + \min((ax_i)^c, 1))x_i = 2x_i.$$

Therefore, we have

$$x_{k_1+\log_2(B)} \geq 2^{\log_2(B/x_{k_1})} x_{k_1} \geq B.$$

This proves that

$$k(B) \leq k_1 + \log_2\left(1 + \frac{B}{x_{k_1}}\right) \leq k_1 + \log_2\left(1 + \frac{B}{x_0}\right) \lesssim (ax_0)^{-c} + \log\left(1 + \frac{B}{x_0}\right).$$

□

**Lemma D.9.** *Given parameters $x_0, a, b > 0$, consider sequence inductively defined by $x_{i+1} = (1+\gamma_i)x_i$, where*

$$\gamma_i = \min\left\{\gamma_i \le 1 \;:\; a\gamma^2 + b\gamma \le 2x_i\right\}.$$

*Given $B$, let $k(B)$ be the minimum integer $i$ such that $x_i \ge B$. Then,*

$$k(B) \lesssim \frac{b}{x_0} + \frac{a}{b} + \log\left(1 + \frac{B}{x_0}\right).$$

*Proof.* We do case analysis.

**Case 1: $x_0 \ge b^2/a$.** We always choose $\gamma_i = \sqrt{x_i/a}$. We can verify that

$$a\left(\frac{x_i}{a}\right) + b\sqrt{\frac{x_i}{a}} \le x_i + \sqrt{\frac{b^2}{a}\cdot x_i} \le 2x_i,$$

and this satisfies the requirement for $\gamma_i$. By applying Lemma D.8, we have that

$$k(B) \lesssim \left(\frac{x_0}{a}\right)^{-1/2} + \log\left(1 + \frac{B}{x_0}\right) \le \frac{a}{b} + \log\left(1 + \frac{B}{x_0}\right).$$

**Case 2: $x_0 \le B \le b^2/a$.** We always choose $\gamma_i = \min(x_i/b, 1)$. We can verify that

$$a\left(\frac{x_i}{b}\right)^2 + b\left(\frac{x_i}{b}\right) \le x_i\left(\frac{ax_i}{b^2}\right) + x_i \le 2x_i,$$

and this satisfies the requirement for $\gamma_i$. By applying Lemma D.8, we have that

$$k(B) \lesssim (x_0/b)^{-1} + \log\left(1 + \frac{B}{x_0}\right).$$

**Case 3: $x_0 \le b^2/a \le B$.** We combine the bound for the first two cases, where we first go from $x_0$ to $b^2/a$, then go from $b^2/a$ to $B$. Then we have

$$k(B) \lesssim \left((x_0/b)^{-1} + \log\left(1 + \frac{B}{x_0}\right)\right) + \left(\frac{a}{b} + \log\left(1 + \frac{B}{x_0}\right)\right) \lesssim \frac{b}{x_0} + \frac{a}{b} + \log\left(1 + \frac{B}{x_0}\right).$$

□

*Proof of Lemma D.4.* Now we describe how we construct an admissible noise schedule. Consider we start from $\eta'_1 = \eta$, and for each $i$, we iteratively choose $\gamma'_i$ to be the maximum $\gamma \le 1$ such that

$$\|A\|^4(f_T^2(\gamma)m + f_T(\gamma)R^2) \le \frac{(\eta'_i)^4}{C\gamma^2},$$

and then set $\eta'_{i+1} = \sqrt{(1+\gamma'_i)(\eta'_i)^2}$. We continue this process until we reach $\eta'_N \ge \frac{\lambda\|A\|}{\varepsilon}\sqrt{\frac{d}{\alpha}}$. It is easy to verify that $(\eta'_N, \eta'_{N-1}, \dots, \eta'_1)$ is an admissible noise schedule. Now we bound the number of iterations $N$.

Since for all $\gamma$, we have $\|A\|^4(f_T^2(\gamma)m + f_T(\gamma)R^2) \le \|A\|^4(\sqrt{m}f_T(\gamma) + \frac{R^2}{2\sqrt{m}})^2$, a sufficient condition for $\|A\|^4(f_T^2(\gamma)m + f_T(\gamma)R^2) \le \frac{(\eta'_i)^4}{C\gamma^2}$ is that

$$\|A\|^4(\sqrt{m}f_T(\gamma) + \frac{R^2}{2\sqrt{m}})^2 \le \frac{(\eta'_i)^4}{C\gamma^2} \iff \|A\|^2(\sqrt{m}f_T(\gamma) + \frac{R^2}{2\sqrt{m}}) \le \frac{(\eta'_i)^2}{C\gamma}.$$

Therefore, fixing $\eta'_i$, we have that $\gamma'_i$ is at least

$$\max\left\{\gamma \le 1 \;:\; \frac{\|A\|^2 m^{1.5}}{\alpha}\gamma^2 + \left(\frac{\|A\|^2\sqrt{m}\log(\lambda/\varepsilon)}{\alpha} + \frac{\|A\|^2 R^2}{\sqrt{m}}\right)\gamma \le \frac{(\eta'_i)^2}{C}\right\}.$$

Now we look at the inductive sequence starting from $x_1 = \eta^2$, and $x_{i+1} = (1 + \widetilde{\gamma}_i)x_i$, where

$$\widetilde{\gamma}_i = \max\left\{\gamma \le 1 \;:\; \frac{\|A\|^2 m^{1.5}}{\alpha}\gamma^2 + \left(\frac{\|A\|^2\sqrt{m}\log(\lambda/\varepsilon)}{\alpha} + \frac{\|A\|^2 R^2}{\sqrt{m}}\right)\gamma \le \frac{x_i}{C}\right\}.$$

By Lemma D.9, we know that for any $\eta_{goal} > 0$, we can achieve $x_N \ge \eta_{goal}^2$ within

$$N \lesssim \frac{\|A\|^2\sqrt{m}\log(\lambda/\varepsilon)}{\alpha\eta^2} + \frac{\|A\|^2 R^2}{\sqrt{m}\eta^2} + \frac{m^2}{m\log(\lambda/\varepsilon) + \alpha R^2} + \log\left(2 + \frac{\eta_{goal}}{\eta}\right).$$

Taking in $\eta_{goal} = \frac{\lambda\|A\|}{\varepsilon}\sqrt{\frac{d}{\alpha}}$, we conclude the lemma. □

# E Theoretical Analysis of Algorithm 1

In this section, we analyze the algorithm presented in Algorithm 1. In Algorithm 1, the algorithm initializes by drawing a sample from the prior distribution $p(x)$ via the diffusion SDE, which introduces sampling error. [CCL+22] demonstrated that this diffusion sampling error is polynomially small, with the exact magnitude depending on the discretization scheme chosen for the diffusion SDE. Since the focus of this paper is on enabling an unconditional diffusion sampling model to perform posterior sampling, the choice of diffusion discretization and its associated error are not not the focus of our analysis. Consequently, we omit the diffusion sampling error in the error analysis presented in this section. This omission does not impact the rigor of the theorems in the main paper, as the error is polynomially small.

We start with the following lemma:

**Lemma E.1.** *Let $C > 0$ be a large enough constant. Let $p$ be a $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned distribution. For every $\delta, \varepsilon \in (0,1)$ and $\lambda > 1$, suppose*

$$R \ge r + C\left(\frac{(m + \log\frac{\lambda}{\varepsilon})\|A\|}{\alpha\eta^2}\left(\|A\|\, r + \eta\sqrt{m + \log(1/\delta)}\right) + \sqrt{\frac{d\log(d/\delta)(m + \log(\lambda/\varepsilon))}{\alpha}}\right).$$

*Then running Algorithm 1 will guarantee that*

$$\Pr_{y_1,\ldots,y_N}\left[\mathrm{TV}(X_N, p(x \mid y)) \lesssim N\left(\varepsilon + \lambda\delta + \lambda\sqrt{\frac{m + \log(\lambda/\varepsilon)}{\alpha}}\cdot(\varepsilon_{dis} + \varepsilon_{score})\right)\right] \ge 1 - \frac{N}{\lambda},$$

*where*

$$\varepsilon_{dis} := \left(\widetilde{L}\alpha + \frac{\|A\|^2}{\eta^2}\right)\left(h\widetilde{L}\alpha R + \frac{h\|A\|^2 R + h\|A\|\sqrt{m}\eta}{\eta^2} + \sqrt{dh}\right).$$

*Proof.* Let $\varepsilon_{\text{step}} := C_0\left(\varepsilon + \lambda\delta + \lambda\sqrt{\frac{m+\log(\lambda/\varepsilon)}{\alpha}}\cdot(\varepsilon_{\text{dis}} + \varepsilon_{\text{score}})\right)$, where $C_0$ is a constant large enough to absorb the implicit constants in Lemma D.3 and Lemma D.2.

We prove by induction that for each $i \in [N]$:

$$\Pr_{y_1,\ldots,y_i}\left[\mathrm{TV}(X_i, p(x \mid y_i)) \le i\cdot\varepsilon_{\text{step}}\right] \ge 1 - \frac{i}{\lambda}. \tag{11}$$

For the base case ($i = 1$), since $X_1 \sim p(x)$, Lemma D.3 gives that $\mathrm{TV}(p(x), p(x \mid y_1)) \le \varepsilon_{\text{step}}$ with probability at least $1 - 1/\lambda$ over $y_1$.

For the inductive step, assume the statement holds for some $i < N$. Let $\mathcal{E}_i$ be the event that $\mathrm{TV}(X_i, p(x \mid y_i)) \le i\cdot\varepsilon_{\text{step}}$, so $\Pr[\mathcal{E}_i^c] \le i/\lambda$.

Let $X_i^* \sim p(x \mid y_i)$ and let $X_{i+1}^*$ be the result of evolving $X_i^*$ for time $T_i$ using the SDE in Equation (2). By Lemma D.2, the event $\mathcal{F}_{i+1}$ that $\mathrm{TV}(X_{i+1}^*, p(x \mid y_{i+1})) \le \varepsilon_{\text{step}}$ has probability at least $1 - 1/\lambda$ over $y_i, y_{i+1}$ and the SDE path.

By the triangle inequality and data processing inequality:

$$\mathrm{TV}(X_{i+1}, p(x \mid y_{i+1})) \le \mathrm{TV}(X_i, p(x \mid y_i)) + \mathrm{TV}(X^*_{i+1}, p(x \mid y_{i+1})). \tag{12}$$

If both $\mathcal{E}_i$ and $\mathcal{F}_{i+1}$ occur, then $\mathrm{TV}(X_{i+1}, p(x \mid y_{i+1})) \le (i+1)\varepsilon_{\text{step}}$. The probability that this bound fails is at most:

$$\begin{aligned}\Pr[\mathcal{E}_i^c \cup \mathcal{F}_{i+1}^c] &\le \Pr[\mathcal{E}_i^c] + \mathbb{E}_{y_1,\dots,y_i}[\mathbf{1}_{\mathcal{E}_i} \Pr[\mathcal{F}_{i+1}^c \mid y_1,\dots,y_i]] \\ &\le \frac{i}{\lambda} + \frac{1}{\lambda} = \frac{i+1}{\lambda}.\end{aligned}$$

Thus, the induction holds for $i+1$, and the lemma follows for $i = N$. □

**Lemma E.2.** *Let $S_1$ and $S_2$ be two random variables such that*

$$\Pr_{y_1,\dots,y_N}[\mathrm{TV}((S_1 \mid y_1,\dots,y_N),(S_2 \mid y_1,\dots,y_N)) \le \varepsilon] \ge 1-\delta.$$

*Then we have*

$$\Pr_{y_N}[\mathrm{TV}((S_1 \mid y_N),(S_2 \mid y_N)) \le 2\varepsilon] \ge 1 - \frac{\delta}{\varepsilon}.$$

*Proof.* Let $E(y_1,\dots,y_N)$ be the event such that $\mathrm{TV}((S_1 \mid E), p((S_2 \mid E)) \le \varepsilon$. Then, we have that

$$\mathrm{TV}((S_1 \mid y_N),(S_2 \mid y_N)) \le \Pr[\overline{E} \mid y_N] + \varepsilon.$$

Since $\Pr[E] \ge 1-\delta$, we apply Markov's inequality, and have

$$\Pr_y\left[\Pr[\overline{E} \mid y_N] \ge \varepsilon\right] \le \frac{\mathbb{E}_y\left[\Pr[\overline{E} \mid y_N]\right]}{\varepsilon} = \frac{\Pr[\overline{E}]}{\varepsilon} \le \frac{\delta}{\varepsilon}.$$

Hence, we have with probability $1 - \frac{\delta}{\varepsilon}$ over $y$,

$$\mathrm{TV}((S_1 \mid y_N),(S_2 \mid y_N)) \le 2\varepsilon.$$

□

Applying Lemma E.2 on Lemma E.1 gives the following corollary.

**Corollary E.3.** *Let $C > 0$ be a large enough constant. Let $p$ be a $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned distribution. For every $\delta, \varepsilon \in (0,1)$ and $\lambda > 1$, suppose*

$$R \ge r + C\left(\frac{(m + \log\frac{\lambda}{\varepsilon})\|A\|}{\alpha\eta^2}\left(\|A\|\, r + \eta\sqrt{m + \log(1/\delta)}\right) + \sqrt{\frac{d\log(d/\delta)(m + \log(\lambda/\varepsilon))}{\alpha}}\right).$$

*Define Then running Algorithm 1 will guarantee that*

$$\Pr_y[\mathrm{TV}(X_N, p(x \mid y)) \le \varepsilon_{error}] \ge 1 - \frac{N}{\lambda\varepsilon_{error}},$$

*with*

$$\varepsilon_{error} \lesssim N\left(\varepsilon + \lambda\delta + \lambda\sqrt{\frac{m + \log(\lambda/\varepsilon)}{\alpha}} \cdot (\varepsilon_{dis} + \varepsilon_{score})\right),$$

*where*

$$\varepsilon_{dis} := \left(\widetilde{L}\alpha + \frac{\|A\|^2}{\eta^2}\right)\left(h\widetilde{L}\alpha R + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta}{\eta^2} + \sqrt{dh}\right).$$

**Lemma E.4** (Main Analysis Lemma for Algorithm 1)**.** *Let $\rho = \frac{\|A\|}{\eta\sqrt{\alpha}}$. For all $0 < \varepsilon, \delta < 1$, there exists*

$$K \le \widetilde{O}\left(\frac{1}{\varepsilon\delta}\left(\frac{\rho^2\left(\left(m^2\rho^4 + 1\right)\widetilde{r}^2 + m^3\rho^2 + dm\right)}{\sqrt{m}} + \frac{m}{d} + \log d\right)\right)$$

*such that: suppose distribution $p$ is a $(\frac{\varepsilon}{K^2}, \widetilde{r}/\sqrt{\alpha}, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned distribution with $R \geq \frac{\sqrt{K\sqrt{m}/\alpha}}{\rho}$, and $\varepsilon_{score} \leq \frac{\sqrt{\alpha/m}}{K^2\delta}$; then Algorithm 1 samples from a distribution $\widehat{p}(x \mid y)$ such that*

$$\Pr_y \left[\mathrm{TV}(\widehat{p}(x \mid y), p(x \mid y)) \leq \varepsilon\right] \geq 1 - \delta.$$

*Furthermore, the total iteration complexity can be bounded by*

$$\widetilde{O}\left(K^3(K^2m^2d + m^3\rho + m^{1.5}(m\rho^2 + 1)\widetilde{r})(\widetilde{L} + \rho^2)^2 + K^3m^2\rho(\widetilde{L} + \rho^2)\right).$$

*Proof.* To distinguish the $\varepsilon$ and $\delta$ in the lemma and the one in Corollary E.3, we will use $\varepsilon_{error}$ and $\delta_{error}$ to denote the $\varepsilon$ and $\delta$ in our lemma statement. We need to set parameters in Corollary E.3. For any given $0 < \delta_{error}, \varepsilon_{error}$, we set

$$\varepsilon = \frac{1}{\lambda\delta_{error}}, \qquad \delta = \frac{\varepsilon_{error}}{\lambda^2},$$

and we set $\lambda$ to be the minimum $\lambda$ that satisfies

$$\rho^2\sqrt{m}\log(\lambda/\varepsilon) + \frac{\rho^2\alpha R^2}{\sqrt{m}} + \frac{m^2}{m\log(\lambda/\varepsilon) + \alpha R^2} + \log\left(2 + \frac{\lambda\sqrt{d}\rho}{\varepsilon}\right) \leq \lambda\delta_{error}\varepsilon_{error}.$$

Now we verify the correctness. Taking in the bound for $N$ in Lemma D.4, we have

$$N \lesssim \rho^2\sqrt{m}\log(\lambda/\varepsilon) + \frac{\rho^2\alpha R^2}{\sqrt{m}} + \frac{m^2}{m\log(\lambda/\varepsilon) + \alpha R^2} + \log\left(2 + \frac{\lambda\sqrt{d}\rho}{\varepsilon}\right) \leq \lambda\delta_{error}\varepsilon_{error}.$$

By the setting of our parameters, we have $N\varepsilon \lesssim \varepsilon_{error}$, $\lambda\delta \lesssim \varepsilon_{error}$, and $N/\lambda\varepsilon_{error} \lesssim \delta_{error}$. This guarantees that

$$\Pr_y\left[\mathrm{TV}(\widetilde{X}_N, p(x \mid y)) \lesssim \varepsilon_{error} + \lambda N\sqrt{\frac{m + \log(\lambda/\varepsilon)}{\alpha}} \cdot (\varepsilon_{dis} + \varepsilon_{score})\right] \geq 1 - \delta_{error}.$$

It is easy to verify our bound on $R$ satisfies the condition in Corollary E.3. Note that if a distribution is $(\delta, r, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned, then it is also $(\delta, r, R', \widetilde{L}, \alpha)$ mode-centered locally well-conditioned for any $R' \leq R$. Therefore, we can set $R$ to be the minimum $R$ that satisfies the condition.

$$\begin{aligned}
\lambda &= \widetilde{O}\left(\frac{1}{\varepsilon_{error}\delta_{error}}\left(\rho^2\sqrt{m} + \frac{\rho^2\alpha R^2}{\sqrt{m}} + \frac{m^2}{m + \alpha R^2} + \log d\right)\right) \\
&= \widetilde{O}\left(\frac{1}{\varepsilon_{error}\delta_{error}}\left(\frac{\rho^2\left(\left(m^2\rho^4 + 1\right)\widetilde{r}^2 + m^3\rho^2 + dm\right)}{\sqrt{m}} + \frac{m}{d} + \log d\right)\right) \\
&\lesssim K.
\end{aligned}$$

Therefore, we only need $\lambda N\sqrt{\frac{m+\log(\lambda/\varepsilon)}{\alpha}}(\varepsilon_{dis} + \varepsilon_{score}) \lesssim \varepsilon_{error}$. This can be satisfied when

$$\varepsilon_{dis} + \varepsilon_{score} \lesssim \frac{1}{\lambda^2\delta_{error}}\sqrt{\frac{\alpha}{\log(\lambda/\varepsilon) + m}} \lesssim \frac{\sqrt{\alpha/m}}{K^2\delta_{error}}.$$

Recall that

$$\begin{aligned}
\varepsilon_{dis} &= \left(\widetilde{L}\alpha + \frac{\|A\|^2}{\eta^2}\right)\left(h\widetilde{L}\alpha R + \frac{h\left\|A\right\|^2 R + h\|A\|\sqrt{m}\eta}{\eta^2} + \sqrt{dh}\right) \\
&\leq \alpha\left(\widetilde{L} + \rho^2\right)\left(h\widetilde{L}\alpha R + h\rho^2\alpha R + h\rho\sqrt{m\alpha} + \sqrt{dh}\right).
\end{aligned}$$

Therefore, we need to set

$$h = \widetilde{\Omega}\left(\min\left\{\frac{1}{K^2\delta_{\text{error}}}\frac{\sqrt{\alpha/m}}{\alpha(\widetilde{L} + \rho^2)[\alpha R(\widetilde{L} + \rho^2) + \rho\sqrt{m\alpha}]}, \frac{1}{K^4\delta_{\text{error}}^2\alpha md(\widetilde{L} + \rho^2)^2}\right\}\right).$$

Note that the bound for the sum of $N$ mixing times can be bounded by

$$\sum_{i=1}^{N-1} T_i \lesssim \frac{N(\log(\lambda/\varepsilon)+m)}{\alpha} \le \widetilde{O}\left(\frac{Km\delta_{error}\varepsilon_{error}}{\alpha}\right).$$

Therefore, the total iteration complexity is bounded by $\widetilde{O}(\frac{Km\delta_{error}\varepsilon_{error}}{\alpha h})$,

$$\widetilde{O}\left(K^3 m(\widetilde{L}+\rho^2)[\alpha R(\widetilde{L}+\rho^2)+\rho\sqrt{m\alpha}]\sqrt{m/\alpha}\,\varepsilon_{\text{error}}^2\delta_{\text{error}} + K^5 m^2 d(\widetilde{L}+\rho^2)^2\varepsilon_{\text{error}}\delta_{\text{error}}^3\right).$$

We can relax it and make the bound be

$$\widetilde{O}\left(K^3(K^2m^2d+\sqrt{m^3\alpha}R)(\widetilde{L}+\rho^2)^2 + K^3m^2\rho(\widetilde{L}+\rho^2)\right).$$

Take in $R$, and we have

$$\widetilde{O}\left(K^3(K^2m^2d+m^3\rho+m^{1.5}(m\rho^2+1)\widetilde{r})(\widetilde{L}+\rho^2)^2 + K^3m^2\rho(\widetilde{L}+\rho^2)\right).$$

□

### E.1 Application on Strongly Log-concave Distributions

By Lemma B.12, any $\alpha$-strongly log-concave distribution that has $L$-Lipschitz score is locally well-conditioned distribution $p$ is $(\delta, 2\sqrt{\frac{d}{\alpha}}+\sqrt{\frac{2\log(1/\delta)}{\alpha}}, \infty, L/\alpha, \alpha)$ mode-centered locally well-conditioned. Therefore, take this into Lemma E.4, we have the following result.

**Lemma E.5.** *Let $p(x)$ be an $\alpha$-strongly log-concave distribution over $\mathbb{R}^d$ with L-Lipschitz score. Let $\rho = \frac{\|A\|}{\eta\sqrt{\alpha}}$. For all $0<\varepsilon,\delta<1$, there exists*

$$K \le \widetilde{O}\left(\frac{1}{\varepsilon\delta}\left(\frac{\rho^2\left(\left(m^2\rho^4+m\right)d+m^3\rho^2\right)}{\sqrt{m}}+\frac{m}{d}+\log d\right)\right)$$

*such that: suppose $\varepsilon_{score} \le \frac{\sqrt{\alpha/m}}{K^2\delta}$, then Algorithm 1 samples from a distribution $\widehat{p}(x \mid y)$ such that*

$$\Pr_y\left[\mathrm{TV}(\widehat{p}(x \mid y), p(x \mid y)) \le \varepsilon\right] \ge 1-\delta.$$

*Furthermore, the total iteration complexity can be bounded by*

$$\widetilde{O}\left(K^3(K^2m^2d+m^3\rho+m^{1.5}(m\rho^2+1)\sqrt{d})(L/\alpha+\rho^2)^2+K^3m^2\rho(L/\alpha+\rho^2)\right).$$

To enhance clarity, we state our result in terms of expectation and established the following theorem:

**Theorem E.6** (Posterior sampling with global log-cancavity)**.** *Let $p(x)$ be an $\alpha$-strongly log-concave distribution over $\mathbb{R}^d$ with L-Lipschitz score. Let $\rho = \frac{\|A\|}{\eta\sqrt{\alpha}}$. For all $0<\varepsilon<1$, there exists*

$$K \le \widetilde{O}\left(\frac{1}{\varepsilon^2}\left(\frac{\rho^2\left(\left(m^2\rho^4+m\right)d+m^3\rho^2\right)}{\sqrt{m}}+\frac{m}{d}+\log d\right)\right)$$

*such that: suppose $\varepsilon_{score} \le \frac{\sqrt{\alpha/m}}{K^2\varepsilon}$, then Algorithm 1 samples from a distribution $\widehat{p}(x \mid y)$ such that*

$$\mathop{\mathbb{E}}_y\left[\mathrm{TV}(\widehat{p}(x \mid y), p(x \mid y))\right] \le \varepsilon.$$

*Furthermore, the total iteration complexity can be bounded by*

$$\widetilde{O}\left(K^3(K^2m^2d+m^3\rho+m^{1.5}(m\rho^2+1)\sqrt{d})(L/\alpha+\rho^2)^2+K^3m^2\rho(L/\alpha+\rho^2)\right).$$

This gives Theorem 1.1.

**Theorem 1.1** (Posterior sampling with global log-concavity)**.** *Let $p(x)$ be an $\alpha$-strongly log-concave distribution over $\mathbb{R}^d$ with $L$-Lipschitz score. For any $0 < \varepsilon < 1$, there exist $K_1 = \mathrm{poly}(d, m, \frac{\|A\|}{\eta\sqrt{\alpha}}, \frac{1}{\varepsilon})$ and $K_2 = \mathrm{poly}(d, m, \frac{\|A\|}{\eta\sqrt{\alpha}}, \frac{1}{\varepsilon}, \frac{L}{\alpha})$ such that: if $\varepsilon_{score} \le \frac{\sqrt{\alpha}}{K_1}$, then there exists an algorithm that takes $K_2$ iterations to sample from a distribution $\widehat{p}(x \mid y)$ with*

$$\mathbb{E}\left[\mathrm{TV}(\widehat{p}(x \mid y), p(x \mid y))\right] \le \varepsilon.$$

**Remark E.7.** *The analysis above is restricted to strongly log-concave distributions, where $\nabla^2 \log p(x) \prec 0$. However, this directly implies that we can use our algorithm to perform posterior sampling on log-concave distributions, for which $\nabla^2 \log p(x) \preceq 0$.*

*Specifically, for any log-concave distribution $p$, we can define a distribution $q(x) \propto p(x) \cdot \exp\left(-\frac{\varepsilon^2\|x-\theta\|^2}{2m_2^2}\right)$, where $\theta$ is the mode of $p$ and $m_2^2$ is the variance of $p$. It is straightforward to verify that $\mathrm{TV}(p, q) \lesssim \varepsilon$, and $q$ is $(\varepsilon^2/m_2^2)$-strongly log-concave. Therefore, by sampling from $q(x \mid y)$, we can approximate $p(x \mid y)$, incurring an additional expected TV error of $\varepsilon$.*

## E.2 Gaussian Measurement

In this section, we prove Theorem 1.2. In Algorithm 2, we describe how to make Algorithm 1 work on the Gaussian case.

We first verify that suppose Assumption 1 holds, we can also have $L^4$-accurate estimates for the smoothed scores of $p_{x_0}$, so this satisfies the requirement of running Algorithm 1. We need to use the following lemma, with proof deferred to Section E.5.

**Lemma E.8.** *Let $X$, $Y$, and $Z$ be random vectors in $\mathbb{R}^d$, where $Y = X + N(0, \sigma_1^2 I_d)$ and $Z = X + N(0, \sigma_2^2 I_d)$. The conditional density of $Z$ given $Y$, denoted $p(Z \mid Y)$, is a multivariate normal distribution with mean*

$$\mu_{Z|Y} = \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1} Y$$

*and covariance matrix*

$$\Sigma_{Z|Y} = \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1}\sigma_1^2.$$

*Then, the gradient of the log-likelihood $\log p(Z \mid Y)$ with respect to $Y$ is given by*

$$\nabla_Y \log p(Z \mid Y) = -\frac{1}{\sigma_1^2}\left(Z - \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1} Y\right).$$

Using this, we can calculate the *smoothed* conditional score given $x_0$:

**Lemma E.9.** *For any smoothing level $t \ge 0$, suppose we have score estimate $\widehat{s}_{t^2}(x)$ of the smoothed distributions $p_{t^2}(x) = p(x) * \mathcal{N}(0, t^2 I_d)$ that satisfies*

$$\mathop{\mathbb{E}}_{p_{t^2}(x)}[\|\widehat{s}_{t^2}(x) - s_{t^2}(x)\|^4] \le \varepsilon_{score}^4.$$

*Then we can calculate a score estimate $\widehat{s}_{x_0,t^2}(x)$ of the distribution $p_{x_0,t^2}(x) = p_{x_0}(x) * \mathcal{N}(0, t^2 I_d)$ such that*

$$\mathop{\mathbb{E}}_{x_0}\left[\mathop{\mathbb{E}}_{p_{x_0,t^2}(x)}[\|\widehat{s}_{x_0,t^2}(x) - s_{x_0,t^2}(x)\|^4]\right] \le \varepsilon_{score}^4.$$

*Proof.* Let $x^{(t)} \sim p_{t^2}$. Then, for any value of $x^{(t)}$, we have

$$\begin{aligned} s_{x_0,t^2}(x^{(t)}) &= \nabla_{x^{(t)}} \log p(x^{(t)} \mid x_0) \\ &= \nabla_{x^{(t)}} \log p(x^{(t)}) + \nabla_{x^{(t)}} \log p(x_0 \mid x^{(t)}) \\ &= s_{t^2}(x^{(t)}) + \nabla_{x^{(t)}} \log p(x_0 \mid x^{(t)}). \end{aligned}$$

Note that the second term is exactly in the form of Lemma E.8, so we can calculate this exaclty. For the first term, we use our score estimate $\widehat{s}_{t^2}(x^{(t)})$ for it. In this way, we have that for any $x$,

$$\|\widehat{s}_{x_0,t^2}(x) - s_{x_0,t^2}(x)\| = \|\widehat{s}_{t^2}(x) - s_{t^2}(x)\|.$$

Therefore,

$$\mathbb{E}_{x_0}\left[\mathbb{E}_{p_{x_0,t^2}(x)}[\|\widehat{s}_{x_0,t^2}(x) - s_{x_0,t^2}(x)\|^4]\right] = \mathbb{E}_{p_{t^2}(x)}[\|\widehat{s}_{x_0,t^2}(x) - s_{x_0,t^2}(x)\|^4] \le \varepsilon_{score}^4.$$

□

Applying Markov's inequality, we have:

**Corollary E.10.** *Suppose Assumption 1 holds for our prior distribution $p$. Then with $1-\delta$ probability over $x_0$: we have smoothed score estimates for $p_{x_0}$ with $L^4$ error bounded by $\varepsilon_{score}^4/\delta$; in other words, Assumption 1 holds for $p_{x_0}$, where $\varepsilon_{score}$ is substituted with $\varepsilon_{score}/\delta^{1/4}$.*

To capture the behavior of a Gaussian measurement more accurately, we first define a relaxed version of mode-centered locally well-conditioned distribution.

**Definition E.11.** *For $\delta \in [0,1)$ and $R, \widetilde{L}, \alpha \in (0,+\infty]$, we say that a distribution $p$ is $(\delta, r, R, \widetilde{L}, \alpha)$ locally well-conditioned if there exists $\theta$ such that*

- $\Pr_{x\sim p}[x \in B(\theta, r)] \ge 1-\delta$.
- *For $x, y \in B(\theta, R)$, we have that $\|s(x) - s(y)\| \le \widetilde{L}\alpha \|x-y\|$.*
- *For $x, y \in B(\theta, R)$, we have that $\langle s(y) - s(x), x - y\rangle \ge \alpha \|x-y\|^2$.*

Note that this definition can still imply that the distribution is mode-centered local well-conditioned, due to the following fact:

**Lemma E.12.** *Let $p$ be a probability density on $\mathbb{R}^d$. Fix $0 < r < R$ and $\theta \in \mathbb{R}^d$ such that*

$$\Pr_{x\sim p}[x \in B(\theta, r)] \ge 0.9, \qquad \nabla^2(-\log p(x)) \succeq \alpha I_d \quad (x \in B(\theta, R)),\ \alpha > 0.$$

*If $R > 4dr$, then there exists $\theta' \in B(\theta, 4dr)$ with $\nabla \log p(\theta') = 0$.*

We defer its proof to Section E.5. This implies the following lemma:

**Lemma E.13.** *Let $p$ be a $(\delta, r, R, \widetilde{L}, \alpha)$ locally well conditioned distribution with $R > 9dr$ and $\delta < 0.1$. Then $p$ is $(\delta, (4d+1)r, R-4dr, \widetilde{L}, \alpha)$ mode-centered locally well conditioned.*

This gives a version of Lemma E.4 for locally well-conditioned distributions as a corollary:

**Lemma E.14.** *Let $\rho = \frac{\|A\|}{\eta\sqrt{\alpha}}$. For all $0 < \varepsilon, \delta < 1$, there exists*

$$K \le \widetilde{O}\left(\frac{1}{\varepsilon\delta}\left(\frac{\rho^2\left(\left(m^2\rho^4+1\right)d^2\widetilde{r}^2 + m^3\rho^2 + dm\right)}{\sqrt{m}} + \frac{m}{d} + \log d\right)\right)$$

*such that: suppose distribution $p$ is a $(\frac{\varepsilon}{K^2}, \widetilde{r}/\sqrt{\alpha}, R, \widetilde{L}, \alpha)$ mode-centered locally well-conditioned distribution with $R \ge \frac{\sqrt{K\sqrt{m}/\alpha}}{\rho}$, and $\varepsilon_{score} \le \frac{\sqrt{\alpha/m}}{K^2\delta}$. Then Algorithm 1 samples from a distribution $\widehat{p}(x \mid y)$ such that*

$$\Pr_y[\mathrm{TV}(\widehat{p}(x\mid y), p(x\mid y)) \le \varepsilon] \ge 1-\delta.$$

*Furthermore, the total iteration complexity can be bounded by*

$$\widetilde{O}\left(K^3(K^2m^2d + m^3\rho + m^{1.5}(m\rho^2+1)\widetilde{r})(\widetilde{L}+\rho^2)^2 + K^3m^2\rho(\widetilde{L}+\rho^2)\right).$$

The reason we want this relaxed notion of locally well-conditioned is that, this captures the behavior of a Gaussian measurement. First note that:

**Lemma E.15.** *Let $p$ be a distribution on $\mathbb{R}^d$. Let $\widetilde{x} = x_{true} + N(0, \sigma^2 I_d)$ be a Gaussian measurement of $x_{true} \sim p$. Let $p_{\widetilde{x}}(x)$ be the posterior distribution of $x$ given $\widetilde{x}$. Then, for any $\delta \in (0,1)$ and $\delta' \in (0,1)$, with probability at least $1-\delta'$ over $\widetilde{x}$,*

$$\Pr_{x\sim p_{\widetilde{x}}}[x \in B(\widetilde{x}, r)] \ge 1-\delta$$

*for $r = \sigma(\sqrt{d} + \sqrt{2\log\frac{1}{\delta\delta'}})$.*

**Algorithm 2** Sampling from $p(x \mid x_0, y)$ given an extra Gaussian measurement $x_0$

```
function GaussianSampler($p : \mathbb{R}^d \to \mathbb{R}$, $x_0 \in \mathbb{R}^d$ , $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times d}$, $\eta, \sigma \in \mathbb{R}$)
    Let $p_{x_0}(x) := p(x \mid x + \mathcal{N}(0, \sigma^2 I_d) = x_0)$.
    Use Algorithm 1, return
        PosteriorSampler($p_{x_0}, y, A, \eta$).
end function
```

Again, we defer its proof to Section E.5. This implies the following lemma.

**Lemma E.16.** *For $\delta \in (0,1)$, suppose $p$ is a distribution over $\mathbb{R}^d$ such that*

$$\Pr_{x' \sim p}\left[\forall x \in B(x', R) : -LI_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2) I_d\right] \geq 1 - \delta.$$

*Given a Gaussian measurement $x_0 = x + \mathcal{N}(0, \sigma^2 I_d)$ of $x \sim p$ with*

$$\sigma \leq \frac{R}{2\sqrt{d} + \sqrt{2\log(1/\delta)} + 2\tau}.$$

*Let $x_0 = x + N(0, \sigma^2 I_d)$, where $x \sim p$. Then, suppose $R$. with probability at least $1 - 3\delta$ probability over $x_0$, $p_{x_0}$ is $(\delta, \sigma(\sqrt{d} + \sqrt{4\log\frac{1}{\delta}}), R/2, 2L\sigma^2 + 2, \frac{1}{2\sigma^2})$ locally well-conditioned.*

*Proof.* Let us check the locally well-conditioned conditions with $\theta = x_0$ one by one. The concentration follows directly from Lemma E.15, incurring an error probability of $\delta$.

By our choice of $\sigma$, we have that

$$\Pr\left[\|x_0 - x\| \leq \frac{R}{2}\right] \geq 1 - \delta.$$

Therefore,

$$\Pr\left[\forall x \in B(x_0, R/2) : -LI_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2) I_d\right] \geq 1 - 2\delta.$$

By direct calculation, we have that

$$-LI_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2) I_d \implies -(L + 1/\sigma^2) I_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2 - \frac{1}{\sigma^2}) I_d$$

By our choice of $\sigma$, we have that whenever $-LI_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2) I_d$,

$$-(2L\sigma^2 + 2)\frac{1}{2\sigma^2} I_d \preceq \nabla^2 \log p(x) \preceq -\frac{1}{2\sigma^2} I_d$$

This satisfies the Lipschitzness and the strong log-concavity condition by giving an additional error probability of $2\delta$. □

This gives us the main lemma for our local log-concavity case:

**Lemma E.17.** *For any $\delta, \varepsilon, \tau, \sigma, R, L > 0$, suppose $p(x)$ is a distribution over $\mathbb{R}^d$ such that*

$$\Pr_{x' \sim p}\left[\forall x \in B(x', R) : -LI_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2) I_d\right] \geq 1 - \delta.$$

*Let $\rho = \frac{\|A\|\sigma}{\eta}$. There exists*

$$K \leq \widetilde{O}\left(\frac{1}{\varepsilon\delta}\left(\frac{\rho^2\left(\left(m^2\rho^4 + 1\right)d^3 + m^3\rho^2 + dm\right)}{\sqrt{m}} + \frac{m}{d} + \log d\right)\right).$$

*such that: suppose $R^2 \geq (\frac{K\sqrt{m}}{\rho^2} + 4\tau)\sigma^2$ and $\varepsilon_{score} \leq \frac{1}{K^2\sqrt{m}\sigma}$, then Algorithm 2 samples from a distribution $\widehat{p}(x \mid x_0, y)$ such that*

$$\Pr_{x_0, y}\left[\mathrm{TV}(\widehat{p}(x \mid x_0, y), p(x \mid x_0, y)) \leq \varepsilon\right] \geq 1 - O(\delta).$$

*Furthermore, the total iteration complexity can be bounded by*

$$\widetilde{O}\left(K^3(K^2m^2d + m^3\rho + m^{1.5}(m\rho^2 + 1)\sqrt{d})(L\sigma^2 + \rho^2 + 1)^2 + K^3m^2\rho(L\sigma^2 + \rho^2 + 1)\right).$$

**Algorithm 3** Competitive Compressed Sensing Algorithm Given a Rough Estimation

```
function CompressedSensing($p : \mathbb{R}^d \to \mathbb{R}$, $x_0 \in \mathbb{R}^d$ , $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times d}$, $\eta, R \in \mathbb{R}$)
    Let $\sigma = R/\delta$.
    Sample $x_0' = x_0 + \mathcal{N}(0, \sigma^2 I_d)$.
    Use Algorithm 2, return
        $\textsc{GaussianSampler}(p, x_0', y, A, \eta, \sigma)$
end function
```

*Proof.* Combining Corollary E.10 with Lemma E.16 enables us to apply Lemma E.14 and proves the lemma. □

Expressing this in expectation, we have the following theorem.

**Theorem E.18** (Posterior sampling with local log-concavity). *For any $\varepsilon, \tau, R, L > 0$, suppose $p(x)$ is a distribution over $\mathbb{R}^d$ such that*

$$\Pr_{x'\sim p}\left[\forall x \in B(x', R) : -LI_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2)I_d\right] \geq 1 - \varepsilon.$$

*Let $\rho = \frac{\|A\|\sigma}{\eta}$. There exists*

$$K \leq \widetilde{O}\left(\frac{1}{\varepsilon\delta}\left(\frac{\rho^2\left(\left(m^2\rho^4+1\right)d^3 + m^3\rho^2 + dm\right)}{\sqrt{m}} + \frac{m}{d} + \log d\right)\right).$$

*such that: given a Gaussian measurement $x_0 = x + \mathcal{N}(0, \sigma^2 I_d)$ of $x \sim p$ with $R^2 \geq (\frac{K\sqrt{m}}{\rho^2} + 4\tau)\sigma^2$, and $\varepsilon_{score} \leq \frac{1}{K^2\sqrt{m}\sigma}$; then Algorithm 2 samples from a distribution $\widehat{p}(x \mid x_0, y)$ such that*

$$\mathop{\mathbb{E}}_{x_0,y}\left[\mathrm{TV}(\widehat{p}(x \mid x_0, y), p(x \mid x_0, y))\right] \lesssim \varepsilon.$$

*Furthermore, the total iteration complexity can be bounded by*

$$\widetilde{O}\left(K^3(K^2m^2d + m^3\rho + m^{1.5}(m\rho^2+1)\sqrt{d})(L\sigma^2+\rho^2+1)^2 + K^3m^2\rho(L\sigma^2+\rho^2+1)\right).$$

This gives us Theorem 1.2:

**Theorem 1.2** (Posterior sampling with local log-concavity). *For any $\varepsilon, \tau, R, L > 0$, suppose $p(x)$ is a distribution over $\mathbb{R}^d$ such that*

$$\Pr_{x'\sim p}\left[\forall x \in B(x', R) : -LI_d \preceq \nabla^2 \log p(x) \preceq (\tau^2/R^2)I_d\right] \geq 1 - \varepsilon.$$

*Then, there exist $K_1, K_2 = \mathrm{poly}(d, m, \frac{\|A\|\sigma}{\eta}, \frac{1}{\varepsilon})$ and $K_3 = \mathrm{poly}(d, m, \frac{\|A\|\sigma}{\eta}, \frac{1}{\varepsilon}, L\sigma^2)$ such that: Given a Gaussian measurement $x_0 = x + \mathcal{N}(0, \sigma^2 I_d)$ of $x \sim p$ with $\sigma \leq R/(K_1 + 2\tau)$. If $\varepsilon_{score} \leq \frac{1}{K_2\sigma}$, then there exists an algorithm that takes $K_3$ iterations to sample from a distribution $\widehat{p}(x \mid x_0, y)$ such that*

$$\mathop{\mathbb{E}}_{y,x_0}\left[\mathrm{TV}(\widehat{p}(x \mid x_0, y), p(x \mid x_0, y))\right] \lesssim \varepsilon.$$

### E.3 Compressed Sensing

In this section, we prove Corollary 1.3. We first describe the sampling procedure in Algorithm 3. Now we verify its correctness.

**Lemma E.19.** *For any $\delta, \tau, R, R', L > 0$, suppose $p(x)$ is a distribution over $\mathbb{R}^d$ such that*

$$\Pr_{x'\sim p}\left[\forall x \in B(x', R') : -LI_d \preceq \nabla^2 \log p(x) \preceq (\tau/R')^2 I_d\right] \geq 1 - \delta.$$

*Let $\rho = \frac{\|A\|R}{\eta}$. There exists*

$$K \leq \widetilde{O}\left(\frac{1}{\delta^2}\left(\frac{\rho^2\left(\left(m^2\rho^4+1\right)d^3 + m^3\rho^2 + dm\right)}{\sqrt{m}} + \frac{m}{d} + \log d\right)\right).$$

*such that: suppose* $(R')^2 \geq (\frac{K\sqrt{m}}{\rho^2} + 4\tau)R^2$ *and* $\varepsilon_{score} \leq \frac{1}{K^2\sqrt{m}R}$*, then conditioned on* $\|x_0 - x\| \leq R$*, Algorithm 3 of Algorithm 3 samples from a distribution* $\widehat{p}$ *(depending on* $x_0'$ *and* $y$*) such that*

$$\Pr_{x_0', y}\left[\mathrm{TV}(\widehat{p}, p(x \mid x + \mathcal{N}(0, \sigma^2 I_d) = x_0', Ax + \xi = y)) \leq \delta\right] \geq 1 - O(\delta).$$

*Furthermore, the total iteration complexity can be bounded by*

$$\widetilde{O}\left(K^3(K^2m^2d + m^3\rho + m^{1.5}(m\rho^2+1)\sqrt{d})(L\sigma^2+\rho^2+1)^2 + K^3m^2\rho(L\sigma^2+\rho^2+1)\right).$$

*Proof.* This is a direct application of Lemma E.17. The sole difference is that $x_0'$ follows $x_0 + \mathcal{N}(0, \sigma^2 I_d)$ instead of $x + \mathcal{N}(0, \sigma^2 I_d)$. Because $\|x_0 - x\| \leq R$, $x_0'$ remains sufficiently close to $x$ for the local Hessian condition to hold, so the proof of Lemma E.17 carries over verbatim. □

Now we explain why we want to sample from $p(x \mid x + \mathcal{N}(0, \sigma^2 I_d) = x_0', Ax + \xi = y)$. Essentially, the extra Gaussian measurement won't hurt the concentration of $p(x \mid y)$ itself. We abstract it as the following lemma:

**Lemma E.20.** *Let* $(X, Y)$ *be jointly distributed random variables with* $X \in \mathbb{R}^d$*. Assume that for some* $r > 0$ *and* $0 < \delta < 1$

$$\Pr_{Y,\ \widehat{X} \sim p(X|Y)}[\|X - \widehat{X}\| \leq r] \ \geq\ 1 - \delta.$$

*Define* $Z = X + \varepsilon$ *where* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ *is independent of* $(X, Y)$*. If*

$$\sigma \ \geq\ \frac{r}{2\delta},$$

*then for* $\widehat{X} \sim p(X \mid Y, Z)$ *one has*

$$\Pr_{Y, Z,\ \widehat{X}}[\|X - \widehat{X}\| \leq r] \ \geq\ 1 - 3\delta.$$

*Proof.* Fix $Y$ and draw an auxiliary point $\widetilde{X} \sim p(X \mid Y)$. Let $Z' = \widetilde{X} + \varepsilon'$ with $\varepsilon' \sim \mathcal{N}(0, \sigma^2 I_d)$ independent of everything else. On the event

$$E = \{\|X - \widetilde{X}\| \leq r\},$$

$Z$ and $Z'$ are Gaussians with the same covariance $\sigma^2 I_d$ and means $X$ and $\widetilde{X}$. Pinsker's inequality combined with the KL divergence between the two Gaussians gives

$$\mathrm{TV}(\mathcal{N}(X, \sigma^2 I_d), \mathcal{N}(\widetilde{X}, \sigma^2 I_d)) \ \leq\ \frac{\|X - \widetilde{X}\|}{2\sigma} \ \leq\ \frac{r}{2\sigma} \ \leq\ \delta.$$

Hence

$$\mathrm{TV}(\mathcal{L}(Y, X, Z), \mathcal{L}(Y, X, Z')) \ \leq\ \Pr[E^c] + \delta \ \leq\ 2\delta,$$

because $\Pr[E^c] \leq \delta$ by the hypothesis on $p(X \mid Y)$.

By construction,

$$p(X \mid Y) = \mathbb{E}_{Z'|Y}[p(X \mid Y, Z')],$$

so

$$\Pr_{Y, Z',\ \widehat{X} \sim p(X|Y,Z')}[\|X - \widehat{X}\| \leq r] \ \geq\ 1 - \delta.$$

For the set $A = \{(Y, Z, \widehat{X}) : \|X - \widehat{X}\| > r\}$ the total-variation bound gives

$$|\Pr_{Y,Z,\widehat{X}}[A] - \Pr_{Y,Z',\widehat{X}}[A]| \leq 2\delta,$$

whence

$$\Pr_{Y, Z,\ \widehat{X}}[\|X - \widehat{X}\| \leq r] \ \geq\ 1 - \delta - 2\delta = 1 - 3\delta.$$ □

This implies the following lemma:

**Lemma E.21.** *Consider the random variables in Algorithm 3. Suppose that*

- *Information theoretically, it is possible to recover $\widehat{x}$ from $y$ satisfying $\|\widehat{x} - x\| \leq r$ with probability $1 - \delta$ over $x \sim p$ and $y$.*
- $\Pr\left[\|x_0 - x\| \leq R\right] \geq 1 - \delta$.

*Then drawing sample $\widehat{x} \sim p(x \mid x + \mathcal{N}(0, \sigma^2 I_d) = x_0', Ax + \xi = y)$ would give that*

$$\Pr\left[\|x - \widehat{x}\| \leq 2r\right] \geq 1 - O(\delta).$$

*Proof.* By [JAD+21], the first condition implies that,

$$\Pr_{x,y,\widehat{x}\sim p(x|y)}\left[\|x - \widehat{x}\| \leq 2r\right] \geq 1 - 2\delta.$$

Then by Lemma E.20, suppose we have $x' = x + \mathcal{N}(0, \sigma^2 I_d)$, then

$$\Pr_{x,y,\widehat{x}\sim p(x|y,x+\mathcal{N}(0,\sigma^2 I_d)=x')}\left[\|x - \widehat{x}\| \leq 2r\right] \geq 1 - 6\delta.$$

Note that whenever $\|x - x_0\| \leq r$, we have

$$\mathrm{TV}(x' \mid x, x_0, x_0' \mid x, x_0) \leq \delta.$$

This proves that

$$\Pr_{x,y,\widehat{x}\sim p(x|y,x+\mathcal{N}(0,\sigma^2 I_d)=x_0')}\left[\|x - \widehat{x}\| \leq 2r\right] \geq 1 - 6\delta.$$

□

**Lemma E.22.** *Consider attempting to accurately reconstruct $x$ from $y = Ax + \xi$. Suppose that:*

- *Information theoretically, it is possible to recover $\widehat{x}$ from $y$ satisfying $\|\widehat{x} - x\| \leq r$ with probability $1 - \delta$ over $x \sim p$ and $y$.*
- *We have access to a "naive" algorithm that recovers $x_0$ from $y$ satisfying $\|x_0 - x\| \leq R$ with probability $1 - \delta$ over $x \sim p$ and $y$.*

*Let $\rho = \frac{\|A\|R}{\eta\delta}$. There exists*

$$K \leq \widetilde{O}\left(\frac{1}{\delta^2}\left(\frac{\rho^2\left(\left(m^2\rho^4 + 1\right)d^3 + m^3\rho^2 + dm\right)}{\sqrt{m}} + \frac{m}{d} + \log d\right)\right).$$

*such that: suppose for $R' = (R/\delta) \cdot \sqrt{\frac{K\sqrt{m}}{\rho^2} + 4\tau}$,*

$$\Pr_{x'\sim p}\left[\forall x \in B(x', R') : -LI_d \preceq \nabla^2 \log p(x) \preceq (\tau/R')^2 I_d\right] \geq 1 - \delta.$$

*Then we give an algorithm that recovers $\widehat{x}$ satisfying $\|\widehat{x} - x\| \leq 2r$ with probability $1 - O(\delta)$, in* $\mathrm{poly}(d, m, \frac{\|A\|R}{\eta}, \frac{1}{\delta})$ *time, under Assumption 1 with $\varepsilon_{score} < \frac{1}{K^2\sqrt{m}(R/\delta)}$.*

*Proof.* By our assumption and Lemma E.19, we have that we are sampling from $p(x \mid x + \mathcal{N}(0, \sigma^2 I_d) = x_0', Ax + \xi = y)$ with $\delta$ TV error with $1 - O(\delta)$ probability. By Lemma E.21, this would recover $x$ within distance $2r$ with $1 - O(\delta)$ probaility. Combining the two gives the result. □

Setting $\tau = 0$ would give Corollary 1.3 as a corollary.

**Corollary 1.3** (Competitive compressed sensing)**.** *Consider attempting to accurately reconstruct $x$ from $y = Ax + \xi$. Suppose that:*

- *Information theoretically (but possibly requiring exponential time or using exact knowledge of $p(x)$), it is possible to recover $\widehat{x}$ from $y$ satisfying $\|\widehat{x} - x\| \leq r$ with probability $1 - \delta$ over $x \sim p$ and $y$.*

- *We have access to a "naive" algorithm that recovers $x_0$ from $y$ satisfying $\|x_0 - x\| \leq R$ with probability $1 - \delta$ over $x \sim p$ and $y$.*
- *For $R' = R \cdot \operatorname{poly}(d, m, \frac{\|A\|R}{\eta}, \frac{1}{\delta})$,*

$$\Pr_{x' \sim p}\left[\forall x \in B(x', R') : -LI_d \preceq \nabla^2 \log p(x) \preceq 0\right] \geq 1 - \delta.$$

*Then we give an algorithm that recovers $\widehat{x}$ satisfying $\|\widehat{x} - x\| \leq 2r$ with probability $1 - O(\delta)$, in* $\operatorname{poly}(d, m, \frac{\|A\|R}{\eta}, \frac{1}{\delta})$ *time, under Assumption 1 with $\varepsilon_{score} < \frac{1}{\operatorname{poly}(d,m,\frac{\|A\|R}{\eta},\frac{1}{\delta},LR^2)R}$.*

### E.4 Ring example

Let $w \in (0, 0.01)$ and let $p_0$ be the uniform probability measure on the unit circle $S^1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$. Define the *circle–Gaussian mixture*

$$p(x) \;=\; (p_0 * \mathcal{N}(0, w^2 I_2))(x) \;=\; \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{2\pi w^2} \exp\Big(-\frac{\|x - (\cos\theta, \sin\theta)\|^2}{2w^2}\Big)\, d\theta, \qquad x \in \mathbb{R}^2.$$

**Lemma E.23.** *For any $x \in \mathbb{R}^2$ with radius $r = \|x\| > 0$, the Hessian of the log–density satisfies*

$$\nabla^2 \log p(x) \;\preceq\; \begin{cases} (\frac{1}{2w^4} - \frac{1}{w^2})I_2, & 0 < r \leq w^2, \\ (\frac{1}{w^2 r} - \frac{1}{w^2})I_2, & w^2 < r \leq 1, \\ 0, & r > 1. \end{cases}$$

*Proof.* Rotational invariance gives $p(x) = p(r)$ with

$$p(r) = \frac{1}{2\pi w^2} \exp\Big(-\frac{r^2+1}{2w^2}\Big) I_0\Big(\frac{r}{w^2}\Big), \qquad r \geq 0.$$

Write $f(r) = \log p(r)$ and set $z = r/w^2 > 0$. Using $I_0'(z) = I_1(z)$, we get the first and second derivatives:

$$f'(r) = \frac{-r + I_1(z)/I_0(z)}{w^2}, \qquad f''(r) = -\frac{1}{w^2} + \frac{I_0(z)I_2(z) - I_1(z)^2}{w^4 I_0(z)^2}.$$

For $r > 0$, the eigenvalues of $\nabla^2 \log p$ are

$$\lambda_r(r) = f''(r), \qquad \lambda_t(r) = \frac{f'(r)}{r}.$$

The Turán inequality $I_1(z)^2 - I_0(z)I_2(z) \geq 0$ implies $\lambda_r(r) \leq -1/w^2$; thus, the largest eigenvalue is $\lambda_t(r)$.

Since $I_1(z)/I_0(z) \leq 1$ for all $z > 0$ and $I_1(z)/I_0(z) \leq z/2$ for $0 < z \leq 1$,

$$\lambda_t(r) = -\frac{1}{w^2} + \frac{1}{w^2 r}\frac{I_1(z)}{I_0(z)} \;\leq\; \begin{cases} -\frac{1}{w^2} + \frac{1}{2w^4}, & 0 < r \leq w^2, \\ -\frac{1}{w^2} + \frac{1}{w^2 r}, & w^2 < r \leq 1, \\ 0, & r > 1. \end{cases}$$

□

**Lemma E.24.** *For every $x \in \mathbb{R}^2$, we have*

$$\nabla^2 \log p(x) \;\succeq\; -\frac{1}{w^2}\, I_2.$$

*Proof.* Write $u = (\cos\theta, \sin\theta)$ and

$$p(x) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2\pi w^2}\, e^{-\|x-u\|^2/(2w^2)}\, d\theta.$$

Differentiating under the integral gives

$$\nabla p(x) = \int \Big(-\frac{x-u}{w^2}\Big) \frac{1}{2\pi}\frac{1}{2\pi w^2} e^{-\|x-u\|^2/(2w^2)}\, d\theta = -\frac{1}{w^2}\, p(x)\,(x - \mathbb{E}[u \mid x]),$$

so

$$\nabla \log p(x) = -\frac{x - \mathbb{E}[u \mid x]}{w^2}.$$

Differentiating once more,

$$\nabla^2 \log p(x) = -\frac{I_2}{w^2} + \frac{1}{w^2}\, \nabla\, \mathbb{E}[u \mid x].$$

A standard score–covariance identity shows

$$\nabla\, \mathbb{E}[u \mid x] = \mathrm{Cov}_{\,u|x}(u, \frac{x-u}{w^2}) = \frac{1}{w^2}\, \mathrm{Cov}_{\,u|x}(u),$$

hence

$$\nabla^2 \log p(x) = \frac{\mathrm{Cov}_{\,u|x}(u)}{w^4} - \frac{I_2}{w^2}.$$

Since $\mathrm{Cov}_{\,u|x}(u) \succeq 0$, it follows that

$$\nabla^2 \log p(x) \succeq -\frac{1}{w^2}\, I_2,$$

as claimed. □

**Lemma E.25.** *For any $w \in (0, 1/2)$, we have that*

$$\Pr_{x' \sim p}\left[\forall x \in B(x', 1/2) : -\frac{1}{w^2} I_d \preceq \nabla^2 \log p(x) \preceq \frac{1}{2w^2} I_d\right] \geq 1 - e^{-\Omega(1/w^2)}.$$

*Proof.* Note that

$$\Pr_{x \sim p}\left[\|x\| > 3/4\right] \geq 1 - e^{-\Omega(1/w^2)}.$$

The rest follows by combining Lemma E.23 and Lemma E.24. □

Hence, we can apply Theorem 1.2 on our ring distribution $p$ and get the following corollary:

**Corollary E.26.** *Let $A \in \mathbb{R}^{C\times 2}$ be a matrix for some constant $C > 0$. Consider $x \sim p$ with two measurements given by*

$$x_0 = x + N(0, \sigma^2 I_2) \quad \textit{and} \quad y = Ax + N(0, \eta^2 I_2).$$

*Suppose $\|A\| w/\eta = O(1)$. Then, if $\sigma \leq cw$ and $\varepsilon_{score} \leq cw^{-1}$ for sufficiently small constant $c > 0$, Algorithm 2 takes a constant number of iterations to sample from a distribution $\widehat{p}(x \mid x_0, y)$ such that*

$$\mathop{\mathbb{E}}_{x_0, y}\left[\mathrm{TV}(\widehat{p}(x \mid x_0, y), p(x \mid x_0, y))\right] < 0.01.$$

## E.5 Deferred Proof

**Lemma E.8.** *Let $X$, $Y$, and $Z$ be random vectors in $\mathbb{R}^d$, where $Y = X + N(0, \sigma_1^2 I_d)$ and $Z = X + N(0, \sigma_2^2 I_d)$. The conditional density of $Z$ given $Y$, denoted $p(Z \mid Y)$, is a multivariate normal distribution with mean*

$$\mu_{Z|Y} = \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1} Y$$

*and covariance matrix*

$$\Sigma_{Z|Y} = \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1}\sigma_1^2.$$

*Then, the gradient of the log-likelihood $\log p(Z \mid Y)$ with respect to $Y$ is given by*

$$\nabla_Y \log p(Z \mid Y) = -\frac{1}{\sigma_1^2}\left(Z - \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1} Y\right).$$

*Proof.* Since $Z \mid Y \sim \mathcal{N}(\mu_{Z|Y}, \Sigma_{Z|Y})$, the log-likelihood function is

$$\log p(Z \mid Y) = -\frac{1}{2}\left((Z-\mu_{Z|Y})^T\Sigma_{Z|Y}^{-1}(Z-\mu_{Z|Y}) + \log\det(\Sigma_{Z|Y}) + d\log(2\pi)\right).$$

To compute the gradient with respect to $Y$, we focus on the term involving $\mu_{Z|Y}$:

$$-\frac{1}{2}\left((Z-\mu_{Z|Y})^T\Sigma_{Z|Y}^{-1}(Z-\mu_{Z|Y})\right).$$

Differentiating with respect to $Y$ gives:

$$\nabla_Y\left[(Z-\mu_{Z|Y})^T\Sigma_{Z|Y}^{-1}(Z-\mu_{Z|Y})\right] = -2\Sigma_{Z|Y}^{-1}(Z-\mu_{Z|Y})\cdot\nabla_Y\mu_{Z|Y}.$$

Since $\mu_{Z|Y} = \sigma_2^2(\sigma_1^2+\sigma_2^2)^{-1}Y$, we have

$$\nabla_Y\mu_{Z|Y} = \sigma_2^2(\sigma_1^2+\sigma_2^2)^{-1}.$$

Thus, the gradient becomes

$$\nabla_Y \log p(Z \mid Y) = -\Sigma_{Z|Y}^{-1}(Z-\mu_{Z|Y})\cdot\sigma_2^2(\sigma_1^2+\sigma_2^2)^{-1}.$$

Substituting the inverse of the covariance matrix $\Sigma_{Z|Y}$, we get

$$\Sigma_{Z|Y}^{-1} = \frac{1}{\sigma_1^2}\left(\sigma_1^2+\sigma_2^2\right),$$

and the final expression for the gradient is

$$\nabla_Y \log p(Z \mid Y) = -\frac{1}{\sigma_1^2}\left(Z - \sigma_2^2(\sigma_1^2+\sigma_2^2)^{-1}Y\right).$$

□

**Lemma E.12.** *Let $p$ be a probability density on $\mathbb{R}^d$. Fix $0 < r < R$ and $\theta \in \mathbb{R}^d$ such that*

$$\Pr_{x\sim p}[x \in B(\theta, r)] \geq 0.9, \qquad \nabla^2(-\log p(x)) \succeq \alpha I_d \quad (x \in B(\theta, R)),\ \alpha > 0.$$

*If $R > 4dr$, then there exists $\theta' \in B(\theta, 4dr)$ with $\nabla \log p(\theta') = 0$.*

*Proof.* By Lemma B.13, there is a normalised density $q$ satisfying $\nabla \log q = \nabla \log p$ on $B(\theta, R)$ and such that $\log q$ is $\alpha$-strongly concave on $\mathbb{R}^d$. The difference $\log p - \log q$ is therefore constant on $B(\theta, R)$; hence

$$p(x) = C\,q(x) \qquad (x \in B(\theta, R))$$

for some $C > 0$.

Let $\mu = \arg\max q$; strong concavity gives $\nabla \log q(\mu) = 0$ and uniqueness of $\mu$. Assume for contradiction that $\|\mu - \theta\| \geq 4dr$. Set $\lambda = 2r/\|\mu-\theta\| \leq 1/(2d)$ and define

$$\tau(x) = (1-\lambda)x + \lambda\mu.$$

Then $\det D\tau = (1-\lambda)^d$ and $\tau(B(\theta, r)) = B(\theta', (1-\lambda)r)$ with $\theta' = \tau(\theta) \subset B(\theta, R)$. Along any ray starting at $\mu$ the function $t \mapsto \log q(\mu + t(x-\mu))$ is strictly decreasing for $t \geq 0$; hence $q(\tau(x)) \geq q(x)$ for every $x$.

A change of variables yields

$$\Pr_q[B(\theta', (1-\lambda)r)] = \int_{B(\theta,r)} q(\tau(x))(1-\lambda)^d\,dx \geq (1-\lambda)^d \Pr_q[B(\theta, r)].$$

Because $\lambda \leq 1/(2d)$, $(1-\lambda)^d \geq e^{-1/2} > 0.6$. Multiplying by $C$ and using $p = Cq$ on $B(\theta, R)$ gives

$$\Pr_p[B(\theta', (1-\lambda)r)] \geq 0.6 \Pr_p[B(\theta, r)] \geq 0.54.$$

The two balls $B(\theta, r)$ and $B(\theta', (1-\lambda)r)$ are disjoint, so $1 \geq 0.9 + 0.54$, a contradiction. Thus $\|\mu - \theta\| < 4dr$.

Because $4dr < R$ we have $\mu \in B(\theta, R)$ and here $\nabla \log p = \nabla \log q$; consequently $\nabla \log p(\mu) = 0$. Putting $\theta' = \mu$ completes the proof. □

**Lemma E.15.** *Let $p$ be a distribution on $\mathbb{R}^d$. Let $\widetilde{x} = x_{true} + N(0, \sigma^2 I_d)$ be a Gaussian measurement of $x_{true} \sim p$. Let $p_{\widetilde{x}}(x)$ be the posterior distribution of $x$ given $\widetilde{x}$. Then, for any $\delta \in (0,1)$ and $\delta' \in (0,1)$, with probability at least $1 - \delta'$ over $\widetilde{x}$,*

$$\Pr_{x \sim p_{\widetilde{x}}}[x \in B(\widetilde{x}, r)] \geq 1 - \delta$$

*for $r = \sigma(\sqrt{d} + \sqrt{2 \log \frac{1}{\delta\delta'}})$.*

*Proof.* Let $Q(\widetilde{x}) = \Pr_{x \sim p_{\widetilde{x}}}[\|x - \widetilde{x}\| > r]$. We want to show that with probability at least $1 - \delta'$ over $\widetilde{x}$, $Q(\widetilde{x}) \leq \delta$. This is equivalent to showing that $\Pr_{\widetilde{x}}[Q(\widetilde{x}) > \delta] \leq \delta'$.

We use Markov's inequality. For any $\delta > 0$:

$$\Pr_{\widetilde{x}}[Q(\widetilde{x}) > \delta] \leq \frac{\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})]}{\delta}.$$

Thus, it suffices to show that $\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})] \leq \delta\delta'$.

Let's compute $\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})]$:

$$\begin{aligned}
\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})] &= \mathbb{E}_{\widetilde{x}}\left[\int_{\|x_1 - \widetilde{x}\| > r} p(x_1 \mid \widetilde{x}) dx_1\right] \\
&= \int p(\widetilde{x}) \left(\int_{\|x_1 - \widetilde{x}\| > r} p(x_1 \mid \widetilde{x}) dx_1\right) d\widetilde{x} \\
&= \int \left(\int_{\|x_1 - \widetilde{x}\| > r} p(x_1, \widetilde{x}) dx_1\right) d\widetilde{x}.
\end{aligned}$$

Using $p(x_1, \widetilde{x}) = p(\widetilde{x} \mid x_1) p(x_1)$, we can change the order of integration:

$$\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})] = \int p(x_1) \left(\int_{\|\widetilde{x} - x_1\| > r} p(\widetilde{x} \mid x_1) d\widetilde{x}\right) dx_1.$$

Given $x_1$, the distribution of $\widetilde{x}$ is $N(x_1, \sigma^2 I_d)$. Let $Z = \widetilde{x} - x_1$. Then $Z \sim N(0, \sigma^2 I_d)$. The inner integral is $\Pr_{Z \sim N(0, \sigma^2 I_d)}[\|Z\| > r]$. Let $W = Z/\sigma$. Then $W \sim N(0, I_d)$. The inner integral becomes $P_G(r/\sigma) = \Pr_{W \sim N(0, I_d)}[\|W\| > r/\sigma]$. So, $\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})] = \int p(x_1) P_G(r/\sigma) dx_1 = P_G(r/\sigma)$.

We need to show $P_G(r/\sigma) \leq \delta\delta'$. We use the standard Gaussian concentration inequality: for $W \sim N(0, I_d)$ and $t \geq 0$,

$$\Pr[\|W\| \geq \sqrt{d} + t] \leq e^{-t^2/2}.$$

We want $P_G(r/\sigma) \leq \delta\delta'$. So we set $e^{-t^2/2} = \delta\delta'$. This implies $t^2/2 = \log(1/(\delta\delta'))$, so $t = \sqrt{2 \log(1/(\delta\delta'))}$. This choice of $t$ is real and non-negative since $\delta, \delta' \in (0,1)$ implies $\delta\delta' \in (0,1)$, so $\log(1/(\delta\delta')) \geq 0$. We set $r/\sigma = \sqrt{d} + t = \sqrt{d} + \sqrt{2 \log(1/(\delta\delta'))}$. Thus, for $r = \sigma\left(\sqrt{d} + \sqrt{2 \log \frac{1}{\delta\delta'}}\right)$, we have $P_G(r/\sigma) \leq \delta\delta'$.

With this choice of $r$, we have $\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})] \leq \delta\delta'$. By Markov's inequality,

$$\Pr_{\widetilde{x}}[Q(\widetilde{x}) > \delta] \leq \frac{\mathbb{E}_{\widetilde{x}}[Q(\widetilde{x})]}{\delta} \leq \frac{\delta\delta'}{\delta} = \delta'.$$

This means that $\Pr_{\widetilde{x}}[Q(\widetilde{x}) \leq \delta] \geq 1 - \delta'$, which is the desired statement:

$$\Pr_{\widetilde{x}}\left[\Pr_{x \sim p_{\widetilde{x}}}[\|x - \widetilde{x}\| \leq r] \geq 1 - \delta\right] \geq 1 - \delta'.$$

□

# F Why Standard Langevin Dynamics Fails

As discussed in Section 3, after we get an initial sample $X_0 \sim p$ on the manifold, a natural attempt to get a sample from $p_y$ is to simply run vanilla Langevin SDE starting from $X_0$:

$$\mathrm{d}\, X_t \;=\; \Big(\widehat{s}(X_t) + \eta^{-2} A^{\mathsf{T}}(y - AX_t)\Big)\, \mathrm{d}\, t + \sqrt{2}\, \mathrm{d}\, B_t, \qquad X_0 \sim p \tag{13}$$

where $\widehat{s}(x)$ is an approximation to the true score $\nabla \log p(x)$. We now show that under any $L^p$ score accuracy assumption, the score error could get exponentially large as the dynamics evolves.

**Averaging over $y$ does not preserve the prior law.** We first consider the simplest one–dimensional Gaussian case of (13). Suppose $p = \mathcal{N}(0,1)$, $A = 1$, and noise $\xi = \mathcal{N}(0, \eta^2)$; so $y \sim \mathcal{N}(0, 1+\eta^2)$. Then with the perfect score estimator $\widehat{s}(X_t) = \nabla \log p(X_t) = -X_t$, (13) reduces to

$$\mathrm{d}\, X_t \;=\; \Big(-X_t + \eta^{-2}(y - X_t)\Big)\, \mathrm{d}\, t + \sqrt{2}\, \mathrm{d}\, B_t, \qquad X_0 \sim \mathcal{N}(0,1). \tag{14}$$

Recall that the hope of guaranteeing the robustness using only an $L^p$ guarantee is that at any time $t$, averaging $X_t$ over $y$ will preserve the original law $p$. We now show that this hope is unfounded even in this simplest case.

**Lemma F.1.** *Let $X_t$ follow (14). Averaging over $y \sim \mathcal{N}(0, 1+\eta^2)$, $X_t$ is Gaussian with mean $0$ and variance*

$$\mathrm{Var}(X_t) = e^{-2\alpha t} + \frac{1 - e^{-2\alpha t}}{\alpha} + \frac{(1 - e^{-\alpha t})^2}{1 + \eta^2} \le 1,$$

*where $\alpha := \frac{1+\eta^2}{\eta^2} > 1$. In particular,* $\mathrm{Var}(X_t) = 1 - \frac{1}{2(1+\eta^2)}$ *at time* $t^\star := \dfrac{\eta^2 \ln 2}{1 + \eta^2}$.

*Proof.* Write the mild solution of (14):

$$X_t = X_0 e^{-\alpha t} + \eta^{-2} y \int_0^t e^{-\alpha(t-s)}\, \mathrm{d}\, s + \sqrt{2} \int_0^t e^{-\alpha(t-s)}\, \mathrm{d}\, B_s = X_0 e^{-\alpha t} + \frac{y}{\eta^2}\frac{1 - e^{-\alpha t}}{\alpha} + \sqrt{2} \int_0^t e^{-\alpha(t-s)}\, \mathrm{d}\, B_s.$$

Because $X_0, B$ are independent of $y$, conditional moments are

$$\mathbb{E}[X_t \mid y] = \frac{y}{\eta^2}\frac{1 - e^{-\alpha t}}{\alpha}, \quad \mathrm{Var}(X_t \mid y) = e^{-2\alpha t} + \frac{1 - e^{-2\alpha t}}{\alpha}.$$

Applying the law of total variance with $\mathrm{Var}(y) = 1 + \eta^2$ gives the stated formula.

Since $X_0$ and $B$ are independent of $y$, conditioning on $y$ gives

$$\mathbb{E}[X_t \mid y] \;=\; \frac{y}{\eta^2}\frac{1 - e^{-\alpha t}}{\alpha}, \qquad \mathrm{Var}(X_t \mid y) \;=\; e^{-2\alpha t} + \frac{1 - e^{-2\alpha t}}{\alpha}.$$

By the law of total variance and $\mathrm{Var}(y) = 1 + \eta^2$,

$$\mathrm{Var}(X_t) \;=\; e^{-2\alpha t} + \frac{1 - e^{-2\alpha t}}{\alpha} + \frac{(1 - e^{-\alpha t})^2}{1 + \eta^2}.$$

Using $\alpha = (1 + \eta^2)/\eta^2$ and simple algebra, this simplifies to

$$\mathrm{Var}(X_t) \;=\; 1 - \frac{2\, e^{-\alpha t}(1 - e^{-\alpha t})}{1 + \eta^2},$$

which is at most 1 and attains $1 - 1/[2(1 + \eta^2)]$ when $e^{-\alpha t} = 1/2$, that is at $t^\star$. □

Thus $\mathrm{Var}(X_t)$ first *shrinks* below 1 (by a constant factor bounded away from 1 when $\eta$ is small) before relaxing back to equilibrium. The phenomenon is harmless in one dimension but is catastrophic in high dimension.

**High-dimensional amplification.** Let $p = \mathcal{N}(0, I_d)$, take $A = I_d$, and set $\eta^2 = 0.1$. Then with the perfect score estimator, (13) reduces to

$$\mathrm{d}\, X_t \;=\; \Big(-X_t + \eta^{-2}(y - X_t)\Big)\,\mathrm{d}\, t + \sqrt{2}\,\mathrm{d}\, B_t, \qquad X_0 \sim \mathcal{N}(0, I_d). \tag{15}$$

By Lemma F.1 applied coordinatewise, at time $t^\star := \dfrac{\eta^2 \ln 2}{1+\eta^2}$, averaging over $y$ yields

$$X_{t^\star} \sim p_{t^\star} := \mathcal{N}(0, \sigma^2 I_d) \quad \text{with} \quad \sigma^2 = 1 - \frac{1}{2(1+\eta^2)} = \frac{6}{11}.$$

Hence $X_{t^\star}$ is exponentially more concentrated in high dimension. We next show that this concentration amplifies score-estimation errors exponentially with the dimension.

**Lemma F.2.** *Let $p = \mathcal{N}(0, I_d)$ and let $p_{t^\star} = \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 = \dfrac{6}{11}$. For any finite $k > 1$ and $0 < \varepsilon < 1$, there exists a score estimate $\widehat{s} : \mathbb{R}^d \to \mathbb{R}^d$ such that*

$$\mathbb{E}_{x\sim p}\big[\|\widehat{s}(x) - \nabla \log p(x)\|^k\big] \le \varepsilon^k,$$

*yet*

$$\Pr_{x\sim p_{t^\star}}\big(\|\widehat{s}(x) - \nabla \log p(x)\| \ge e^{cd}\,\varepsilon\big) \;\ge\; 1 - 2e^{-cd}$$

*for some constant $c > 0$ depending only on $k$.*

*Proof.* Fix $k > 1$ and $0 < \varepsilon < 1$. Let $\sigma^2 = \dfrac{6}{11} \in (0,1)$ and choose $\rho \in (0, \min\{1/2,\, 1/\sigma^2 - 1\})$. Define the shell

$$S_\rho := \Big\{x \in \mathbb{R}^d :\; \big|\,\|x\|^2 - \sigma^2 d\,\big| \le \rho\,\sigma^2 d\Big\}.$$

Write $m := \Pr_{x\sim p}[x \in S_\rho]$ and $q := \Pr_{x\sim p_{t^\star}}[x \in S_\rho]$. Since $\|X\|^2/\sigma^2 \sim \chi^2_d$ under $p_{t^\star}$, the chi-square concentration inequality Lemma A.11 gives

$$q \;\ge\; 1 - 2\exp\!\left(-\frac{\rho^2 d}{8}\right).$$

Since $(1+\rho)\sigma^2 < 1$, the Chernoff left-tail bound for $\chi^2_d$ yields

$$m \;\le\; \Pr_{x\sim p}\big[\|x\|^2 \le (1+\rho)\sigma^2 d\big] \;\le\; \exp(-Id)\,, \qquad I := \frac{1}{2}\Big((1+\rho)\sigma^2 - 1 - \ln((1+\rho)\sigma^2)\Big) > 0.$$

Choose any unit vector $u$ and set

$$e(x) := M\,\mathbf{1}_{S_\rho}(x)\,u, \qquad \widehat{s}(x) := \nabla \log p(x) + e(x), \qquad M := \varepsilon\, m^{-1/k}.$$

Then

$$\mathop{\mathbb{E}}_{x\sim p}\big[\|\widehat{s}(x) - \nabla \log p(x)\|^k\big] = \mathop{\mathbb{E}}_{x\sim p}\big[\|e(x)\|^k\big] = M^k m = \varepsilon^k.$$

Moreover $\|e(x)\| \equiv M$ on $S_\rho$, hence

$$\Pr_{x\sim p_{t^\star}}[\|\widehat{s}(x) - \nabla \log p(x)\| \ge M] = \Pr_{x\sim p_{t^\star}}[x \in S_\rho] \ge 1 - 2e^{-\rho^2 d/8}.$$

Using $m \le e^{-Id}$ we have $M = \varepsilon\, m^{-1/k} \ge \varepsilon\, e^{(I/k)d}$. Setting

$$c := \min\left\{\frac{I}{k},\, \frac{\rho^2}{8}\right\} > 0,$$

which depends only on $\sigma$ and $k$, gives

$$\Pr_{x\sim p_{t^\star}}\big[\|\widehat{s}(x) - \nabla \log p(x)\| \ge e^{cd}\,\varepsilon\big] \;\ge\; 1 - 2e^{-cd}.$$

This completes the proof. □

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims are all reflected in the scope and contributions, as discussed in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations of our algorithm and theorems in the introduction.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The main paper contains proof overviews, and the appendices contain the full rigorous versions of all proofs of all stated lemmas and theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details for reproducing the experiments in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the code. The datasets we use are open access.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all details for the experiments in the paper

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars where relevant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information about the compute needed in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors read through and reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our contribution is domain-agnostic foundational research, so a detailed societal-impact discussion would necessarily be high-level and speculative. We do not believe it would be a useful use of reviewer time.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We are not releasing data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: There is no LLM usage in the core methods in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.