Fisher information lower bounds for sampling

Sinho Chewi[*]   SCHEWI@MIT.EDU
MIT

Patrik Gerber   PRGERBER@MIT.EDU
MIT

Holden Lee   HLEE283@JHU.EDU
Johns Hopkins University

Chen Lu   CHENL819@MIT.EDU
MIT

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We prove two lower bounds for the complexity of non-log-concave sampling within the framework of Balasubramanian et al. (2022), who introduced the use of Fisher information (FI) bounds as a notion of approximate first-order stationarity in sampling. Our first lower bound shows that averaged Langevin Monte Carlo (LMC) is optimal for the regime of large FI by reducing the problem of finding stationary points in non-convex optimization to sampling. Our second lower bound shows that in the regime of small FI, obtaining a FI of at most ε^2 from the target distribution requires $\mathrm{poly}(1/\varepsilon)$ queries, which is surprising as it rules out the existence of high-accuracy algorithms (e.g., algorithms using Metropolis–Hastings filters) in this context.

Keywords: Fisher information, gradient descent, Langevin Monte Carlo, non-log-concave sampling, sampling lower bound, stationary point

1. Introduction

What is the query complexity of sampling from a β-log-smooth but possibly non-log-concave target distribution π on $\mathbb{R}^d$? Until recently, this question was only investigated from an upper bound perspective, and only for restricted classes of distributions, such as distributions satisfying functional inequalities (Vempala and Wibisono, 2019; Wibisono, 2019; Ma et al., 2021; Chewi et al., 2022b), distributions with tail decay conditions (Durmus and Moulines, 2017; Cheng et al., 2018a; Xu et al., 2018; Li et al., 2019; Majka et al., 2020; Erdogdu and Hosseinzadeh, 2021; Zou et al., 2021; He et al., 2022), or mixtures of log-concave distributions (Lee et al., 2018).

Recently Balasubramanian et al. (2022) developed a general framework to investigate non-log-concave sampling. Motivated by stationary point analysis in non-convex optimization (see, e.g., Nesterov, 2018) and the interpretation of sampling as optimization over the space of probability measures (Jordan et al., 1998; Wibisono, 2018), Balasubramanian et al. proposed to call any measure μ satisfying $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \le \varepsilon$ an ε-stationary point for sampling, where $\mathsf{FI}(\mu\|\pi) := \mathbb{E}_\mu[\|\nabla \log \frac{\mu}{\pi}\|^2]$ denotes the *relative Fisher information* of μ from π. They explained the interpretation of this condition via the classical phenomenon of *metastability* (Bovier et al., 2002, 2004, 2005); in particular, for a multimodal distribution, small Fisher information means that the distribution locally approximates the shape at each mode, but not necessarily the relative weights between the modes. They further

---

showed that averaged Langevin Monte Carlo (LMC) can find an ε-stationary point in $\mathcal{O}(\beta^2 dK_0/\varepsilon^4)$ iterations, where $K_0 := \mathsf{KL}(\mu_0 \parallel \pi)$ is the initial Kullback–Leibler (KL) divergence to the target π.

In the field of optimization, however, there are also corresponding *lower bounds* on the complexity of finding stationary points (Vavasis, 1993; Nesterov, 2012; Bubeck and Mikulincer, 2020; Carmon et al., 2020, 2021; Chewi et al., 2022a). Such lower bounds are important for identifying optimal algorithms and understanding the fundamental difficulty of the task at hand. For example, the work of Carmon et al. (2020) shows that the standard gradient descent algorithm is optimal for finding stationary points of smooth functions, at least in high dimension.

In this work, we establish the first lower bounds for Fisher information guarantees for sampling, resolving an open question posed in Balasubramanian et al. (2022). As we discuss further below, our results also reveal a surprising equivalence between the task of obtaining a sample which has moderate Fisher information relative to a target distribution and the task of finding an approximate stationary point of a smooth function, thereby strengthening the connection between the fields of non-convex optimization and non-log-concave sampling.

Our contributions. We now informally describe our main results. Details on notation, our oracle model, and the definition query complexity for sampling (Definition 5) are given in Section 2. Precise statements of our results are given in Sections 3 and 4. For a density $\pi \propto \exp(-U)$ the function $U : \mathbb{R}^d \to \mathbb{R}$ is called the *potential*. Throughout, our notion of complexity is the number of queries made to an oracle that returns the value of U (up to an additive constant) and its gradients. For a 1-smooth function $V : \mathbb{R}^d \to \mathbb{R}$ and $\beta > 0$ let us define the density $\pi_\beta \propto \exp(-\beta V)$, assuming it is well-defined (i.e. $\int \exp(-\beta V) < \infty$).

Our first result connects the task of obtaining Fisher information guarantees with finding stationary points in non-convex optimization, for a particular regime of large smoothness β.

Theorem 1 (equivalence, informal) *The following two problems are equivalent.*

1. *Output an ε-stationary point of V.*

2. *Output a sample from a measure μ such that $\mathsf{FI}(\mu \parallel \pi_\beta) \lesssim \beta d$, where $\beta \asymp d/\varepsilon^2$.*

By combining this equivalence with the lower bound of Carmon et al. (2020) for finding ε-stationary points, we obtain:

Theorem 2 (first lower bound, informal) *The number of queries required to obtain a sample from a measure μ satisfying $\sqrt{\mathsf{FI}(\mu \parallel \pi_\beta)} \lesssim \sqrt{\beta d}$, starting from an initial distribution μ_0 with KL divergence $K_0 := \mathsf{KL}(\mu_0 \parallel \pi_\beta)$, is at least $\Omega(K_0/d)$. The lower bound is attained by averaged LMC (Langevin Monte Carlo) as given in Balasubramanian et al. (2022).*

To our knowledge an optimality result for LMC was not previously known in any setting.

The first lower bound addresses the regime of large Fisher information, $\mathsf{FI}(\mu \parallel \pi_\beta) \lesssim \beta d$. In order to target the regime of small Fisher information, we give a construction based on hiding a bump of large mass and prove the following:

Theorem 3 (second lower bound, informal) *The number of queries required to obtain a sample from a measure μ satisfying $\sqrt{\mathsf{FI}(\mu \parallel \pi_\beta)} \leq \varepsilon$, starting from an initial distribution μ_0 with KL divergence $K_0 := \mathsf{KL}(\mu_0 \parallel \pi_\beta) \leq 1$, is at least $(\sqrt{\beta}/\varepsilon)^{2d/(d+2)-o(1)}$ as $\varepsilon \to 0$.*

We give a more precise form of our lower bound in Section 4. In infinite dimension (actually, $d \geq \widetilde{\Omega}(\sqrt{\log(\beta/\varepsilon^2)})$ suffices, see Section 4), the lower bound reads $\widetilde{\Omega}(\beta/\varepsilon^2)$, which can be compared to the averaged LMC upper bound of $\mathcal{O}(\beta^2 d/\varepsilon^4)$. It is an open question to close this gap.

In terms of technical novelty, we note that the difficulty of showing the first lower bound lies mainly in establishing the equivalence between optimization and sampling, after which lower bounds from optimization apply; on the other hand, the second lower bound requires significant technical work to establish.

We next discuss implications of our results.

- **Towards a theory of lower bounds for sampling.** The problem of obtaining sampling lower bounds is a notorious open problem raised in many prior works (see, e.g., Cheng et al., 2018b; Ge et al., 2020; Chatterji et al., 2022; Lee et al., 2021). So far, unconditional lower bounds have only been obtained in restricted settings such as in dimension 1; see Chewi et al. (2022c) and the discussion therein, as well as the reduction to optimization in Gopi et al. (2022). Our lower bounds are the first of their kind for Fisher information guarantees, and are some of the *only* lower bounds for sampling in general. Hence, our work takes a significant step towards a better understanding of the complexity of sampling. In particular, our first lower bound identifies a regime in which (averaged) LMC is *optimal*, which was not previously known in any setting.

- **Stronger connections between non-convex optimization and non-log-concave sampling.** The equivalence in Theorem 1 provides compelling evidence that Fisher information guarantees are the correct analogue of stationary point guarantees in non-convex optimization, thereby supporting the framework of Balasubramanian et al. (2022).

- **Obtaining an approximate stationary point in sampling is strictly harder for non-log-concave targets.** Ignoring the dependence on other parameters besides the accuracy, our second lower bound yields a $\mathrm{poly}(1/\varepsilon)$ lower bound for the Fisher information task for non-log-concave targets. In contrast, it is morally possible to solve this task in $\mathrm{polylog}(1/\varepsilon)$ queries for *log-concave* targets; see Appendix A for justification. This exhibits a stark separation between log-concave and non-log-concave sampling. Note that the analogous separation does not exist in the context of optimization, because there is a $\mathrm{poly}(1/\varepsilon)$ lower bound for finding an ε-stationary point of a convex and smooth function (Carmon et al., 2021).

- **A separation between optimization and sampling.** Finally, our second lower bound yields a $\mathrm{poly}(1/\varepsilon)$ lower bound, even in dimension one. In contrast, for the analogous question in optimization of finding an ε-stationary point of a univariate function, the recent work of Chewi et al. (2022a) exhibits an algorithm with $\mathcal{O}(\log(1/\varepsilon))$ complexity. To our knowledge, this is one of the first instances in which sampling is provably harder than optimization.

2. Notation and setting

Notation. Given a probability measure π on $\mathbb{R}^d$ which admits a density w.r.t. the Lebesgue measure, we abuse notation by identifying π with its density.

The class of distributions that we wish to sample from are the β-log-smooth distributions on $\mathbb{R}^d$, defined as follows:

Definition 4 (log-smooth distributions) *The class of β-log-smooth distributions consists of distributions π_β supported on $\mathbb{R}^d$ whose densities are of the form $\pi \propto \exp(-U_\beta)$, for potential functions $U_\beta : \mathbb{R}^d \to \mathbb{R}$ that are twice continuously differentiable, and satisfy*

$$\|\nabla U_\beta(x) - \nabla U_\beta(y)\| \leq \beta \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Oracle model. We work under the following oracle model. The algorithm is given access to a target distribution π in our class via two oracles: initialization and local information. The initialization oracle outputs samples from some distribution μ_0 for which $\mathsf{KL}(\mu_0 \| \pi) \leq K_0$. The local oracle for π, given a query point $x \in \mathbb{R}^d$, returns the value of the potential (up to an additive constant) and its gradient at the query point x, i.e., the tuple $(U_\beta(x), \nabla U_\beta(x))$. Algorithms can access samples from μ_0 for free, and we care about the number of local information queries needed. The query complexity is defined as follows.

Definition 5 (query complexity) *Let $\mathscr{C}(d, K_0, \varepsilon; \beta)$ be the largest number $n \in \mathbb{N}$ such that any algorithm which works in the oracle model described above and outputs a sample from a measure μ_β satisfying $\sqrt{\mathsf{FI}(\mu_\beta \| \pi_\beta)} \leq \varepsilon$, for any β-log-smooth target π_β and any valid initialization oracle for π_β, requires at least n queries to the local oracle for π_β.*

The upper bound of [Balasubramanian et al. (2022)](#) shows that using averaged LMC,

$$\mathscr{C}(d, K_0, \varepsilon; \beta) \lesssim 1 \vee \frac{\beta^2 d K_0}{\varepsilon^4}. \tag{1}$$

We also note the following rescaling lemma.

Lemma 6 (rescaling) *It holds that*

$$\mathscr{C}(d, K_0, \varepsilon; \beta) = \mathscr{C}\left(d, K_0, \frac{\varepsilon}{\sqrt{\beta}}; 1\right).$$

Proof Suppose that $U_\beta : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and that $\pi_\beta \propto \exp(-U_\beta)$ is a density. Define the rescaled potential $U : \mathbb{R}^d \to \mathbb{R}$ via $U(x) := U_\beta(x/\sqrt{\beta})$, and let $\pi \propto \exp(-U)$. (Note that the relationship between π and π_β is different from that in Section [1](#).) Note that U is 1-smooth; moreover, if $Z \sim \pi_\beta$ then $\sqrt{\beta} Z \sim \pi$. Suppose $\mathsf{KL}(\mu_\beta \| \pi_\beta) = K_0$ and that $X_\beta \sim \mu_\beta$ is a sample from μ_β, and let $\mu := \mathrm{law}(\sqrt{\beta} X_\beta)$. Since the KL divergence is invariant under bijective transformations, we have $\mathsf{KL}(\mu \| \pi) = K_0$, which shows that we can simulate an initialization oracle for π given an initialization oracle for π_β. We can also simulate the local oracle for π given a local oracle for π_β, as $\nabla U(x) = \frac{1}{\sqrt{\beta}} \nabla U_\beta(x/\sqrt{\beta})$. Finally, let $\hat{\mu}$ satisfy $\sqrt{\mathsf{FI}(\hat{\mu} \| \pi)} \leq \varepsilon/\sqrt{\beta}$ and write $\hat{\mu}_\beta := \mathrm{law}(\hat{X}/\sqrt{\beta})$ where $\hat{X} \sim \hat{\mu}$. A straightforward calculation shows that $\sqrt{\mathsf{FI}(\hat{\mu}_\beta \| \pi_\beta)} \leq \varepsilon$. This proves the upper bound $\mathscr{C}(d, K_0, \varepsilon; \beta) \leq \mathscr{C}(d, K_0, \varepsilon/\sqrt{\beta}; 1)$, and the reverse bound follows because this reduction is reversible. ∎

From here on, we abbreviate $\mathscr{C}(d, K_0, \varepsilon) := \mathscr{C}(d, K_0, \varepsilon; 1)$.

3. Reduction to optimization and the first lower bound

In this section, we show a perhaps surprising equivalence between obtaining Fisher information guarantees in sampling and finding stationary points of smooth functions in optimization. The formal statement of the equivalence is as follows.

Theorem 7 (equivalence) *Let $V : \mathbb{R}^d \to \mathbb{R}$ be a 1-smooth function such that for any $\beta > 0$, the function $\exp(-\beta V)$ is integrable. Let π_β be the probability measure with density $\pi_\beta \propto \exp(-\beta V)$, where $\beta = d/\varepsilon^2$.*

1. *Suppose that $x \in \mathbb{R}^d$ is a point with $\|\nabla V(x)\| \leq \varepsilon$. Then, for $\mu_\beta := \mathsf{normal}(x, \beta^{-1} I_d)$, it holds that $\mathsf{FI}(\mu_\beta \,\|\, \pi_\beta) \leq 10\beta d$.*

2. *Conversely, suppose that μ is a probability measure on $\mathbb{R}^d$ such that $\mathsf{FI}(\mu \,\|\, \pi_\beta) \leq \beta d$. Let $X \sim \mu$ be a sample. Then, $\|\nabla V(X)\| \leq 3\varepsilon$ with probability at least $1/2$.*

Proof See Appendix B.1. ∎

Note that an oracle for βV can be simulated from an oracle for V, so that the above theorem provides an exact equivalence between a sampling problem and an optimization problem within the oracle model, up to universal constants.

As a first application of this equivalence, we observe that averaged LMC yields an nearly optimal algorithm for finding stationary points of smooth functions. We recall the LMC algorithm for sampling from a density $\pi \propto \exp(-U)$. We fix a step size $h > 0$, initialize at $X_0 \sim \mu_0$, and for $t \in [kh, (k+1)h]$, we set

$$X_t = X_{kh} - (t - kh)\,\nabla U(X_{kh}) + \sqrt{2}\,(B_t - B_{kh}), \tag{2}$$

where $(B_t)_{t \geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. Let $\mu_t := \mathrm{law}(X_t)$ denote the law of the algorithm at time t. Then, the *averaged* LMC algorithm at iteration N outputs a sample from the law of $\bar{\mu}_{Nh} := (Nh)^{-1} \int_0^{Nh} \mu_t \, dt$. This is obtained algorithmically as follows: first, we sample a time $t \in [0, Nh]$ uniformly at random (independently of all other random variables). Let k denote the largest integer such that $kh \leq t$. We then compute $X_0, X_h, X_{2h}, \ldots, X_{kh}$ using the LMC recursion, and then output X_t which is obtained via the partial LMC update (2).

Corollary 8 (averaged LMC is nearly optimal for finding stationary points) *Let $V : \mathbb{R}^d \to \mathbb{R}$ be 1-smooth and satisfy $V(0) - \inf V \leq \Delta$. Let $\varepsilon > 0$ be such that $\Delta/\varepsilon^2 \geq 1$. Assume that for $\beta = d/\varepsilon^2$, the probability measure with density $\pi_\beta \propto \exp(-\beta V)$ is well-defined and that $\int \|\cdot\|^2 \, d\pi_\beta \leq \mathrm{poly}(\Delta, d, 1/\varepsilon)$. Consider running averaged LMC with step size $h = \widetilde{\Theta}(1/\beta)$, initial distribution $\mu_0 = \mathsf{normal}(0, \beta^{-1} I_d)$, and target π_β, with*

$$N \geq \widetilde{\Omega}\Big(\frac{\Delta}{\varepsilon^2}\Big) \qquad \textit{iterations}\,.$$

Then, we obtain a sample X such that with probability at least $1/2$, it holds that $\|\nabla V(X)\| \lesssim \varepsilon$.

Proof We combine Theorem 7 with the analysis of averaged LMC in Balasubramanian et al. (2022); see Appendix B.2. ∎

This matches the usual $\mathcal{O}(\Delta/\varepsilon^2)$ complexity for the standard gradient descent algorithm to find an ε-stationary point (see, e.g., Bubeck, 2015; Nesterov, 2018). On its own, this observation is not terribly surprising because as $\beta \to \infty$, the LMC iteration (2) recovers the gradient descent algorithm. However, it is remarkable that the analysis of Balasubramanian et al. (2022) of averaged LMC in Fisher information nearly recovers the gradient descent guarantee.

This observation also suggests that the lower bound of Carmon et al. (2020), which establishes optimality of gradient descent for finding stationary points in high dimension, also implies optimality of averaged LMC in a certain regime. We obtain the following theorem.

Theorem 9 (first lower bound) *Suppose that the dimension d satisfies $\widetilde{\mathcal{O}}(K_0) \geq d \geq \widetilde{\Omega}(K_0^{2/3})$. Then, it holds that*

$$\mathscr{C}\big(d, K_0, \varepsilon = \sqrt{\beta d}; \beta\big) \gtrsim \frac{K_0}{d}\,.$$

Proof In the lower bound of Carmon et al. (2020), the authors construct a family of functions $\mathscr{F}$ such that each $f \in \mathscr{F}$ is β-smooth and satisfies $f(0) - \inf f \leq \Delta$. Moreover, any randomized algorithm which, for any $f \in \mathscr{F}$, makes queries to a local oracle for f and outputs an δ-stationary point of f with probability at least $1/2$, requires at least $\Omega(\beta\Delta/\delta^2)$ queries. The dimension of the functions in the construction is $d = \widetilde{\Theta}(\beta^2\Delta^2/\delta^4)$. Setting $\beta V = f$ and using the equivalence from Theorem 7 completes the proof. Details are given in Appendix B.3. ∎

The lower bound of Theorem 9 is matched by averaged LMC, see (1). In the theorem, the restriction $d \geq \widetilde{\Omega}(K_0^{2/3})$ arises because the lower bound construction of Carmon et al. (2020) for finding a ε-stationary point of a smooth function requires a large dimension $d \geq \widetilde{\Omega}(1/\varepsilon^4)$. If, as conjectured in Bubeck and Mikulincer (2020) and Chewi et al. (2022a), the lower bound construction can be embedded in dimension $d \gtrsim \log(1/\varepsilon)$, then the restriction in Theorem 9 would instead become $d \gtrsim \log K_0$.

4. Bump construction and the second lower bound

The main drawback of the first lower bound (Theorem 9) is that it only provides a lower bound on the Fisher information for a specific value of the target accuracy, $\varepsilon = \sqrt{\beta d}$. To complement this result, we provide the following lower bound for the query complexity of sampling to high accuracy in Fisher information; recall that it suffices to consider $\beta = 1$ by the rescaling lemma (Lemma 6).

Theorem 10 (second lower bound) *For the class of 1-log-smooth distributions on $\mathbb{R}^d$, there exist universal constants $c, c' > 0$, such that for all $\varepsilon < \exp(-c'd)$, we have*

$$\mathscr{C}(d, K_0 = 1, \varepsilon) \gtrsim \Big(\frac{cd}{\log(1/\varepsilon)}\Big)^{d/2} \frac{1}{\varepsilon^{2d/(d+2)}}\,. \tag{3}$$

Proof Here we sketch the main ideas of the proof. We construct a family of distributions in our class which put a constant fraction of their mass on disjoint bumps. Specifically, let B_r denote the ball of radius r in $\mathbb{R}^d$, and let $\mathscr{P}_{2r,R}$ be a maximal $2r$-packing of B_{R-r}. For any $\omega \in \mathscr{P}_{2r,R}$, let $\tilde{\pi}_\omega$ denote the unnormalized density

$$\tilde{\pi}_\omega(x) \coloneqq \exp\bigg(r^2\phi\Big(\frac{\|x - \omega\|}{r}\Big) - \frac{1}{2}\,(\|x\| - R)_+^2\bigg) =: \exp\big(-V_\omega(x)\big)\,, \tag{4}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a decreasing, twice continuously differentiable function supported on $[0, 1]$ with bounded second derivative, chosen such that $\tilde{\pi}_\omega$ is 1-log-smooth. We see that the mass of the distribution π_ω will be concentrated on B_R. Moreover, by a careful choice of r we can ensure that exactly half of the mass of π_ω is in the set $\omega + B_r$.

The key idea is the following reduction: being able to sample from π_ω within small Fisher information means that we can estimate $\omega \in \mathscr{P}_{2r,R}$. To make this reduction work, note that if we make a query within $\omega + B_r$, then we can immediately identify ω. Because π_ω puts half of its mass on $\omega + B_r$ by construction, if we can sample from a distribution within total variation distance less than $1/2$ from π_ω then we will sample a point in $\omega + B_r$ with constant probability. The last ingredient is to note that sampling close to π_ω in Fisher information implies that we are close in total variation distance due to the following functional inequality (see Guillin et al. (2009)): for any probability measure μ,

$$\mathsf{TV}(\mu, \pi_\omega)^2 \leq \frac{1}{4} C_{\mathsf{PI}}(\pi_\omega) \, \mathsf{FI}(\mu \parallel \pi_\omega) \,,$$

where $C_{\mathsf{PI}}(\pi_\omega)$ is the Poincaré constant of π_ω.

As a result, a query complexity lower bound on sampling in Fisher information directly follows from a lower bound on the query complexity of estimating ω, which by standard information-theoretic arguments takes $\Omega(|\mathcal{P}_{2r,R}|)$ queries.

Although the scheme of the argument is straightforward, the actual proof requires careful balancing of the parameters r, R, d and ε and some delicate calculations to satisfy all of the desired properties. The full details are given in Appendix C. ∎

The lower bound in Theorem 10 deteriorates in high dimension; note that due to the restriction $\varepsilon \leq \exp(-c'd)$, the first factor in (3) is exponentially small in d. However, we can remedy this by noting that a d-dimensional construction can be embedded into $\mathbb{R}^{d'}$ for any $d' \geq d$, and hence

$$\mathscr{C}(d, K_0 = 1, \varepsilon) \gtrsim \max_{d_\star \leq d} \left[\left(\frac{cd_\star}{\log(1/\varepsilon)} \right)^{d_\star/2} \varepsilon^{4/(d_\star+2)} \right] \frac{1}{\varepsilon^2} \,.$$

By optimizing over $d_\star$, we show (Appendix C.8) that if $\varepsilon \leq 1/C$, then

$$\mathscr{C}(d, 1, \varepsilon) \gtrsim \begin{cases} \dfrac{1}{\varepsilon^{2d/(d+2)} \exp(C\sqrt{\log(1/\varepsilon) \log\log(1/\varepsilon)})} \,, & \text{for all } d \lesssim \sqrt{\dfrac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}} \,, \\[4ex] \dfrac{1}{\varepsilon^2 \exp(C\sqrt{\log(1/\varepsilon) \log\log(1/\varepsilon)})} \,, & \text{for all } d \gtrsim \sqrt{\dfrac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}} \,, \end{cases}$$

$$= \frac{1}{\varepsilon^{\min\{2d/(d+2),2\}-o(1)}} \,, \qquad\qquad \text{for all } d \geq 1 \,,$$

as $\varepsilon \to 0$, where $C > 0$ is universal. Noting $2d/(d+2) < 2$, this yields the simplified bound in Theorem 3.

For $d = 1$, the lower bound of Theorem 10 reads $\mathscr{C}(1, 1, \varepsilon) \gtrsim 1/(\varepsilon^{2/3}\sqrt{\log(1/\varepsilon)})$. However, for the one-dimensional case we can in fact obtain better bounds on the Poincaré constants of the measures in our lower bound construction, leading to an improvement of the exponent from $2/3$ to 1. This result is stated below.

Theorem 11 (second lower bound, univariate case) *For the class of 1-log-smooth distributions on $\mathbb{R}$, there exists a universal constant $c > 0$, such that for all $\varepsilon < c$, we have*

$$\mathscr{C}(d = 1, K_0 = 1, \varepsilon) \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}} \,.$$

Proof See Appendix C.9. ■

The univariate setting also provides a convenient setting in order to compare our lower bounds with algorithms such as rejection sampling, so we include a detailed discussion in Appendix D. We highlight a few interesting conclusions of the discussion here.

- Although rejection sampling can indeed obtain Fisher information guarantees with complexity $\mathcal{O}(\log(1/\varepsilon))$ (Proposition 30), this does not contradict our lower bounds because rejection sampling cannot be directly implemented within our oracle model. Instead of an initialization μ_0 satisfying $\mathsf{KL}(\mu_0 \parallel \pi) \leq K_0$, rejection sampling requires the stronger assumption $\max\{\sup \log(\mu_0/\pi), \sup \log(\pi/\mu_0)\} \leq M_0$. Under this stronger initialization oracle, the complexity guarantee for rejection sampling is $\mathcal{O}(\exp(3M_0)\log(1/\varepsilon))$.

- In the model with the stronger initialization oracle (i.e., bounded M_0), any algorithm which has $\mathrm{polylog}(1/\varepsilon)$ dependence on the accuracy ε necessarily incurs exponential dependence on M_0 (Corollary 32). This demonstrates a fundamental trade-off between high accuracy (e.g., rejection sampling) and polynomial dependence on M_0 (e.g., averaged LMC).

- The initialization oracle with bounded M_0 is strictly stronger than the one with bounded K_0. In other words, sampling is strictly easier in the presence of an initialization with bounded density ratio to the target (i.e., a *warm start*) than an initialization with bounded KL divergence. This is consistent with intuition from prior work on the complexity of the Metropolis-adjusted Langevin algorithm (see Chewi et al., 2021; Lee et al., 2021; Wu et al., 2022).

- The effective radius R of our lower bound construction scales with $1/\varepsilon$. This is in fact necessary: if R is fixed then there is an algorithm with $\mathcal{O}(\log(1/\varepsilon))$ complexity (Proposition 33).

5. Conclusion

In this work, we have provided the first lower bounds for the query complexity of obtaining Fisher information guarantees for sampling. Due to the scarcity of general sampling lower bounds, our bounds are in fact some of the *only* known lower bounds for sampling. Our results have a number of interesting implications, which we discussed thoroughly in previous sections, and they advance our understanding of the fundamental task of non-log-concave sampling.

To conclude, we highlight a few problems left open in our work. Most notably, our lower bound in Theorem 10 does not match the upper bound of averaged LMC, and it is an important question to close this gap. We also note that our lower bounds in Theorems 10 and 11 do not capture the dependence of K_0, and this is also left for future work.

References

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

Krishna Balasubramanian, Sinho Chewi, Murat A. Erdogdu, Adil Salim, and Matthew Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2896–2923. PMLR, 02–05 Jul 2022.

Sergey G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27(4):1903–1921, 1999.

Sergey G. Bobkov. Spectral gap and concentration for some spherically symmetric probability measures. In *Geometric aspects of functional analysis*, volume 1807 of *Lecture Notes in Math.*, pages 37–43. Springer, Berlin, 2003.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability and low lying spectra in reversible Markov chains. *Comm. Math. Phys.*, 228(2):219–255, 2002.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc. (JEMS)*, 6(4):399–424, 2004.

Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues. *J. Eur. Math. Soc. (JEMS)*, 7(1):69–99, 2005.

Sébastien Bubeck and Dan Mikulincer. How to trap a gradient flow. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 940–960. PMLR, 09–12 Jul 2020.

Sébastien Bubeck. Convex optimization: algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Math. Program.*, 184(1-2, Ser. A):71–120, 2020.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: first-order methods. *Math. Program.*, 185(1-2, Ser. A):315–355, 2021.

Niladri S. Chatterji, Peter L. Bartlett, and Philip M. Long. Oracle lower bounds for stochastic gradient sampling algorithms. *Bernoulli*, 28(2):1074–1092, 2022.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2984–3014. PMLR, 02–05 Jul 2022.

Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv e-prints*, art. arXiv:1805.01648, 2018a.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018b.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1260–1300. PMLR, 15–19 Aug 2021.

Sinho Chewi, Sébastien Bubeck, and Adil Salim. On the complexity of finding stationary points of smooth functions in one dimension. *arXiv e-prints*, art. arXiv:2209.07513, 2022a.

Sinho Chewi, Murat A. Erdogdu, Mufan (Bill) Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1–2. PMLR, 02–05 Jul 2022b.

Sinho Chewi, Patrik R. Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2041–2059. PMLR, 02–05 Jul 2022c.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Murat A. Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: the interplay between tail growth and smoothness. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1776–1822. PMLR, 15–19 Aug 2021.

Rong Ge, Holden Lee, and Jianfeng Lu. Estimating normalizing constants for log-concave distributions: algorithms and lower bounds. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 579–586, 2020.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on*

Learning Theory, volume 178 of *Proceedings of Machine Learning Research*, pages 1948–1989. PMLR, 02–05 Jul 2022.

Arnaud Guillin, Christian Léonard, Liming Wu, and Nian Yao. Transportation-information inequalities for Markov processes. *Probab. Theory Related Fields*, 144(3-4):669–695, 2009.

Ye He, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Heavy-tailed sampling via transformed unadjusted Langevin algorithm. *arXiv e-prints*, art. arXiv:2201.08349, 2022.

Richard Holley and Daniel W. Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. 1986.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona, editors, *Algorithmic Learning Theory*. PMLR, 2023.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Lower bounds on Metropolized sampling methods for well-conditioned distributions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18812–18824. Curran Associates, Inc., 2021.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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? *Bernoulli*, 27(3): 1942 – 1992, 2021.

Mateusz B. Majka, Aleksandar Mijatović, and Łukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *Ann. Appl. Probab.*, 30(4):1534–1581, 2020.

Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Stephen A. Vavasis. Black-box complexity of local minimization. *SIAM J. Optim.*, 3(1):60–80, 1993.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8094–8106. Curran Associates, Inc., 2019.

Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

Andre Wibisono. Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 06–09 Jul 2018.

Andre Wibisono. Proximal Langevin algorithm: rapid convergence under isoperimetry. *arXiv e-prints*, art. arXiv:1911.01469, 2019.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270): 1–63, 2022.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1152–1162. PMLR, 27–30 Jul 2021.

Appendix A. Separation between log-concave and non-log-concave sampling

We show that $\mathcal{O}(\log \frac{1}{\varepsilon})$ Fisher information query complexity is attainable for log-concave densities, by giving a generic post-processing method to turn χ^2-error guarantees into Fisher information guarantees.

A.1. Post-processing lemma

Let Q_t denote heat flow for time t (i.e., convolution with a Gaussian of variance t). We aim to bound $\mathsf{FI}(\mu Q_t \,\|\, \pi)$, where π is the distribution that we wish to sample from, and μ is the output of a sampling algorithm with chi-squared error guarantees.

Lemma 12 (Fisher information guarantee from a chi-squared guarantee) *Suppose that μ and π are two probability measures on $\mathbb{R}^d$, that π is β-log-smooth, and that $\chi^2(\mu \,\|\, \pi) \le \varepsilon_\chi^2 \le 1$. Then, if $t \lesssim 1/\beta$ for a small enough implied constant, it holds that*

$$\mathsf{FI}(\mu Q_t \,\|\, \pi) \lesssim \frac{\varepsilon_\chi \left(d + \log(1/\varepsilon_\chi)\right)}{t} + \beta^2 dt \,.$$

To prove Lemma 12, we start with

$$\mathsf{FI}(\mu Q_t \,\|\, \pi) := \int \|\nabla \log(\mu Q_t)(x) - \nabla \log \pi(x)\|^2 \, \mu Q_t(\mathrm{d}x)$$

$$\le 2\,\mathsf{FI}(\mu Q_t \,\|\, \pi Q_t) + 2 \int_{\mathbb{R}^d} \|\nabla \log(\pi Q_t)(x) - \nabla \log \pi(x)\|^2 \, \mu Q_t(\mathrm{d}x) \,. \quad (5)$$

For the first term in (5), we use the following lemma on error in the score function (gradient of the log-density).

Lemma 13 (score error under heat flow, Lee et al. (2023, Lemma 6.2)) *Let μ and π be probability measures on $\mathbb{R}^d$, and let Q_t denote the heat semigroup at time t. In addition, we assume that $\chi^2(\mu \,\|\, \pi) \le \varepsilon_\chi^2 \le 1$. Then,*

$$\mathsf{FI}(\mu Q_t \,\|\, \pi Q_t) = \int_{\mathbb{R}^d} \|\nabla \log(\mu Q_t)(x) - \nabla \log(\pi Q_t)(x)\|^2 \, \mu Q_t(\mathrm{d}x) \lesssim \frac{\varepsilon_\chi \left(d + \log \frac{1}{\varepsilon_\chi}\right)}{t} \,.$$

For the second term in (5), we use the following score perturbation lemma.

Lemma 14 (Lee et al. (2022, Lemma C.11)) *Suppose that $\pi \propto \exp(-V)$ is a probability density on $\mathbb{R}^d$, where V is β-smooth. Then for $\beta \le \frac{1}{2t}$,*

$$\left\|\nabla \log \frac{\pi(x)}{(\pi Q_t)(x)}\right\| \le 6\beta d^{1/2} t^{1/2} + 2\beta t \,\|\nabla V(x)\| \,.$$

We are now ready to prove Lemma 12.

Proof [Proof of Lemma 12] For the second term in (5), Lemma 14 yields

$$\mathbb{E}_{\mu Q_t}[\|\nabla \log(\pi Q_t) - \nabla \log \pi\|^2] \lesssim \beta^2 dt + \beta^2 t^2 \,\mathbb{E}_{\mu Q_t}[\|\nabla V\|^2] \,.$$

On the other hand, Lemma 16 below yields

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

Hence, from (5) and Lemma 13,

$$\mathsf{FI}(\mu Q_t \| \pi) \lesssim \mathsf{FI}(\mu Q_t \| \pi Q_t) + \mathbb{E}_{\mu Q_t}[\|\nabla \log(\pi Q_t) - \nabla \log \pi\|^2]$$
$$\lesssim \frac{\varepsilon_\chi \, (d + \log(1/\varepsilon_\chi))}{t} + \beta^2 dt + \beta^2 t^2 \, \mathsf{FI}(\mu Q_t \| \pi) \,.$$

If $t \lesssim 1/\beta$ for a small enough implied constant, it implies

$$\mathsf{FI}(\mu Q_t \| \pi) \lesssim \frac{\varepsilon_\chi \, (d + \log(1/\varepsilon_\chi))}{t} + \beta^2 dt$$

as desired. ∎

A.2. High-accuracy Fisher information guarantees for log-concave targets

We now apply the post-processing lemma (Lemma 12). We recall the following high-accuracy guarantee for sampling from log-concave targets in chi-squared divergence, based on the proximal sampler.

Theorem 15 (Chen et al. (2022, Corollary 7)) *Suppose that the target distribution $\pi \propto \exp(-V)$ is β-log-smooth and satisfies a Poincaré inequality with constant C_{PI}. Then, the proximal sampler, with rejection sampling implementation of the restricted Gaussian oracle (RGO) and initialized at μ_0, outputs a sample from a measure μ with $\chi^2(\mu \| \pi) \leq \varepsilon_\chi^2$ using N queries to π in expectation, where N satisfies*

$$N \leq \widetilde{\mathcal{O}}\Big(C_{\mathsf{PI}}\beta d \left(\log(1 + \chi^2(\mu_0 \| \pi)) \vee \log \frac{1}{\varepsilon_\chi}\right)\Big) \,.$$

We now briefly justify why this morally leads to an $\mathcal{O}(\log(1/\varepsilon))$ complexity guarantee in Fisher information, omitting details for brevity. Assume that $\beta = 1$ and that π is log-concave. If we set $t \asymp \varepsilon^2/d$ in Lemma 12, then we can ensure that $\mathsf{FI}(\mu Q_t \| \pi) \leq \varepsilon^2$, where μ is the output of the proximal sampler, provided that $\varepsilon_\chi \leq \widetilde{\mathcal{O}}(\varepsilon^4/d^2)$. Applying Theorem 15, this can be achieved using

$$N = \widetilde{\mathcal{O}}\Big(C_{\mathsf{PI}}d \left(\log(1 + \chi^2(\mu_0 \| \pi)) \vee \log \frac{\sqrt{d}}{\varepsilon}\right)\Big)$$

queries in expectation. Let us give crude bounds for these terms. First, let $\mathfrak{m}_2^2 := \mathbb{E}_\pi[\|\cdot\|^2]$ denote the second moment of π. Then, we know that the Poincaré constant of π is bounded because π is log-concave, and in fact $C_{\mathsf{PI}} \lesssim \mathfrak{m}_2^2$ (see, e.g., Bobkov, 1999). Also, if $\nabla V(0) = 0$, then we can initialize with $\log(1 + \chi^2(\mu_0 \| \pi)) \leq \widetilde{\mathcal{O}}(d)$ (see Chewi et al., 2022b, Lemma 29). Putting this together, we see that $N = \mathrm{poly}(d, \mathfrak{m}_2, \log(\sqrt{d}/\varepsilon))$ queries suffice in expectation in order to obtain the guarantee $\sqrt{\mathsf{FI}(\mu Q_t \| \pi)} \leq \varepsilon$. This is in contrast with our lower bound in Theorem 10, which shows that $\mathrm{poly}(1/\varepsilon)$ queries are necessary to obtain Fisher information guarantees for *non-log-concave* targets, thereby establishing a separation between log-concave and non-log-concave sampling in this context.

The astute reader will observe that there are some holes in this argument when comparing the lower and upper bounds. Namely, the upper bound uses further properties about the target distribution (e.g., $\nabla V(0) = 0$) and does not strictly hold in the oracle model that we describe in Section 2; the upper bound is in terms of the expected number of queries made, because the number of queries made by the algorithm is random; and the upper bound depends on other parameters such as $\mathfrak{m}_2$ which do not appear in the lower bound. In particular, the third point requires some consideration because in our lower bound construction for Theorem 10, the effective radius R of the distributions depends on $1/\varepsilon$. We claim, however, that if we set $d, R = \mathrm{polylog}(1/\varepsilon)$, then the upper bound for log-concave targets is $\mathrm{polylog}(1/\varepsilon)$ (with the caveats just discussed) and the lower bound for non-log-concave targets is $\mathrm{poly}(1/\varepsilon)$. As this is not the focus of our work, we do not attempt to make this reasoning more rigorous; rather, we leave it as the sketch of an argument showing that non-log-concave sampling is fundamentally harder than log-concave sampling. We also note that our argument in fact shows that $\mathrm{polylog}(1/\varepsilon)$ query complexity is possible for distributions satisfying a Poincaré inequality, which form a strict superclass of log-concave distributions.

Appendix B. Proofs for the first lower bound

B.1. Proof of the equivalence

In order to prove the equivalence in Theorem 7, we recall the following useful lemma from Chewi et al. (2022b).

Lemma 16 (Chewi et al. (2022b, Lemma 16)) *Let $\pi \propto \exp(-V)$ be a β-log-smooth density on $\mathbb{R}^d$. Then, for any probability measure μ,*

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

With the lemma in hand, we are ready to prove Theorem 7.

Proof [Proof of Theorem 7]

1. We can explicitly compute

$$\mathsf{FI}(\mu_\beta \,\|\, \pi_\beta) = \int \|\nabla \log \mu_\beta - \nabla \log \pi_\beta\|^2 \, \mathrm{d}\mu_\beta = \int \|\beta\,(z - x) - \beta\,\nabla V(z)\|^2 \, \mathrm{d}\mu_\beta(z)$$

$$\leq 2\beta^2 \int \|z - x\|^2 \, \mathrm{d}\mu_\beta(z) + 2\beta^2 \int \|\nabla V(z)\|^2 \, \mathrm{d}\mu_\beta(z)$$

$$\leq 2\beta^2 \int \|z - x\|^2 \, \mathrm{d}\mu_\beta(z) + 4\beta^2 \int \{\|z - x\|^2 + \|\nabla V(x)\|^2\} \, \mathrm{d}\mu_\beta(z)$$

$$\leq 6\beta^2 \int \|z - x\|^2 \, \mathrm{d}\mu_\beta(z) + 4\beta^2 \underbrace{\|\nabla V(x)\|^2}_{\leq \varepsilon^2} \,,$$

where we used the Lipschitzness of ∇V. Also, $\int \|z - x\|^2 \, \mathrm{d}\mu_\beta(z) = d/\beta$. Hence,

$$\mathsf{FI}(\mu_\beta \,\|\, \pi_\beta) \leq 6\beta d + 4\beta^2 \varepsilon^2 = 10\beta d \,,$$

provided $\beta = d/\varepsilon^2$.

2. Conversely, since $\nabla \log(1/\pi_\beta) = \beta \nabla V$ is β-Lipschitz, then Lemma 16 yields

$$\mathbb{E}_\mu[\|\nabla V\|^2] = \frac{1}{\beta^2} \mathbb{E}_\mu[\|\nabla(\beta V)\|^2] \le \frac{1}{\beta^2} \{\mathsf{FI}(\mu \parallel \pi_\beta) + 2\beta d\} \le \frac{3d}{\beta}.$$

If we take $\beta = d/\varepsilon^2$, then $\mathbb{E}_\mu[\|\nabla V\|^2] \le 3\varepsilon^2$. By Chebyshev's inequality, $X \sim \mu$ satisfies $\|\nabla V(X)\| \le \sqrt{6}\,\varepsilon$ with probability at least $1/2$.

$\blacksquare$

B.2. Proof of the averaged LMC guarantee

In order to apply (Balasubramanian et al., 2022, Theorem 4), we need a bound on the KL divergence at initialization. Such bounds are standard; however, since (Chewi et al., 2022b, Lemma 30) assumes that we start at a stationary point of V (contrary to the present setting), we present an adapted version.

Lemma 17 (KL divergence at initialization) *Suppose that $U : \mathbb{R}^d \to \mathbb{R}$ is a function such that $U(0) - \inf U \le \Delta$, ∇U is β-Lipschitz, and $\mathfrak{m} := \int \|\cdot\|\, d\pi < \infty$ where $\pi \propto \exp(-U)$. Then, for $\mu_0 = \mathsf{normal}(0, \beta^{-1} I_d)$, we have the bound*

$$\mathsf{KL}(\mu_0 \parallel \pi) \lesssim \Delta + d \left(1 \vee \log(\beta \mathfrak{m}^2)\right).$$

Proof Write

$$\frac{\mu_0}{\pi} = \exp\left(U - \frac{\beta}{2} \|\cdot\|^2\right) \frac{\int \exp(-U)}{\int \exp(-U - \delta \|\cdot\|^2)} \frac{\int \exp(-U - \delta \|\cdot\|^2)}{(2\pi/\beta)^{d/2}},$$

where $\delta > 0$ is chosen later.

For the first term, by smoothness and Young's inequality,

$$U(x) - \frac{\beta}{2} \|x\|^2 \le U(0) + \langle \nabla U(0), x \rangle \le U(0) + \frac{\|\nabla U(0)\|^2}{2\beta} + \frac{\beta \|x\|^2}{2}.$$

Plugging in $x = -\frac{1}{\beta} \nabla U(0)$,

$$U\left(-\frac{1}{\beta} \nabla U(0)\right) - U(0) \le -\frac{1}{2\beta} \|\nabla U(0)\|^2$$

or

$$\|\nabla U(0)\|^2 \le 2\beta \left(U(0) - U\left(-\frac{1}{\beta} \nabla U(0)\right)\right) \le 2\beta \left(U(0) - \inf U\right) \le 2\beta\Delta.$$

Hence, for any x,

$$U(x) - \frac{\beta}{2} \|x\|^2 \le U(0) + \Delta + \frac{\beta \|x\|^2}{2}.$$

For the second term, Markov's inequality yields

$$\frac{\int \exp(-U - \delta \|\cdot\|^2)}{\int \exp(-U)} = \int \exp(-\delta \|\cdot\|^2) \, \mathrm{d}\pi \geq \exp(-4\delta\mathfrak{m}^2) \, \pi\{\|\cdot\| \leq 2\mathfrak{m}\}$$

$$\geq \frac{1}{2} \exp(-4\delta\mathfrak{m}^2) \, .$$

For the third term,

$$\frac{\int \exp(-U - \delta \|\cdot\|^2)}{(2\pi/\beta)^{d/2}} \leq \frac{\exp(-\inf U) \int \exp(-\delta \|\cdot\|^2)}{(2\pi/\beta)^{d/2}} = \exp(-\inf U) \left(\frac{\beta}{2\delta}\right)^{d/2} \, .$$

Combining these bounds,

$$\mathsf{KL}(\mu_0 \parallel \pi) = \mathbb{E}_{\mu_0} \log \frac{\mu_0}{\pi} \leq U(0) - \inf U + \Delta + \frac{\beta}{2} \mathbb{E}_{\mu_0}[\|\cdot\|^2] + \log 2 + 4\delta\mathfrak{m}^2 + \frac{d}{2} \log \frac{\beta}{2\delta} \, .$$

Now we set $\delta = \frac{1}{4\mathfrak{m}^2}$ to obtain

$$\mathsf{KL}(\mu_0 \parallel \pi) \lesssim \Delta + d \left(1 \vee \log(\beta\mathfrak{m}^2)\right)$$

as claimed. ∎

Proof [Proof of Corollary 8] Let V be 1-smooth and apply the above lemma to $U = \beta V$, which is β-smooth and satisfies $U(0) - \inf U \leq \beta\Delta$, so that

$$K_0 := \mathsf{KL}(\mu_0 \parallel \pi_\beta) \lesssim \beta\Delta + d \left(1 \vee \log(\beta \, \mathbb{E}_{\pi_\beta}[\|\cdot\|^2])\right) = \widetilde{\mathcal{O}}(\beta\Delta + d) \, . \tag{6}$$

The main result of Balasubramanian et al. (2022) says that after N steps of averaged LMC, with an appropriate choice of step size h, we output a sample from μ satisfying

$$\mathsf{FI}(\mu \parallel \pi_\beta) \lesssim \frac{\beta\sqrt{K_0 d}}{\sqrt{N}} \, .$$

To apply this result, we find N such that this inequality implies $\mathsf{FI}(\mu\|\pi_\beta) \leq \beta d$, where we recall that $\beta = d/\varepsilon^2$; this requires $N \gtrsim K_0/d$. From (6), it suffices to have $N \geq \widetilde{\Omega}(\Delta/\varepsilon^2)$, provided $\Delta/\varepsilon^2 \geq 1$. The result for finding stationary points via averaged LMC now follows from the equivalence in Theorem 7. ∎

B.3. Proof of the first lower bound

Proof [Proof of Theorem 9] Let $\mathscr{F}$ be the family of functions constructed in the lower bound of Carmon et al. (2020), and let $f \in \mathscr{F}$. Recall that $\mathscr{F}$ satisfies the following properties: each $f \in \mathscr{F}$ is β-smooth and satisfies $f(0) - \inf f \leq \Delta$, any randomized algorithm which, for any $f \in \mathscr{F}$, makes queries to a local oracle for f and outputs an δ-stationary point of f with probability at least $1/2$, requires at least $\Omega(\beta\Delta/\delta^2)$ queries.

We set $\delta := 4\sqrt{\beta d}$. From the Fisher information lemma (Lemma 16), if we can obtain a sample from a measure μ such that for $\pi_f \propto \exp(-f)$, it holds that $\mathsf{FI}(\mu \parallel \pi_f) \leq \beta d$, then a sample from μ is a δ-stationary point of f with probability at least $1/2$.

We set the initialization oracle to simply output samples from $\mu_0 := \mathsf{normal}(0, \beta^{-1}I_d)$. We need to compute the value of $K_0 := \sup_{f \in \mathscr{F}} \mathsf{KL}(\mu_0 \parallel \pi_f)$, and for this we use Lemma 17. First, we must bound the second moment $\mathbb{E}_{\pi_f}[\|\cdot\|^2]$. Since we only care about polynomial dependencies for this calculation, let poly denote any positive quantity for which both the quantity and its inverse are bounded above by polynomials in β, Δ, d, and $1/\delta$. Inspecting the proof of Carmon et al. (2020) and using the notation therein, each $f \in \mathscr{F}$ is of the form

$$f(x) = \mathsf{poly} \cdot \tilde{f}_{T,U}\big(\rho(x/\mathsf{poly})\big) + \frac{1}{2\tau^2}\|x\|^2, \qquad \text{where } \tau = \mathsf{poly}.$$

Also, $\tilde{f}_{T,U}$ is bounded; thus, $\pi_f \propto \exp(-f)$ is well-defined. To bound the second moment of π_f, we can use the Donsker–Varadhan variational principle to write, for any $\lambda > 0$,

$$\mathbb{E}_{\pi_f}[\|\cdot\|^2] \leq \frac{1}{\lambda}\left\{\mathsf{KL}(\pi_f \parallel \nu) + \log \mathbb{E}_\nu \exp(\lambda\|\cdot\|^2)\right\},$$

where $\nu := \mathsf{normal}(0, \tau I_d)$. By choosing $\lambda = 1/\mathsf{poly}$, we can ensure that $\log \mathbb{E}_\nu \exp(\lambda\|\cdot\|^2) \leq 1$. Next, since ν satisfies a log-Sobolev inequality with constant poly, we obtain

$$\mathbb{E}_{\pi_f}[\|\cdot\|^2] \leq \mathsf{poly} \cdot \big(1 + \mathsf{FI}(\pi_f \parallel \nu)\big).$$

The Fisher information is computed to be

$$\mathsf{FI}(\pi_f \parallel \nu) = \mathsf{poly} \cdot \mathbb{E}_{\pi_f}\big[\big\|\nabla\big(\tilde{f}_{T,U}\big(\rho(\cdot/\mathsf{poly})\big)\big)\big\|^2\big].$$

Here, $\tilde{f}_{T,U} : \mathbb{R}^d \to \mathbb{R}$ and $\rho : \mathbb{R}^d \to \mathbb{R}^d$ are poly-Lipschitz, and hence

$$\big\|\nabla\big(\tilde{f}_{T,U}\big(\rho(\cdot/\mathsf{poly})\big)\big)\big\| \leq \mathsf{poly}.$$

Putting everything together, we deduce that $\mathbb{E}_{\pi_f}[\|\cdot\|^2] \leq \mathsf{poly}$.

From Lemma 17, we can take $K_0 \lesssim \Delta + \widetilde{\mathcal{O}}(d)$. If $K_0 \geq \widetilde{\Omega}(d)$, then this shows that $\Delta \gtrsim K_0$. From the lower bound of Carmon et al. (2020), we obtain

$$\mathscr{C}(d, K_0, \sqrt{\beta d}; \beta) \gtrsim \frac{\beta\Delta}{\delta^2} \gtrsim \frac{\beta K_0}{\beta d} = \frac{K_0}{d}.$$

Finally, in order for the construction of Carmon et al. (2020) to be valid, the functions must be defined in dimension $d \geq \widetilde{\Omega}((K_0/d)^2)$, which is satisfied provided $d \geq \widetilde{\Omega}(K_0^{2/3})$. ∎

Appendix C. Proofs for the second lower bound

C.1. Proof of Theorem 10

Throughout the proof, we will often work with unnormalized densities. For a distribution π, which we identify with its density, we denote by $\tilde{\pi}$ an unnormalized density, where $\pi = \frac{\tilde{\pi}}{Z}$ and the normalizing constant is given by $Z := \int_{\mathbb{R}^d} \tilde{\pi}(x)\,\mathrm{d}x$.

We reduce the task of estimating the distribution from queries to the task of sampling. Namely, we construct a set of distributions π that are 1-log-smooth, such that if we can sample well from π in Fisher information, then we can estimate its identity. Let B_r denote the ball of radius r in $\mathbb{R}^d$; let

$V_d := \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ denote the volume of B_1, and let $A_{d-1} = dV_d$ denote the surface area of ∂B_1. Let $\mathscr{P}_{2r,R}$ be a maximal $2r$-packing of B_{R-r}, for some $R \geq r$ to be specified. By standard volume arguments (see, e.g., Vershynin, 2018, §4.2), we know that $|\mathscr{P}_{2r,R}| \geq \left(\frac{R-r}{2r}\right)^d$. For any $\omega \in \mathscr{P}_{2r,R}$, let $\tilde{\pi}_\omega$ denote the unnormalized density

$$\tilde{\pi}_\omega(x) := \exp\left(r^2\phi\left(\frac{\|x-\omega\|}{r}\right) - \frac{1}{2}\left(\|x\| - R\right)_+^2\right) =: \exp\left(-V_\omega(x)\right), \tag{7}$$

where $(x)_+ := \max(0, x)$, and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a bump function with the following properties[1] :

(ϕ.1) ϕ is continuous, decreasing, supported on $[0, 1]$, and twice continuously differentiable on the open interval $(0, \infty)$.

(ϕ.2) $\phi(x) = \phi(0) - \frac{1}{2}x^2$ for all $x \in [0, \alpha]$ for some $\alpha > 0$.

(ϕ.3) $\sup_{x>0} |\phi''(x)| \leq 1$.

The above implies that on $\mathbb{R}^d$, $x \mapsto \phi(\|x\|)$ is 1-smooth (see Lemma 27), and hence $\tilde{\pi}_\omega$ is 1-log-smooth. For a measurable set A, we will write $\tilde{\pi}_\omega(A) := \int_A \tilde{\pi}_\omega(x)\, dx$ and we let $Z_\omega := \tilde{\pi}_\omega(\mathbb{R}^d)$ denote the normalizing constant for $\tilde{\pi}_\omega$.

We also define the null probability measure π_{init} to have unnormalized density

$$\tilde{\pi}_{\mathsf{init}}(x) := \exp\left(-\frac{1}{2}\left(\|x\| - R\right)_+^2\right),$$

with normalizing constant $Z_{\mathsf{init}} := \tilde{\pi}_{\mathsf{init}}(\mathbb{R}^d)$.

The distribution π_ω is the combination of a flat, uniform distribution on B_R, fast decaying tails outside of B_R, and a bump of radius r around the point $\omega \in \mathscr{P}_{2r,R}$. The following lemma summarizes the properties that we need for the lower bound construction. Together, Properties (P.1) and (P.2) imply that if an algorithm outputs a sample X from a distribution which is close in Fisher information to π_ω, then X is likely to lie in the set $\omega + B_r$. Hence, an algorithm for sampling from π_ω can be used to *estimate* ω. Property (P.4) is then used to prove a lower bound on the number of queries to solve the estimation task. Finally, Property (P.3) is needed in order to ensure that there is a valid initialization oracle with $K_0 = 1$.

Lemma 18 (lower bound construction)   *There exist universal constants $c_\varepsilon, c > 0$ such that for every $d \geq 1$ and $\varepsilon \leq \exp(-c_\varepsilon d)$ we can choose r, R such that the following properties hold.*

(P.1) (most of the mass lies in the bump) For any $\omega \in \mathscr{P}_{2r,R}$,

$$\pi_\omega(\omega + B_r) = \frac{1}{2}.$$

(P.2) (FI guarantees imply TV guarantees) For any $\omega \in \mathcal{P}_{2r,R}$ and any probability measure μ,

$$\sqrt{\mathsf{FI}(\mu \,\|\, \pi_\omega)} \leq \varepsilon \implies \mathsf{TV}(\mu, \pi_\omega) \leq \frac{1}{3}.$$

---

1. One such function is $\phi(x) = \begin{cases} \frac{11}{64} - \frac{1}{2}x^2, & \text{for } x \in [0, 1/4], \\ \frac{1}{27}\left(4 + 8x - 48x^2 + 56x^3 - 20x^4\right), & \text{for } x \in [1/4, 1], \\ 0, & \text{otherwise}. \end{cases}$

(P.3) (initial KL divergence) There exists a probability measure π_{init} that satisfies

$$\max_{\omega \in \mathscr{P}_{2r,R}} \mathsf{KL}(\pi_{\mathsf{init}} \parallel \pi_\omega) \le \log 2 \,.$$

(P.4) (lower bound on the packing number) It holds that

$$|\mathscr{P}_{2r,R}| \ge \Big(\frac{cd}{\log(1/\varepsilon)}\Big)^{d/2} \frac{1}{\varepsilon^{2d/(d+2)}} \,.$$

Proof Property **(P.1)** holds by the definition of the parameters r and R, see (15) and Lemma 22. We prove Property **(P.2)** in Appendix C.4, Property **(P.3)** in Appendix C.5, and Property **(P.4)** in Appendix C.6. ■

Remark 19 *In order for the bound in Property **(P.4)** to be non-trivial, i.e., $|\mathscr{P}_{2r,R}| \gtrsim 1$, we require $\varepsilon^{-2d/(d+2)} \gtrsim (\sqrt{\log(1/\varepsilon)/(cd)})^d$. Taking logarithms, we want*

$$\frac{2d}{d+2} \log \frac{1}{\varepsilon} \overset{!}{\ge} \frac{d}{2} \log \log \frac{1}{\varepsilon} - \frac{d}{2} \log d + \Omega(d) \,.$$

Let γ be such that $\log(1/\varepsilon) = \gamma d$. Substituting this in, we require

$$\frac{2\gamma d^2}{d+2} \overset{!}{\ge} \frac{d}{2} \log \gamma + \Omega(d) \,.$$

This holds as long as γ is larger than a universal constant, which is equivalent to $\varepsilon \le \exp(-c_\varepsilon d)$ for a sufficiently large absolute constant $c_\varepsilon > 0$.

Using the lemma, we can now apply a standard information theoretic argument. We recall the statement of Fano's inequality, see (Cover and Thomas, 2006, §2) for background.

Theorem 20 (Fano's inequality) *Let $\omega \sim \mathsf{uniform}(\mathfrak{X})$, where $\mathfrak{X}$ is a finite set. Then, for any estimator $\widehat{\omega}$ which is measurable w.r.t. the data ξ, it holds that*

$$\mathbb{P}\{\widehat{\omega} \ne \omega\} \ge 1 - \frac{I(\xi; \omega) + \log 2}{\log |\mathfrak{X}|} \,,$$

where I denotes the mutual information.

With this theorem in hand, we are ready to prove Theorem 10.

Proof [Proof of Theorem 10] Let $\omega \sim \mathsf{uniform}(\mathscr{P}_{2r,R})$ and consider the task of estimating ω with randomized algorithms that have query access to π_ω. We first show that a sampling algorithm can solve this estimation task. Suppose that there is an algorithm that works under the oracle model specified in Section 2, with initialization oracle outputting samples from $\mu_0 = \pi_{\mathsf{init}}$ given in Property **(P.3)**, which guarantees that $\mathsf{KL}(\mu_0 \parallel \pi_\omega) \le \log 2$. For any $\omega \in \mathscr{P}_{2r,R}$ and target π_ω, the algorithm makes at most N queries to the local oracle, and outputs a sample from the measure μ_N with $\sqrt{\mathsf{FI}(\mu_N \parallel \pi_\omega)} \le \varepsilon$. We can then estimate ω as follows: let $X \sim \mu_N$, and set

$$\widehat{\omega} := \underset{\omega \in \mathscr{P}_{2r,R}}{\arg\min} \|X - \omega\| \,.$$

Because the initialization oracle μ_0 is independent of the choice of ω, the estimator $\widehat{\omega}$ is the output of a randomized algorithm that only uses the query information to π_ω to estimate ω.

The probability that $\widehat{\omega}$ succeeds can be calculated as follows. By Property (**P.2**), we have $\mathsf{TV}(\mu_N, \pi_\omega) \leq 1/3$. Let $X \sim \mu_N$; then,

$$\mathbb{P}\{X \in \omega + B_r\} = \mu_N(\omega + B_r) \geq \pi_\omega(\omega + B_r) - \mathsf{TV}(\mu_N, \pi_\omega) \geq \frac{1}{2} - \frac{1}{3} = \frac{1}{6},$$

where we used Property (**P.1**). Hence we see that

$$\mathbb{P}\{\widehat{\omega} = \omega\} \geq \frac{1}{6}. \tag{8}$$

Now we prove a lower bound for the estimation task for any algorithm that succeeds with probability at least $\frac{1}{6}$. Let $x_1, \ldots, x_N$ denote the query points made by the algorithm. We first prove a lower bound for deterministic algorithms, where each query point x_i is a deterministic function of the previous queries and oracle outputs $(x_{i'}, V_\omega(x_{i'}), \nabla V_\omega(x_{i'}) : i' = 1, \ldots, i - 1)$. Since the initialization oracle is independent of ω, the data available to the algorithm is

$$\xi_N := \big(x_i, V_\omega(x_i), \nabla V_\omega(x_i) : i = 1, \ldots, N\big).$$

We assume that the algorithm has made at most $N \leq M/2$ queries where $M := |\mathscr{P}_{2r,R}|$ (otherwise, $N \geq M/2$ and this is our desired lower bound).

Applying Fano's inequality (Theorem 20):, we therefore have

$$\mathbb{P}\{\widehat{\omega} \neq \omega\} \geq 1 - \frac{I(\xi_N; \omega) + \log 2}{\log M}. \tag{9}$$

Applying the chain rule for the mutual information,

$$I(\xi_N; \omega) \leq \sum_{i=1}^{N} I\big(x_i, V_\omega(x_i), \nabla V_\omega(x_i); \omega \mid \xi_{i-1}\big).$$

Given ξ_{i-1}, the query point x_i is deterministic. We can bound the mutual information via the conditional entropy,

$$I\big(x_i, V_\omega(x_i), \nabla V_\omega(x_i); \omega \mid \xi_{i-1}\big) \leq H\big(V_\omega(x_i), \nabla V_\omega(x_i) \mid \xi_{i-1}\big).$$

If one of the query points $x_1, \ldots, x_{i-1}$ landed in the ball $\omega + B_r$, then ω is fully known and the conditional entropy is zero. Otherwise, given the history ξ_{i-1}, the random variable ω is uniformly distributed on the set

$$\mathscr{P}_{2r,R}(i) := \{\omega' \in \mathscr{P}_{2r,R} \mid x_{i'} \notin \omega' + B_r \text{ for } i = 1, \ldots, i-1\}.$$

If x_i does not belong to $\omega' + B_r$ for some $\omega' \in \mathscr{P}_{2r,R}(i)$, then the query point is useless and the conditional entropy is again zero. Otherwise, conditionally on ξ_{i-1}, the tuple $(V_\omega(x_i), \nabla V_\omega(x_i))$ can take on two possible values with probability $1/|\mathscr{P}_{2r,R}(i)|$ and $1 - 1/|\mathscr{P}_{2r,R}(i)|$ respectively, depending on whether or not $x_i \in \omega + B_r$. The conditional entropy is thus bounded by

$$H\big(V_\omega(x_i), \nabla V_\omega(x_i) \mid \xi_{i-1}\big) \leq h\Big(\frac{1}{|\mathscr{P}_{2r,R}(i)|}\Big) \leq h\Big(\frac{2}{M}\Big),$$

where $h(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ is the binary entropy function. Assuming that $M \geq 4$ (which can be ensured thanks to Remark 19),

$$h\Big(\frac{2}{M}\Big) \leq \frac{4}{M} \log \frac{M}{2} \,.$$

Substituting this into (9),

$$\mathbb{P}\{\widehat{\omega} \neq \omega\} \geq 1 - \frac{\frac{4N}{M} \log(M/2) + \log 2}{\log M} \,. \tag{10}$$

If $M \geq 4$, and $N \leq \frac{1}{12} M$, we would obtain $\mathbb{P}\{\widehat{\omega} \neq \omega\} > 5/6$, contradicting (8). Hence, we deduce that $N \gtrsim M$.

In general, if the algorithm is randomized, it can depend on a random seed ζ that is independent of ω. Then we can apply (10) conditional on ζ, and obtain

$$\mathbb{P}\{\widehat{\omega} \neq \omega \mid \zeta\} \geq 1 - \frac{\frac{4N}{M} \log(M/2) + \log 2}{\log M} \,.$$

Taking expectation over ζ, we see that the lower bound (10), and hence $N \gtrsim M$, holds for randomized algorithms as well.

The proof of Theorem 10 is concluded by noting that the estimation lower bound gives a lower bound on sampling, and that Property (**P.4**) provides us with a lower bound on M. ∎

In the remaining sections, we focus on establishing Lemma 18.

C.2. Estimates for integrals

In this section we provide useful estimates for integrals that appear in the normalizing constants for our lower bound construction. Notice that since $\tilde{\pi}_\omega = 1$ on $B_R \setminus (\omega + B_r)$,

$$Z_\omega = \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + \tilde{\pi}_\omega(B_R) = \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + (R^d - r^d)\, V_d + \int_{B_r} \exp\Big(r^2 \phi\big(\frac{\|x\|}{r}\big)\Big) \, \mathrm{d}x$$

$$= \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + (R^d - r^d)\, V_d + r^d\, I_r \,,$$

where we define $I_r := \int_{B_1} \exp(r^2 \phi(\|x\|)) \, \mathrm{d}x$. We record some useful properties of the quantities defined thus far that will be used throughout the proof of Lemma 18.

Lemma 21 (main estimates) *For any number $c > 0$ there exists $c_r(c) > 0$ depending only on c such that for all $r \geq c_r(c)\sqrt{d}$, the following hold:*

1. *(asymptotics of I_r)*

$$\frac{1}{2} \leq \frac{r^d\, I_r}{(2\pi)^{d/2} \exp(r^2 \phi(0))} \leq 2 \,. \tag{11}$$

2. *(mass outside B_R) There is a universal constant $c_0 > 2$, independent of c, such that*

$$\sqrt{\frac{\pi}{2}}\, V_d\, d R^{d-1} \leq \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) \leq V_d c_0^d\, (d R^{d-1} + d^{(d+1)/2}) \,. \tag{12}$$

3. *(mass on the bump)*

$$\log \frac{I_r}{V_d} \geq cd \,. \tag{13}$$

Proof Because we have chosen ϕ to be a quadratic function on the range $[0, \alpha]$ (see (ϕ.**2**)), we can decompose I_r as follows:

$$
\begin{aligned}
I_r &:= \int_{B_1} \exp\bigl(r^2 \phi(\|x\|)\bigr) \,\mathrm{d}x \\
&= \underbrace{\int_{B_1 \backslash B_\alpha} \exp\bigl(r^2 \phi(\|x\|)\bigr) \,\mathrm{d}x}_{\mathsf{A}} + \underbrace{\exp\bigl(r^2 \phi(0)\bigr) \int_{B_\alpha} \exp\Bigl(-\frac{r^2 \|x\|^2}{2}\Bigr) \,\mathrm{d}x}_{\mathsf{B}} \,.
\end{aligned}
$$

As ϕ is decreasing by (ϕ.**1**), clearly $\mathsf{A} \leq V_d \exp(r^2 \phi(\alpha))$, and the second term is given by

$$\mathsf{B} = \frac{(2\pi)^{d/2}}{r^d} \exp\bigl(r^2 \phi(0)\bigr) \, \mathbb{P}(\|X\| \leq \alpha r) \,,$$

where X is a standard Gaussian in $\mathbb{R}^d$. By standard concentration inequalities (e.g., Markov's inequality suffices), there exists a universal constant c_1 such that the above probability is at least $1/2$ provided $r \geq c_1 \sqrt{d}$. Recall that $\log \Gamma(\frac{d}{2} + 1) = \frac{d}{2} \log d + \mathcal{O}(d)$. Thus, for $r \geq c_1 \sqrt{d}$ we have

$$
\begin{aligned}
\log \frac{\mathsf{A}}{\mathsf{B}} &\leq \log \frac{2 V_d \exp(r^2 \phi(\alpha))}{(2\pi)^{d/2} \, r^{-d} \exp(r^2 \phi(0))} = \log \frac{2 \exp(r^2 \phi(\alpha))}{2^{d/2} \, r^{-d} \exp(r^2 \phi(0)) \, \Gamma(\frac{d}{2} + 1)} \\
&= \mathcal{O}(d) - r^2 \left(\phi(0) - \phi(\alpha) \right) + d \log r - \frac{d}{2} \log d \\
&= \mathcal{O}(d) - d \left(\bigl(\frac{r}{\sqrt{d}}\bigr)^2 \left(\phi(0) - \phi(\alpha) \right) - \log \bigl(\frac{r}{\sqrt{d}}\bigr) \right) .
\end{aligned}
$$

From the above it is clear that there is a universal constant c_2 such that $r \geq c_2 \sqrt{d}$ implies that $\mathsf{A} \leq \mathsf{B}$. Thus, for $r \geq (c_1 \vee c_2) \sqrt{d}$ the following holds:

$$\mathsf{B} \leq I_r \leq 2\mathsf{B} \,, \tag{14}$$

proving (11). We now turn to the proof of (12). By integrating in polar coordinates, and taking X to be a standard Gaussian on $\mathbb{R}$,

$$
\begin{aligned}
\tilde{\pi}_\omega(\mathbb{R}^d \backslash B_R) &= A_{d-1} \int_R^\infty s^{d-1} \exp\Bigl(-\frac{1}{2} (s - R)^2\Bigr) \,\mathrm{d}s \\
&= A_{d-1} \int_0^\infty (s + R)^{d-1} \exp\Bigl(-\frac{s^2}{2}\Bigr) \,\mathrm{d}s \\
&\leq \sqrt{2\pi} \, A_{d-1} \, \mathbb{E}[|X + R|^{d-1}] \\
&\leq \sqrt{2\pi} \, A_{d-1} 2^d \left(R^{d-1} + \mathbb{E}[|X|^{d-1}] \right) \\
&\leq A_{d-1} c_0^d \left(R^{d-1} + (d - 1)^{(d-1)/2} \right) \\
&\leq V_d c_0^d \left(d R^{d-1} + d^{(d+1)/2} \right)
\end{aligned}
$$

for some universal constant $c_0 > 2$. For the other direction we can simply write

$$\tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) = A_{d-1} \int_0^\infty (s + R)^{d-1} \exp\left(-\frac{s^2}{2}\right) \mathrm{d}s$$

$$\geq \sqrt{\frac{\pi}{2}} \, A_{d-1} R^{d-1} \,.$$

Finally, we prove (13). We again use the fact $\log \Gamma(\frac{d}{2} + 1) = \frac{d}{2} \log d + \mathcal{O}(d)$. Therefore, for $r \geq (c_1 \vee c_2)\sqrt{d}$ and using (11) we obtain

$$\log \frac{I_r}{V_d} \geq \log \frac{(2\pi)^{d/2} \, r^{-d} \exp(r^2 \phi(0))/2}{\pi^{d/2}/\Gamma(\frac{d}{2} + 1)}$$

$$= d\left(\left(\frac{r}{\sqrt{d}}\right)^2 \phi(0) - \log\left(\frac{r}{\sqrt{d}}\right)\right) + \mathcal{O}(d) \,.$$

Clearly, there exists a constant c_3 (depending only on c) such that $r \geq c_3\sqrt{d}$ implies that the RHS is at least linear in d with a positive constant. Taking $c_r = c_1 \vee c_2 \vee c_3$ concludes the proof. ∎

C.3. Proof of Property (P.1)

We choose r, R such that (P.1) holds, i.e., $\pi_\omega(\omega + B_r) = 1/2$. This holds provided that

$$f(r) := (I_r + V_d)\, r^d \overset{!}{=} \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d R^d =: g(R) \,. \tag{15}$$

Lemma 22 (choice of r, R) *For any value of $d \geq 1$ and $R \geq 0$, there exists a corresponding value of r such that (15) holds. Moreover, there is a universal constant $c_R \geq 1$ such that for any $R \geq c_R\sqrt{d}$, the corresponding r solving (15) satisfies*

$$r \geq c_r\big(\log(6c_0)\big)\sqrt{d} \,, \tag{16}$$

$$R/r \geq 2 \,, \tag{17}$$

where $c_r(\cdot)$ is the function defined in Lemma 21.

The argument $\log(6c_0)$ to $c_r(\cdot)$ in Lemma 22 is chosen for later convenience.

Proof Notice that f and g are continuous and increasing in r, R respectively. Moreover, we check that $f(0) = 0$, $g(0) = (2\pi)^{d/2}$, and $f(\infty) = g(\infty) = \infty$. This tells us that for any value of $d \geq 1$ and $R \geq 0$, there exists a value of $r \geq 0$ for which $f(r) = g(R)$.

For the rest of the proof, we abbreviate $c_r := c_r(\log(6c_0))$.

First, we prove (16). Note that since (16) is a hypothesis of Lemma 21, we cannot invoke Lemma 21 during the proof of (16) in order to avoid a circular argument.

By the definitions of r and R,

$$(I_r + V_d)\, r^d \geq V_d \, R^d \,.$$

Taking logarithms and using the definition of I_r, this rewrites as

$$d \log \frac{R}{r} \leq \log\big(1 + \frac{I_r}{V_d}\big) = \log\Big(1 + \frac{\int_{B_1} \exp(r^2 \phi(\|x\|))\, \mathrm{d}x}{V_d}\Big) \leq \log\big(1 + \exp\big(r^2 \phi(0)\big)\big) \,.$$

24

Suppose, for the sake of contradiction, that $r < c_r\sqrt{d}$. Then, we have

$$d\log\frac{R}{r} \leq c_r^2 d\phi(0) + \log 2 \,.$$

Rearranging,

$$R \leq \exp\Big(c_r^2\phi(0) + \frac{\log 2}{d}\Big)\, r \leq \exp\Big(c_r^2\phi(0) + \frac{\log 2}{d}\Big)\, c_r\sqrt{d} \,.$$

Hence, if $R \geq c_R\sqrt{d}$ for a large enough universal constant c_R, then we arrive at the desired contradiction. For later convenience we choose c_R to always be at least 1. This proves (16).

Next, we prove (17). We use the fact that $R \geq c_R\sqrt{d}$; so that in particular $c_R \geq 1$ and thus $\sqrt{d} \leq R$. Then, using (12) from Lemma 21,

$$(I_r + V_d)\, r^d \leq V_d\,(c_0^d dR^{d-1} + c_0^d d^{(d+1)/2} + R^d) \leq V_d\,(c_0^d\sqrt{d}R^d + c_0^d\sqrt{d}R^d + R^d)$$
$$\leq V_d \cdot 3c_0^d\sqrt{d}\,R^d \,.$$

Taking logarithms, rearranging, and using (13) from Lemma 21,

$$d\log\frac{R}{r} \geq \log\big(1 + \frac{I_r}{V_d}\big) - d\log c_0 - \log(3\sqrt{d}) \geq \big(c - \log c_0 - \underbrace{\frac{\log(3\sqrt{d})}{d}}_{\leq \log 3}\big)\, d \,.$$

Taking $c = \log c_0 + \log 3 + \log 2 = \log(6c_0)$, this implies $R/r \geq 2$ as desired. ∎

C.4. Proof of Property (P.2)

The proof of Property (P.2) requires an upper bound on the Poincaré constant of π_ω. We recall that the Poincaré constant of a probability measure π is the smallest constant $C_{\mathsf{PI}}(\pi) > 0$ such that for all smooth and bounded test functions $f : \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\mathrm{var}_\pi(f) \leq C_{\mathsf{PI}}(\pi)\, \mathbb{E}_\pi[\|\nabla f\|^2] \,.$$

We begin with a Poincaré inequality for π_{init}.

Lemma 23 (Poincaré inequality for π_{init}) *If $R \geq \sqrt{d}$, then the probability measure π_{init} has Poincaré constant at most $c_{\mathsf{PI}}R^2/d$ for a universal constant c_{PI}.*

Proof From Bobkov (2003) and the fact that π_{init} is a radially symmetric log-concave measure, the Poincaré constant of π_{init} is bounded by

$$C_{\mathsf{PI}}(\pi_{\mathsf{init}}) \leq \frac{13\,\mathbb{E}_{\pi_{\mathsf{init}}}[\|\cdot\|^2]}{d} \,.$$

The second moment is

$$\mathbb{E}_{\pi_{\mathsf{init}}}[\|\cdot\|^2] = \frac{\int_{B_R}\|\cdot\|^2}{Z_{\mathsf{init}}} + \frac{\int_{\mathbb{R}^d\setminus B_R}\|\cdot\|^2\exp(-\frac{1}{2}\,(\|\cdot\| - R)^2)}{Z_{\mathsf{init}}}$$
$$\leq \frac{\int_{B_R}\|\cdot\|^2}{V_dR^d} + \frac{A_{d-1}\int_0^\infty (r+R)^{d+1}\exp(-r^2/2)\,\mathrm{d}r}{A_{d-1}\int_0^\infty (r+R)^{d-1}\exp(-r^2/2)\,\mathrm{d}r} \leq R^2 + \int (r+R)^2\,\nu(\mathrm{d}r)\,,$$

where ν is the probability measure on $\mathbb{R}_+$ with density

$$\nu(r) \propto (r + R)^{d-1} \exp\left(-\frac{r^2}{2}\right). \tag{18}$$

Note that ν is 1-strongly-log-concave. Hence, by Durmus and Moulines (2019, Proposition 1),

$$\int (r + R)^2 \,\nu(\mathrm{d}r) \lesssim R^2 + \int r^2 \,\nu(\mathrm{d}r) \lesssim R^2 + r_\star^2 + \int (r - r_\star)^2 \,\nu(\mathrm{d}r) \leq R^2 + r_\star^2 + 1\,,$$

where $r_\star$ is the mode of ν. To find the mode, (18) and elementary calculus show that $r_\star$ satisfies $r_\star (r_\star + R) = d - 1$, which implies $r_\star \leq (d-1)/R$. If $R \geq \sqrt{d}$, then $r_\star \lesssim R$. Combining the bounds, we obtain $C_{\mathsf{PI}}(\pi_{\mathsf{init}}) \lesssim R^2/d$. ∎

Next, we recall the statement of the Holley–Stroock perturbation principle.

Theorem 24 (Holley–Stroock perturbation principle, Holley and Stroock (1986)) *Let π be a probability measure which satisfies a Poincaré inequality. Suppose that μ is another probability measure such that*

$$0 < c \leq \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leq C < \infty\,.$$

Then, μ also satisfies a Poincaré inequality, with

$$C_{\mathsf{PI}}(\mu) \leq \frac{C}{c}\, C_{\mathsf{PI}}(\pi)\,.$$

Proof See Bakry et al. (2014, Lemma 5.1.7). ∎

Corollary 25 (Poincaré inequality for π_ω) *Assume that $R \geq \sqrt{d}$. Then, for each $\omega \in \mathscr{P}_{2r,R}$,*

$$C_{\mathsf{PI}}(\pi_\omega) \leq \frac{2c_{\mathsf{PI}} R^2}{d} \exp\left(r^2 \phi(0)\right).$$

Proof By (ϕ.1), we know that $\tilde{\pi}_\omega \geq \tilde{\pi}_{\mathsf{init}}$ and hence $Z_\omega \geq Z_{\mathsf{init}}$. It follows that

$$\frac{Z_{\mathsf{init}}}{Z_\omega} \leq \frac{\pi_\omega}{\pi_{\mathsf{init}}} = \frac{\tilde{\pi}_\omega}{\tilde{\pi}_{\mathsf{init}}} \frac{Z_{\mathsf{init}}}{Z_\omega} \leq \frac{\tilde{\pi}_\omega}{\tilde{\pi}_{\mathsf{init}}} \leq \exp\left(r^2 \phi(0)\right).$$

Also, by (15),

$$Z_\omega = \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d\, R^d + (I_r - V_d)\, r^d \leq \tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d\, R^d + (I_r + V_d)\, r^d$$
$$= 2\left(\tilde{\pi}_\omega(\mathbb{R}^d \setminus B_R) + V_d\, R^d\right) = 2\, Z_{\mathsf{init}}\,.$$

Hence, $Z_{\mathsf{init}}/Z_\omega \geq 1/2$. The result now follows from Lemma 23 and the Holley–Stroock perturbation principle (Theorem 24). ∎

To translate Fisher information guarantees into total variation guarantees, we use the following consequence of the Poincaré inequality.

Proposition 26 (Fisher information controls total variation) *Suppose that a probability measure π satisfies a Poincaré inequality. Then, for any probability measure μ,*

$$\mathsf{TV}(\mu, \pi)^2 \le \frac{C_{\mathsf{PI}}(\pi)}{4} \, \mathsf{FI}(\mu \parallel \pi) \,.$$

Proof See Guillin et al. (2009). ■

We are finally ready to prove Property **(P.2)**. More specifically, we will show that there is a universal constant $c_\varepsilon > 0$ such that if $\varepsilon \le \exp(-c_\varepsilon d)$, then we can choose r and R (depending on ε) such that: (i) r and R are related according to (15), which is necessary for Property **(P.1)**; (ii) $R \ge c_R \sqrt{d}$, which is necessary for Lemma 22; and (iii) Property **(P.2)** holds.

Proof [Proof of Property **(P.2)**] For any $\omega \in \mathscr{P}_{2r,R}$, suppose that μ satisfies $\sqrt{\mathsf{FI}(\mu \parallel \pi_\omega)} \le \varepsilon$. Then, by Corollary 25 and Proposition 26, we have

$$\mathsf{TV}^2(\mu, \pi_\omega) \le \frac{C_{\mathsf{PI}}(\pi_\omega)}{4} \, \mathsf{FI}(\mu \parallel \pi) \le \frac{c_{\mathsf{PI}} R^2 \exp(r^2 \phi(0))}{2d} \, \varepsilon^2 \,. \tag{19}$$

Hence, if we choose

$$R^2 = \frac{2d}{9 c_{\mathsf{PI}} \varepsilon^2 \exp(r^2 \phi(0))} \tag{20}$$

then $\sqrt{\mathsf{FI}(\mu \parallel \pi_\omega)} \le \varepsilon$ implies $\mathsf{TV}(\mu, \pi_\omega) \le 1/3$, i.e., Property **(P.2)** holds.

To justify (20), note that thus far we have shown that for any choice of R, there exists a choice of r which depends on R, which we temporarily denote by $r(R)$, such that (15) holds. Also, $r(\cdot)$ is an increasing function. In order for (20) to hold, it is equivalent to require

$$R^2 \exp\bigl(r(R)^2 \, \phi(0)\bigr) = \frac{2d}{9 c_{\mathsf{PI}} \varepsilon^2} \tag{21}$$

where the left-hand side is an increasing function of R. We also want R to satisfy $R \ge c_R \sqrt{d}$, where c_R is the universal constant in Lemma 22. From Lemma 22, for the choice of $R = c_R \sqrt{d}$,

$$r(c_R \sqrt{d}) \le \frac{c_R \sqrt{d}}{2} \,.$$

Therefore, for this choice of R, the left-hand side of (21) is bounded by

$$c_R^2 d \exp\Bigl(\frac{c_R^2 d}{4} \, \phi(0)\Bigr) \,.$$

If it holds that

$$\varepsilon^2 \le \frac{2}{9 c_{\mathsf{PI}} c_R^2 \exp(c_R^2 d \phi(0)/4)} \tag{22}$$

then the R satisfying (20) necessarily satisfies $R \ge c_R \sqrt{d}$. In turn, (22) holds if $\varepsilon \le \exp(-c_\varepsilon d)$ for a universal constant $c_\varepsilon > 0$. ■

C.5. Proof of Property (P.3)

Proof [Proof of Property (P.3)] In the proof of Corollary 25, we showed that $Z_\omega \leq 2\,Z_{\mathsf{init}}$. The KL divergence is bounded by

$$\mathsf{KL}\big(\pi_{\mathsf{init}} \parallel \pi_\omega\big) = \mathbb{E}_{\pi_{\mathsf{init}}} \log\Big(\underbrace{\frac{\tilde\pi_{\mathsf{init}}}{\tilde\pi_\omega}}_{\leq 1}\underbrace{\frac{Z_\omega}{Z_{\mathsf{init}}}}_{\leq 2}\Big) \leq \log 2\,,$$

which is what we wanted to show. ∎

C.6. Proof of Property (P.4)

Proof [Proof of Property (P.4)] We choose r and R to satisfy (15) and (20). If $\varepsilon \leq \exp(-c_\varepsilon d)$, then we showed in the proof of Property (P.2) that $R \geq c_R \sqrt{d}$ and hence Lemmas 21 and 22 apply.

As in the proof of (17) in Lemma 22, $R \geq \sqrt{d}$ implies

$$(I_r + V_d)\,r^d \leq V_d \cdot 3c_0^d \sqrt{d}\,R^d\,.$$

Taking logarithms in (11) from Lemma 21 and using the above inequality, we obtain

$$r^2 \phi(0) \leq \log\frac{2r^d I_r}{(2\pi)^{d/2}} \leq \mathcal{O}(d) + \log V_d + d \log R\,.$$

From (20), we have

$$\log R = \frac{1}{2}\log d + \log\frac{1}{\varepsilon} - \frac{1}{2}\,r^2\phi(0) + \mathcal{O}(1)\,.$$

Substituting this in and using $\log V_d = -\frac{d}{2}\log d + \mathcal{O}(d)$,

$$r^2\phi(0) \leq d\log\frac{1}{\varepsilon} - \frac{d}{2}\,r^2\phi(0) + \mathcal{O}(d)$$

which is rearranged to yield

$$r^2\phi(0) \leq \frac{2d}{d+2}\log\frac{1}{\varepsilon} + \mathcal{O}(1)\,.$$

Then, the packing number is lower bounded by

$$\begin{aligned}
|\mathscr{P}_{2r,R}| &\geq \Big(\frac{R-r}{2r}\Big)^d \geq \Big(\frac{R}{4r}\Big)^d \\
&\geq \Big(c\,\frac{\sqrt{d}\exp(-\frac{d}{d+2}\log(1/\varepsilon))}{\varepsilon\sqrt{\log(1/\varepsilon)}}\Big)^d \\
&\geq \Big(c\,\sqrt{\frac{d}{\log(1/\varepsilon)}}\Big)^d\frac{1}{\varepsilon^{2d/(d+2)}}\,,
\end{aligned}$$

for some universal constant c. ∎

C.7. Auxiliary lemmas

Lemma 27 *Suppose that $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies (ϕ.1), (ϕ.2), and (ϕ.3). Then, the map $x \mapsto \phi(\|x\|)$ is 1-smooth on $\mathbb{R}^d$.*

Proof First, we claim that $|\phi'(x)|/x \leq 1$ for all $x > 0$. This follows from (ϕ.3) because (ϕ.2) implies that the right derivative $\phi'(0+)$ exists and equals 0.

Next, we have for $x \neq 0$

$$\frac{\partial^2 \phi(\|x\|)}{\partial x_j \, \partial x_i} = \frac{\partial}{\partial x_j} \, \phi'(\|x\|) \, \frac{x_i}{\|x\|}$$

$$= \phi''(\|x\|) \, \frac{x_i x_j}{\|x\|^2} - \phi'(\|x\|) \, \frac{x_i x_j}{\|x\|^3} + \delta_{i,j} \, \phi'(\|x\|) \, \frac{1}{\|x\|} \, .$$

Thus, in matrix form we have

$$\nabla_x^2 \phi(\|x\|) = \frac{\phi'(\|x\|)}{\|x\|} \, I_d + \Big(\frac{\phi''(\|x\|)}{\|x\|^2} - \frac{\phi'(\|x\|)}{\|x\|^3} \Big) \, xx^\mathsf{T} .$$

In particular, the eigenvalues are always $\frac{\phi'(\|x\|)}{\|x\|}$ with multiplicity $d-1$ and $\phi''(\|x\|)$ with multiplicity 1. The fact that $\phi(\|\cdot\|)$ is 1-smooth follows. $\blacksquare$

C.8. Optimization of the bound

We wish to find d which maximizes

$$\Big(\frac{cd}{\log(1/\varepsilon)} \Big)^{d/2} \varepsilon^{4/(d+2)} ,$$

or after taking logarithms, we wish to maximize

$$f(d) := \frac{d}{2} \log d - \frac{4}{d+2} \log \frac{1}{\varepsilon} - \frac{d}{2} \log \log \frac{1}{\varepsilon} - \frac{d}{2} \log \frac{1}{c} \, .$$

Rather than maximizing this expression exactly, we shall ignore the last two terms and pick d to be the smallest integer such that the sum of the first two terms is non-negative, i.e.,

$$\frac{d \, (d+2) \log d}{8} \geq \log \frac{1}{\varepsilon} \, .$$

It suffices to find d such that $g(d) := d^2 \log d \geq 8 \log(1/\varepsilon)$. In order to invert g, let y be sufficiently large and consider finding x such that $g(x) = y$. We make the choice $x = \alpha \sqrt{y/(\log y)}$ and plug this into the expression for g in order to obtain

$$\log g\Big(\alpha \sqrt{\frac{y}{\log y}} \Big) = 2 \log \alpha + \log y - \log \log y + \log \log \Big(\alpha \sqrt{\frac{y}{\log y}} \Big)$$

$$= 2 \log \alpha + \log y + \log \underbrace{\frac{\frac{1}{2} \log y - \frac{1}{2} \log \log y + \log \alpha}{\log y}}_{\to \log(1/2) \text{ as } y \to \infty} .$$

29

From this expression, we see that provided y is sufficiently large, this expression is less than $\log y$ for $\alpha = 0$ and greater than $\log y$ for $\alpha = 3$. We conclude that $g^{-1}(y) \asymp \sqrt{y/(\log y)}$, and therefore that our choice of d satisfies

$$d \asymp \sqrt{\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}}\,.$$

In particular, since $d = o(\log(1/\varepsilon))$, then the condition $\varepsilon \le \exp(-c_\varepsilon d)$ holds for all sufficiently small ε, and Theorem 10 holds. Then,

$$f(d) \ge -\frac{d}{2}\log\log\frac{1}{\varepsilon} - \frac{d}{2}\log\frac{1}{c} \asymp -\sqrt{\left(\log\frac{1}{\varepsilon}\right)\left(\log\log\frac{1}{\varepsilon}\right)}\,.$$

This verifies the expression in Section 4.

To justify the simplified expression of the bound that we gave in the informal statement of Theorem 3, note that in dimension

$$d \lesssim \sqrt{\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}} \tag{23}$$

we have

$$\log\left(\left(\frac{cd}{\log(1/\varepsilon)}\right)^{d/2}\right) = \frac{d}{2}\underbrace{\left(\log d - \log\log\frac{1}{\varepsilon} - \log\frac{1}{c}\right)}_{\text{negative as } \varepsilon \searrow 0} \gtrsim -\sqrt{\left(\log\frac{1}{\varepsilon}\right)\left(\log\log\frac{1}{\varepsilon}\right)}\,.$$

In other words, we can simplify our bound as follows. For all $d \ge 1$ and all ε smaller than a universal constant, if the condition (23) holds, then we have the lower bound

$$\mathscr{C}(d, 1, \varepsilon) \gtrsim \frac{1}{\varepsilon^{2d/(d+2)}\exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\,.$$

Otherwise, if the condition (23) fails, then we instead have the bound

$$\mathscr{C}(d, 1, \varepsilon) \gtrsim \frac{1}{\varepsilon^2 \exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})} \ge \frac{1}{\varepsilon^{2d/(d+2)}\exp(C\sqrt{\log(1/\varepsilon)\log\log(1/\varepsilon)})}\,.$$

In either case, we have $\mathscr{C}(d, 1, \varepsilon) \ge (1/\varepsilon)^{2d/(d+2)-o(1)}$. Together with Theorem 11 on the univarate case and Lemma 6 on rescaling, it yields Theorem 3.

C.9. Proof of Theorem 11

In the univarate case, we can sharpen Theorem 10 by obtaining a better bound on the Poincaré constant of π_ω. We use the following result.

Theorem 28 (Muckenhoupt's criterion) *Let π be a probability density on $\mathbb{R}$ and let m be a median of π. Then,*

$$C_{\mathsf{PI}}(\pi) \asymp \max\left\{\sup_{x<m}\pi\big((-\infty, x]\big)\int_x^m\frac{1}{\pi}, \ \sup_{x>m}\pi\big([x, +\infty)\big)\int_m^x\frac{1}{\pi}\right\}.$$

Proof See Bakry et al. (2014, Theorem 4.5.1). ∎

Lemma 29 (improved Poincaré inequality for π_ω) *Suppose that $d = 1$ and $R \geq 1$. Then, for all $\omega \in \mathscr{P}_{2r,R}$,*

$$C_{\mathsf{PI}}(\pi_\omega) \lesssim R^2 \,.$$

Proof We use Muckenhoupt's criterion (Theorem 28). First, we note that by Property (P.1), it holds that $\pi_\omega(\omega + B_r) = \frac{1}{2}$ which implies that $\omega - r \leq m \leq \omega + r$. We proceed to check that

$$\sup_{x > m} \pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \lesssim R^2 \,.$$

The other condition is verified in the same way due to symmetry.

We split into three cases. First, suppose that $m < x < \omega + r$. Then, as in the proof of Corollary 25, we have $Z_\omega \leq 2 Z_{\mathsf{init}} = 2\,\tilde{\pi}_{\mathsf{init}}(\mathbb{R} \setminus B_R) + 4R \leq 2\sqrt{2\pi} + 4R \lesssim R$. Then,

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \leq Z_\omega\,(x - m) \lesssim Rr \leq R^2 \,.$$

Next, suppose that $\omega + r < x < R$. Then,

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} = \tilde{\pi}_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\tilde{\pi}_\omega} \leq \Big(R - x + \sqrt{\frac{\pi}{2}}\Big)(x - m) \lesssim R^2 \,.$$

Finally, suppose that $x > R$. Then, using standard Gaussian tail bounds,

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} = \tilde{\pi}_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\tilde{\pi}_\omega}$$

$$\leq \Big[\sqrt{2\pi}\,\big(\tfrac{1}{2} \wedge \tfrac{1}{x - R}\big) \exp\Big(-\frac{(x - R)^2}{2}\Big)\Big] \Big[R - m + (x - R)\exp\Big(\frac{(x - R)^2}{2}\Big)\Big] \,.$$

If $x - R \leq 1$, then this yields

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \lesssim R \,.$$

Otherwise, if $x - R \geq 1$, then we obtain

$$\pi_\omega\big([x, +\infty)\big) \int_m^x \frac{1}{\pi_\omega} \lesssim \frac{1}{x - R} \exp\Big(-\frac{(x - R)^2}{2}\Big) \Big[R + (x - R)\exp\Big(\frac{(x - R)^2}{2}\Big)\Big] \lesssim R \,.$$

This completes the proof. ∎

We now use the improved Poincaré inequality in order to establish Theorem 11.

Proof [Proof of Theorem 11] We follow the proof of Theorem 10. The proofs of Properties (P.1) and (P.3) remain unchanged.

In the proof of Property (**P.2**), the equation (19) is replaced by

$$\mathsf{TV}^2(\mu, \pi_\omega) \le c_{\mathsf{PI}} R^2 \varepsilon^2$$

for a different universal constant $c_{\mathsf{PI}} > 0$, using Lemma 29. Hence, we choose $R^2 = 1/(9c_{\mathsf{PI}}\varepsilon^2)$ in order to verify Property (**P.2**). Since we require $R \ge c_R$ for a universal constant $c_R \ge 1$, this requires $\varepsilon \le \exp(-c_\varepsilon)$ for a universal constant $c_\varepsilon > 0$.

Next, we turn towards the sharpened statement of Property (**P.4**). From (15), r is chosen so that

$$(I_r + 2)\, r = \tilde{\pi}_\omega(\mathbb{R} \setminus B_R) + 2R\,.$$

Using (11) from Lemma 21, we have

$$r I_r \asymp \exp\!\big(r^2 \phi(0)\big) \gtrsim r\,.$$

This implies that

$$\exp\!\big(r^2 \phi(0)\big) \gtrsim (I_r + 2)\, r \gtrsim R\,,$$

or $r \gtrsim \sqrt{\log R} \asymp \sqrt{\log(1/\varepsilon)}$. Hence,

$$|\mathscr{P}_{2r,R}| \ge \frac{R}{4r} \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\,.$$

By substituting this new bound on the packing number into the information theoretic argument of Theorem 10 (see (10), where $M = |\mathscr{P}_{2r,R}|$), we obtain Theorem 11. ∎

Appendix D. Further discussion of the univariate case

In this section, we provide further discussion of algorithms for the univariate case.

Rejection sampling. First of all, we note that the $\mathrm{poly}(1/\varepsilon)$ lower bounds of Theorems 10 and 11 may come as a surprise due to the existence of the rejection sampling algorithm. We briefly recall rejection sampling here. Let $\tilde{\pi}$ be an unnormalized density, let $Z_\pi := \int \tilde{\pi}$ denote the normalizing constant, and let $\pi := \tilde{\pi}/Z$ denote the target distribution. Rejection sampling requires knowledge of an upper envelope $\tilde{\mu}$ for $\tilde{\pi}$, i.e., a function $\tilde{\mu}$ satisfying $\tilde{\mu} \ge \tilde{\pi}$ pointwise. The algorithm proceeds by repeatedly drawing samples from the density $\mu := \tilde{\mu}/Z_\mu$, where $Z_\mu := \int \tilde{\mu}$; each sample X is accepted with probability $\tilde{\pi}(X)/\tilde{\mu}(X)$.

It is standard to show (see, e.g., Chewi et al., 2022c) that the accepted samples are drawn exactly from the target π, and that the number of queries made to $\tilde{\pi}$ until the first accepted sample is geometrically distributed with mean Z_μ/Z_π. To translate this into a total variation guarantee, we run the algorithm for N iterations and output "FAIL" if we have not accepted a sample by iteration N. The probability of failure is at most $(1 - Z_\pi/Z_\mu)^N$, so the number of iterations required for the output of the algorithm to be ε-close to the target π in total variation distance is $N \ge \log(1/\varepsilon)/\log(1 - Z_\pi/Z_\mu)$.

Although this is a total variation guarantee, rather than a Fisher information guarantee, it suggests (similarly to Appendix A) that $\log(1/\varepsilon)$ rates are attainable using rejection sampling. The

reason why this does not contradict our lower bounds in Theorems 10 and 11 is that the initialization oracle we consider, which provides a measure μ_0 such that $\mathsf{KL}(\mu_0 \parallel \pi) \leq K_0$, is not sufficient to construct an upper envelope of the unnormalized density $\tilde{\pi}$.

Indeed, consider instead a stronger initialization oracle which outputs a measure μ_0 such that

$$\max\Big\{\sup \log \frac{\mu_0}{\pi}, \ \sup \log \frac{\pi}{\mu_0}\Big\} \leq M_0 < \infty \, .$$

Denote the complexity of obtaining $\sqrt{\mathsf{FI}(\mu \parallel \pi)} \leq \varepsilon$ over the class of 1-log-smooth distributions on $\mathbb{R}^d$ with this stronger initialization oracle by $\mathscr{C}_\infty(d, M_0, \varepsilon)$. Then, the rejection sampling algorithm can be implemented within this new oracle model. It yields the following.

Proposition 30 (Fisher information guarantees via rejection sampling) *It holds that*

$$\mathscr{C}_\infty(d, M_0, \varepsilon) \leq \widetilde{\mathcal{O}}\Big(\exp(3M_0) \log \frac{\sqrt{d}}{\varepsilon}\Big) \, .$$

Proof For the algorithm, we use rejection sampling, which requires producing an upper envelope. Recall that in our oracle model, we can query the value of an unnormalized version $\tilde{\pi}$ of π. By replacing $\tilde{\pi}$ with $\tilde{\pi}/\tilde{\pi}(0)$, we can assume that $\tilde{\pi}(0) = 1$. Then,

$$\tilde{\pi} = \frac{\tilde{\pi}}{\tilde{\pi}(0)} = \frac{\pi}{\pi(0)} \leq \frac{\exp(M_0) \, \mu_0}{\exp(-M_0) \, \mu_0(0)} = \underbrace{\frac{\exp(2M_0)}{\mu_0(0)}}_{:=Z_{\mu_0}} \mu_0 \, .$$

This shows that $\tilde{\mu}_0 := Z_{\mu_0} \, \mu_0$ is an upper envelope for $\tilde{\pi}$. Also, using $\pi(0) = 1/Z_\pi$,

$$\frac{Z_{\mu_0}}{Z_\pi} = \exp(2M_0) \, \frac{\pi(0)}{\mu_0(0)} \leq \exp(3M_0) \, .$$

Hence, we can run rejection sampling, where we output a sample from μ_0 if the algorithm exceeds N iterations. Therefore, the law of the output of rejection sampling is $\mu = (1-p) \, \pi + p \, \mu_0$, where $p = (1 - Z_\pi/Z_{\mu_0})^N \leq \exp(-N Z_\pi/Z_{\mu_0})$ is the probability of failure. We calculate

$$1 + \chi^2(\mu \parallel \pi) = \mathbb{E}_\mu\Big(\frac{\mu}{\pi}\Big) = 1 - p + p \, \mathbb{E}_\mu\Big(\frac{\mu_0}{\pi}\Big) \leq 1 + p \exp(M_0) \, .$$

Applying Lemma 12 with $\varepsilon_\chi^2 = p \exp(M_0)$ (assuming that $p \leq \exp(-M_0)$) and $t \lesssim 1$, we obtain

$$\mathsf{FI}(\mu Q_t \parallel \pi) \lesssim \frac{p \exp(M_0) \, (d + \log(1/p) - M_0)}{t} + dt \, .$$

We set $t \lesssim \varepsilon^2/d$ so that

$$\mathsf{FI}(\mu Q_t \parallel \pi) \lesssim \frac{d^2 \exp(M_0) \, p \log(1/p)}{\varepsilon^2} + \varepsilon^2 \, .$$

In order to make the first term at most $\varepsilon^2/2$, we take $p = \widetilde{\Theta}(\varepsilon^4/(d^2 \exp(M_0)))$. In turn, this is satisfied provided

$$N \geq \frac{Z_{\mu_0}}{Z_\pi} \log \frac{1}{p} \asymp \exp(3M_0) \log \frac{d^2 \exp(M_0)}{\varepsilon^4} \, ,$$

which proves the desired result. ∎

Hence, under the stronger oracle model, $\log(1/\varepsilon)$ rates are indeed possible (albeit with exponential dependence on M_0). To see why this does not contradict the lower bound construction of Theorem 11, observe that if we take the initialization oracle to be π_{init}, then our construction satisfies $M_0 = r^2 \phi(0)$. By inspecting the proof of Theorem 11, one sees that $r \asymp \sqrt{\log(1/\varepsilon)}$. Hence, our construction does not provide a lower bound for $\mathscr{C}_\infty(1, M_0, \varepsilon)$ for constant M_0. Instead, we obtain the following lower bound.

Corollary 31 (lower bound for the stronger initialization oracle) *There exists a universal constant $c > 0$ such that for all $\varepsilon \leq 1/c$, it holds that*

$$\mathscr{C}_\infty\big(1, c\log(1/\varepsilon), \varepsilon\big) \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}} \,.$$

Note also the following corollary.

Corollary 32 (high-accuracy Fisher information requires exponential dependence on M_0)
Suppose that there exists an algorithm which works within the stronger oracle model and which, for any 1-log-smooth distribution π on $\mathbb{R}$, outputs a measure μ with $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \leq \varepsilon$ using N queries, where the query complexity satisfies

$$N \leq f(M_0) \operatorname{polylog}\Big(\frac{1}{\varepsilon}\Big)$$

for some increasing function $f : [1, \infty) \to \mathbb{R}_+$. Then, there is a universal constant $c' > 0$ such that

$$f(M_0) \geq \widetilde{\Omega}\big(\exp(c'M_0)\big) \,.$$

Proof Using Corollary 31 with $M_0 = c\log(1/\varepsilon)$, we have

$$f\Big(c\log\frac{1}{\varepsilon}\Big)\operatorname{polylog}\Big(\frac{1}{\varepsilon}\Big) \geq N \gtrsim \frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}} \,,$$

or

$$f\Big(c\log\frac{1}{\varepsilon}\Big) \geq \frac{1}{\varepsilon\operatorname{polylog}(1/\varepsilon)} \,.$$

Writing this in terms of $M_0 = c\log(1/\varepsilon)$, or $\varepsilon = \exp(-M_0/c)$,

$$f(M_0) \geq \frac{\exp(M_0/c)}{(M_0/c)^{\mathcal{O}(1)}} = \widetilde{\Omega}\Big(\exp\big(\frac{M_0}{c}\big)\Big)$$

which establishes the result. ∎

Hence, we see that there is a fundamental trade-off in the stronger oracle model: any algorithm must either incur polynomial dependence on $1/\varepsilon$ (e.g., averaged LMC), or exponential dependence on M_0 (e.g., rejection sampling, see Proposition 30).

The stronger oracle model is strictly stronger. We also observe the following consequence of these observations. On one hand, our lower bound in Thoerem 11 shows that

$$\mathscr{C}(1, K_0 = 1, \varepsilon) \geq \Omega\Big(\frac{1}{\varepsilon\sqrt{\log(1/\varepsilon)}}\Big)\,.$$

On the other hand, for constant M_0, rejection sampling (Proposition 30) yields

$$\mathscr{C}_\infty(1, M_0, \varepsilon) \leq \widetilde{\mathcal{O}}\Big(\exp(3M_0)\log\frac{1}{\varepsilon}\Big)\,.$$

Hence, the stronger oracle model is indeed stronger: *obtaining Fisher information guarantees is strictly easier with access to an oracle with bounded M_0, rather than an oracle with bounded K_0.*

On the effect of the radius of the effective support. In our lower bound construction, the distributions are "effectively" supported on a ball of radius R, where R scales with $1/\varepsilon$. Here, we show that this is in fact necessary, by showing that for any *fixed d and R*, it is possible to sample from such a distribution in Fisher information using $\mathcal{O}(\log(1/\varepsilon))$ queries. The algorithm involves uses a simple grid search.

Proposition 33 (sampling from bounded effective support) *Suppose that the target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ has the following properties:*

1. *$V(0) = 0$.*

2. *$V(x) = \frac{1}{2}\left(\|x\| - R\right)_+^2$, for $\|x\| \geq R$.*

3. *V is 1-smooth.*

Then, there is an algorithm which outputs μ with $\sqrt{\mathsf{FI}(\mu \,\|\, \pi)} \leq \varepsilon$ using N queries to $(V, \nabla V)$, where the number of queries satisfies

$$N \lesssim (cR)^d + \log\frac{\sqrt{d}}{\varepsilon}\,,$$

where $c > 0$ is a universal constant.

Proof We use function approximation to build an upper envelope for $\tilde{\pi} := \exp(-V)$, and then apply rejection sampling. Namely, let $\mathscr{N}$ be a 1-net of B_R, and for each $x \in B_R$ let $x_{\mathscr{N}}$ denote a closest point of $\mathscr{N}$ to x. Define the approximation

$$\widehat{V}(x) := \begin{cases} \frac{1}{2}\left(\|x\| - R\right)_+^2\,, & \|x\| \geq R\,, \\ V(x_{\mathscr{N}}) + \langle\nabla V(x_{\mathscr{N}}), x - x_{\mathscr{N}}\rangle - \frac{1}{2}\|x - x_{\mathscr{N}}\|^2\,, & \|x\| < R\,. \end{cases}$$

By 1-smoothness of V, we have $V \geq \widehat{V}$, so that if we let $\tilde{\mu}_0 := \exp(-\widehat{V})$, then $\tilde{\mu}_0 \geq \tilde{\pi}$. Also, for $\|x\| < R$, we have the bound

$$\begin{aligned} \tilde{\mu}_0(x) &= \exp\Big(-V(x_{\mathscr{N}}) - \langle\nabla V(x_{\mathscr{N}}), x - x_{\mathscr{N}}\rangle + \frac{1}{2}\|x - x_{\mathscr{N}}\|^2\Big) \\ &\leq \exp\big(-V(x) + \|x - x_{\mathscr{N}}\|^2\big) = \tilde{\pi}(x)\exp(\|x - x_{\mathscr{N}}\|^2) \leq \exp(1)\,\tilde{\pi}(x)\,, \end{aligned}$$

35

so that $Z_{\mu_0}/Z_\pi \lesssim 1$. We now perform rejection sampling using N' iterations with upper envelope $\tilde\mu_0$, outputting a sample from μ_0 if N' iterations are exceeded. Tracing through the proof of Proposition 30, one can show that for the output μ of rejection sampling, it holds that $\mathsf{FI}(\mu Q_t \parallel \pi) \le \varepsilon^2$ for an appropriate choice of t. Moreover, the number of iterations of rejection sampling required to achieve this satisfies $N' \lesssim \log(\sqrt{d}/\varepsilon)$. Finally, since $|\mathscr{N}| \le (cR)^d$ for a universal constant $c > 0$, it requires $\mathcal{O}((cR)^d)$ queries in order to build the upper envelope $\tilde\mu_0$, which proves the result. ∎

To summarize the situation, if the effective radius R is known and fixed, then it is possible to obtain $\mathcal{O}(\log(1/\varepsilon))$ complexity. However, if there is no a priori upper bound on the radius R, then the lower bounds of Theorem 11 and Corollary 31 apply.