The Adaptive Complexity of Minimizing Relative Fisher Information

Huanjian Zhou
Graduate School of Frontier Sciences
The University of Tokyo
Center for Advanced Intelligence Project
RIKEN
zhou@ms.k.u-tokyo.ac.jp

Masashi Sugiyama
Center for Advanced Intelligence Project
RIKEN
Graduate School of Frontier Sciences
The University of Tokyo
sugi@k.u-tokyo.ac.jp

Abstract

Non-log-concave sampling from an unnormalized density is fundamental in machine learning and statistics. As datasets grow larger, computational efficiency becomes increasingly important, particularly in reducing adaptive complexity, namely the number of sequential rounds required for sampling algorithms. In this work, we initiate the study of the adaptive complexity of non-log-concave sampling within the framework of relative Fisher information introduced by Balasubramanian et al. in 2022. To obtain a relative Fisher information of at most ε^2 from the target distribution, we propose a novel algorithm that reduces the adaptive complexity from $\mathcal{O}(d^2/\varepsilon^4)$ to $\mathcal{O}(d/\varepsilon^2)$ by leveraging parallelism. Furthermore, we show our algorithm is optimal for a specific regime of large ε. Our algorithm builds on a diagonally parallelized Picard iteration, while the lower bound is based on a reduction from the problem of finding stationary points.

1 Introduction

We study the problem of adaptive sampling from a target distribution over $\mathbb{R}^d$ given query access to its unnormalized density, a fundamental task in areas such as Bayesian inference, randomized algorithms, and machine learning [MR$^+$07, NWS19, RCC99]. Recently, significant progress has been made in developing sequential algorithms for this problem, drawing inspiration from the extensive optimization toolkit, particularly when the target distribution is log-concave [JKO98, DMM19, MCC$^+$21]. Typically, when access to the function value is available, many high-accuracy samplers[1] have been designed based on Metropolis–Hastings filters or a proximal sampler [DCWY19, CDWY20, LST20, ALPW24, LST21, AC24, FYC23].

However, in many practical applications, such as energy-based models and Markov Decision Processes, evaluating the log-likelihood is often computationally intractable [LCH$^+$06, SB13]. In such scenarios, an alternative approach to designing high-accuracy samplers involves leveraging parallelism [SL19, YD24, ACV24, ZS24]. These algorithms effectively leverage contemporary parallel computing resources, such as multi-core central processing units (CPUs) and many-core graphics processing units (GPUs), especially since log-likelihood gradient evaluations often admit parallelization [HLB$^+$21, HLFS21]. In particular, the authors [SL19, YD24, ACV24, ZS24] proposed samplers that find an ε-accurate solution within $\mathcal{O}(\mathsf{poly}\log(d/\varepsilon^2))$ iterations, significantly improving upon the sub-polynomial complexity of $\mathcal{O}(d^a/\varepsilon^b)$ for log-concave distributions for some constant $a, b \in (0, 1)$.

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

[1] Samplers with complexity $\mathsf{poly}(d, \log(K_0/\varepsilon))$, assuming constant smoothness and condition number and K_0 as initial KL divergence.

In contrast, there are comparatively few works which study the adaptive complexity[2] when the target distribution is not strongly log-concave or are multimodal, such as mixtures of Gaussians.

To study the complexity of sampling from non-log-concave distributions, a general framework inspired by stationary point analysis in non-convex optimization (see, e.g., [N$^+$18]) has been developed. Specifically, Balasubramanian et al. [BCE$^+$22] proposed defining an ε-stationary point for sampling as any measure μ satisfying $\sqrt{\mathsf{FI}(\mu\|\pi)} \leq \varepsilon$, where $\mathsf{FI}(\mu\|\pi) = \mathbb{E}_\mu\left[\|\nabla \log(\mu/\pi)\|^2\right]$ denotes the *relative Fisher information* between μ and the target distribution π. They demonstrated that averaged Langevin Monte Carlo finds an ε-stationary point within $\mathcal{O}(dK_0/\varepsilon^4)$ iterations, where $K_0 := \mathsf{KL}(\mu_0\|\pi)$ represents the initial Kullback–Leibler divergence from the initial measure μ_0 to π. Furthermore, Chewi et al. [CGLL23] established an $\Omega(1/\varepsilon^2)$ query complexity lower bound for this setting. Existing studies offer only a few bounds on the query complexity of non-log-concave sampling, and our understanding of the adaptive complexity remains critically limited. This gap motivates our investigation into the question:

How many <u>sequential rounds</u> are needed to <u>minimize FI</u> for <u>non-log-concave sampling</u>?

1.1 Our Contribution

In this paper, we establish the *first* upper and lower bounds for the parallel runtime complexity of sampling. We now informally describe our main results. We assume access to an initial point $\boldsymbol{x}^0 \sim \mu_0$, where the KL divergence from the target distribution is $K_0 = \mathsf{KL}(\mu_0\|\pi)$ and the target distribution π is L-log-smooth.

New parallelized algorithm with improved complexity. By parallelizing the averaged Langevin Monte Carlo [BCE$^+$22], which has optimal query complexity for a specific regime of large FI ($\varepsilon = \sqrt{Ld}$) [CGLL23], we improve the adaptive complexity from $\mathcal{O}(\frac{L^2dK_0}{\varepsilon^4})$ to $\mathcal{O}(\frac{LK_0}{\varepsilon^2} + \log(\frac{Ld}{\varepsilon^2}))$ (Theorem 3.1). When all parameters except the dimension d are treated as constants, our algorithm improves the adaptive complexity from $\mathcal{O}(d)$ to $\mathcal{O}(\log d)$, matching the parallel speedup known for strongly log-concave sampling [ZS24]. Moreover when $K_0 = \mathcal{O}(d)$, which is common assumption and the analog of the optimal gap ($f(0) - \min f(\boldsymbol{x}) \lesssim d$) in non-convex optimization [CEL$^+$24, Appendix A], our algorithm achieves an adaptive complexity of $\widetilde{\mathcal{O}}(\frac{Ld}{\varepsilon^2})$, improving over the prior complexity of $\mathcal{O}(\frac{L^2d^2}{\varepsilon^4})$.

Lower bound. We further prove our parallelized algorithm is optimal for a specific regime of large FI ($\varepsilon = \sqrt{Ld}$) by showing when $\widetilde{\mathcal{O}}(K_0) \geq d \geq \widetilde{\Omega}(K_0^{2/3})$, the adaptive complexity is $\Omega(\frac{K_0}{d})$ (Theorem 4.1). For the accuracy level $\varepsilon = \sqrt{Ld}$, the adaptive complexity of the parallelized averaged Langevin Monte Carlo matches that of its sequential counterpart. Therefore, the sequential algorithm proposed by Balasubramanian et al. [BCE$^+$22] is also adaptively optimal in this regime. We summarize the comparison between existing bounds and our results in Table 1.

Moreover, although this lower bound only applies to a specific accuracy regime, it rules out the possibility of a general high-accuracy sampler via parallelism. This highlights a stark separation between log-concave and non-log-concave sampling, whereas Zhou et al.[ZS24] developed a general high-accuracy sampler via parallelism for strongly log-concave distributions, along with a tight lower bound on the accuracy [ZWS24].

A separation between optimization and sampling. Our work also highlights a fundamental difference between sampling and optimization: Unlike in sampling, no analogous separation exists between convex and non-convex optimization, as parallelism fails to accelerate gradient descent for either class in high-dimensional settings [BS18b, DG19, ZHTS25]. In contrast, for sampling, parallelism can accelerate Langevin Monte Carlo or its averaged version for both the strongly log-concave case [ZS24] and non-log-concave case (Theorem 3.1).

[2] Adaptive complexity refers to the minimal number of sequential rounds required for an algorithm to achieve a desired accuracy, assuming polynomially many queries can be executed in parallel at each round [BS18a, ZWS24].

Table 1: Comparisons of our lower bounds and upper bounds. Here, $\widetilde{\Omega}$ and $\widetilde{\mathcal{O}}$ omit logarithmic factors. K_0 denotes the initial KL divergence, defined as $K_0 = \mathsf{KL}(\mu_0 \| \pi)$, where the initial point is drawn from the distribution μ_0.

Works	Adaptive Complexity	Queries per Iteration
Sequential averaged Langevin Monte Carlo [BCE+22, Theorem 2]	$\mathcal{O}(\frac{L^2 d K_0}{\varepsilon^4})$	1
Parallelized averaged Langevin Monte Carlo Theorem 3.1	$\mathcal{O}(\frac{L K_0}{\varepsilon^2} + \log(\frac{Ld}{\varepsilon^2}))$	$\widetilde{\mathcal{O}}(\frac{L^2 d K_0}{\varepsilon^4})$
Lower bound for $\varepsilon = \sqrt{Ld}$ [CGLL23, Theorem 9]	$\Omega(\frac{K_0}{d})$	1
Lower bound for $\varepsilon = \sqrt{Ld}$ Theorem 4.1	$\Omega(\frac{K_0}{d})$	$\mathsf{poly}(d)$

1.2 Related works

Related works for minimizing FI**.** The relative Fisher information (FI) between two distributions quantifies the score matching error, which is the expected squared distance between their score functions (the gradients of their log-densities). In contrast, KL divergence (KL) compares the ratio of their density functions. When the reference distribution satisfies a log-Sobolev inequality or Poincaré inequality, small FI implies small KL or small total variation (TV). However, it is possible for KL or TV to remain small even when FI becomes arbitrarily large (see [Wib25, Appendix D]). Moreover, when the reference distribution is not log-concave, FI can be made arbitrarily small while TV remains bounded away from zero [BCE+22, Proposition 1].

A line of works in the literature investigates the mixing of FI for the sequential methods. For non-log-concave case, Balasubramanian et al. [BCE+22] analyzed the averaged Langevin Monte Carlo and Chewi et al. [CGLL23] proved two query complexity lower bound. For log-concave case, Chewi et al. [CGLL23, Appendix A] established a high-accuracy mixing time for proximal sampler with post-processing via the heat flow. For strongly log-concave case, Wibisono [Wib25] proved exponential convergence of the proximal sampler using the strong data processing inequality.

Related works for minimizing TV **or** KL **for non-log-concave case.** Another line of works study the complexity of minimizing TV or KL for non-log-concave sampling. For mixture Gaussian distributions with components sharing the same shape, Ge et al. [LRG18] showed polynomially many queries are sufficient to minimize TV for simulated tempering Langevin Monte Carlo. For general non-log-concave distributions, Guo et al. [GTC24] analyzed annealed Langevin Monte Carlo and established polynomial query complexity in terms of the action associated with a curve of probability measures interpolating the target distribution and a readily sampleable distribution. Notably, when the components of a mixture of Gaussians share the same shape and mode norm, the corresponding action grows only polynomially with respect to the dimension. Recently, He et al. [HZ25] proved that the optimal query complexity scales as $(\frac{L\mathsf{m}_2}{d\varepsilon})^{\Theta(d)}$ where m_2 is the second moment of the target distribution.

2 Preliminaries

2.1 Problem setting

Given the potential function $V : \mathcal{D} \to \mathbb{R}$, the goal of the sampling task is to draw a sample from the density $\pi_V = Z_V^{-1} \exp(-V)$, where $Z_V := \int_{\mathcal{D}} \exp(-V) \mathrm{d}\mathbf{x}$ is the normalizing constant.

Distribution class and assumption. If V is twice-differentiable and $\nabla^2 V \preceq LI$ with $L > 0$ (where $\preceq$ denotes the Loewner order and I is the identity matrix), we say the distribution π_V is *L-log-smooth*.

Oracle. Given the potential function V, and a query $\boldsymbol{x} \in \mathcal{D}$, the 0-th order oracle answers the function value $V(\boldsymbol{x})$ and the 1-st order oracle answers both $V(\boldsymbol{x})$ and its gradient value $\nabla V(\boldsymbol{x})$. We denote the oracle as Or. We also extend the oracle to parallel case, with input as $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \in \mathbb{R}^{dk}$ and return multiple answers $\{V(\boldsymbol{x}_1), \nabla V(\boldsymbol{x}_1), \ldots, V(\boldsymbol{x}_k), \nabla V(\boldsymbol{x}_k)\} \in \mathbb{R}^{2dk}$ with $k = \mathsf{poly}(d)$.

The adaptive algorithm class The class of *adaptive* algorithms is formally defined as follows [DG19]. For any dimension d, an adaptive algorithm A takes $V : \mathbb{R}^d \to \mathbb{R}$ and a (possibly random) initial point $\boldsymbol{x}^0$ and iteration number r as input and returns an output $\boldsymbol{x}^{r+1}$, which is denoted as $\mathsf{A}[V, \boldsymbol{x}^0, r] = \boldsymbol{x}^{r+1}$. At iteration $i \in [r] := \{1, \ldots, r\}$, A performs a batch of queries

$$Q^i = \{\boldsymbol{x}^{i,1}, \ldots, \boldsymbol{x}^{i,k_i}\}, \quad \text{with } \boldsymbol{x}^{i,j} \in \mathcal{D},\ \ j \in [k_i],\ \ k_i = \mathsf{poly}(d),$$

such that for any $m, n \in [k_i]$, $\boldsymbol{x}^{i,m}$ and $\boldsymbol{x}^{i,n}$ are *conditionally independent* given all existing queries $\{Q^j\}_{j\in[i-1]}$ and $\boldsymbol{x}^0$. Give queries set Q_i, the oracle returns a batch of answers: $\mathsf{Or}(Q_i) = \{\mathsf{Or}\boldsymbol{x}^{i,1}), \ldots, \mathsf{Or}(\boldsymbol{x}^{i,k_i})\}$.

An adaptive algorithm A is *deterministic* if in every iteration $i \in \{0, \ldots, r\}$, A operates with the form

$$Q^{i+1} = \mathsf{A}^i(Q^0, \mathsf{Or}(Q^0), \ldots, Q^i, \mathsf{Or}(Q^i)),$$

where A^i is mapping into $\mathbb{R}^{dk_{i+1}}$ with $Q^{r+1} = \boldsymbol{x}^{r+1}$ as output and $Q^0 = \boldsymbol{x}^0$ as an initial point. We denote the class of adaptive deterministic algorithms by $\mathcal{A}_{\text{det}}$.

An adaptive *randomized* algorithm has the form

$$Q^{i+1} = \mathsf{A}^i(\xi_i, Q^0, \mathsf{Or}(Q^0), \ldots, Q^i, \mathsf{Or}(Q^i)),$$

given access to a random uniform variable on $[0, 1]$ (i.e., infinitely many random bits), where A^i is mapping into $\mathbb{R}^{dk_{i+1}}$. We denote the class of adaptive randomized algorithms by $\mathcal{A}_{\text{rand}}$.

Measure of the output Consider the joint distribution of all involved points $\{\mathbf{x} : \mathbf{x} \in Q^i, i = 0, \ldots, r+1\}$ and the random bits ξ_i. Let the marginal distribution of the output $\mathbf{x}^{r+1}$ be ρ. We say the output to be ε^2-accurate in relative Fisher information (FI) if $\mathsf{FI}(\rho, \pi_V) := \mathbb{E}_\rho\left[\left\|\nabla \log(\rho/\pi)\right\|^2\right] \le \varepsilon^2$.

Initialization. We assume access to an initialization oracle that returns a sample from a distribution μ_0 satisfying $\mathsf{KL}(\mu_0\|\pi) \le K_0$ since it suffice to find a stationary point which lies in a ball of radius $\mathcal{O}(\sqrt{d})$, centered at the minimizer of f [CEL+24]. And such a stationary point can be found fast in both strongly-convex or non-convex cases [BV04, BM20].

Notion of complexity Given $\varepsilon > 0$, $V \in \mathcal{F}$, and some algorithm A, define the running iteration $\mathsf{T}(\mathsf{A}, V, K_0, \varepsilon)$ as the minimum number of rounds such that given a initial point with initial KL divergence upper bounded as K_0, algorithm A outputs a solution $\boldsymbol{x}$ whose marginal distribution ρ satisfies $\mathsf{FI}(\rho, \pi_V) \le \varepsilon$, i.e., $\mathsf{T}(\mathsf{A}, V, K_0, \varepsilon) = \sup\left\{\boldsymbol{x}^0 \sim \rho_0,\ s.t.\ \mathsf{KL}(\rho_0\|\pi) \le K_0 : \inf\left\{t : \mathsf{FI}\left(\rho(\mathsf{A}[V, \boldsymbol{x}^0, t]), \pi_f\right) \le \varepsilon\right\}\right\}$[3]. We define the *worst case* complexity as

$$\mathsf{Comp}_{\mathsf{WC}}(\mathcal{F}, \varepsilon, K_0) := \inf_{\mathsf{A}\in\mathcal{A}_{\text{det}}} \sup_{V\in\mathcal{F}} \mathsf{T}(\mathsf{A}, V, K_0, \varepsilon).$$

For some randomized algorithm $\mathsf{A} \in \mathcal{A}_{\text{rand}}$, we define the *randomized* complexity as[4]

$$\mathsf{Comp}_{\mathsf{R}}(\mathcal{F}, \varepsilon, K_0) := \inf_{\mathsf{A}\in\mathcal{A}_{\text{rand}}} \sup_{V\in\mathcal{F}} \mathsf{T}(\mathsf{A}, V, K_0, \varepsilon).$$

By definition, we have $\mathsf{Comp}_{\mathsf{WC}}(\mathcal{F}, \varepsilon, K_0) \ge \mathsf{Comp}_{\mathsf{R}}(\mathcal{F}, \varepsilon, K_0)$. In the rest of this paper, we only consider the randomized complexity and we lower-bound it by considering the *distributional* complexity:

$$\mathsf{Comp}_{\mathsf{D}}(\mathcal{F}, \varepsilon, K_0) := \sup_{F\in\Delta(\mathcal{F})} \inf_{\mathsf{A}\in\mathcal{A}_{\text{rand}}} \mathop{\mathbb{E}}_{V\sim F} \mathsf{T}(\mathsf{A}, V, K_0, \varepsilon),$$

where $\Delta(\mathcal{F})$ is the set of probability distributions over the class of functions $\mathcal{F}$.

[3] We note that in sampling, the iteration complexity is determined by the output of the last iteration, which is analogous to last-iteration properties in optimizations [ALW19].

[4] We note that in sampling, we cannot define the randomized complexity as the expected running iteration over mixtures of deterministic algorithms as in the case of optimization [BGP17], since the intrinsic randomness ξ_i will affect the marginal distribution of output. Furthermore, Yao's minimax principle [AB09] cannot be applied, since the different definition of randomized complexity. We acknowledge that another possible option not discussed in this paper is the "Las Vegas" algorithm, which can return "failure," as described in [AC24].

2.2 Averaged Langevin Monte Carlo

One of the most commonly-used dynamics for sampling is Langevin dynamics [Che23], which is the solution to the following SDE, $\mathrm{d}\boldsymbol{x} = -\nabla V(\boldsymbol{x})\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t$, where $(\boldsymbol{B}_t)_{t\in[0,T]}$ is a standard Brownian motion in $\mathbb{R}^d$. When π is α strongly log-concave or V is α strongly convex, the time derivative of the relative Fisher information satisfies

$$\partial_t \mathsf{FI}(\mu_t\|\pi) \le -2\alpha \mathsf{FI}(\mu_t\|\pi),$$

where μ_t is the law at time t, (see Section 2.5 [Wib25]). To the best of our knowledge, when π_V is non-log-concave, its contraction properties remain unknown. However, a discrete-time analog of the de Bruijn identity holds for the Langevin Monte Carlo with step size $h \lesssim \frac{1}{L}$ [BCE$^+$22, Appendix B]:

$$\partial_t \mathsf{KL}(\mu_t\|\pi) \le -\frac{1}{2}\mathsf{FI}(\mu_t\|\pi) + \mathcal{O}(L^2dh).$$

By integrating and summing, the averaged FI along Langevin Monte Carlo can be bounded as

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t\|\pi)\mathrm{d}t \le \frac{2\mathsf{KL}(\mu_0\|\pi)}{Nh} + \mathcal{O}(L^2dh).$$

By the convexity of the Fisher information, it is sufficient to output a sample from the averaged distribution $\bar{\mu}_{Nh} = \frac{1}{Nh}\int_0^{Nh}\mu_t\mathrm{d}t$.

2.3 Parallelized Langevin Monte Carlo

The main idea of parallelized Langevin Monte Carlo is to regroup the discrete grids along time horizon and update all grids in same group simultaneously [SL19, ACV24, YD24, ZS24]. Specifically, taking Picard iteration [Cle57, ACV24] as example, to approximate the difference $\boldsymbol{x}_{t_{n+1}} - \boldsymbol{x}_{t_n}$ over time slice $[t_n, t_{n+1}]$ as

$$\begin{aligned}\boldsymbol{x}_{t_{n+1}} - \boldsymbol{x}_{t_n} &= \int_{t_n}^{t_{n+1}} V(\boldsymbol{x}_s)\mathrm{d}s + \sqrt{2}(\boldsymbol{B}_{t_{n+1}} - \boldsymbol{B}_{t_n}) \\ &\approx \sum_{i=1}^{M} w_i V(\boldsymbol{x}_{t_n+\tau_{n,i}})\mathrm{d}s + \sqrt{2}(\boldsymbol{B}_{t_{n+1}} - \boldsymbol{B}_{t_n}),\end{aligned}$$

with a discrete grid of M collocation points as $t_n = t_n + \tau_{n,0} \le t_n + \tau_{n,1} \le t_n + \tau_{n,2} \le \cdots \le t_n + \tau_{n,M} = t_{n+1}$. We update the points in a wave-like fashion, which inherently allows for parallelization: for $m' = 1, \ldots, M$, $p = 0, 1, \ldots, K-1$,

$$\boldsymbol{x}^{p+1}_{t_n+\tau_{n,m}} = \boldsymbol{x}_{t,n} + \sum_{m=1}^{M-1} w_m V(\boldsymbol{x}^p_{t_n+\tau_{n,m}}) + \sqrt{2}(\boldsymbol{B}_{t_n+\tau_{n,m}} - \boldsymbol{B}_{t_n}).$$

With such regrouping, as long as the total time length of each group scales as $\mathcal{O}(1/L)$, the grids will converge exponentially fast. Given a sufficiently accurate starting point at time t_n, the initial error scales as $\mathcal{O}(d)$. Therefore, $K = \mathcal{O}\left(\log\left(\frac{d}{\varepsilon^2}\right)\right)$ steps suffice for the convergence of each group. In the strongly log-concave case, it suffices to simulate Langevin dynamics over the time interval $[0, \mathcal{O}(\log(\frac{\mathsf{KL}(\mu_0\|\pi)}{\varepsilon^2}))]$. Therefore, $N = \mathcal{O}(\log(\frac{\mathsf{KL}(\mu_0\|\pi)}{\varepsilon^2}))$ groups are sufficient, and the total number of steps scales as $KN = \mathcal{O}(\log^2(\frac{d}{\varepsilon^2}))$, assuming $\mathsf{KL}(\mu_0\|\pi) = \mathcal{O}(d)$. Recently, Zhou et al. [ZS24] showed the sequential update over each group is not necessary and proposed a diagonal style update with $\mathcal{O}(\log(\frac{d}{\varepsilon^2}))$ total steps (See Figure 1).

3 Parallel Picard method for minimizing relative Fisher information

In this section, we present parallel Picard methods for minimizing relative Fisher information in non-log-concave case (Algorithm 1) and show it holds improved convergence rate (Theorem 3.1). We illustrate the algorithm in Section 3.1, and give a proof sketch in Section 3.2. All the missing proofs can be found in Appendix A.

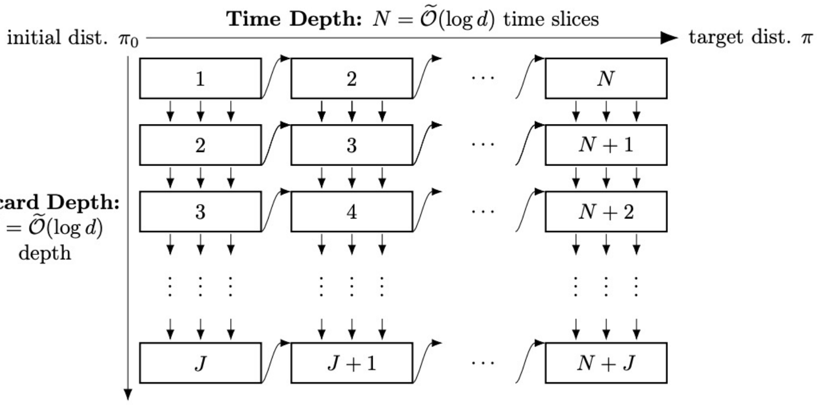
Time Depth: $N = \widetilde{\mathcal{O}}(\log d)$ time slices
initial dist. π_0
target dist. π
Picard Depth:
$J = \widetilde{\mathcal{O}}(\log d)$
depth
1
2
N
2
3
$N+1$
3
4
$N+2$
J
$J+1$
$N+J$

Figure 1: Illustration of the parallel Picard method: each rectangle represents an update, and the number within each rectangle indicates the index of the Picard iteration. The approximate time complexity is $N + J = \widetilde{\mathcal{O}}(\log d)$.

Theorem 3.1. *Given two integer parameters $J \geq N$ and M and an access to the gradient oracle of ∇V, there is an algorithm that runs $N + J$ iterations with at most $M(N + J)$ queries per iteration and outputs a sample with marginal distribution ρ such that*

$$\mathsf{FI}(\rho\|\pi) \lesssim \underbrace{\frac{L\mathsf{KL}(\mu_0\|\pi)}{N}}_{convergence\ of\ averaged\ LMC} + \underbrace{\frac{Ld}{M}}_{discretization\ error} + \underbrace{\left(\frac{L\mathsf{KL}(\mu_0\|\pi)}{N} + Ld\right)0.5^{J-N}}_{parallization\ error}. \tag{1}$$

Furthermore, by setting $N = \frac{L\mathsf{KL}(\mu_0\|\pi)}{\varepsilon^2}$ and $M = \frac{Ld}{\varepsilon^2}$ and $J - N = \mathcal{O}(\log(\frac{Ld}{\varepsilon^2}))$, the algorithm runs within $\mathcal{O}(\frac{LK_0}{\varepsilon^2} + \log\left(\frac{Ld}{\varepsilon^2}\right))$ iterations with at most $\mathcal{O}(\frac{L^2dK_0}{\varepsilon^4} + \frac{Ld}{\varepsilon}\log\left(\frac{Ld}{\varepsilon^2}\right))$ queries per iteration, and returns a ε^2-accurate sample $\boldsymbol{x}$ in FI.

Remark 3.2. *We can usually take $\mathsf{KL}(\mu_0\|\pi) = \mathcal{O}(d)$. Then taking*

$$N = \mathcal{O}\left(\frac{Ld}{\varepsilon^2}\right), \quad M = \mathcal{O}\left(\frac{Ld}{\varepsilon^2}\right), J = N + \mathcal{O}\left(\log\frac{Ld}{\varepsilon^2}\right).$$

the algorithm runs $\mathcal{O}\left(\frac{Ld}{\varepsilon^2}\right)$ iterations with $\mathcal{O}\left(\frac{L^2d^2}{\varepsilon^4}\right)$ queries per iteration and return a sample having ε accuracy in terms of the Fisher information w.r.t. the target.

Remark 3.3. *Compared to the bound for the sequential method presented in [BCE+22], our upper bound (right-hand side of Eq. (1)) for the parallel method also include one converge term $\frac{L\mathsf{KL}(\mu_0\|\pi)}{N}$ and a discretization error term $\frac{Ld}{M}$, and an additional exponentially decaying error term by parallelism $\left(\frac{L\mathsf{KL}(\mu_0\|\pi)}{N} + Ld\right)0.5^{J-N}$.*

Remark 3.4 (Tradeoff between query per round and adaptive complexity). *When the number of computation cores, denoted by W, is limited, the adaptive complexity of our algorithm is $\widetilde{\mathcal{O}}\left(\frac{d}{\varepsilon^2} + \frac{d^2}{\varepsilon^4W}\log\left(\frac{d}{\varepsilon^2}\right)\right)$. When $W = 1$, this recovers the sequential method; when $W = \widetilde{\mathcal{O}}\left(\frac{d^2}{\varepsilon^4}\right)$ with assumption $K_0 = \mathcal{O}(d)$, it recovers our fully parallel method. Another interesting intermediate case arises when applying averaged Langevin Monte Carlo to the algorithm in [ACV24], which updates time slices sequentially. This corresponds to $W = \frac{d}{\varepsilon^2}$, yielding an adaptive complexity of $\mathcal{O}\left(\frac{d}{\varepsilon^2}\log(\frac{d}{\varepsilon^2})\right)$ which fits naturally within the above tradeoff curve.*

3.1 Parallel Picard method

We adopt the parallel Picard method [ZS24] which achieve nearly tight result for log-concave sampling. In Lines 5–9, we apply the averaged Langevin Monte Carlo [BCE+22] to initialize the vector value at all grids with Fisher information bounded by $\mathcal{O}(d)$. Specifically, we initialize $\boldsymbol{x}_{n,m}^0$ at all points along the time horizon using the output of the averaged Langevin Monte Carlo procedure in Line 9. In Lines 1–4, we generate the random noises and fixed them. Subsequently, we seek to construct an approximate path of the true Langevin dynamics by means of parallel computation. Specifically, in Lines 10–16, we apply parallel Picard method with forward Euler-Maruyama Method in diagonal style as illustrated in Figure 1. In j-th each update, for m-th grid in n-th time slices, we perform

$$\boldsymbol{x}^{j}_{n,m} = \boldsymbol{x}^{j}_{n,0} - \frac{h}{M}\sum_{m'=0}^{m-1} \nabla V(\boldsymbol{x}^{j-1}_{n,m'}) + \sqrt{2}(B_{nh+mh/M} - B_{nh}),$$

where $\boldsymbol{x}^{\cdot}_{n,m}$ corresponds to time $nh + \frac{m}{M}h$. In Lines 17, we return the average point along the interpolation path of $\{\boldsymbol{x}^{j}_{n,m} : n \in [N], m \in [M]\}$.

Algorithm 1: Parallel Picard method for non-log-concave sampling

```
Input : $x^0 \sim \mu_0$, gradient oracle of $\nabla V$, the number of the iterations in outer loop $J$, the number
        of time slices $N$, the length of time slices $h$, the number of points on each time slices $M$.
1  for $n = 0, \dots, N-1$, $m = 0, \dots, M$ (in parallel) do
2      $\xi_{n,m} = \mathcal{N}(0, (h/M)\boldsymbol{I}_d)$                          ▷ generate the noise
3  for $n = 0, \dots, N-1$, $m = 0, \dots, M$ (in parallel) do
4      $B_{nh+(m+1)h/M} = \sum_{n'=\{0,\dots,n-1\}} \sum_{m' \in [M]} \xi_{n',m'} + \sum_{m' \in [m]} \xi_{n,m'}$
5  $\boldsymbol{x}^{j}_{-1,0} = \boldsymbol{x}^0$, for $j = -1, \dots, J$
6  for $k = 1, \dots, N$ do
7      $\boldsymbol{x}^{-1}_{kh} = \boldsymbol{x}^{-1}_{(k-1)h} - h\nabla V(\boldsymbol{x}^{-1}_{(k-1)h}) + \sqrt{2}(B_{kh} - B_{(k-1)h})$,     ▷ initialization
8  Pick a time $t \in [0, Nh]$ uniformly at random,
9  Let $k$ be the largest integer such that $kh \le t$, for all $n = 0, \dots, N-1$ and $m \in [M]$, set
       $\boldsymbol{x}^{0}_{n,m} = \boldsymbol{x}^{-1}_{kh} - (t - kh)\nabla V(\boldsymbol{x}^{-1}_{kh}) + \sqrt{2}(B_t - B_{kh})$.
   for $k = 1, \dots, N$ do
10     for $j = 1, \dots, \min\{k-1, J\}$ and $m = 1, \dots, M$ (in parallel) do
11         let $n = k - j$ and $\boldsymbol{x}^{j}_{n,0} = \boldsymbol{x}^{j}_{n-1,M}$,
12         $\boldsymbol{x}^{j}_{n,m} = \boldsymbol{x}^{j}_{n,0} - \frac{h}{M}\sum_{m'=0}^{m-1} \nabla V(\boldsymbol{x}^{j-1}_{n,m'}) + \sqrt{2}(B_{nh+mh/M} - B_{nh})$,
13 for $k = N+1, \dots, N+J-1$ do
14     for $n = \max\{0, k-J\}, \dots, N-1$ and $m = 1, \dots, M$ (in parallel) do
15         let $j = k - n$ and $\boldsymbol{x}^{j}_{n,0} = \boldsymbol{x}^{j}_{n-1,M}$,
16         $\boldsymbol{x}^{j}_{n,m} = \boldsymbol{x}^{j}_{n,0} - \frac{h}{M}\sum_{m'=0}^{m-1} \nabla V(\boldsymbol{x}^{j-1}_{n,m'}) + \sqrt{2}(B_{nh+mh/M} - B_{nh})$,
17 Pick a time $t \in [0, Nh]$ uniformly at random, let $k$ be the largest integer such that $kh/M \le t$,
       $\boldsymbol{x}_t = \boldsymbol{x}^{J}_{\lfloor k/M \rfloor, k - \lfloor k/M \rfloor M} - (t - kh/M)\nabla V(\boldsymbol{x}^{J}_{\lfloor k/M \rfloor, k - \lfloor k/M \rfloor M}) + \sqrt{2}(B_t - B_{\lfloor kh/M \rfloor})$.
   return $\boldsymbol{x}_t$.
```

3.2 Proof sketch of Theorem 3.1: analysis of Algorithm 1

Following [VW19, ACV24, ZS24], we use interpolation method to obtain a discrete-time analog of de Bruijn identity:

$$\partial_t \mathsf{KL}(\mu_t \| \pi) \le -\frac{3}{4}\mathsf{FI}(\mu_t \| \pi) + \mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t) - \nabla V(\boldsymbol{x}^{J-1}_{n,m})\right\|^2\right].$$

By Lipschitz condition, Integrating over time $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$ and summing, we can upper bound the averaged FI along the discrete trajectory by

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \| \pi)\mathrm{d}t \le \frac{2K_0}{Nh} + \frac{2Ld}{M} + 3L^2\mathcal{E},$$

where $\mathcal{E} = \max_{n\in[N],\, m\in[M]} \mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^{J} - \boldsymbol{x}_{n,m}^{J-1}\right\|^2\right]$ which represents the convergence error of Picard iteration. It remains to prove the convergence of Picard iteration. The key is to decompose the error during the diagonal update and upper bound the initial error. Let $\mathcal{E}_n^j := \max_{m\in[M]} \mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^{j} - \boldsymbol{x}_{n,m}^{j-1}\right\|^2\right]$. By definition of Euler-Maruyama scheme, we can decompose the error as

$$\mathcal{E}_n^j \leq 2\mathcal{E}_{n-1}^j + 2h^2L^2\mathcal{E}_{n-1}^j,$$

where h is the length of time slices, and L is Lispchitz constant of the gradient. Thus by choosing the time length h sufficiently small relative to the Lipschitz constant L, we ensure convergence along the Picard direction. As for the initial error, we decompose it as

$$\mathcal{E}_n^1 \leq 2\mathcal{E}_{n-1}^1 + 2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh,$$

where $\mathsf{FI}(\mu^0\|\pi)$ is controlled by the initialization procedure using averaged Langevin Monte Carlo with a large step size $h = \mathcal{O}(1)$.

The details of the proof can be found in Appendix A.

4 Lower bound

In this section, we establish our lower bound (Theorem 4.1) by combining a reduction from minimizing the relative Fisher information to the problem of finding a stationary point with the optimal adaptive complexity for this task, as established by Zhou et al. [ZHTS25].

Theorem 4.1. *Let the dimension d satisfies $\widetilde{\mathcal{O}}(K_0) \geq d \geq \widetilde{\Omega}(K_0^{2/3})$. There exists a function class $\mathcal{F}$, consisting of L-smooth functions, such that for any $\varepsilon \geq \sqrt{Ld}$ and $\boldsymbol{x}^0 \sim \rho_0$ with $\mathsf{KL}(\rho_0\|\pi_V) \leq K_0$ for any $\{\pi_V\}_{V\in\mathcal{F}}$,*

$$\mathsf{Comp}_{\mathsf{R}}(\mathcal{F}, \varepsilon, K_0) \gtrsim \frac{K_0 L}{\varepsilon^2} \gtrsim \frac{K_0}{d}.$$

Remark 4.2. *This lower bound matches the upper bound in Theorem 3.1 for specific regime of $\varepsilon = \sqrt{Ld}$. The condition $d \geq \widetilde{\Omega}(K_0^{2/3})$ arises because the lower bound construction in [ZHTS25] lies in high dimensional regime $d \geq \widetilde{\Omega}(\varepsilon^{-4})$.*

Proof of Theorem 4.1. To prove Theorem 4.1, we will reduce the problem to that of finding a stationary point in parallel, and then verify the initialization condition. We first recall the reduction lemma from non-log-concave sampling to non-convex optimization.

Lemma 4.3 ([CEL$^+$24, Lemma 16]). *Let $\pi \propto \exp(-V)$ be a β-log-smooth density on $\mathbb{R}^d$. Then, for any probability measure μ,*

$$\mathbb{E}_\mu\left[\|\nabla V\|^2\right] \leq \mathsf{FI}(\mu\|\pi) + 2\beta d.$$

To apply this reduction, we recall the adaptive complexity of finding stationary point which scales as $O(\Delta L \cdot \varepsilon^{-2})$ for high dimensional regime ($d = \widetilde{\Omega}\left(\varepsilon^{-4}\right)$).

Theorem 4.4 (The adaptive complexity of finding stationary points [ZHTS25]). *Assume $d = \widetilde{\Omega}\left(\varepsilon^{-4}\right)$. There exits a function class $\mathcal{F}$ consisting of some L-smooth function with given initial point $\boldsymbol{x}^0$, such that $V(\boldsymbol{x}^0) - \min_{\boldsymbol{x}} V(\boldsymbol{x}) \leq \Delta$, and the following holds: any (possible randomized) algorithm running within $O(\Delta L \cdot \varepsilon^{-2})$ iterations with $\mathsf{poly}(d)$ queries per iteration fails to find ε-approximate point for any $V \in \mathcal{F}$ with probability $1 - d^{-\omega(1)}$.*

We set $\varepsilon = 4\sqrt{Ld}$. From the reduction lemma (Lemma 4.3), if we can obtain a sample from a measure μ such that for $\pi_V \propto \exp(-V)$, it holds that $\mathsf{FI}(\mu\|\pi_V) \leq Ld$, then a sample from μ is a ε-stationary point of f with probability at least $1/2$.

In the following, we check the initialization condition. We set initialization oracle to output a sampler from $\mu \sim \mathcal{N}(0, L^{-1}I_d)$. Now we need to compute the value of $K_0 := \sup_{V\in\mathcal{F}} \mathsf{KL}(\mu_0\|\pi_V)$. To do so, we use the following lemma.

Lemma 4.5 (KL divergence at initialization [CGLL23, Lemma 17]). *Suppose that $V : \mathbb{R}^d \mapsto \mathbb{R}$ is a function such that $V(0) - \inf V \leq \Delta$, ∇V is L-Lipschitz, and $\mathfrak{m} := \int \|\cdot\| \, \mathrm{d}\pi < \infty$ where $\pi \propto \exp(-V)$. Then, for $\mu_0 = \mathcal{N}(0, L^{-1} I_d)$, we have the bound*

$$\mathsf{KL}(\mu_0 \| \pi) \lesssim \Delta + d(1 \vee \log(L\mathfrak{m}^2)).$$

We remind the hardness function in [ZHTS25] takes form as

$$V(\boldsymbol{x}) = \mathsf{poly} \cdot (g(\rho(\boldsymbol{x}/\mathsf{poly}))) + \frac{1}{2\tau^2} \|\boldsymbol{x}\|^2 ,$$

where poly denote any positive quantity for which both the quantity and its inverse are bounded above by polynomials in L, Δ, d, and $1/\varepsilon$, and $\tau = \mathsf{poly}$. Furthermore $g : \mathbb{R}^d \mapsto \mathbb{R}$ and $\rho : \mathbb{R}^d \mapsto \mathbb{R}^d$ are poly-Lipschitz, then

$$\|\nabla g(\rho(\cdot/\mathsf{poly}))\| \leq \mathsf{poly}.$$

Thus $\mathsf{FI}(\pi_V \| \nu) = \mathsf{poly} \cdot \mathbb{E}_{\pi_V} [\|\|\| \nabla g(\rho(\cdot/\mathsf{poly}))] \leq \mathsf{poly}$ where $\nu = \mathcal{N}(0, \tau I_d)$. By the Donsker–Varadhan variational principle [PW25, Theorem 4.6] and the fact that ν satisfies the log-Sobolev inequality with poly, we have

$$\begin{aligned}
\mathbb{E}_{\pi_V} \left[\|\cdot\|^2 \right] &\leq \frac{1}{\lambda} \left\{ \mathsf{KL}(\pi_V \| \nu) + \log \mathbb{E}_\nu \exp(\lambda \| \cdot \|^2) \right\} \\
&\leq \mathsf{poly} \left\{ \mathsf{FI}(\pi_V \| \nu) + 1 \right\} \\
&\leq \mathsf{poly},
\end{aligned}$$

with $\lambda = \frac{1}{\mathsf{poly}}$ and $\mathbb{E}_\nu \exp(\lambda \| \cdot \|^2) \leq 1$. Thus $K_0 = \mathsf{KL}(\mu_0 \| \pi) \lesssim \Delta + \widetilde{\mathcal{O}}(d)$. Thus if $K_0 \geq \widetilde{\Omega}(d)$, $\Delta \gtrsim K_0$. Since $\varepsilon = \sqrt{Ld}$, the number of the required iteration satisfies

$$\#\text{iteration} \gtrsim \frac{\Delta L}{\varepsilon^2} \gtrsim \frac{K_0}{d}.$$

Finally we check the requirement of dimension, $d \geq \widetilde{\Omega}(K_0/d)^2$, which is satisfied provided $d \geq \widetilde{\Omega}(K_0^{2/3})$. □

5 Discussion and Conclusion

In this work, we initialize the studying of parallelize minimizing the relative Fisher information for non-log-concave sampling by showing (1) averaged Langevin Monte Carlo can be accelerated by parallelism and (2) offer a tight lower bound for specific accuracy regime. Our results rule out the possibility of designing general high-accuracy relative Fisher information minimizer via parallelism in the non-log-concave setting, contrasting with the log-concave case. Furthermore, our results offer a new understanding for the theme of "sampling versus optimization" by revealing the distinct role parallelism plays in separating the two.

We believe there are several intriguing directions for future work exploring the role of parallelism in sampling versus optimization, and we conclude by highlighting a few of them.

1. **(Constant-dimensional case).** In the constant-dimensional setting, it is possible to find a stationary point in $k = \mathcal{O}(\log(1/\varepsilon))$ rounds using $\mathcal{O}\left(\varepsilon^{-\frac{d-1}{2}(1+\mathcal{O}(2^{-k}))}\right)$ queries per round by leveraging gradient flow trapping [BM20, HZ23, ZHTS25]. This raises the question: can one similarly minimize the relative Fisher information within $\mathcal{O}(\log(1/\varepsilon))$ rounds by trapping the Langevin dynamics?
2. **(Lower bounds beyond specific large accuracy).** Although our lower bound is tight, it applies only to a specific high-accuracy regime ($\varepsilon = \sqrt{Ld}$). In contrast, for parallel non-convex optimization, tight lower bounds on adaptive complexity, namely, $\Omega(\varepsilon^{-2})$, are known for finding stationary points when $\varepsilon \leq \widetilde{\mathcal{O}}(d^{1/4})$ [ZHTS25]. A natural question is whether similar $\mathsf{poly}(\varepsilon^{-1})$ lower bounds can be established beyond this setting, particularly in the low-accuracy regime ($\varepsilon = o(1)$), as by the bump function construction used for query complexity in Section 4 of [CGLL23].

3. **(Functional inequality case).** For strongly log-concave distributions, it is possible to design high-accuracy samplers leveraging parallelism [YD24, ACV24, ZS24]. In contrast, for non-log-concave distributions, designing general high-accuracy samplers via parallelism becomes impossible. This raises a natural question: what is the boundary between these two cases? A much weaker question is whether high-accuracy samplers can be designed via parallelism for distributions satisfying functional inequalities such as the log-Sobolev or Poincaré inequality.

Acknowledgment

The authors thank Sinho Chewi for very helpful conversations. HZ was supported by International Graduate Program of Innovation for Intelligent World and Next Generation Artificial Intelligence Research Center. MS was supported by JST ASPIRE Grant Number JPMJAP2405.

References

[AB09] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[AC24] Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 2024.

[ACV24] Nima Anari, Sinho Chewi, and Thuy-Duong Vuong. Fast parallel sampling under isoperimetry. *arXiv preprint arXiv:2401.09016*, 2024.

[ALPW24] Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Explicit convergence bounds for metropolis markov chains: isoperimetry, spectral gaps and profiles. *The Annals of Applied Probability*, 34(4):4022–4071, 2024.

[ALW19] Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

[BCE$^+$22] Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

[BGP17] Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7):4709–4724, 2017.

[BM20] Sébastien Bubeck and Dan Mikulincer. How to trap a gradient flow. In *Conference on Learning Theory*, pages 940–960. PMLR, 2020.

[BS18a] Eric Balkanski and Yaron Singer. The adaptive complexity of maximizing a submodular function. In *Proceedings of the 50th annual ACM SIGACT Symposium on Theory of Computing*, 2018.

[BS18b] Eric Balkanski and Yaron Singer. Parallelization does not accelerate convex optimization: Adaptivity lower bounds for non-smooth convex minimization. *arXiv preprint arXiv:1808.03880*, 2018.

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

[CDWY20] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 2020.

[CEL$^+$24] Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. *Foundations of Computational Mathematics*, pages 1–51, 2024.

[CGLL23] Sinho Chewi, Patrik Gerber, Holden Lee, and Chen Lu. Fisher information lower bounds for sampling. In *International Conference on Algorithmic Learning Theory*, 2023.

[Che23] Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho. github. io*, 9:17–18, 2023.

[Cle57] CW Clenshaw. The numerical solution of linear differential equations in Chebyshev series. In *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press, 1957.

[DCWY19] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 2019.

[DG19] Jelena Diakonikolas and Cristóbal Guzmán. Lower bounds for parallel and randomized convex optimization. In *Conference on Learning Theory*, 2019.

[DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via Convex Optimization. *The Journal of Machine Learning Research*, 2019.

[FYC23] Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In *Conference on Learning Theory*, 2023.

[GTC24] Wei Guo, Molei Tao, and Yongxin Chen. Provable benefit of annealed langevin monte carlo for non-log-concave sampling. *arXiv preprint arXiv:2407.16936*, 2024.

[HLB$^+$21] Andrew J Holbrook, Philippe Lemey, Guy Baele, Simon Dellicour, Dirk Brockmann, Andrew Rambaut, and Marc A Suchard. Massive parallelization boosts big bayesian multidimensional scaling. *Journal of Computational and Graphical Statistics*, 30(1):11–24, 2021.

[HLFS21] Andrew J Holbrook, Charles E Loeffler, Seth R Flaxman, and Marc A Suchard. Scalable bayesian inference for self-excitatory stochastic processes applied to big american gunfire data. *Statistics and computing*, 31:1–15, 2021.

[HZ23] Alexandros Hollender and Emmanouil Zampetakis. The computational complexity of finding stationary points in non-convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5571–5572. PMLR, 2023.

[HZ25] Yuchen He and Chihao Zhang. On the query complexity of sampling from non-log-concave distributions. *arXiv preprint arXiv:2502.06200*, 2025.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 1998.

[LCH$^+$06] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[LRG18] Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. *Advances in neural information processing systems*, 31, 2018.

[LST20] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo. In *Conference on learning theory*, 2020.

[LST21] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

[MCC$^+$21] Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of nesterov acceleration for gradient-based mcmc? 2021.

[MR$^+$07] Jean-Michel Marin, Christian P Robert, et al. *Bayesian Core: A Practical Approach to Computational Bayesian statistics*, volume 268. Springer, 2007.

[N$^+$18] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[NWS19] Shinichi Nakajima, Kazuho Watanabe, and Masashi Sugiyama. *Variational Bayesian learning theory*. Cambridge University Press, 2019.

[PW25] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge university press, 2025.

[RCC99] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[SB13] Olivier Sigaud and Olivier Buffet. *Markov decision processes in artificial intelligence*. John Wiley & Sons, 2013.

[SL19] Ruoqi Shen and Yin Tat Lee. The Randomized Midpoint Method for Log-Concave Sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

[VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[Wib25] Andre Wibisono. Mixing time of the proximal sampler in relative fisher information via strong data processing inequality. *arXiv preprint arXiv:2502.05623*, 2025.

[YD24] Lu Yu and Arnak Dalalyan. Parallelized midpoint randomization for langevin monte carlo. *arXiv preprint arXiv:2402.14434*, 2024.

[ZHTS25] Huanjian Zhou, Andi Han, Akiko Takeda, and Masashi Sugiyama. The adaptive complexity of finding a stationary point. In *Conference on Learning Theory*, 2025.

[ZS24] Huanjian Zhou and Masashi Sugiyama. Parallel simulation for sampling under isoperimetry and score-based diffusion models. *arXiv preprint arXiv:2412.07435*, 2024.

[ZWS24] Huanjian Zhou, Baoxiang Wang, and Masashi Sugiyama. Adaptive complexity of log-concave sampling. *arXiv preprint arXiv:2408.13045*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and summarize the key results and methodologies, providing a true overview of the research and its significance.

   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 discussed the limitations in Section 5 and Appendix B.

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: We present the problem setting and assumptions in Section 2.1. For our first main result (Theorem 3.1), a proof sketch is provided in Section 3.2, with the full proof given in Appendix A. The second main result (Theorem 4.1) is proved in Section 4.

   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: [NA]

   Justification: This paper focuses on theoretical results.

   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: [NA]

   Justification: This paper focuses on theoretical results.

   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: [NA]

   Justification: This paper focuses on theoretical results.

   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: [NA]

   Justification: This paper focuses on theoretical results.

   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: [NA]

   Justification: This paper focuses on theoretical results.

   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: [NA]

   Justification: I confirmed it.

   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: [Yes]

    Justification: We discussed in Section C.

    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: This paper focuses on theoretical results.

    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: [NA]

    Justification: This paper focuses on theoretical results. It does not use existing assets.

    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: This paper focuses on theoretical results and we do 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: This paper focuses on theoretical results. It 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: This paper focuses on theoretical results.

    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: LLM is used only for grammar checking.

    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.

A Proof of Theorem 3.1

A.1 Useful facts

In this section, we first recall several useful lemmas.

Lemma A.1 (Differential inequality of KL along interpolation [BCE$^+$22, Lemma 12]). *Consider the stochastic process defined by*

$$\boldsymbol{x}_t := \boldsymbol{x}_0 - t\boldsymbol{g}_0 + \sqrt{2}\,B_t\,, \qquad \textit{for } t \geq 0\,,$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$ which is independent of $(\boldsymbol{x}_0, \boldsymbol{g}_0)$. Let μ_t for the law of $\boldsymbol{x}_t$. Then

$$\begin{aligned}\partial_t \mathsf{KL}(\mu_t \| \pi) &\leq -\frac{3}{4}\mathsf{FI}(\mu_t\|\pi) + \mathbb{E}\big[\|\nabla V(\boldsymbol{x}_t) - \mathbb{E}[\boldsymbol{g}_0 \mid x_t]\|^2\big] \\ &\leq -\frac{3}{4}\mathsf{FI}(\mu_t\|\pi) + \mathbb{E}[\|\nabla V(\boldsymbol{x}_t) - \boldsymbol{g}_0\|^2]\,.\end{aligned}$$

Lemma A.2 (Initialization [BCE$^+$22, Theorem 2]). *For all $n = 0, \ldots, N-1$ and $m \in [M]$, let $\mu^0_{n,m}$ be the law of $\boldsymbol{x}^0_{n,m}$, then we have*

$$\mathsf{FI}(\mu^0_{n,m}\|\pi) \leq \frac{2\mathsf{KL}(\mu_0\|\pi)}{Nh} + 8L^2dh.$$

Lemma A.3 ([CEL$^+$24, Lemma 16]). *Assume that ∇V is L-Lipschitz. For any probability measure μ, it holds that*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \leq \mathsf{FI}(\mu\|\pi) + 2dL\,.$$

A.2 Decomposition via interpolation method

We denote $\mathsf{KL}_{n,m} = \mathsf{KL}(\mu^J_{n,m}\|\pi)$ where $\mu^J_{n,m}$ represents the law of $\boldsymbol{x}^J_{n,m}$. Let $\boldsymbol{x}_t$ be the linear interpolation between $\boldsymbol{x}^J_{n,m}$ and $\boldsymbol{x}^J_{n,m+1}$, *i.e.*, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)hh}{M}\right]$, let

$$\boldsymbol{x}_t = \boldsymbol{x}^J_{n,m} - \left(t - nh - \frac{mh}{M}\right)\nabla V(\boldsymbol{x}^{J-1}_{n,m}) + \sqrt{2}(B_t - B_{nh+mh/M}).$$

Then Lemma A.1 yields

$$\partial_t \mathsf{KL}(\mu_t\|\pi) \leq -\frac{3}{4}\mathsf{FI}(\mu_t\|\pi) + \mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t) - \nabla V(\boldsymbol{x}^{J-1}_{n,m})\right\|^2\right].$$

For the second term, by smooth of V,

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t) - \nabla V(\boldsymbol{x}^{J-1}_{n,m})\right\|^2\right] \\ &\leq L^2\mathbb{E}\left[\left\|\boldsymbol{x}_t - \boldsymbol{x}^{J-1}_{n,m}\right\|^2\right] \\ &\leq 2L^2\mathbb{E}\left[\left\|\boldsymbol{x}_t - \boldsymbol{x}^{J}_{n,m}\right\|^2\right] + 2L^2\mathbb{E}\left[\left\|\boldsymbol{x}^J_{n,m} - \boldsymbol{x}^{J-1}_{n,m}\right\|^2\right] \\ &\leq 4L^2\left(t - nh - \frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}^{J-1}_{n,m})\right\|^2\right] + 8L^2\mathbb{E}\left[\left\|B_t - B_{nh+mh/M}\right\|^2\right] + 2L^2\mathbb{E}\left[\left\|\boldsymbol{x}^J_{n,m} - \boldsymbol{x}^{J-1}_{n,m}\right\|^2\right] \\ &\leq 8L^2\left(t - nh - \frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t) - \nabla V(\boldsymbol{x}^{J-1}_{n,m})\right\|^2\right] + 8L^2\left(t - nh - \frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t)\right\|^2\right] \\ &\quad + 8L^2\mathbb{E}\left[\left\|B_t - B_{nh+mh/M}\right\|^2\right] + 2L^2\mathbb{E}\left[\left\|\boldsymbol{x}^J_{n,m} - \boldsymbol{x}^{J-1}_{n,m}\right\|^2\right].\end{aligned}$$

Taking $Lh \leq 0.1$, we have

$$\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t) - \nabla V(\boldsymbol{x}^{J-1}_{n,m})\right\|^2\right]$$

$$\leq 9L^2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t)\right\|^2\right]+9L^2\mathbb{E}\left[\left\|B_t-B_{nh+mh/M}\right\|^2\right]+3L^2\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right].$$

For the first term, by Lemma A.3 and $Lh \leq 0.1$, we have

$$\begin{aligned}\partial_t\mathsf{KL}(\mu_t\|\pi) \;&\leq\; -\frac{3}{4}\mathsf{FI}(\mu_t\|\pi)+\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_t)-\nabla V(\boldsymbol{x}_{n,m}^{J-1})\right\|^2\right]\\ &\leq\; -\left(\frac{3}{4}-\frac{9L^2h^2}{M^2}\right)\mathsf{FI}(\mu_t\|\pi)+18L^3d\left(t-nh-\frac{mh}{M}\right)^2+9L^2d\left(t-nh-\frac{mh}{M}\right)\\ &\quad+3L^2\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right]\\ &\leq\; -\frac{1}{2}\mathsf{FI}(\mu_t\|\pi)+18L^3d\left(t-nh-\frac{mh}{M}\right)^2+9L^2d\left(t-nh-\frac{mh}{M}\right)\\ &\quad+3L^2\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right].\end{aligned}$$

Integrating it over $t\in\left[nh+\frac{mh}{M}, nh+\frac{(m+1)hh}{M}\right]$, we obtain

$$\begin{aligned}\mathsf{KL}_{n,m+1}-\mathsf{KL}_{n,m} \;&\leq\; -\frac{1}{2}\int_{nh+\frac{mh}{M}}^{nh+\frac{(m+1)hh}{M}}\mathsf{FI}(\mu_t\|\pi)\mathrm{d}t+6L^3d\frac{h^3}{M^3}+5L^2d\frac{h^2}{M^2}+3L^2\frac{h}{M}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right]\\ &\leq\; -\frac{1}{2}\int_{nh+\frac{mh}{M}}^{nh+\frac{(m+1)hh}{M}}\mathsf{FI}(\mu_t\|\pi)\mathrm{d}t+6\frac{L^2dh^2}{M^2}+\frac{3L^2h}{M}\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right].\end{aligned}$$

Now we assume there is a uniform upper bound for $\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right]$ for any $n\in[N]$ and $m\in[M]$, which represents the convergence error of Picard iteration. Specifically, we assume $\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right]\leq\mathcal{E}$ for any $n\in[N]$ and $m\in[M]$. Then by summing, we have

$$\frac{1}{Nh}\int_0^{Nh}\mathsf{FI}(\mu_t\|\pi)\mathrm{d}t\leq\frac{2\mathsf{KL}_{0,0}}{Nh}+\frac{2Ld}{M}+3L^2\mathcal{E}. \tag{2}$$

A.3 Convergence of Picard iteration

We will end the prove by show the uniform upper bound for $\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^J-\boldsymbol{x}_{n,m}^{J-1}\right\|^2\right]$. To bound it, we define $\mathcal{E}_n^j := \max_{m=1,\ldots,M}\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^j-\boldsymbol{x}_{n,m}^{j-1}\right\|^2\right]$.

Lemma A.4 (Decomposition of $\mathcal{E}_n^j$). *Assume $Lh=\frac{1}{10}$ and let μ^0 is the law of $\boldsymbol{x}_{n,0}^0$. We have the following decompositions and initialization estimations:*

1. $\mathcal{E}_n^j\leq 2\mathcal{E}_{n-1}^j+0.02\mathcal{E}_n^{j-1}$, *for any $j=2,\ldots,J$ and $n=1,\ldots,N-1$;*
2. $\mathcal{E}_n^1\leq 2\mathcal{E}_{n-1}^1+2h^2\mathsf{FI}(\mu^0\|\pi)+5dh$, *for $j=1$ and $n=1,\ldots,N-1$;*
3. $\mathcal{E}_0^j\leq 0.01\mathcal{E}_0^{j-1}$, *for any $j=2,\ldots,J$ and $n=0$;*
4. $\mathcal{E}_0^1\leq 2h^2\mathsf{FI}(\mu^0\|\pi)+5dh$*;*

Proof. For $j\in[J]$, $n=1,\ldots,N-1$, $m=0,\ldots,M-1$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{x}_{n,m}^j-\boldsymbol{x}_{n,m}^{j-1}\right\|^2\right]\\ &\leq 2\mathbb{E}\left[\left\|\boldsymbol{x}_{n,0}^j-\boldsymbol{x}_{n,0}^{j-1}\right\|^2\right]+2\frac{h^2}{M^2}\mathbb{E}\left[\left\|\sum_{m'=0}^{m-1}\nabla V(\boldsymbol{x}_{n,m'}^{j-1})-\nabla V(\boldsymbol{x}_{n,m'}^{j-2})\right\|^2\right]\\ &\leq 2\mathbb{E}\left[\left\|\boldsymbol{x}_{n,0}^j-\boldsymbol{x}_{n,0}^{j-1}\right\|^2\right]+2h^2\max_{m'=1,\ldots,m}\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}_{n,m'}^{j-1})-\nabla V(\boldsymbol{x}_{n,m'}^{j-2})\right\|^2\right]\end{aligned}$$

$$\le 2\mathcal{E}^j_{n-1} + 0.02\mathcal{E}^{j-1}_n.$$

For $j = 1$, $n = 1, \dots, N-1$, $m = 0, \dots, M-1$, and $p = 1, \dots, P$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}^1_{n,m} - \boldsymbol{x}^0_{n,m}\right\|^2\right] \\
&\le 2\mathbb{E}\left[\left\|\boldsymbol{x}^1_{n,0} - \boldsymbol{x}^0_{n,0}\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{h}{M}\sum_{m'=0}^{m-1}\nabla V(\boldsymbol{x}^0_{n,m'}) + \sqrt{2}(B_{nh+mh/M} - B_{nh})\right\|^2\right] \\
&\le 2\mathbb{E}\left[\left\|\boldsymbol{x}^1_{n,0} - \boldsymbol{x}^0_{n,0}\right\|^2\right] + 2h^2\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}^0_{n,m'})\right\|^2\right] + 4dh \\
&\le 2\mathbb{E}\left[\left\|\boldsymbol{x}^1_{n,0} - \boldsymbol{x}^0_{n,0}\right\|^2\right] + 2h^2(\mathsf{FI}(\mu^0\|\pi) + 2dL) + 4dh \\
&\le 2\mathcal{E}^1_{n-1} + 2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh.
\end{aligned}$$

When $n = 0$, $j \ge 2$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}^j_{0,m} - \boldsymbol{x}^{j-1}_{0,m}\right\|^2\right] \\
&\le \frac{h^2}{M^2}\mathbb{E}\left[\left\|\sum_{m'=0}^{m-1}\nabla V(\boldsymbol{x}^{j-1}_{0,m'}) - \nabla V(\boldsymbol{x}^{j-2}_{0,m'})\right\|^2\right] \\
&\le h^2 \max_{m'=1,\dots,m}\mathbb{E}\left[\left\|\nabla V(\boldsymbol{x}^{j-1}_{0,m'}) - \nabla V(\boldsymbol{x}^{j-2}_{0,m'})\right\|^2\right] \\
&\le L^2h^2 \max_{m'=1,\dots,m}\mathbb{E}\left[\left\|\boldsymbol{x}^{j-1}_{0,m'} - \boldsymbol{x}^{j-2}_{0,m'}\right\|^2\right].
\end{aligned}$$

When $n = 0$, $j = 1$ we have

$$\mathbb{E}\left[\left\|\boldsymbol{x}^1_{0,m} - \boldsymbol{x}^0_{0,m}\right\|^2\right] \le 2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh.$$

□

Lemma A.5. *If $J \ge N$, then for $n = 0, \dots, N-1$ we have*

$$\mathcal{E}^J_n \le 200(2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.5^{J-N}.$$

Proof. By Lemma A.4, we can recursively bound $\mathcal{E}^j_n$. Specifically, for $n \ge 1$ and $j = 1$, we have $\mathcal{E}^1_n \le 2\mathcal{E}^1_{n-1} + 2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh \le 2^n\mathcal{E}^1_0 + (2^n - 1)(2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh) \le 2^{n+1}(2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)$, and for $n = 0$, $j \ge 2$, $\mathcal{E}^j_0 \le 0.01^{j-1}(2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)$. Furthermore, for $n \ge 1$ and $j \ge 2$, by $\binom{m}{n} \le (\frac{em}{n})^n$, if $j \ge n$, we have

$$\begin{aligned}
\mathcal{E}^j_n &\le \sum_{a=2}^{n} 0.02^{j-1}2^{n-a}\binom{n-a+j-2}{j-2}\mathcal{E}^1_a + \sum_{b=2}^{j}\binom{n+j-b}{j-b}0.02^{j-b}2^n\mathcal{E}^b_0 \\
&\le (2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)\left(\sum_{a=2}^{n} 0.02^{j-1}2^{n-a}\binom{n-a+j-2}{j-2}2^{a+1} + \sum_{b=2}^{j}\binom{n+j-b}{j-b}0.02^{j-b}2^n 0.01^{b-1}\right) \\
&\le (2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.02^{j-1}2^{n+1}\left(\sum_{a=2}^{n}\binom{n-a+j-2}{j-2} + \sum_{b=2}^{j}\binom{n+j-b}{j-b}\right) \\
&\le (2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.02^{j-1}2^{n+1}\left(\sum_{a=2}^{n} e^{j-2}\left(\frac{n-a+j-2}{j-2}\right)^{j-2} + \sum_{i=0}^{j-2} e^{j-2}\left(\frac{n+i}{i}\right)^i\right) \\
&\le (2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.02^{j-1}2^{n+1}\left(\sum_{a=2}^{n} e^{j-2}2^{j-2} + \sum_{i=0}^{j-2} e^{j-2}e^n\right) \\
&\le (2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.02^{j-1}2^{n+1}e^{2j}(2j) \\
&\le 200(2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.5^j 2^n,
\end{aligned}$$

where the last equation holds since $0.02^{j-1}e^{2j}(2j) \le 100 \cdot 0.5^j$. □

A.4 Overall bound

Combine Equation (2) and Lemma A.5, we have

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t\|\pi)\mathrm{d}t \le \frac{2\mathsf{KL}_{0,0}}{Nh} + \frac{2Ld}{M} + 600L^2(2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.5^{J-N}.$$

By the convexity of the Fisher information, the averaged distribution $\bar{\mu} := (Nh)^{-1}\int_0^{Nh}\mu_t\mathrm{d}t$ also satisfies

$$\mathsf{FI}(\bar{\mu}\|\pi)\mathrm{d}t \le \frac{2\mathsf{KL}_{0,0}}{Nh} + \frac{2Ld}{M} + 600L^2(2h^2\mathsf{FI}(\mu^0\|\pi) + 5dh)0.5^{J-N}.$$

We also observe the output is sampled from $\bar{\mu}$. By Lemma A.2, and $Lh = 0.1$ we have

$$\begin{aligned}\mathsf{FI}(\bar{\mu}\|\pi)\mathrm{d}t \ &\le \frac{2L\mathsf{KL}_{0,0}}{N} + \frac{2Ld}{M} + 600L^2\left(2h^2\left(\frac{2\mathsf{KL}(\mu_0\|\pi)}{Nh} + 8L^2dh\right) + 5dh\right)0.5^{J-N}\\ &\le \frac{2L\mathsf{KL}_{0,0}}{N} + \frac{2Ld}{M} + \left(\frac{24L\mathsf{KL}(\mu_0\|\pi)}{N} + 320Ld\right)0.5^{J-N}.\end{aligned}$$

B Limitations

As an initial exploration of the adaptive complexity of minimizing relative Fisher information, our work still leaves a significant gap between the upper and lower bounds, particularly in the small-accuracy regime.

C Social Impacts

We present several theoretical results for minimizing relative Fisher information. While we do not see any immediate societal impact, there may be potential indirect consequences of our work that are not apparent at this time.