EFFICIENT MASS MATRIX ESTIMATION WITH GAUSSIAN COOLING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce an algorithm for efficiently preconditioning log-concave and log-smooth distributions that scales logarithmically with the condition number of the underlying distribution. Based on Gaussian cooling, our multistage method approximately samples from a sequence of well-conditioned distributions to construct a preconditioner for each subsequent stage of the cooling schedule. This method is motivated by existing practical approaches for mass matrix estimation in Markov chain Monte Carlo (MCMC), in which an effective preconditioner for the target distribution is estimated from samples produced during a warm-up period. However, our approach is conceptually distinct from existing approaches and moreover permits theoretical analysis. Our algorithmic framework is flexible, allowing essentially any log-concave sampling algorithm to act as the subroutine within each cooling stage. For instance, using first-order rejection sampling (FORS) as the sampler, the first-order query complexity of our method is $\tilde{O}(d^{4/3} \log \kappa)$ for a large class of log-concave and log-smooth distributions with condition number κ and $\tilde{O}(d^{1/3} \log \kappa)$ for translation-invariant distributions. In the translation-invariant setting, we provide numerical experiments for the lattice ϕ^4 model, a simple lattice field theory.

1 INTRODUCTION

Given a potential $U : \mathbb{R}^d \to \mathbb{R}$, sampling from the probability distribution specified by the density $\pi(x) \propto \exp(-U(x))$ is a ubiquitous task in scientific computing, arising in generative modeling, molecular dynamics, and a variety of other applications across physics, chemistry, computer science, and computational mathematics. Most modern approaches to this sampling problem involve some form of *Markov chain Monte Carlo* (MCMC), in which a Markov chain that admits (an approximation of) π as an invariant distribution is constructed. Given query access to the potential U and, if necessary, its gradient, MCMC practitioners then simulate a prescribed chain, ultimately obtaining a set of approximate samples from π. Popular examples of such methods include the Metropolis-adjusted Langevin algorithm (MALA), Hamiltonian Monte Carlo (HMC) and its adjusted variant, and ensemble MCMC samplers involving populations of interacting particles. From an algorithmic perspective, the mixing times of MCMC methods are typically bottlenecked by three factors: (i) dimension dependence, (ii) multimodality, and (iii) condition number dependence.

Significant recent progress has been made towards understanding the optimal dimension dependence of MALA (Chewi et al., 2021), underdamped Langevin Monte Carlo (ULMC) (Altschuler et al., 2026), and HMC (Zhang et al., 2026) for sampling from strongly log-concave distributions. Multimodal distributions, which do not fall under the umbrella of (strong) log-concavity, pose many challenges to MCMC methods. While many adaptive methods attempt to address the problem of sampling from multimodal distributions, we do not focus on this setting in our work.

Closely related to the question of optimal condition number dependence is the notion of *affine invariance*, popularized by Goodman & Weare (2010). Loosely speaking, the mixing time of an affine invariant sampler does not depend on affine changes of coordinates, and such samplers should therefore be unfazed by problems with extremely large condition numbers. While the affine invariant ensemble samplers proposed in Goodman & Weare (2010), and more recently, in Chen (2025), are distinct from the discretized diffusion processes (e.g., ULMC) discussed above, several algorithms exist that attempt to unify the two approaches. For example, Affine Invariant Langevin Dynamics

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(ALDI) is an ensemble sampler that encodes affine invariance into a coupled system of SDEs, each of which governs the evolution of a particle according to a preconditioned version of overdamped Langevin dynamics (Garbuno-Inigo et al., 2020). The entire system of SDEs is coupled by the ensemble's sample covariance, which appears as a preconditioner of the drift term in each individual particle's dynamics. The performance of all affine invariant ensemble samplers still depends on the initialization of the ensemble, and a poorly initialized ensemble can lead to degraded performance for ill-conditioned distributions (Garbuno-Inigo et al., 2020).

In the HMC literature, a linear preconditioner is typically called the *mass matrix* (Betancourt & Stein, 2011; Betancourt, 2018), and selecting a good mass matrix is of paramount importance in modern implementations of HMC such as Stan (Carpenter et al., 2017). In such software packages, one often estimates the mass matrix from samples during a warmup period. Although it may take a long time to sample along *all* eigendirections specified by the covariance of the target distribution, the mass matrix estimation approach implemented by Stan and other MCMC packages can still work quite well in practice. At least heuristically, the success of this method appears to be explained by its ability to mix along different eigendirections at different timescales. The directions corresponding to the covariance's smallest eigenvalues are explored first (with a very small step size), and during the warmup period, the mass matrix estimate is then updated to accurately reflect the true covariance along larger and larger length scales. In practice, this estimation takes place over a handful of warmup stages in which the mass matrix is updated several times (Carpenter et al., 2017). Although we are inspired by this multistage approach and view it as a strong practical benchmark, to the best of our knowledge there is no theoretical analysis available that captures its advantages.

In this work, we aim to provide a strong theoretical foundation for mass matrix estimation and a route towards logarithmic condition number dependence for log-concave sampling, at least for distributions that are amenable to linear preconditioning. With an eye towards sampling efficiently from distributions from Bayesian posteriors to lattice field theories, we develop: (i) a natural condition (Equation 2) that encodes the idea that, in some (a priori unknown) coordinate system, the target distribution is well-conditioned, (ii) an implementable algorithm (Algorithm 1) with logarithmic condition number dependence for mass matrix estimation and sampling in the log-concave setting, and (iii) a detailed theoretical framework (Theorem 3.2) for analyzing our algorithm.

1.1 Contributions

Suppose that π is a distribution on $\mathbb{R}^d$ that admits a density $\pi(x) \propto \exp(-U(x))$ with respect to the Lebesgue measure, where $U : \mathbb{R}^d \to \mathbb{R}$ is a C^2 potential. We begin with the standard assumption that the target distribution is μ-strongly log concave and L-smooth, in the sense that

$$\mu I \preceq \nabla^2 U(x) \preceq LI, \quad \forall x \in \mathbb{R}^d \tag{1}$$

for some $L > \mu > 0$. Even in this well-studied setting, the condition number $\kappa = L/\mu$ may be extremely large, resulting in a computational bottleneck. However, it may be the case that

$$\alpha B \preceq \nabla^2 U(x) \preceq B, \quad \forall x \in \mathbb{R}^d \tag{2}$$

for some positive definite matrix $B \in \mathbb{R}^{d\times d}$ and a parameter $\alpha \in (0, 1)$. Note that any distribution that satisfies Equation 1 also satisfies Equation 2 with $B = LI$ and $\alpha = 1/\kappa$. However, as we shall see in several examples in Section 4, it is possible for κ to be extremely large but for Equation 2 to hold with $\alpha = \Theta(1)$ for some matrix $B \succ 0$. This discrepancy can arise naturally when a well-conditioned distribution is transformed by an ill-conditioned affine mapping, a setting that affine invariant samplers are designed to address.

After transforming to the coordinate system specified by the matrix B in Equation 2, to which we do *not* assume access in our algorithms, the issue of a large condition number may be partially or fully alleviated. In particular, we provide an algorithm that, without assuming the structure of the matrix B or the value of the parameter α, allows for efficient sampling from distributions whose potential satisfy Equation 2. On top of assuming that Equation 1 and Equation 2 hold for the target distribution π, we also assume access to $x^* := \arg\min_{x\in\mathbb{R}^d} U(x)$, the global minimizer of the target potential. Note that this minimizer is unique due to our strong convexity assumption.

We then define a sequence of perturbed potentials by

$$U_k(x) := U(x) + \frac{\gamma^k L}{2}\|x - x^*\|_2^2, \quad k \in \mathbb{N}, \tag{3}$$

where $\gamma \in (0, 1)$ is a parameter chosen by the user. By construction, x^* is also the unique global minimizer of U_k for each $k \in \mathbb{N}$. At each stage of Gaussian cooling, we aim to sample from the distribution $\pi_k(x) \propto \exp(-U_k(x))$, whose potential approaches that of the target distribution as cooling progresses. Although U_k may still specify a distribution whose condition number is unacceptably large, our scheme uses the sample covariance from each stage of cooling to precondition the next stage. We ensure that, at each stage, the resulting preconditioned distribution $\widetilde{\pi}_k$ is well-conditioned. Once k is sufficiently large, the sample covariance from stage k of Gaussian cooling provides an effective preconditioner for the target potential itself.

Our method does *not* depend on the choice of underlying sampling algorithm. In particular, at each stage, the sampler that produces the samples used to estimate the sample covariance can be drawn from any one of the options described in Appendix B. In Section 3, we first present our theoretical guarantees in a sampler-agnostic manner. Later, we use two explicit choices of sampling algorithm: first-order rejection sampling (FORS) (Chen et al., 2026c) yields the best possible query complexity guarantees at the time of writing, while underdamped Langevin Monte Carlo (ULMC) (Altschuler & Chewi, 2023) is straightforward to implement and hence our choice of sampler in the experiments in Section 4. With this in mind, we may present an informal version of our query complexity result based on the first-order query complexity afforded by FORS.

Theorem 1.1 (Informal version of Corollary 3.1)**.** *Suppose that $U : \mathbb{R}^d \to \mathbb{R}$ satisfies Equation 1 and Equation 2. Then, with high probability, Algorithm 1 yields a preconditioner $\widetilde{\Sigma}$ for the Hessian $\nabla^2 U(x)$ of the target potential that satisfies*

$$\frac{\alpha}{4} I \preceq \widetilde{\Sigma}^{1/2} \nabla^2 U(x) \widetilde{\Sigma}^{1/2} \preceq \frac{2}{\alpha} I, \quad \forall x \in \mathbb{R}^d \tag{4}$$

with $\tilde{\Theta}(d^{4/3} \log(\kappa))$ queries to U and its gradient. In particular, the distribution specified by the potential $\widetilde{U}(x) = U(\widetilde{\Sigma}^{1/2} x)$ has condition number $O(1/\alpha^2)$.

We often anticipate $\alpha = \Theta(1)$ for applications of interest, meaning that Algorithm 1 provides an efficient pathway to sampling from a well-conditioned distribution. Moreover, when the target distribution has translation-invariant covariance, the dimension dependence of the query complexity presented in Theorem 1.1 can be improved even further. In Section 4 we consider sampling from a lattice ϕ^4 field theory, a natural example where translation invariance holds. This application is motivated by recent numerical efforts in mass matrix estimation for more complicated lattice field theories (Lunts et al., 2023; Patel et al., 2024; 2025). In the translation-invariant setting, the query complexity in Theorem 1.1 can be further reduced to $\widetilde{\Theta}(d^{1/3} \log(\kappa))$.

1.2 Related Work

Gaussian Cooling, Volume Estimation, and Log-Concave Sampling. Our algorithm is partially inspired by Gaussian cooling (Cousins & Vempala, 2018), an algorithm for estimating the volume of convex bodies. The algorithm samples from a series of distributions, beginning with a sharply concentrated Gaussian that "cools" according to schedule that gradually increases the variance of the Gaussian until the resulting samples are approximately uniform from the target convex body. Although our algorithm is conceptually similar, we borrow little from the analysis and specific algorithm from Cousins & Vempala (2018). Other works that use the notion of Gaussian cooling to analyze the complexity of volume estimation and log-concave sampling include Lee & Vempala (2018) and Kook & Vempala (2024). In a similar vein, our method builds upon a large body of recent work aimed at characterizing the query complexity of (strongly) log-concave sampling, both with respect to the underlying dimension d and the condition number κ of the target distribution. We provide a brief, non-exhaustive summary of some of these results in Appendix B, split into two categories: low and high accuracy samplers. For a more in-depth survey on log-concave sampling algorithms and their query complexity, see Chewi (2023).

Affine Invariant Sampling. In contrast with the setting of log-concave sampling, the literature on affine invariant sampling places less emphasis on first-order query complexity as the guiding standard for algorithmic performance. In fact, for the "stretch move" and other affine invariant ensemble samplers introduced in Goodman & Weare (2010), the question of ergodicity remains open (Chen, 2025). This is not the case for ALDI, for which both non-degeneracy of the empirical covariance

used to evolve the ensemble and ergodicity have been established in Garbuno-Inigo et al. (2020) and for various adjusted versions of ALDI as in Sprungk et al. (2025). A number of techniques also attempt to circumvent condition number dependence by instead discretizing a diffusion process on an appropriate Riemannian manifold (Girolami & Calderhead, 2011). Riemannian HMC, for example, yields a state-of-the-art algorithm for polytope volume computation (Lee & Vempala, 2018) and other constrained sampling problems (Kook et al., 2023). Similarly, the mirror Langevin algorithm presents an alternative route towards condition number-free sampling (Ahn & Chewi, 2021; Gatmiry & Vempala, 2022).

Learning Linear Preconditioners. Several recent results address the question of efficiently learning a preconditioner *prior* to running a high-accuracy sampler. For example, Hird & Livingstone (2025) establishes that there exist distributions with arbitrarily large condition number κ such that no linear preconditioner can improve the conditioning of the sampling problem, akin to the example we provide in Appendix J. Hird & Livingstone (2025) then characterizes distributions for which preconditioning with a diagonal preconditioner, the Fisher information, or the sample covariance can reduce the condition number of the underlying problem. On the other hand, the authors show that preconditioning with a diagonal covariance can sometimes *increase* the condition number of the problem, even for Gaussian targets. The same group of authors quantify the complexity of estimating the sample covariance and Fisher information directly from MCMC samples in Hird et al. (2026).

This work is perhaps most conceptually similar to our method, and at a technical level, some of the tools used in Hird et al. (2026) are quite similar to the tools that we use in Appendix F. However, our algorithmic approach is fundamentally different from that of Hird et al. (2026), which focuses on learning preconditioners directly from samples during warm-up. Moreover, the results in Hird et al. (2026) require polynomial dependence on the condition number of the underlying distribution in order to estimate a good preconditioner. For the class of distributions that satisfy Equation 2 with $\alpha = \Theta(1)$, our method reduces this dependence to logarithmic rather than polynomial.

Sampling from Lattice Field Theories. Our work is partially motivated by the goal of efficiently sampling from lattice field theories, a difficult problem with consequences in condensed matter and high-energy physics. This direction has been explored in some detail in Lunts et al. (2023), which uses HMC to sample from the $O(3)$ spin-fermion model. Although the action for this model and the exact application of HMC are quite complicated, the authors crucially leverage translation invariance to accelerate HMC, estimating the mass matrix in Fourier space. We draw upon this idea in our application to lattice ϕ^4 field theory in Section 4.2. A subset of the same group of authors continued this line of work in Patel et al. (2024) and Patel et al. (2025), utilizing HMC in a similar manner to investigate other spin-boson and spin-fermion models. We emphasize that while our theoretical guarantees are limited to the setting of strong log-concavity, we hope that our algorithm itself may see practical use in more complicated settings such as the above.

2 Gaussian Cooling Algorithm

In this section, we describe our primary algorithmic contribution in detail. Although the structure of the Gaussian cooling algorithm is independent of the choice of sampler used at each stage, its query complexity is dependent on the choice of sampler for each stage of Gaussian cooling. Before presenting two concrete choices for the sampler, we present a general analysis that holds for a broad class of log-concave sampling algorithms, including both low-accuracy samplers (Table 1) and high-accuracy samplers (Table 2).

Given an evaluation and gradient oracle for the target potential $U : \mathbb{R}^d \to \mathbb{R}$, one can form an evaluation and gradient oracle for any potential of the form in Equation 3. In fact, if $M \succ 0$ is any positive definite matrix, then we can also form an evaluation and gradient oracle for the augmented potential $x \mapsto U_k(M^{1/2}x)$. At each stage of Gaussian cooling, we will query such an oracle, where M is the preconditioner formed from samples at the previous stage of cooling. Below, $\widehat{S}_n(x_1, \dots, x_n)$ is the (biased) sample covariance formed from n independent samples; see Appendix A. Finally, we write $\mathcal{A}(V, x^*, N, n)$ to denote a call to a generic log-concave sampling algorithm $\mathcal{A}$ that takes as input a potential V, a global minimizer x^*, the number of steps N to take before returning a sample, and the total number of samples n to produce in parallel.

Algorithm 1 Preconditioning via Gaussian Cooling

Require: Sampling algorithm $\mathcal{A}$, smoothness parameter $L > 0$, global minimizer $x^* \in \mathbb{R}^d$, potential $U : \mathbb{R}^d \to \mathbb{R}$ with evaluation and gradient oracle, number of Gaussian cooling stages K, number of samples per stage n, number of sampler steps per sample N.

Ensure: A preconditioner $\widetilde{\Sigma}_n^{(K)}$ satisfying Equation 6 with high probability.

$\widetilde{\Sigma}_n^{(-1)} \leftarrow L^{-1} I$
for $k = 0, \dots, K$ **do**
  Form $\widetilde{U}_k(x) = U_k((\widetilde{\Sigma}_n^{(k-1)})^{1/2} x)$
  $x_1^{(k)}, \dots, x_n^{(k)} \leftarrow \mathcal{A}\left(\widetilde{U}_k, (\widetilde{\Sigma}_n^{(k-1)})^{-1/2} x^*, N, n\right)$
  $\widehat{\Sigma}_n^{(k)} \leftarrow \widehat{S}_n(x_1^{(k)}, \dots, x_n^{(k)})$
  $\widetilde{\Sigma}_n^{(k)} \leftarrow (\widetilde{\Sigma}_n^{(k-1)})^{1/2} \widehat{\Sigma}_n^{(k)} (\widetilde{\Sigma}_n^{(k-1)})^{1/2}$
end for
return $\widetilde{\Sigma}_n^{(K)}$

As written above, Algorithm 1 returns a preconditioner $\widetilde{\Sigma}_n^{(K)}$ for the target distribution in a sense made precise below. If the desired use case is mass matrix tuning in HMC, for example, this output is all that is needed. Of course, if the user's end goal is a high accuracy sample from the target distribution, one additional sampling step is needed.

Recall that we successively sample from distributions $\pi(x) \propto \exp(-U_k(x))$, with potentials U_k defined in Equation 3. Throughout, we let $\pi_k(x) \propto \exp(-U_k(x))$ be the corresponding distribution and take $\Sigma^{(k)} := \mathrm{Cov}(\pi_k)$. We first reduce the analysis of Algorithm 1 to establishing an appropriate concentration guarantee for the preconditioner produced at each stage of cooling.

Proposition 2.1. *Suppose that, at each stage of Algorithm 1, the sampling algorithm $\mathcal{A}$ outputs a preconditioner $\widetilde{\Sigma}_n^{(k-1)}$ satisfying*

$$(1+\delta)^{-1}\Sigma^{(k-1)} \preceq \widetilde{\Sigma}_n^{(k-1)} \preceq (1+\delta)\Sigma^{(k-1)}, \tag{5}$$

for some $\delta > 0$. Then, after $K = O\left(\frac{\log(\kappa)}{\log(1/\gamma)}\right)$ stages of Gaussian cooling, Algorithm 1 will yield a preconditioner that satisfies

$$\frac{\alpha}{(1+\alpha)(1+\delta)} I \preceq (\widetilde{\Sigma}_n^{(K)})^{1/2} \nabla^2 U(x) (\widetilde{\Sigma}_n^{(K)})^{1/2} \preceq \frac{1+\delta}{\alpha} I. \tag{6}$$

Due to Proposition 2.1, whose proof is in Appendix E, our task reduces to showing that $\mathcal{A}$ can, with high probability, ensure that Equation 5 holds at every stage of Gaussian cooling.

3 Query Complexity of Gaussian Cooling

We place the following assumption on the sampling algorithm $\mathcal{A}$. For specific choices of $\mathcal{A}$ such as ULMC or FORS, the parameters in the following assumption are made explicit.

Assumption 3.1. *We say $\mathcal{A}$ satisfies a low accuracy guarantee if, given a target distribution $\pi(x) \propto \exp(-U(x))$ satisfying Equation 1 and with global minimizer x^*, there exist constants $c_1, c_2, c_3 > 0$ such that for any $\varepsilon > 0$, the sampling algorithm $\mathcal{A} = \mathcal{A}(U, x^*, N, 1)$ can make at most*

$$N = \widetilde{O}\left(\kappa^{c_1} d^{c_2} \varepsilon^{-c_3}\right) \tag{7}$$

queries to U and its gradient and output a sample with distribution $\widehat{\pi}_N$ that satisfies $\mathcal{R}_2(\widehat{\pi}_N \| \pi) \leq \varepsilon^2$, where $\mathcal{R}_2(\cdot\|\cdot)$ is the Rényi 2-divergence defined in Appendix A. If we can instead take

$$N = \widetilde{O}\left(\kappa^{c_1} d^{c_2} (\log(\varepsilon^{-1}))^{c_3} + \log(1/\eta)\right), \tag{8}$$

to obtain the same guarantee in Rényi-2 divergence with probability at least $1 - \eta$, we say that $\mathcal{A}$ satisfies a high accuracy guarantee.

At each stage of Gaussian cooling, a sampling algorithm that satisfies Assumption 3.1 maintains Equation 5 across all stages of cooling, with high probability.

Theorem 3.2. *Fix $p \in (0,1)$ and suppose that $\mathcal{A}$ satisfies Assumption 3.1. Take K as in Proposition 2.1 and*

$$n = \Theta\left(\frac{d + \log(K/p)}{\widetilde{\alpha}^4 \delta^2}\right)$$

samples per stage, where $\widetilde{\alpha} := \min\{\alpha, \gamma\}$. Then, with probability at least $1-p$, Algorithm 1 returns a preconditioner $\widetilde{\Sigma}_n^{(K)}$ satisfying Equation 6, and each of the nK calls to $\mathcal{A}$ uses at most

$$N = \begin{cases} \widetilde{\Theta}\left((\widetilde{\alpha})^{-2c_1}\, d^{c_2} \delta^{-c_3}\right), & \mathcal{A} \text{ is a low accuracy sampler}, \\ \widetilde{\Theta}\left((\widetilde{\alpha})^{-2c_1}\, d^{c_2} (\log(\delta^{-1}))^{c_3}\right) + O\left(\log(nK/p)\right), & \mathcal{A} \text{ is a high accuracy sampler} \end{cases}$$

queries to U and its gradient oracle.

The proof of this theorem leverages the method developed in Appendix F to establish that Equation 5 holds at each stage of cooling with the given choice of parameters. Once this is established, Proposition 2.1 yields the desired preconditioning guarantee, and a union bound ensures that the failure probability of Gaussian cooling across all K stages remains controlled.

3.1 TRANSLATION-INVARIANT CASE

In many applications of interest, the target covariance will be *translation-invariant*, in the sense that Σ has block circulant structure. This is the case in the lattice ϕ^4 model that we consider in Section 4 and for more complicated lattice field theories (Lunts et al., 2023). Leaving the exact details of the lattice ϕ^4 model to the next section, we simply note that if $\Sigma \in \mathbb{R}^{d\times d}$ is block circulant with $\sqrt{d}\times\sqrt{d}$ circulant blocks, then it is exactly diagonalized by a discrete Fourier transform (DFT), which we denote by $\mathcal{F} \in \mathbb{C}^{d\times d}$. In particular, there exists a vector $\sigma \in \mathbb{R}^d$ such that $\Sigma = \mathcal{F}\,\mathrm{diag}(\sigma)\mathcal{F}^*$, where $\mathcal{F}^*$ denotes the adjoint of the discrete Fourier transform. Thus, in the translation-invariant case, our task is reduced from estimating the covariance matrix to estimating the vector $\sigma \in \mathbb{R}^d$. In this setting, we replace Line 5 of Algorithm 1 with

$$\widehat{\sigma}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{F}(x_i^{(k)} - \widehat{m}_n(x_1^{(k)}, \ldots, x_n^{(k)})) \odot \overline{\mathcal{F}(x_i^{(k)} - \widehat{m}_n(x_1^{(k)}, \ldots, x_n^{(k)}))}, \tag{9}$$

$$\widehat{\Sigma}_n^{(k)} = \mathcal{F}\,\mathrm{diag}(\widehat{\sigma}^{(k)})\mathcal{F}^*, \tag{10}$$

where $\widehat{m}_n(x_1, \ldots, x_n)$ is the sample mean. With this change in place, Corollary F.3 then yields the following query complexity guarantee.

Theorem 3.3. *Taking n as in Theorem 3.2, with $d + \log(K/p)$ replaced by $\log(dK/p)$, and N and K as above, Algorithm 1 returns a preconditioner $\widetilde{\Sigma}_n^{(K)}$ for the target distribution satisfying Equation 6 with probability at least $1 - p$.*

The above result follows from essentially the same reasoning as the argument preceding Theorem 3.2. In particular, the total number of Gaussian cooling stages k and the number of first-order queries per stage N remain unchanged. In fact, the only change in the analysis is the fact that, as in Corollary F.3, we can replace all instances of $d^{c_2}(d + \log(1/p))$ that arise in the general case with $d^{c_2}\log(d/p)$. The factor d^{c_2} comes from the number of steps (per sample) at each stage of cooling required by $\mathcal{A}$, while the logarithmic factor comes from the sample complexity estimate in Corollary F.3.

Choosing a sampler. To demonstrate the utility of Gaussian cooling, we now present explicit query complexity bounds for two samplers: ULMC and FORS. The necessary background on FORS is in Appendix C, while the corresponding background on ULMC can be found in Appendix D. In practice, ULMC is simple to implement and parallelize over many chains, making it an attractive choice for estimating preconditioners for high-dimensional distributions. In the numerical experiments in Section 4, we therefore use ULMC. However, very recent progress on the optimal dimension and condition-number dependence of log-concave sampling via FORS enables us to present even better theoretical guarantees. The FORS algorithm is equipped with the *high accuracy* mixing guarantee in Theorem C.2, so Theorem 3.2 yields the following corollary.

Corollary 3.1. *FORS (Algorithm 2) is a high accuracy sampler satisfying Assumption 3.1 with constants $c_1 = 2/3$, $c_2 = 1/3$, and $c_3 = 5/3$. Consequently, Gaussian cooling with FORS taking the place of $\mathcal{A}$ returns a preconditioner satisfying Equation 6 with probability at least $1 - p$ using*

$$NnK = \begin{cases} \widetilde{O}\left(\frac{d^{1/3}[d+\log(1/p)]\log(\kappa)}{\widetilde{\alpha}^{16/3}\delta^2\log(1/\gamma)}\right) & \textit{general target}, \\ \widetilde{O}\left(\frac{d^{1/3}\log(1/p)\log(\kappa)}{\widetilde{\alpha}^{16/3}\delta^2\log(1/\gamma)}\right) & \textit{translation-invariant target} \end{cases}$$

queries to U and its gradient.

For ULMC, Theorem D.2 presents a low accuracy guarantee, allowing us to apply Theorem 3.2 with parameters $c_1 = 3/2$, $c_2 = 1/2$, and $c_3 = 1$. The resulting query complexity is $\widetilde{O}(d^{3/2}\log(\kappa))$ in the general case and $\widetilde{O}(d^{1/2}\log(\kappa))$ in the translation-invariant case; see Corollary D.1 for a full statement of the ULMC analogue of Corollary 3.1.

4 Examples and Applications

In this section, we discuss two examples to which our framework naturally applies. In Section 4.1, we discuss a generic problem in which a well-conditioned problem is transformed under a potentially ill-conditioned, unknown affine transformation. This setting can arise, for example, in physical problems when the underlying coordinate system changes. In Section 4.2, we discuss applications to a simple discretized field theory, the lattice ϕ^4 model. Throughout, we measure the success of various preconditioner estimation strategies in terms of their relative condition number. In particular, if Σ is the covariance of the target distribution and $\widehat{\Sigma}$ is an estimate, then we define $\kappa_{\mathrm{rel}}(\widehat{\Sigma}, \Sigma) := \kappa(\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2})$. The closer that this quantity is to one, the better the estimate of the true covariance. When it is clear from context, we write κ_{rel} instead of $\kappa_{\mathrm{rel}}(\widehat{\Sigma}, \Sigma)$.

We compare the performance of Algorithm 1, and in the translation-invariant setting, the performance of Algorithm 4, to two baselines. First, we compare to a "raw" ULMC baseline. Given the same number of queries as Algorithm 1, we use Algorithm 3 to produce approximate samples from the target distribution, without any preconditioning. We then estimate the target covariance from the resulting samples. In the translation-invariant setting, estimating the dense empirical covariance is impractical, so we compute the unpreconditioned baseline in Fourier space using Equation 9 when the dimension of the underlying distribution is greater than 400. We refer to the translation-invariant version of this baseline as *translation-averaged ULMC*.

We also compare to a stagewise empirical preconditioner estimate akin to the procedure that MCMC packages such as Carpenter et al. (2017) use in practice. Given the same number of chains n and samples per chain N, this procedure carries out K stages of preconditioner estimation. At the end of stage k, the sample covariance $\widetilde{\Sigma}_n^{(k)}$ is computed and then used as a preconditioner. In all cases, we ensure that any methods that we compare receive the same number of first-order queries. For reference preconditioners (i.e., no adaptive preconditioning or cooling takes place), this simply means that we run n chains for KN steps each.

4.1 Well-Conditioned Problems Under an Unknown Affine Transformation

The most straightforward application of the algorithm developed above is to well-conditioned problems transformed by unknown affine maps. Suppose that $\pi(x) \propto \exp(-U(x))$, where the potential $U : \mathbb{R}^d \to \mathbb{R}$ satisfies Equation 1 with condition number $\kappa = L/\mu = O(1)$. Now, given an arbitrary positive definite matrix $B \succ 0$, we consider the transformed distribution $\tilde{\pi}(x) \propto \exp(-U(L^{-1/2}B^{1/2}x))$. The potential $\widetilde{U}(x) := U(L^{-1/2}B^{1/2}x)$ then satisfies $\frac{\mu}{L}B \preceq \nabla^2\widetilde{U}(x) \preceq B$ for all $x \in \mathbb{R}^d$. Hence, for $\alpha = \mu/L = \Theta(1)$, the transformed distribution $\tilde{\pi}$ satisfies Equation 2, with the geometry specified by the transformation B.

In the following numerical experiments, we take π to be a well-conditioned Gaussian measure on $\mathbb{R}^d$. Then, with only black-box access to the quadratic potential U and its gradient, Algorithm 1 provides an efficient method for estimating the covariance of the transformed distribution, with no dependence on the conditioning of the transformation B. In this sense, our method allows one to sample efficiently from well-conditioned distributions that have been corrupted by an unknown and

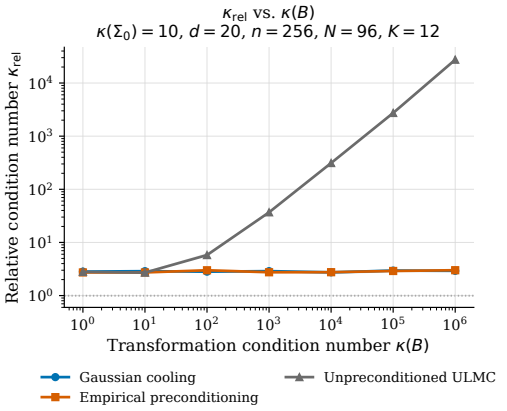


Figure 1: As the condition number $\kappa(B)$ of the transformation applied to the well-conditioned distribution increases, empirical covariance from unpreconditioned ULMC degrades. Gaussian cooling as specified in Algorithm 1 and stagewise empirical preconditioning perform comparably.

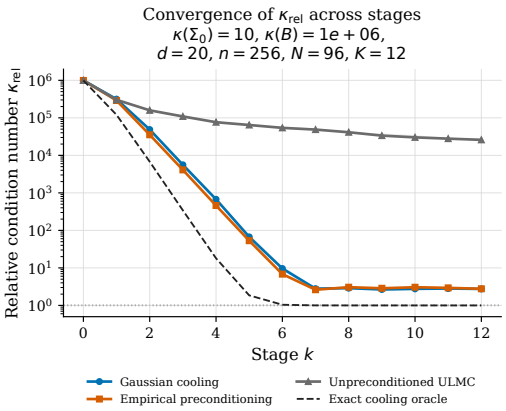


Figure 2: When $d = 20$ and $\kappa(B) = 10^6$, both Gaussian cooling and stagewise empirical preconditioning can attain a constant relative condition number within approximately seven stages. The full empirical covariance fails to push the relative condition number below 10^4.

potentially ill-conditioned affine transformation. In Figure 1, we demonstrate that the quality of the preconditioner formed by Gaussian cooling does not degrade as $\kappa(B)$ increases. Similarly, Figure 2 shows that for $\kappa(B) = 10^6$, Gaussian cooling is able to produce a preconditioner whose relative condition number is $\Theta(1)$ in approximately seven iterations.

4.2 Lattice ϕ^4 Theory

The lattice ϕ^4 model, a standard discretization of the classical ϕ^4 field theory (Bauerschmidt et al., 2019, Chapter 1.6), provides a more specialized testbed for our algorithm. Its quartic term results in a large condition number, even after truncating the quartic term past some radius to ensure smoothness. In practice, one typically estimates the covariance (or a diagonal approximation of it) from samples produced during the warm-up period of an HMC chain, as in Lunts et al. (2023).

Let $\Lambda = \{(x_1, x_2) \in \mathbb{Z}^2 : 0 \le x_1, x_2 \le d - 1\}$ be the periodic square lattice of side length d, and write $x \sim y$ if x and y are neighbors, with periodic boundary. To each pair of sites we associate a symmetric coupling constant $\beta_{xy} = \beta_{yx} \ge 0$, with $\beta_{xy} = 0$ unless $x \sim y$, and following Bauerschmidt et al. (2019), we define the weighted Laplacian

$$(\Delta_\beta \phi)_x := \sum_{y \in \Lambda} \beta_{xy}(\phi_x - \phi_y)$$

for each scalar field $\phi : \Lambda \to \mathbb{R}$. The Hamiltonian

$$H(\phi) := \frac{1}{2}\langle \phi, (\Delta_\beta + mI)\phi \rangle + \frac{\lambda}{4} \sum_{x \in \Lambda} \phi_x^4 \tag{11}$$

induces a distribution $\pi(\phi) \propto \exp(-H(\phi))$ over $\mathbb{R}^{d^2}$; see Appendix G for further background. This model has translation-invariant covariance, so Theorem 3.3 applies.

We compare translation-invariant Gaussian cooling (Algorithm 4) to translation-invariant empirical preconditioning, which estimates the covariance in stages via Equation 9. Figure 3 shows that translation-invariant Gaussian cooling provides an effective preconditioner on lattices of sidelength up to $d = 512$ and is comparable to empirical preconditioning across all regimes. Note that for a lattice with sidelength d, the target distribution is a probability measure over $\mathbb{R}^{d^2}$, so our largest experiments take the target dimension to be $d^2 = 262,144$. In the small mass regime $m = 0.001$, Figure 4 shows that both methods begin to outperform the unpreconditioned chain (with the covariance estimated in Fourier space) given the same number of gradient queries. Figure 5 shows that our method continues to perform well on the true quartic potential, which is not log-smooth.

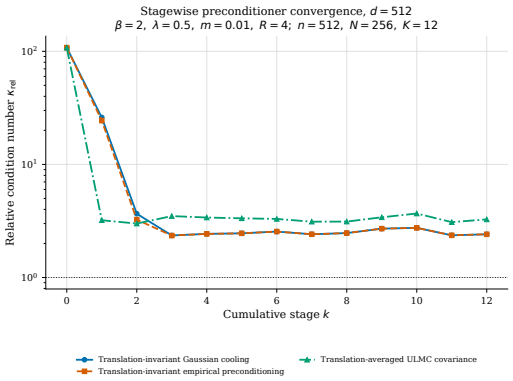


Figure 3: On lattices of sidelength up to $d = 512$ with mass $m = 0.01$, both translation-invariant Gaussian cooling and translation-invariant empirical preconditioning obtain a constant relative condition number within several steps.

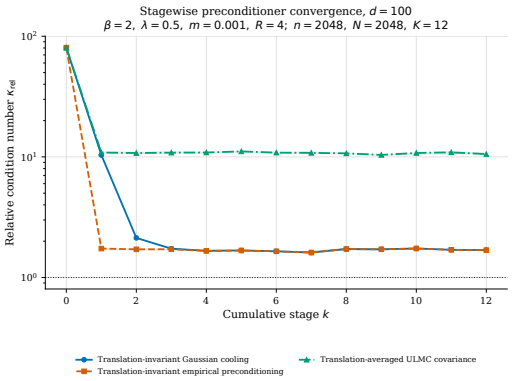


Figure 4: For sidelength $d = 100$, decreasing the mass to $m = 0.001$ begins to distinguish Gaussian cooling and empirical preconditioning from translation-averaged ULMC.

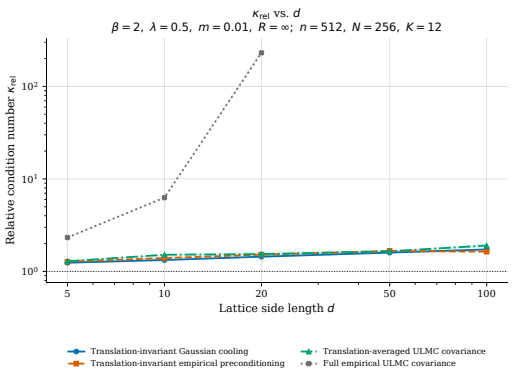


Figure 5: For the genuinely quartic (untruncated) potential, all three translation-invariant methods perform comparably up to $d = 100$, while dense ULMC fails for $d = 5, 10, 20$ with n and N held fixed.

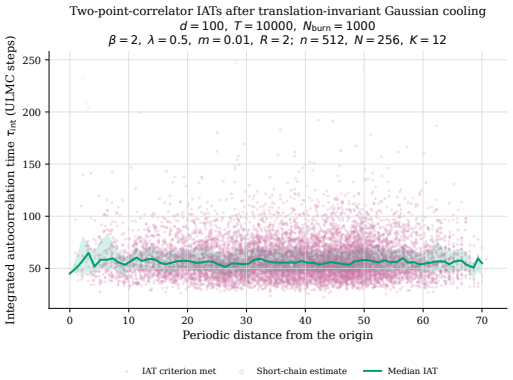


Figure 6: With the translation-invariant Gaussian cooling preconditioner at $d = 100$, the two-point correlators between $(0, 0)$ and all other sites have median IAT $\tau_{\text{int}} \approx 50$ for a ULMC chain of length $10,000$.

In Figure 6, we use `emcee` (Foreman-Mackey et al., 2013) to compute the integrated autocorrelation time (IAT) τ_{int} of the two-point correlators between the origin and all other sites. The convergence criterion $T \geq 50\tau_{\text{int}}$ is met for nearly all sites with chain length $T = 10000$ and burn-in $N_{\text{burn}} = 1000$. Appendix G studies the effect of N and n on the performance of Algorithm 1.

5 Conclusion

Gaussian cooling provides a theoretically grounded method for mass matrix estimation across a broad class of sampling problems. The first-order query complexity of our algorithm, as presented in Theorem 3.2, will only continue to improve as the theory of log-concave sampling advances. For instance, Chen et al. (2026a) and Chen et al. (2026b) indicate that one may be able to improve upon the first-order query complexity of FORS, obtaining a high accuracy query complexity of $\widetilde{O}(\sqrt{\kappa}d^{1/4}\text{polylog}(1/\varepsilon))$ to obtain a sample that is within ε distance of a log-concave target in total variation distance. On the practical side, our algorithm provides a compelling, easy-to-implement alternative for preconditioning high-dimensional distributions, matching the performance of widely-used warmup strategies (Carpenter et al., 2017) while permitting rigorous guarantees.

AI USE STATEMENT

In this work, we used generative AI tools for the parts of the implementation of Algorithm 1, with Algorithm 3 as the underlying sampler. For each implementation, detailed scaffolds of the desired algorithms and functionality were provided in the form of pseudocode. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, support qualitative and thematic data analysis, or interpret results. The remaining required disclosure tasks in the ICLR author guide are not applicable to this work.

In addition to the above tasks, we used generative AI tools to find and summarize relevant literature, brainstorm ideas for mathematical proofs, and proofread our final paper for readability and minor errors. All code and figures produced using generative AI tools were checked by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix H outlines our experimental setup, including all relevant implementation details. All code used to produce the figures in Section 4 can be found in the uploaded supplementary material.

The majority of the mathematical content of our work can be found in the remainder of the appendix, structured as follows. First, Appendix A outlines the notation and technical assumptions that underlie our theoretical contributions. Then, Appendix B gives a more in-depth overview of the first-order query complexity of log-concave sampling. After this, Appendices C and D describe the two samplers, FORS and ULMC, that we use for our theoretical results and numerical experiments in Section 3 and Section 4, respectively. Once this background information has been provided, Appendix E provides rigorous proofs of Theorem 3.2 and its translation-invariant counterpart. The bulk of the technical work regarding concentration of the sample covariance and the corresponding preconditioning guarantee that this concentration provides is then given in Appendix F. Appendix G gives additional background and experiments on the lattice ϕ^4 model. We conclude with Appendix I, describing various extensions and variants of Algorithm 1 that may be more suitable for various applications, and Appendix J, a simple example of a distribution that satisfies Assumptions A.4– A.6 but has $\alpha = O(1/\kappa)$, where κ is the condition number of the target distribution.

REFERENCES

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vh7qBSDZW3G.

Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 2169–2176, 2023. doi: 10.1109/FOCS57990.2023.00134.

Jason M. Altschuler, Sinho Chewi, and Matthew S. Zhang. Shifted composition iv: Toward ballistic acceleration for log-concave sampling, 2026. URL https://arxiv.org/abs/2506.23062.

Roland Bauerschmidt, David C. Brydges, and Gordon Slade. *Introduction to a Renormalisation Group Method*. Springer Singapore, 2019. ISBN 9789813295933. doi: 10.1007/978-981-32-9593-3. URL http://dx.doi.org/10.1007/978-981-32-9593-3.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo, 2018. URL https://arxiv.org/abs/1701.02434.

Michael Betancourt and Leo C. Stein. The geometry of hamiltonian monte carlo, 2011. URL https://arxiv.org/abs/1112.4118.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic

programming language. *Journal of Statistical Software*, 76(1), 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL http://dx.doi.org/10.18637/jss.v076.i01.

Fan Chen, Sinho Chewi, Jianfeng Lu, and Matthew S Zhang. Accelerated high-accuracy sampling from a warm start via the proximal bouncy particle sampler, 2026a. URL https://arxiv.org/abs/2609.06905.

Fan Chen, Sinho Chewi, Jianfeng Lu, and Matthew S Zhang. Smoothed picard hamiltonian monte carlo, 2026b. URL https://arxiv.org/abs/2609.06906.

Fan Chen, Sinho Chewi, Alexander Rakhlin, and Matthew S. Zhang. Exact simulation of diffusions and improved algorithms for log-concave sampling, 2026c. URL https://arxiv.org/abs/2608.05022.

Yifan Chen. New affine invariant ensemble samplers and their dimensional scaling, 2025. URL https://arxiv.org/abs/2505.02987.

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

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 (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1260–1300. PMLR, 2021. URL https://proceedings.mlr.press/v134/chewi21a.html.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1–2. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/chewi22a.html.

Ben Cousins and Santosh Vempala. Gaussian Cooling and $O^*(n^3)$ Algorithms for Volume and Gaussian Volume. *SIAM Journal on Computing*, 47(3):1237–1273, 2018. doi: 10.1137/15M1054250. URL https://doi.org/10.1137/15M1054250.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *J. Mach. Learn. Res.*, 20(1):2666–2711, January 2019. ISSN 1532-4435.

Yam Eitan. The centered convex body whose marginals have the heaviest tails. *Studia Mathematica*, 274(3):201–215, 2024. ISSN 1730-6337. doi: 10.4064/sm211027-20-11. URL http://dx.doi.org/10.4064/sm211027-20-11.

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1473–1521. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/fan23a.html.

Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The mcmc hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, March 2013. ISSN 1538-3873. doi: 10.1086/670067. URL http://dx.doi.org/10.1086/670067.

Alfredo Garbuno-Inigo, Nikolas Nüsken, and Sebastian Reich. Affine invariant interacting langevin dynamics for bayesian inference. *SIAM Journal on Applied Dynamical Systems*, 19 (3):1633–1658, January 2020. ISSN 1536-0040. doi: 10.1137/19m1304891. URL http://dx.doi.org/10.1137/19M1304891.

Khashayar Gatmiry and Santosh S. Vempala. Convergence of the riemannian langevin algorithm, 2022. URL https://arxiv.org/abs/2204.10818.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2): 123–214, March 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2010.00765.x. URL http://dx.doi.org/10.1111/j.1467-9868.2010.00765.x.

Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1):65–80, January 2010. ISSN 1559-3940. doi: 10.2140/camcos.2010.5.65. URL http://dx.doi.org/10.2140/camcos.2010.5.65.

Max Hird and Samuel Livingstone. Quantifying the effectiveness of linear preconditioning in markov chain monte carlo. *Journal of Machine Learning Research*, 26(119):1–51, 2025. URL http://jmlr.org/papers/v26/23-1633.html.

Max Hird, Florian Maire, and Jeffrey Negrea. A non-asymptotic analysis for learning and applying a preconditioner in mcmc, 2026. URL https://arxiv.org/abs/2602.10714.

Yunbum Kook and Santosh S. Vempala. Gaussian cooling and Dikin walks: The interior-point method for logconcave sampling. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3137–3240. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/kook24b.html.

Yunbum Kook, Yin Tat Lee, Ruoqi Shen, and Santosh Vempala. Condition-number-independent convergence rate of riemannian hamiltonian monte carlo with numerical integrators. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4504–4569. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/kook23a.html.

Yin Tat Lee and Santosh S. Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC '18, pp. 1115–1121. ACM, June 2018. doi: 10.1145/3188745.3188774. URL http://dx.doi.org/10.1145/3188745.3188774.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Lower bounds on metropolized sampling methods for well-conditioned distributions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=napaTaDQ0lY.

Peter Lunts, Michael S. Albergo, and Michael Lindsey. Non-hertz-millis scaling of the antiferromagnetic quantum critical metal via scalable hybrid monte carlo. *Nature Communications*, 14(1), May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-37686-4. URL http://dx.doi.org/10.1038/s41467-023-37686-4.

Daniel Murawski. Comparing moments of real log-concave random variables. *Bernoulli*, 32(3), August 2026. ISSN 1350-7265. doi: 10.3150/25-bej1958. URL http://dx.doi.org/10.3150/25-BEJ1958.

Aavishkar A. Patel, Peter Lunts, and Subir Sachdev. Localization of overdamped bosonic modes and transport in strange metals. *Proceedings of the National Academy of Sciences*, 121(14), March 2024. ISSN 1091-6490. doi: 10.1073/pnas.2402052121. URL http://dx.doi.org/10.1073/pnas.2402052121.

Aavishkar A. Patel, Peter Lunts, and Michael S. Albergo. Strange metals and planckian transport in a gapless phase from spatially random interactions. *Physical Review X*, 15(3), September 2025. ISSN 2160-3308. doi: 10.1103/611k-yxb9. URL http://dx.doi.org/10.1103/611k-yxb9.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, December 2024. ISBN 9781108832908. doi: 10.1017/9781108966351. URL http://dx.doi.org/10.1017/9781108966351.

Björn Sprungk, Simon Weissmann, and Jakob Zech. Metropolis-adjusted interacting particle sampling. *Statistics and Computing*, 35(3), March 2025. ISSN 1573-1375. doi: 10.1007/s11222-025-10595-w. URL http://dx.doi.org/10.1007/s11222-025-10595-w.

P. Stavrakakis and P. Valettas. *On the Geometry of Log-Concave Probability Measures with Bounded Log-Sobolev Constant*, pp. 359–380. Springer New York, 2013. ISBN 9781461464068. doi: 10.1007/978-1-4614-6406-8_17. URL http://dx.doi.org/10.1007/978-1-4614-6406-8_17.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, September 2018. ISBN 9781108415194. doi: 10.1017/9781108231596. URL http://dx.doi.org/10.1017/9781108231596.

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. URL http://jmlr.org/papers/v23/21-1184.html.

Matthew S. Zhang, Jason M. Altschuler, and Sinho Chewi. Algorithmic warm starts for hamiltonian monte carlo, 2026. URL https://arxiv.org/abs/2603.22741.

A Notation and Technical Assumptions

In this section, we provide notation and well-established results that we will use throughout the analysis of our algorithm. Throughout, we generally use the notation $\|\cdot\|$ for two purposes: the Euclidean norm on $\mathbb{R}^d$ and the spectral norm on matrices in $\mathbb{R}^{d\times d}$. Throughout, we also denote $[n] := \{1, \dots, n\}$ for concision.

We begin with some preliminaries on the information-theoretic divergences that will aid in our analysis. Throughout, $\mathcal{P}(\mathbb{R}^d)$ denotes the space of probability measures on $\mathbb{R}^d$ and

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{\mu \in \mathcal{P}(\mathbb{R}^d) : \int \|x\|^2 d\mu(x) < \infty\right\}$$

is the space of probability measures with finite second moment.

Definition 1. *If $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ are two probability measures such that $\mu \ll \nu$, then the χ^2-divergence of μ relative to ν is defined by*

$$\chi^2(\mu\|\nu) := \int_{\mathbb{R}^d} \left(\frac{d\mu}{d\nu} - 1\right)^2 d\nu = \int_{\mathbb{R}^d} \left(\frac{d\mu}{d\nu}\right)^2 d\nu - 1. \tag{12}$$

More generally, the Rényi divergence of order q between μ and ν is defined by

$$\mathcal{R}_q(\mu\|\nu) := \frac{1}{q-1} \log\left(\int_{\mathbb{R}^d} \left(\frac{d\mu}{d\nu}\right)^q d\nu\right). \tag{13}$$

Note that

$$\mathcal{R}_2(\mu\|\nu) = \log(1 + \chi^2(\mu\|\nu)),$$

with both sides infinite if μ is not absolutely continuous with respect to ν.

The χ^2-divergence admits the following concise variational characterization.

Theorem A.1 (Theorem 7.26 and Example 7.4 in Polyanskiy & Wu (2024))**.** *Suppose $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ are such that $\mu \ll \nu$ and that $\mathcal{G}$ denotes the set of measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ such that $g \in L^2(\nu)$ and $\mathrm{Var}_{x\sim\nu}(g(x)) > 0$. Then, we can express*

$$\chi^2(\mu\|\nu) = \sup_{g\in\mathcal{G}} \frac{(\mathbb{E}_{x\sim\mu}[g(x)] - \mathbb{E}_{x\sim\nu}[g(x)])^2}{\mathrm{Var}_{x\sim\nu}(g(x))}.$$

Beyond this point, $\mathbb{E}_\mu[g(x)] := \mathbb{E}_{x\sim\mu}[g(x)]$ for measurable $g : \mathbb{R}^d \to \mathbb{R}$, with identical notation for variances and covariances.

We also aim to relate the condition in Equation 2 to a condition on the covariance matrix of the target distribution π. Using a combination of a Brascamp–Lieb inequality and a Cramér–Rao bound, if U satisfies the condition in Equation 2, then we have that

$$\alpha B \preceq \Sigma^{-1} \preceq B, \tag{14}$$

where Σ is the covariance of the equilibrium measure $\pi(x) \propto \exp(-U(x))$. To obtain this inequality, we recall the Brascamp–Lieb and Cramér–Rao inequalities, respectively.

Theorem A.2 (Theorem 2.2.9 in Chewi (2023))**.** *If $\pi \propto \exp(-V)$, where V is strictly convex in $\mathbb{R}^d$ and twice continuously differentiable, then for all $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Var}_\pi f \le \mathbb{E}_\pi \langle \nabla f, (\nabla^2 V)^{-1} \nabla f\rangle.$$

The Cramér–Rao inequality then provides the accompanying lower bound, albeit in a bit more generality.

Theorem A.3 (Theorem 3.5.8 in Chewi (2023))**.** *If $\pi \propto \exp(-V)$ is a probability measure on $\mathbb{R}^d$, then for any twice continuously differentiable $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Var}_\pi f \ge \langle \mathbb{E}_\pi \nabla f, (\mathbb{E}_\pi \nabla^2 V)^{-1} \mathbb{E}_\pi \nabla f\rangle.$$

Equation 14 then follows immediately upon applying Theorem A.2 and Theorem A.3 with linear test functions of the form $f_v(x) = \langle x, v\rangle$, where $v \in \mathbb{R}^d$ is arbitrary. More specifically,

Corollary A.1. *Suppose that $\pi \propto \exp(-V)$ is a strictly log-concave measure on $\mathbb{R}^d$, with covariance Σ. Then,*

$$(\mathbb{E}_\pi \nabla^2 V)^{-1} \preceq \Sigma \preceq \mathbb{E}_\pi[(\nabla^2 V)^{-1}].$$

Equation 14 is then an immediate consequence of Corollary A.1 under the assumption in Equation 2.

We impose the following assumptions throughout our analysis, unless otherwise specified.

Assumption A.4. *Assume that there exist a positive-definite matrix $B \in \mathbb{R}^{d\times d}$ and $\alpha \in (0,1)$ such that*

$$\alpha B \preceq \nabla^2 U(x) \preceq B, \quad \forall x \in \mathbb{R}^d. \tag{15}$$

Assumption A.5. *Assume that there exist parameters $0 < \mu \leq L$ such that*

$$\mu I \preceq \nabla^2 U(x) \preceq LI, \quad \forall x \in \mathbb{R}^d. \tag{16}$$

In line with the literature on log-concave sampling, if $\pi(x) \propto \exp(-U(x))$ for some potential $U : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumption A.5, we say that π is μ-strongly log-concave and L-smooth. Of course, Assumption A.5 implies that the potential U is μ-strongly convex. In particular, it admits a unique global minimizer $x^* \in \mathbb{R}^d$, leading us to our third and final assumption.

Assumption A.6. *We assume oracle access to the global minimizer x^*, e.g., via precomputation using a standard convex optimization algorithm.*

Note that the global minimizer may be computed once using a second-order algorithm such as the Newton–Raphson method, for example. As above, we define $\kappa := L/\mu$ to be the *global* condition number of π.

To ease notation below, we define two functions that return the empirical mean and (biased) empirical covariance of a set of samples, respectively. In particular, $\widehat{m}_n : (\mathbb{R}^d)^n \to \mathbb{R}^d$ and $\widehat{S}_n : (\mathbb{R}^d)^n \to \mathbb{R}^{d\times d}$ are given by

$$\begin{aligned}
\widehat{m}_n(x_1, \ldots, x_n) &:= \frac{1}{n}\sum_{i=1}^n x_i, \\
\widehat{S}_n(x_1, \ldots, x_n) &:= \frac{1}{n}\sum_{i=1}^n (x_i - \widehat{m}_n(x_1, \ldots, x_n))(x_i - \widehat{m}_n(x_1, \ldots, x_n))^\top.
\end{aligned} \tag{17}$$

When the set of samples $\{x_1, \ldots, x_n\}$ in question is clear from context, we abuse notation slightly and write $\widehat{m}_n$ and $\widehat{S}_n$.

B First-Order Query Complexity of Log-Concave Sampling

In this appendix, we provide a high-level overview of recent results on the first-order query complexity of log-concave sampling. We distinguish between low accuracy samplers, whose first-order query complexities depend polynomially on the target error $1/\varepsilon$, and high accuracy samplers, whose dependence on $1/\varepsilon$ is instead polylogarithmic. For the first category of samplers, see Table 1. Importantly, ULMC belongs to the family of low accuracy samplers. On the other hand, Table 2 contains analogous results for high accuracy samplers, including FORS.

C First-Order Rejection Sampling

In order to minimize the dependence on the underlying dimension d, we turn to first-order rejection sampling (FORS), as presented in Chen et al. (2026c). In particular, our theoretical results utilize a particular form of FORS that comes equipped with a strong mixing time guarantee in Rényi divergence; see Theorem C.2 below. For the time being, we only need to provide the formulation of FORS that our algorithm requires.

Method	Metric	Query complexity	Ref.
LMC	$\sqrt{\mu}W_2$	$\widetilde{O}\left(\frac{\kappa d}{\varepsilon^2}\right)$	Durmus et al. (2019)
LMC	$\mathcal{R}_q$	$\widetilde{O}\left(\frac{\kappa^2 dq}{\varepsilon^2}\right)$	Chewi et al. (2022)
ULMC	$\mathcal{R}_q$	$\widetilde{O}\left(\frac{\kappa^{3/2} d^{1/2} q^{1/2}}{\varepsilon}\right)$	Altschuler & Chewi (2023)
RM–ULMC	KL	$\widetilde{O}\left(\frac{\kappa d^{1/3}}{\varepsilon^{2/3}}\right)$	Altschuler et al. (2026)
RM–ULMC, low friction	KL	$\widetilde{O}\left(\frac{\kappa^{5/6} d^{5/3}}{\varepsilon^{2/3}}\right)$	Altschuler et al. (2026)

Table 1: First-order query complexities for a handful of low-accuracy samplers. RM-ULMC refers to the randomized midpoint version of ULMC, analyzed in Altschuler et al. (2026). Although the query complexity in order to obtain ε error in $\sqrt{\mathsf{KL}}$ divergence scales better with κ and d, we require a stronger χ^2-divergence guarantee (equivalently, a Rényi divergence guarantee) for our theoretical analysis.

Method	Query complexity	Ref.
Metropolized HMC / MALA, feasible start	$\widetilde{\Omega}(\kappa\, d)$	Lee et al. (2021)
MALA, warm start	$\widetilde{O}(\kappa\, d^{1/2})$	Wu et al. (2022)
Prox. sampler $\to$ rejection sampling	$\widetilde{O}(\kappa\, d^{1/2})$	Fan et al. (2023)
ULMC $\to$ MALA $\to$ prox. sampler	$\widetilde{O}(\kappa\, d^{1/2})$	Altschuler & Chewi (2023)
Metropolized HMC	$\widetilde{O}\big((\kappa^{3/2} + \kappa^{1/2}\kappa_{\mathrm{H}_2}^{1/2} + \kappa_{\mathrm{H}_2}^{2/3})\, d^{1/4}\big)$	Zhang et al. (2026)
FORS	$\widetilde{O}\big(\kappa^{2/3} d^{1/3}\big)$	Chen et al. (2026c)

Table 2: First-order query complexities for a collection of high-accuracy samplers whose complexity scales as $\mathrm{polylog}(1/\varepsilon)$ with respect to the target error ε. All complexities refer to mixing in TV distance, although the last two rows also apply in $\sqrt{\chi^2}$, $\sqrt{\mu}W_2$, and $\sqrt{\mathsf{KL}}$. The guarantee for metropolized HMC in the penultimate row only holds under the stronger assumption that the Hessian is Lipschitz with respect to the Frobenius norm, with Lipschitz constant β_{H_2} and $\kappa_{\mathrm{H}_2} := \beta_{\mathrm{H}_2}/\mu^{3/2}$.

First, recall the *underdamped Langevin diffusion* with friction parameter $\tilde{\gamma} > 0$ and $V : \mathbb{R}^d \to \mathbb{R}$ is a sufficiently smooth potential:

$$\begin{cases} dX_t = P_t dt, \\ dP_t = -\nabla V(X_t)dt - \tilde{\gamma} P_t dt + \sqrt{2\tilde{\gamma}} dB_t. \end{cases} \tag{18}$$

The invariant measure of this diffusion process is $\pi(x, p) \propto \exp(-V(x) - \frac{1}{2}\|p\|_2^2)$, meaning that the x marginal of the invariant measure is precisely the target distribution. Throughout, we abbreviate $z = (x, p) \in \mathbb{R}^d \times \mathbb{R}^d$ to denote the coordinates of the diffusion in phase space. To simulate the dynamics in Equation 18, we first fix a base point $z_0 = (x_0, p_0)$. Then, the gradient term in Equation 18 is frozen at $X_0 = x_0$, yielding a linear SDE of the form

$$\begin{cases} d\widetilde{X}_t = \widetilde{P}_t dt, \\ d\widetilde{P}_t = -\nabla V(X_0)dt - \tilde{\gamma}\widetilde{P}_t dt + \sqrt{2\tilde{\gamma}} dB_t, \end{cases} \tag{19}$$

with initial condition $\widetilde{X}_0 = x$ and $\widetilde{P}_0 = p$. Denote by $\boldsymbol{Q} := \boldsymbol{Q}_{z_0}$ the path measure corresponding to $\widetilde{\boldsymbol{Z}} := (\widetilde{X}_t, \widetilde{P}_t)_{t\in[0,T]}$, where $T > 0$ is fixed. Now, let

$$\mu(x) := \frac{1}{\sqrt{2\tilde{\gamma}}}(\nabla V(x) - \nabla V(x_0)).$$

Applying Girsanov's theorem, there exists a path measure $\boldsymbol{P} = \boldsymbol{P}_{z_0}$ defined by

$$\frac{d\boldsymbol{P}}{d\boldsymbol{Q}} = \exp\left(-\int_0^T \mu(\widetilde{X}_t) \cdot dB_t - \frac{1}{2}\int_0^T \|\mu(\widetilde{X}_t)\|^2 dt\right) \tag{20}$$

such that the process $\widetilde{B}_t := B_t + \int_0^t \mu(\widetilde{X}_s)ds$ is a $\boldsymbol{P}$-Brownian motion and the process $\widetilde{\boldsymbol{Z}}$ follows the dynamics in Equation 18, driven by $\widetilde{B}_t$. To implement FORS, we design an estimator for the logarithm of Radon–Nikodym derivative in Equation 20. For the second term in the exponent, note that such an estimator can be constructed easily by sampling $t \sim \mathrm{Unif}([0, T])$. On the other hand, to handle the stochastic integral, we first must discretize. Given an integer $N \in \mathbb{N}$, take $h = T/N$ to be the corresponding step size. Then, we can approximate the stochastic integral as

$$\begin{aligned} \sum_{j=0}^{N-1} \mu(\widetilde{X}_{jh}) \cdot (B_{(j+1)h} - B_{jh}) &= \sum_{j=1}^{N} (\mu(\widetilde{X}_{jh}) - \mu(\widetilde{X}_{(j-1)h})) \cdot (B_T - B_{jh}) \\ &= \frac{1}{\sqrt{2\tilde{\gamma}}} \sum_{j=1}^{N} (\nabla V(\widetilde{X}_{jh}) - \nabla V(\widetilde{X}_{(j-1)h})) \cdot (B_T - B_{jh}). \end{aligned}$$

In particular, to give an unbiased estimate Equation 20, discretized with step size h, we sample $t \sim \mathrm{Unif}([0, T])$. If $t \in [(j-1)h, jh)$, the estimate in question is then

$$-W(t; \widetilde{\boldsymbol{Z}}) = \frac{T}{\sqrt{2\tilde{\gamma}h}}(\nabla V(\widetilde{X}_{jh}) - \nabla V(\widetilde{X}_{(j-1)h})) \cdot (B_T - B_{jh}) + \frac{T}{2}\|\mu(\widetilde{X}_t)\|^2. \tag{21}$$

Using $O(1)$ unbiased samples of the form Equation 21, FORS approximates the change-of-measure in Equation 20, allowing us to approximately sample from the path measure of Equation 18. In order to accomplish this, FORS proceeds using the following rejection sampling process. Below, $\mathrm{Clip}_{[0,1]}(x) := \min\{1, \max\{0, x\}\}$.

Although the above algorithm may appear mysterious at a glance, its output satisfies a rather concrete guarantee.

Theorem C.1 (Theorem 2.3 in Chen et al. (2026c))**.** *Algorithm 2 outputs a random point $\widetilde{Z}_T$ whose distribution $\widehat{\mu}$ is the marginal of a path measure $\widehat{\boldsymbol{P}}$, satisfying*

$$\frac{d\widehat{\boldsymbol{P}}}{d\boldsymbol{Q}} \propto \exp(\widehat{w}(\widetilde{\boldsymbol{Z}})), \quad \widehat{w}(\widetilde{\boldsymbol{Z}}) := \frac{1}{T}\int_0^T \mathrm{Clip}_{[-B,B]}(W(t; \widetilde{\boldsymbol{Z}}))dt,$$

and with probability at least $1 - \delta$, the number of queries of W is bounded by $3Be^{2B}\log(2/\delta)$.

Algorithm 2 First-order rejection sampling (FORS)

Require: Parameter $B > 0$, proposal distribution $\boldsymbol{Q}$, unbiased estimator W.
1: **for** $i = 1, 2, \ldots$ **do**
2: Sample $J \sim \text{Poisson}(2B)$.
3: Sample $t_1, \ldots, t_J \sim \text{Unif}([0, T])$.
4: Sample $(W(t_1; \widetilde{\boldsymbol{Z}}), \ldots, W(t_J; \widetilde{\boldsymbol{Z}}), \widetilde{Z}_T)$ following the proposal $\widetilde{\boldsymbol{Z}} \sim \boldsymbol{Q}$.
5: $p_J \leftarrow \prod_{j=1}^{J} \text{Clip}_{[0,1]} \left(\frac{B + W(t_j; \widetilde{\boldsymbol{Z}})}{2B} \right)$
6: Output $\widetilde{Z}_T$ with probability p_J.
7: **end for**

The clipping parameter $B > 0$ directly controls the number of queries to W, defined in Equation 21, that must be made during a single call to Algorithm 2. Although taking this parameter larger will decrease the error incurred by FORS, the exponential dependence of the query complexity of Algorithm 2 on B means that we must take B to be a small constant in practice. For example, the authors of Chen et al. (2026c) take $B \in [1, 2]$. Fortunately, even for this choice of B, the error incurred by FORS can be controlled precisely. The following theorem is then adapted from Theorem 3.2(ii) in Chen et al. (2026c), specialized to the case of a distribution satisfying Assumption A.5.

Theorem C.2. *Suppose that $\pi(x) \propto \exp(-V(x))$ satisfies Assumption A.5. For any $q \geq 2$, set $\Delta_q := d + \mathcal{R}_{2q}(\nu \| \pi)$, where $(X_0, P_0) \sim \nu$ and π is the invariant measure of the underdamped Langevin diffusion in Equation 18 with friction $\widetilde{\gamma} = \Theta(\sqrt{\mu})$. Denoting*

$$M_q := q + \log \left(\frac{q \kappa^{2/3} \Delta_q^{1/3} \log(q \Delta_q / \varepsilon^2)}{\varepsilon^2} \right),$$

we may take

$$T = \Theta \left(\mu^{-1/2} \kappa^{-2/3} \Delta_q^{-1/3} M_q^{-2/3} \right)$$

and

$$N = \Theta \left(\kappa^{2/3} \Delta_q^{1/3} M_q^{2/3} \log(q \Delta_q / \varepsilon^2) \right).$$

so that, with probability at least $1 - p$, the output distribution $\widehat{\nu}$ of N steps of Algorithm 2 with timestep T satisfies $\mathcal{R}_q(\widehat{\nu} \| \pi) \leq \varepsilon^2$ with $O(N + \log(1/p))$ first-order queries.

In our setting, we can control the initialization error for approximately sampling from the invariant measure of the underdamped Langevin diffusion in Equation 18 by taking an appropriate Gaussian initialization. If the target distribution satisfies Assumption A.5, and Algorithm 2 is initialized with $(X_0, P_0) \sim \nu_0 := \mathcal{N}(x^*, L^{-1} I_d) \otimes \mathcal{N}(0, I_d)$, then the following technical lemma allows us to bound the initialization error $\mathcal{R}_{2q}(\nu_0 \| \pi)$, where $\pi(x, p) \propto \exp(-V(x) - \frac{1}{2}\|p\|_2^2)$ is the target measure. For any *any* $q \geq 1$, we can prove the following initialization guarantee.

Lemma C.3. *Borrowing notation from Chen et al. (2026c), if $\pi_V(x) = Z^{-1} \exp(-V(x))$ is the first marginal of the invariant measure of the underdamped Langevin diffusion in Equation 18, then for all $q > 1$*

$$\mathcal{R}_{2q}(\nu_0 \| \pi) = \mathcal{R}_{2q}(\mathcal{N}(x^*, L^{-1} I_d) \| \pi_V) \leq \frac{d}{2} \log \kappa.$$

Proof. The first equality above holds because the initialization ν_0 and the invariant measure π are both product measures share the same marginal in p. For the remaining inequality, observe that the density ratio in question can be bounded pointwise as

$$Z \left(\frac{L}{2\pi} \right)^{d/2} \exp \left(V(x) - \frac{L}{2} \|x - x^*\|_2^2 \right) \leq Z \exp(V(x^*)) \left(\frac{L}{2\pi} \right)^{d/2},$$

as Assumption A.5 ensures that

$$V(x) \leq V(x^*) + \frac{L}{2} \|x - x^*\|_2^2,$$

for all $x \in \mathbb{R}^d$. Meanwhile, strong convexity implies that

$$Z = \int_{\mathbb{R}^d} \exp(-V(x))dx \le \exp(-V(x^*)) \int_{\mathbb{R}^d} \exp\left(-\frac{\mu}{2}\|x - x^*\|_2^2\right) dx$$
$$= \exp(-V(x^*)) \left(\frac{2\pi}{\mu}\right)^{d/2}.$$

Thus, the density ratio that we must evaluate is bounded pointwise $\kappa^{d/2}$. Hence,

$$\mathcal{R}_{2q}(\mathcal{N}(x^*, L^{-1}I_d)\|\pi_V) \le \log(\kappa^{d/2}) \le d\log\kappa$$

for all $q > 1$, proving the lemma. □

D Underdamped Langevin Monte Carlo

Although the query complexity guarantee of FORS is impressive, the algorithm is somewhat difficult to implement in practice. Instead, practitioners may prefer a more standard discretization of the underdamped Langevin diffusion, described in detail in Altschuler & Chewi (2023). Rather than using the Euler–Maruyama discretization of Equation 18 to obtain a practical algorithm, we fix the nonlinearity that manifests in the gradient term at each time step, obtaining a piecewise linear SDE that is exactly integrable. In particular, if $h > 0$ is a timestep parameter, then we may consider SDE with piecewise linear coefficients given by

$$\begin{cases} d\widehat{X}_t = \widehat{P}_t dt, \\ d\widehat{P}_t = -\nabla V(\widehat{X}_{kh})dt - \tilde{\gamma}\widehat{P}_t dt + \sqrt{2\tilde{\gamma}}dB_t \end{cases} \tag{22}$$

on the time interval $t \in [kh, (k+1)h]$, where $k \in \mathbb{N}$. Conditioned on the value of $(\widehat{X}_{kh}, \widehat{P}_{kh})$, we have the following characterization of the next pair of iterates $(\widehat{X}_{(k+1)h}, \widehat{P}_{(k+1)h})$.

Proposition D.1 (Lemma 5.3.8 in Chewi (2023)). *Conditioned on the value of $(\widehat{X}_{kh}, \widehat{P}_{kh})$, the next iterate of Equation 22 is distributed as $(\widehat{X}_{(k+1)h}, \widehat{P}_{(k+1)h}) \sim \mathcal{N}(M_{(k+1)h}, \Sigma_{\tilde{\gamma},h})$, where*

$$M_{(k+1)h} := \begin{bmatrix} \widehat{X}_{kh} + \tilde{\gamma}^{-1}\left(1 - e^{-\tilde{\gamma}h}\right)\widehat{P}_{kh} - \tilde{\gamma}^{-1}\left(h - \tilde{\gamma}^{-1}\left(1 - e^{-\tilde{\gamma}h}\right)\right)\nabla V(\widehat{X}_{kh}) \\ e^{-\tilde{\gamma}h}\widehat{P}_{kh} - \tilde{\gamma}^{-1}\left(1 - e^{-\tilde{\gamma}h}\right)\nabla V(\widehat{X}_{kh}) \end{bmatrix}$$

and

$$\Sigma_{\tilde{\gamma},h} := \begin{bmatrix} \frac{2}{\tilde{\gamma}}\left[h - \frac{2}{\tilde{\gamma}}\left(1 - e^{-\tilde{\gamma}h}\right) + \frac{1}{2\tilde{\gamma}}\left(1 - e^{-2\tilde{\gamma}h}\right)\right] & \frac{1}{\tilde{\gamma}}\left(1 - e^{-\tilde{\gamma}h}\right)^2 \\ \frac{1}{\tilde{\gamma}}\left(1 - e^{-\tilde{\gamma}h}\right)^2 & 1 - e^{-2\tilde{\gamma}h} \end{bmatrix} \otimes I_d.$$

This result follows straightforwardly upon noting that, given $(\widehat{X}_{kh}, \widehat{P}_{kh})$, the SDE describing the evolution of $\widehat{P}_t$ on the interval $[kh, (k+1)h]$ is simply the Ornstein–Uhlenbeck process, which can be integrated exactly. In turn, the evolution for $\widehat{X}_t$ is then a linear ODE that depends only on $\widehat{P}_t$.

Algorithm 3 forms the backbone of Gaussian cooling, allowing us to sample efficiently at each stage of the algorithm. In practice, we run many ULMC chains in parallel. More concretely, if $n \in \mathbb{N}$, then we may produce n approximate samples from a target distribution with potential V by running Algorithm 3 n times in parallel. In shorthand, we refer to this procedure by

$$x_1, \dots, x_n = \mathrm{ULMC}(V, x^*, \tilde{\gamma}, L, h, N, n) \tag{23}$$

when Algorithm 3 is utilized as a subprocess of Gaussian cooling. For simplicity, we summarize all of the parameters needed for an implementation of ULMC:

- Relative preconditioning error $\delta > 0$.
- Conditioning parameter $\alpha \in (0, 1)$.
- Cooling parameter $\gamma \in (0, 1)$.
- Augmented conditioning parameter $\tilde{\alpha} = \min\{\alpha, \gamma\}$.

Algorithm 3 Underdamped Langevin Monte Carlo (ULMC)

Require: Potential $V : \mathbb{R}^d \to \mathbb{R}$ with gradient oracle ∇V, global minimizer x^*, friction parameter $\tilde{\gamma} > 0$, smoothness parameter $L > 0$, step size $h > 0$, number of iterations N

1: Draw $(\widehat{X}_0, \widehat{P}_0) \sim \mathcal{N}(x^*, L^{-1} I_d) \otimes \mathcal{N}(0, I_d)$
2: **for** $k = 0, 1, \ldots, N-1$ **do**
3: $\quad \widehat{V}_{kh} \leftarrow \nabla V(\widehat{X}_{kh})$
4: $\quad$ Sample $(\eta_k, \xi_k) \sim \mathcal{N}(0, \Sigma_{\tilde{\gamma},h})$
5: $\quad \widehat{X}_{(k+1)h} \leftarrow \widehat{X}_{kh} + \frac{1 - e^{-\tilde{\gamma}h}}{\tilde{\gamma}} \widehat{P}_{kh} - \frac{\tilde{\gamma}h - 1 + e^{-\tilde{\gamma}h}}{\tilde{\gamma}^2} \widehat{V}_{kh} + \eta_k$
6: $\quad \widehat{P}_{(k+1)h} \leftarrow e^{-\tilde{\gamma}h} \widehat{P}_{kh} - \frac{1 - e^{-\tilde{\gamma}h}}{\tilde{\gamma}} \widehat{V}_{kh} + \xi_k$
7: **end for**
8: **return** $\widehat{X}_{Nh}$

- Smoothness parameter $L = \tilde{\alpha}^{-1}(1 + \delta)$.

The choice of $L = \widetilde{\alpha}^{-1}(1 + \delta)$ comes from Equation 26, in which the preconditioned potential is L-smooth with this choice of L. We replace α with $\tilde{\alpha} \leq \alpha$ because while α itself may not be known, in all applications we consider in this work, a reasonable lower bound on α is known and used as the cooling parameter γ.

We invoke an intermediate result on the optimal scaling of ULMC, based on the analysis in (Altschuler & Chewi, 2023, Theorem 4.1). The specific discretization that we take for ULMC is as in (Chewi, 2023, Section 5.3.2). Below, V is a potential from which we want to sample. We then form a piecewise system of linear SDEs, discretizing the underdamped Langevin dynamics, that can be integrated *exactly* as in (Chewi, 2023, Lemma 5.3.8):

$$\begin{cases} d\widehat{X}_t = \widehat{P}_t dt & t \geq 0, \\ d\widehat{P}_t = -(\nabla V(\widehat{X}_{nh}) + \widetilde{\gamma}\widehat{P}_t)dt + \sqrt{2\widetilde{\gamma}} dB_t & t \in [nh, (n+1)h]. \end{cases} \tag{24}$$

Above, $\widetilde{\gamma}$ is a friction parameter and h is a step size parameter, both of which we specify below. For this discretization of ULMC, we have the following Rényi divergence guarantee, which naturally translates to a guarantee in χ^2-divergence.

Theorem D.2 (Theorem 6.3.6 in Chewi (2023))**.** *Assume that V is μ-strongly convex and L-smooth and that $\pi(x) \propto \exp(-V(x))$ for $x \in \mathbb{R}^d$, with $x^* \in \mathbb{R}^d$ the unique minimizer of V. For each $n \in \mathbb{N}$, let $\widehat{\pi}_{nh} = \mathrm{Law}(\widehat{X}_{nh})$ in Equation 24, with friction parameter $\widetilde{\gamma} = \sqrt{2L}$ and initialization $\widehat{\pi}_0 = \mathcal{N}(x^*, L^{-1} I_d) \otimes \mathcal{N}(0, I_d)$. Then, for any $q \geq 2$ and step size $h = \widetilde{\Theta}(\epsilon / \sqrt{L\kappa dq})$, we obtain $\mathcal{R}_q(\widehat{\pi}_{Nh} \| \pi) \leq \epsilon^2$ after*

$$N = \widetilde{\Theta}\left(\frac{\kappa^{3/2} d^{1/2} q^{1/2}}{\epsilon}\right)$$

iterations of Equation 24.

Using this as our sampler at each stage, we can fully characterize the query complexity of Gaussian cooling as in Corollary D.1. To be as explicit as possible, we quantify all the parameters that are required to use ULMC as the sampler for Gaussian cooling in Table 3.

We can now provide a precise analogue of Corollary 3.1 for ULMC:

Corollary D.1. *ULMC, as presented in Algorithm 3, is a low accuracy sampler satisfying Assumption 3.1 with constants $c_1 = 3/2$, $c_2 = 1/2$, and $c_3 = 1$. Consequently, Gaussian cooling with ULMC taking the place of $\mathcal{A}$ satisfies Equation 6 with probability at least $1 - p$ using*

$$NnK = \begin{cases} \widetilde{O}\left(\frac{d^{1/2}[d + \log(1/p)]\log(\kappa)}{\tilde{\alpha}^7 \delta^3 \log(1/\gamma)}\right) & \textit{general target}, \\ \widetilde{O}\left(\frac{d^{1/2}\log(1/p)\log(\kappa)}{\tilde{\alpha}^7 \delta^3 \log(1/\gamma)}\right) & \textit{translation-invariant target} \end{cases}$$

queries to U and its gradient.

Quantity	Symbol	Value
ULMC step size	h	$\widetilde{\Theta}\left(\frac{\delta\widetilde{\alpha}^{3/2}}{(1+\delta)^{3/2}d^{1/2}}\right)$
ULMC friction parameter	$\widetilde{\gamma}$	$\sqrt{2\widetilde{\alpha}^{-1}(1+\delta)}$
ULMC steps per sample	N	$\widetilde{\Theta}\left(\frac{(1+\delta)^3 d^{1/2}}{\widetilde{\alpha}^3\delta}\right)$
Samples per cooling stage	n	$\Theta\left(\frac{d(1+\delta)^4}{\widetilde{\alpha}^4\delta^2}\right)$
Stages of Gaussian cooling	K	$O\left(\frac{\log\kappa}{\log(1/\gamma)}\right)$

Table 3: Parameters passed into ULMC and the number of iterations required for each subprocess of the Gaussian cooling algorithm.

E Query Complexity Proofs

We begin with the proof of Proposition 2.1, which inductively establishes the preconditioning guarantee afforded by Algorithm 1 and gives a bound on the total number of cooling stages required to obtain a good preconditioner of the target distribution, as specified in Equation 6.

Proof of Proposition 2.1. By construction,

$$\alpha B + \gamma^k LI \preceq \nabla^2 U_k(x) \preceq B + \gamma^k LI$$

for all $x \in \mathbb{R}^d$, so by Corollary A.1, it follows that

$$\alpha B + \gamma^k LI \preceq (\Sigma^{(k)})^{-1} \preceq B + \gamma^k LI.$$

Now, in the first step of the algorithm, we have that

$$LI \preceq \nabla^2 U_0(x) \preceq 2LI, \quad x \in \mathbb{R}^d,$$

so U_0 is strongly convex with condition number at most 2. We now proceed inductively, showing that if

$$(1+\delta)^{-1}\Sigma^{(k-1)} \preceq \widetilde{\Sigma}_n^{(k-1)} \preceq (1+\delta)\Sigma^{(k-1)}, \tag{25}$$

where $\widetilde{\Sigma}_n^{(k-1)}$ is the preconditioner formed in Line 6 of Algorithm 1 from n samples approximately drawn from π_{k-1}, then $\widetilde{\Sigma}_n^{(k-1)}$ is a good preconditioner for π_k. Indeed, we have that

$$\nabla^2 U_k(x) \preceq \alpha^{-1}(\alpha B + \alpha\gamma^k LI) \preceq \alpha^{-1}(\Sigma^{(k-1)})^{-1} \preceq \alpha^{-1}(1+\delta)(\widetilde{\Sigma}_n^{(k-1)})^{-1}$$

and

$$\nabla^2 U_k(x) \succeq \widetilde{\alpha}(B + \gamma^{k-1} LI) \succeq \widetilde{\alpha}(\Sigma^{(k-1)})^{-1} \succeq \widetilde{\alpha}(1+\delta)^{-1}(\widetilde{\Sigma}_n^{(k-1)})^{-1},$$

In particular,

$$\widetilde{\alpha}(1+\delta)^{-1}(\widetilde{\Sigma}_n^{(k-1)})^{-1} \preceq \nabla^2 U_k(x) \preceq \alpha^{-1}(1+\delta)(\widetilde{\Sigma}_n^{(k-1)})^{-1}, \quad x \in \mathbb{R}^d.$$

If we take

$$\widetilde{U}_k(x) := U_k((\widetilde{\Sigma}_n^{(k-1)})^{1/2}x),$$

it therefore follows that

$$\widetilde{\alpha}(1+\delta)^{-1} I \preceq \nabla^2 \widetilde{U}_k(x) \preceq \alpha^{-1}(1+\delta) I, \quad \forall x \in \mathbb{R}^d. \tag{26}$$

Hence, $\widetilde{\pi}_k(x) \propto \exp(-\widetilde{U}_k(x))$ is strongly log-concave with condition number at most $\frac{(1+\delta)^2}{\alpha\widetilde{\alpha}}$.

Finally, suppose that $\gamma^k L \leq \mu$, which holds once $k \geq \log(\kappa)/\log(1/\gamma)$. By Corollary A.1, Assumption A.4, and the fact that $\mu I \preceq \nabla^2 U(x)$, this choice of k allows us to write

$$(\Sigma^{(k)})^{-1} \preceq B + \gamma^k LI \preceq \alpha^{-1}\nabla^2 U(x) + \nabla^2 U(x) = \frac{1+\alpha}{\alpha}\nabla^2 U(x)$$

and

$$(\Sigma^{(k)})^{-1} \succeq \alpha B + \gamma^k L I \succeq \alpha B \succeq \alpha \nabla^2 U(x).$$

Combining these with the inequality $(1+\delta)^{-1}\Sigma^{(k)} \preceq \widetilde{\Sigma}_n^{(k)} \preceq (1+\delta)\Sigma^{(k)}$ gives

$$\frac{\alpha}{(1+\alpha)(1+\delta)}(\widetilde{\Sigma}_n^{(k)})^{-1} \preceq \nabla^2 U(x) \preceq \frac{1+\delta}{\alpha}(\widetilde{\Sigma}_n^{(k)})^{-1}$$

for all $x \in \mathbb{R}^d$. In turn, $\widetilde{U}(x) := U((\widetilde{\Sigma}_n^{(k)})^{1/2}x)$ has condition number at most

$$\frac{(1+\delta)^2(1+\alpha)}{\alpha^2} \leq \frac{2(1+\delta)^2}{\alpha^2},$$

allowing us to sample from the target distribution with a good preconditioner. □

We now prove Theorem 3.2, distinguishing between low accuracy and high accuracy samplers as in Assumption 3.1.

Proof of Theorem 3.2 for Low Accuracy $\mathcal{A}$. In both the low accuracy and high accuracy settings, stage $k = 0$ remains the same. The target distribution is induced by the potential

$$\widetilde{U}_0(x) := U_0(L^{-1/2}x).$$

Because $\nabla^2 \widetilde{U}_0(x) = L^{-1}\nabla^2 U_0(L^{-1/2}x)$, it follows that

$$I \preceq \nabla^2 \widetilde{U}_0(x) \preceq 2I, \quad \forall x \in \mathbb{R}^d,$$

so that the corresponding measure $\widetilde{\pi}^{(0)} \in \mathcal{P}_2(\mathbb{R}^d)$ is 1-strongly log-concave and 2-smooth. By Theorem F.7, if we take $n = \Theta\left(\frac{d}{\delta^2}\right)$ approximate samples from $\widetilde{\pi}^{(0)}$, then with high probability, the resulting sample covariance $\widetilde{\Sigma}_n^{(0)}$ satisfies

$$(1+\delta)^{-1}\mathrm{Cov}(\widetilde{\pi}^{(0)}) \preceq \widetilde{\Sigma}_n^{(0)} \preceq (1+\delta)\mathrm{Cov}(\widetilde{\pi}^{(0)}).$$

In the low accuracy case, each of the samples can be produced with

$$N = \widetilde{O}\left(d^{c_2}\delta^{-c_3}\right)$$

first-order queries. Now, because

$$\Sigma^{(0)} = L^{-1}\mathrm{Cov}(\widetilde{\pi}^{(0)}),$$

it follows that

$$(1+\delta)^{-1}\Sigma^{(0)} \preceq L^{-1}\widetilde{\Sigma}_n^{(0)} \preceq (1+\delta)\Sigma^{(0)}.$$

Taking $\widehat{\Sigma}_n^{(0)} := L^{-1}\widetilde{\Sigma}_n^{(0)}$, it follows that $\widehat{\Sigma}_n^{(0)}$ is a good preconditioner for π_0.

At stage k of Gaussian cooling, assume that we have formed a sample covariance matrix $\widehat{\Sigma}_n^{(k-1)}$ such that

$$(1+\delta)^{-1}\Sigma^{(k-1)} \preceq \widehat{\Sigma}_n^{(k-1)} \preceq (1+\delta)\Sigma^{(k-1)},$$

allowing us to form a potential

$$\widetilde{U}_k(x) := U_k((\widehat{\Sigma}_n^{(k-1)})^{1/2}x)$$

with corresponding measure $\widetilde{\pi}^{(k)} \in \mathcal{P}_2(\mathbb{R}^d)$ that is $\widetilde{\alpha}(1+\delta)^{-1}$-strongly log-concave and $\widetilde{\alpha}^{-1}(1+\delta)$-smooth.

Note that if $\mathcal{R}_2(\widehat{\pi}\|\pi) \leq \varepsilon^2$, then

$$\chi^2(\widehat{\pi}\|\pi) \leq \exp(\varepsilon^2) - 1.$$

Leveraging the guarantee in Theorem F.7, we take

$$c(\delta) := \min\left\{\left(\sqrt{3-(1+\delta)^{-1/2}}-\sqrt{2}\right)^2, \frac{1}{16}\right\}$$

In turn, let $\kappa = \frac{(1+\delta)^2}{\widetilde{\alpha}^2}$, $L = \widetilde{\alpha}^{-1}(1+\delta)$, and $\varepsilon = \sqrt{\log(c(\delta)+1)}$. With these parameters, after

$$N = \widetilde{\Theta}\left(\kappa^{c_1} d^{c_2} (\log(c(\delta)+1))^{-c_3/2}\right) = \widetilde{\Theta}\left(\left(\frac{1+\delta}{\widetilde{\alpha}}\right)^{2c_1} d^{c_2}\delta^{-c_3}\right)$$

first-order queries, Assumption 3.1 ensures that the resulting samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \widehat{\pi}_N^{(k)}$ are such that

$$\chi^2(\widehat{\pi}_N^{(k)} \| \widetilde{\pi}^{(k)}) \le c(\delta).$$

As long as we take

$$n = \Theta\left(\frac{d(1+\delta)^4}{\widetilde{\alpha}^4\delta^2}\right)$$

samples, Theorem F.7 then implies that, with high probability,

$$(1+\delta)^{-1}\mathrm{Cov}(\widetilde{\pi}^{(k)}) \preceq \widetilde{\Sigma}_n^{(k)} \preceq (1+\delta)\mathrm{Cov}(\widetilde{\pi}^{(k)}), \tag{27}$$

where $\widetilde{\Sigma}_n^{(k)}$ is the sample covariance formed from n samples drawn from $\widehat{\pi}_N^{(k)}$. As

$$\Sigma^{(k)} := \mathrm{Cov}(\pi^{(k)}) = (\widehat{\Sigma}_n^{(k-1)})^{1/2}\mathrm{Cov}(\widetilde{\pi}^{(k)})(\widehat{\Sigma}_n^{(k-1)})^{1/2},$$

it follows that

$$(1+\delta)^{-1}\Sigma^{(k)} \preceq (\widehat{\Sigma}_n^{(k-1)})^{1/2}\widetilde{\Sigma}_n^{(k)}(\widehat{\Sigma}_n^{(k-1)})^{1/2} \preceq (1+\delta)\Sigma^{(k)},$$

allowing us to take $\widehat{\Sigma}_n^{(k)} := (\widehat{\Sigma}_n^{(k-1)})^{1/2}\widetilde{\Sigma}_n^{(k)}(\widehat{\Sigma}_n^{(k-1)})^{1/2}$ as a preconditioner for the next stage of cooling.

Finally, we perform a simple union bound over the probability that the concentration in Theorem F.7 fails at each stage of Gaussian cooling. Specifically, across $K = O\left(\frac{\log(\kappa)}{\log(1/\gamma)}\right)$ stages of cooling, there exists constants $C_1, C_2 > 0$ such that as long as we take

$$n = \frac{C_1 d(1+\delta)^4}{\widetilde{\alpha}^4\delta^2}$$

samples at each stage, then the probability that our concentration inequality fails at each stage is at most $4\exp\left(\frac{-C_2\widetilde{\alpha}^4\delta^2 n}{(1+\delta)^4}\right)$. By a union bound, the total probability of failure is thus at most

$$4C_3\exp\left(-C_2\frac{\widetilde{\alpha}^4\delta^2 n}{(1+\delta)^4}\right)\cdot\frac{\log(\kappa)}{\log(1/\gamma)}$$

for some constant $C_3 > 0$. Rearranging to find the dependence of n on the remaining parameters, we find that in order for the above quantity to be at most p for some arbitrary $p \in (0,1)$, we must take

$$n = \Theta\left(\frac{(1+\delta)^4\left[d+\log(K/p)\right]}{\widetilde{\alpha}^4\delta^2}\right).$$

samples at each stage. In turn, taking n as above will ensure that Gaussian cooling succeeds with probability at least $1-p$. This validates the query complexity reported above, as we have only incurred an additional sublogarithmic factor in κ to ensure that the algorithm succeeds with high probability. □

In the high accuracy case, the above proof only needs to be modified slightly. Namely, the dependence on δ improves slightly due to the polylogarithmic dependence of the query complexity of a high accuracy sampler on the target accuracy. However, we must perform an additional union bound to account for the nonzero probability that the number of first-order queries at each stage exceeds the threshold in Assumption 3.1. We remark that this additional consideration is specifically due to the fact that FORS relies on rejection sampling.

Proof of Theorem 3.2 for high accuracy $\mathcal{A}$. With $c(\delta)$, κ, L, and ε as in the low accuracy case, the high accuracy setting in Assumption 3.1 means that a single call to $\mathcal{A}$ produces a sample from a distribution $\widehat{\pi}^{(k)}$ with $\chi^2(\widehat{\pi}^{(k)} \| \widetilde{\pi}^{(k)}) \le c(\delta)$, and that, with probability at least $1-\eta$, the call uses at most

$$N := N_0 + O(\log(1/\eta)), \qquad N_0 := \widetilde{\Theta}\left(\left(\frac{1+\delta}{\widetilde{\alpha}}\right)^{2c_1} d^{c_2} (\log(\delta^{-1}))^{c_3}\right)$$

first-order queries. Across all K stages of Gaussian cooling, the sampler $\mathcal{A}$ is called nK times. Performing a union bound, the probability that any call uses more than N queries is at most $nK\eta$. We therefore choose

$$\eta := \frac{p}{2nK},$$

so that this probability is at most $p/2$. We can then directly compute, due to our choice of n,

$$\log(1/\eta) = \log\left(\frac{2nK}{p}\right) = O\left(\log d + \log(\widetilde{\alpha}^{-1}) + \log(\delta^{-1}) + \log K + \log(1/p)\right),$$

which is polylogarithmic in all problem parameters. Now, the concentration argument and the union bound over the K stages of cooling from the low accuracy case apply verbatim, with $p/2$ in place of p. In particular, taking

$$n = \Theta\left(\frac{(1+\delta)^4 \left[d + \log(K/p)\right]}{\widetilde{\alpha}^4 \delta^2}\right)$$

samples per stage ensures that, with probability at least $1 - p/2$, the preconditioning guarantee holds at every stage. By a union bound over these two possible options for failure (i.e., the number of queries exceeds the desired threshold or concentration of the sample covariance fails), with probability at least $1-p$ the preconditioning guarantee holds at every stage and every call to $\mathcal{A}$ uses at most N first-order queries, for a total of at most

$$nKN = nKN_0 + O\left(nK \log(nK/p)\right)$$

first-order queries. □

In the translation-invariant case, instead suppose that we have an estimate $\widehat{\sigma}$ of σ such that

$$\| \operatorname{diag}(\sigma)^{-1/2} \operatorname{diag}(\widehat{\sigma}) \operatorname{diag}(\sigma)^{-1/2} - I\|_2 \le C(\delta), \tag{28}$$

where $C(\delta) := 1 - (1+\delta)^{-1/2}$. Because $\mathcal{F}$ is unitary, assuming that it is scaled appropriately, we have that

$$\|\mathcal{F} \operatorname{diag}(\sigma)^{-1/2} \mathcal{F}^* \mathcal{F} \operatorname{diag}(\widehat{\sigma}) \mathcal{F}^* \mathcal{F} \operatorname{diag}(\sigma)^{-1/2} \mathcal{F}^* - I\|_2 \le C(\delta),$$

which holds if and only if

$$\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I\|_2 \le C(\delta),$$

where

$$\widehat{\Sigma} = \mathcal{F} \operatorname{diag}(\widehat{\sigma}) \mathcal{F}^*.$$

As seen in Section F, this is sufficient to conclude that

$$(1+\delta)^{-1/2} \Sigma \preceq \widehat{\Sigma} \preceq (1+\delta)^{1/2} \Sigma,$$

which is precisely the guarantee that we require for our analysis in Section F. Moreover, for Equation 28 to hold, it suffices to show that

$$\|\widehat{\sigma} \odot \sigma^{-1} - \mathbf{1}\|_\infty \le C(\delta), \tag{29}$$

where $v \odot w \in \mathbb{R}^d$ denotes the Hadamard (entrywise) product of two vectors $v, w \in \mathbb{R}^d$ and $\mathbf{1} \in \mathbb{R}^d$ is the vector of all ones.

Given a set of samples $x_1, \dots, x_n \overset{i.i.d.}{\sim} \mu$, where μ is a distribution whose covariance $\Sigma = \mathrm{Cov}(\mu)$ we wish to estimate, the natural empirical counterpart to σ is

$$\widehat{\sigma} := \frac{1}{n}\sum_{i=1}^{n} \mathcal{F}(x_i - \widehat{m}_n) \odot \overline{\mathcal{F}(x_i - \widehat{m}_n)}, \tag{30}$$

where the complex conjugate on the second term is taken entrywise. To see where this estimate comes from, note that $\mathcal{F}^*\Sigma\mathcal{F} = \mathrm{diag}(\sigma)$. If we similarly conjugate by $\mathcal{F}$ the (biased) sample covariance $\widehat{\Sigma}_n$ formed from the samples $x_1, \dots, x_n$, we obtain

$$\begin{aligned}\mathcal{F}^*\widehat{\Sigma}_n\mathcal{F} &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{F}^*(x_i - \widehat{m}_n)(x_i - \widehat{m}_n)^\top \mathcal{F} \\ &= \frac{1}{n}\sum_{i=1}^{n} ((x_i - \widehat{m}_n)^\top \mathcal{F})^*(x_i - \widehat{m}_n)^\top \mathcal{F}.\end{aligned}$$

Extracting the diagonal from the above expression, our estimate $\widehat{\sigma}$ of σ is thus the vector with entries

$$\widehat{\sigma}_j = \frac{1}{n}\sum_{i=1}^{n} \overline{(x_i - \widehat{m}_n)^\top \mathcal{F} e_j}(x_i - \widehat{m}_n)^\top \mathcal{F} e_j, \quad \forall j \in [d].$$

Finally, because the discrete Fourier transform is symmetric, we have that

$$(x_i - \widehat{m}_n)^\top \mathcal{F} e_j = (\mathcal{F}(x_i - \widehat{m}_n))_j,$$

which yields the estimate proposed in Equation 30. In Section F, we analyze the number of samples required in Equation 30 to ensure that Equation 29 holds with high probability, after which the analysis in both the low and high accuracy cases can be repeated.

F Approximating the Sample Covariance

In the proofs below, we will require the following standard definitions pertaining to subgaussian and subexponential distributions, taken from (Vershynin, 2018, Chapter 2).

Definition 2. *A real-valued random variable X is subgaussian if there exists $K_1 > 0$ such that*

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2), \quad t \geq 0.$$

Equivalently, X is subgaussian if there exists $K_2 > 0$ such that

$$\mathbb{E}[\exp(X^2/K_2^2)] \leq 2,$$

and we define the corresponding subgaussian norm by

$$\|X\|_{\psi_2} := \inf\{K > 0 : \mathbb{E}[\exp(X^2/K^2)] \leq 2\}.$$

The definition of a *subexponential* random variable is entirely analogous, so we only mention the relevant subexponential norm, given by

$$\|X\|_{\psi_1} := \inf\{K > 0 : \mathbb{E}[\exp(|X|/K)] \leq 2\}.$$

The proof of our main result relies on three technical tools. First, we leverage the variational characterization of χ^2-divergence provided in Theorem A.1 to describe the relative distance, in spectral norm, between the covariance matrices of two distributions that are close in χ^2-divergence. Then, we show that closeness in χ^2-divergence preserves subgaussianity of log-concave measures. Finally, subgaussianity of the sampled distribution will allow us to show that the sample covariance concentrates well, providing a good preconditioner for the next stage of Gaussian cooling. At a high level, the primary technical contribution of this section is an answer to the following question.

Question 1. *Suppose that $\{x_1, \dots, x_n\}$ is a set of approximate samples from a log-concave measure μ, in the sense that $x_i \overset{i.i.d}{\sim} \widehat{\mu}$, where*

$$\mathcal{D}(\widehat{\mu}\|\mu) \leq \varepsilon$$

for some divergence $\mathcal{D}$ and all $i \in [n]$. If $\delta > 0$ is fixed, then how many samples do we need to ensure that the empirical covariance $\widehat{\Sigma}_n$ of $\{x_1, \dots, x_n\}$ satisfies

$$(1+\delta)^{-1}\Sigma \preceq \widehat{\Sigma}_n \preceq (1+\delta)\Sigma, \tag{31}$$

where Σ is the covariance of μ?

By answering Question 1, we guarantee that stage k of Gaussian cooling provides a good preconditioner for stage $k+1$, as specified by the parameter $\delta > 0$.

F.1 Preconditioning Under Closeness in χ^2-Divergence

Perhaps the most natural characterization of χ^2-divergence for our purposes is its variational formulation, elucidated in Appendix A. For completeness, we recall it here:

$$\chi^2(\mu\|\nu) = \sup_{g:\mathbb{R}^d\to\mathbb{R}} \frac{(\mathbb{E}_\mu[g(X)] - \mathbb{E}_\nu[g(Y)])^2}{\mathrm{Var}_\nu(g(Y))}, \tag{32}$$

where the supremum is taken over all functions $g : \mathbb{R}^d \to \mathbb{R}$ such that $g \in L^2(\nu)$ and $\mathrm{Var}_{Y\sim\nu}(g(Y)) > 0$. Note that for distributions with all moments finite, such as subgaussian distributions, this class of functions includes all polynomial functions of the form $g(x) = \langle v, x - w\rangle^p$, where $p \in \mathbb{N}$ and $v, w \in \mathbb{R}^d$ are arbitrary vectors.

To begin, one may attempt to take $g(x) = \langle v, x\rangle$ for some arbitrary unit vector $v \in \mathbb{R}^d$, which yields

$$\frac{(\mathbb{E}_\mu[\langle v, X\rangle] - \mathbb{E}_\nu[\langle v, Y\rangle])^2}{\mathrm{Var}_\nu(\langle v, Y\rangle)} = \frac{(\langle v, \mathbb{E}[X] - \mathbb{E}[Y]\rangle)^2}{\langle v, \Sigma_2 v\rangle}.$$

Hence, if $\chi^2(\mu\|\nu) < \varepsilon$ for some $\varepsilon > 0$, then

$$|\langle v, \mathbb{E}[X] - \mathbb{E}[Y]\rangle| < \sqrt{\varepsilon\langle v, \Sigma_2 v\rangle}. \tag{33}$$

Although this doesn't give immediate control over the difference between the covariance matrices of μ and ν respectively, Equation 33 will prove to be useful later.

If we instead take $m_1 := \mathbb{E}_\mu[X]$ and $\Sigma_1 = \mathrm{Cov}_\mu(X)$, and similarly let $m_2 := \mathbb{E}_\nu[X]$ and $\Sigma_2 := \mathrm{Cov}_\nu(X)$, we may instead consider the test function

$$g_v(x) := \langle v, x - m_2\rangle^2.$$

Then, we have that

$$\begin{aligned}
\mathbb{E}_\mu[g_v(X)] &= \mathbb{E}_\mu[\langle v, X - m_2\rangle^2] \\
&= \mathbb{E}_\mu[(\langle v, X - m_1\rangle + \langle v, m_1 - m_2\rangle)^2] \\
&= \mathbb{E}_\mu[\langle v, X - m_1\rangle^2] + \langle v, m_1 - m_2\rangle^2 + 2\langle v, m_1 - m_2\rangle\mathbb{E}_\mu[\langle v, X - m_1\rangle] \\
&= \langle v, \Sigma_1 v\rangle + \langle v, m_1 - m_2\rangle^2.
\end{aligned}$$

On the other hand,

$$\mathbb{E}_\nu[g_v(Y)] = \mathbb{E}_\nu[\langle v, Y - m_2\rangle^2] = \langle v, \Sigma_2 v\rangle.$$

Hence, the numerator of the above expression becomes

$$(\mathbb{E}_\mu[g(X)] - \mathbb{E}_\nu[g(Y)])^2 = (\langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle + \langle v, m_1 - m_2\rangle^2)^2 \tag{34}$$

On the other hand, we compute

$$\begin{aligned}
\mathrm{Var}_\nu(g(Y)) &= \mathbb{E}_\nu[\langle v, Y - m_2\rangle^4] - (\mathbb{E}_\nu[\langle v, Y - m_2\rangle^2])^2 \\
&= \mathbb{E}_\nu[\langle v, Y - m_2\rangle^4] - \langle v, \Sigma_2 v\rangle^2.
\end{aligned} \tag{35}$$

To motivate the computation that is to come, observe that if ν is the law of a centered Gaussian $Y \sim \mathcal{N}(0, I_d)$, then

$$\begin{aligned}
\mathbb{E}_\nu[\langle v, Y - m_2\rangle^4] = \mathbb{E}_{g_i\sim\mathcal{N}(0,1)}\left[\left(\sum_{i=1}^d v_i g_i\right)^4\right] &= \mathbb{E}_{g_i\sim\mathcal{N}(0,1)}\left[\sum_{1\le i\ne j\le d} v_i^2 v_j^2 g_i^2 g_j^2 + \sum_{i=1}^d v_i^4 g_i^4\right] \\
&= 6\sum_{1\le i<j\le d} v_i^2 v_j^2 + 3\sum_{i=1}^d v_i^4 \\
&= 3\left(\sum_{i=1}^d v_i^2\right)^2 \\
&= 3\mathbb{E}_{g_i\sim\mathcal{N}(0,1)}\left[\left(\sum_{i=1}^d v_i g_i\right)^2\right]^2 \\
&= 3\mathbb{E}_\nu[\langle v, Y - m_2\rangle^2]^2 \\
&= 3\langle v, \Sigma_2 v\rangle^2.
\end{aligned}$$

Motivated by this simple example, we anticipate that if ν is subgaussian, then we can control the fourth moments of their one-dimensional marginals by the corresponding second moments.

F.2 L^4–L^2 Equivalence for Log-Concave Measures

Several recent results concerning one-dimensional moment comparison bounds come to the rescue (Eitan, 2024; Murawski, 2026). To reduce to the one-dimensional case, observe that if Y is a centered log-concave random vector, then $\langle v, Y\rangle$ is a centered log-concave random variable for any $v \in \mathbb{R}^d$. Indeed, if $U \in \mathbb{R}^{d\times d}$ is an orthogonal matrix such that $Uv = e_1$, where e_1 is the first standard basis vector in $\mathbb{R}^d$, then $\langle v, Y\rangle = \langle e_1, UY\rangle$. Because log-concavity is preserved under affine transformations and marginalization, the above claim follows.

Hence, to obtain the L^4–L^2 norm equivalence that we require to continue our analysis in χ^2-divergence, it suffices to find the optimal constant $C > 0$ such that

$$\mathbb{E}[X^4] \leq C\mathbb{E}[X^2]^2 \tag{36}$$

for all centered, real-valued log-concave distributions. As one might suspect, the optimal constant in Equation 36 is attained by a centered exponential distribution. Precisely, we have the following result from Eitan (2024).

Proposition F.1 (Theorem 4 in Eitan (2024))**.** *Let X be a centered, real-valued log-concave random variable. Then, for all $q \in \mathbb{N}$, we have that*

$$\frac{\|X\|_{2q}}{\|X\|_2} \leq (!(2q))^{1/(2q)},$$

with equality if and only if X has the law of $\pm\Gamma$, where $!n$ is the number of derangements of n and Γ has density $f(x) = e^{-(x+1)}$ for $x \geq -1$.

Above, another formula for the number of derangements of n (i.e., permutations with no fixed points) is

$$!n = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!},$$

so for $2q = 4$, we have that $!(2q) = 9$. Hence, Proposition F.1 implies that for any centered, real-valued log-concave random variable X,

$$\mathbb{E}[X^4] = \|X\|_4^4 \leq 9\|X\|_2^4 = 9\mathbb{E}[X^2]^2,$$

and that the constant $C = 9$ is attained by the centered exponential distribution. Returning to the multivariate setting, we immediately obtain the following corollary.

Corollary F.1. *Suppose that μ is a log-concave distribution on $\mathbb{R}^d$. Then, for any $v \in \mathbb{R}^d$, we have that*

$$\mathbb{E}[\langle v, X - \mathbb{E}[X]\rangle^4] \leq 9\langle v, \Sigma v\rangle^2.$$

This follows immediately upon applying the following result from Stavrakakis & Valettas (2013) to $Y = X - \mathbb{E}[X]$, whose law is still log-concave under μ, and considering the seminorm $f_v(x) = |\langle v, x\rangle|$.

Proposition F.2. *Suppose that μ is a log-concave probability measure on $\mathbb{R}^d$ and that $f : \mathbb{R}^d \to \mathbb{R}_+$ is a seminorm. Then, for any $1 \leq p < q$, we have that*

$$\|f\|_{L^q(\mu)} \leq \frac{cq}{p}\|f\|_{L^p(\mu)},$$

where $c > 0$ is a universal constant.

By restricting our attention to seminorms of the form $f_v(x) = |\langle v, x\rangle|$, where $v \in \mathbb{R}^d$, the L^4–L^2 comparison bound that we require for our χ^2-divergence analysis becomes apparent.

Now, from the variational characterization of χ^2-divergence in Equation 32, if $\chi^2(\mu\|\nu) < \epsilon$, then Equation 34 and Equation 35 combine into

$$(\langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle + \langle v, m_1 - m_2\rangle^2)^2 < \epsilon(\mathbb{E}_\nu[\langle v, Y - m_2\rangle^4] - \langle v, \Sigma_2 v\rangle^2).$$

If ν is also log-concave (as will be the case for us), then it follows that

$$(\langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle + \langle v, m_1 - m_2\rangle^2)^2 \leq 8\epsilon \langle v, \Sigma_2 v\rangle^2,$$

applying Corollary F.1 Taking square roots of both sides,

$$\langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle \leq \langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle + \langle v, m_1 - m_2\rangle^2 \leq \sqrt{8\epsilon}\langle v, \Sigma_2 v\rangle,$$

and similarly, that

$$\langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle + \langle v, m_1 - m_2\rangle^2 \geq -\sqrt{8\epsilon}\langle v, \Sigma_2 v\rangle.$$

Additionally, we already showed that in Equation 33 that if $\chi^2(\mu\|\nu) < \epsilon$, then $\langle v, m_1 - m_2\rangle^2 < \epsilon\langle v, \Sigma_2 v\rangle$. Hence, we have that

$$\langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle \geq -\left(\sqrt{8\epsilon} + \epsilon\right)\langle v, \Sigma_2 v\rangle.$$

To summarize:

Lemma F.3. *If $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ are two distributions such that ν is log-concave and $\chi^2(\mu\|\nu) < \epsilon$ for some $\epsilon > 0$, then*

$$|\langle v, \Sigma_1 v\rangle - \langle v, \Sigma_2 v\rangle| \leq \left(\sqrt{8\epsilon} + \epsilon\right)\langle v, \Sigma_2 v\rangle,$$

where $C > 0$ is the universal constant from Corollary F.1 and $v \in \mathbb{R}^d$ is an arbitrary unit vector.

The following corollary is immediate upon carrying out some straightforward algebraic manipulations.

Corollary F.2. *Suppose that $\mu, \widehat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ are such that $\mathrm{Cov}(\mu) = \Sigma$ and $\mathrm{Cov}(\hat{\mu}) = \widehat{\Sigma}$, with μ log-concave. If*

$$\chi^2(\widehat{\mu}\|\mu) < \left(\sqrt{3 - (1+\delta)^{-1/2}} - \sqrt{2}\right)^2$$

for some $\delta > 0$, then

$$(1+\delta)^{-1/2}\Sigma \preceq \widehat{\Sigma} \preceq (1+\delta)^{1/2}\Sigma.$$

For example, for $\delta = 1$, we require that $\chi^2(\widehat{\mu}\|\mu) \approx 0.01$. For a more precise characterization of the first-order contribution, note that because $\sqrt{1+x} = 1 + \frac{x}{2} + O(x^2)$ for $|x| < 1$, we have that

$$\left(\sqrt{3 - (1+\delta)^{-1/2}} - \sqrt{2}\right)^2 = O(\delta^2)$$

for $|\delta| < 1$. Hence, for small $\delta > 0$, we only require that $\chi^2(\widehat{\mu}\|\mu) = O(\delta^2)$ in order to obtain the desired preconditioning guarantee.

F.3 Preconditioning with the Sample Covariance

At a high level, we require a strong matrix concentration bound to show that the sample covariance is a viable preconditioner at each stage of Gaussian cooling, and closeness in χ^2-divergence preserves subgaussianity, a sufficient property to apply such a concentration inequality. The argument in the next section then proceeds in two steps:

(1) If the target distribution μ at step k α-strongly log-concave and $(1/\alpha)$-smooth, then the approximate distribution, furnished by the user's sampler of choice after sufficiently many steps to satisfy the hypothesis of Corollary F.2, will be subgaussian, with subgaussian norm that depends on α.

(2) Having characterized the subgaussianity of the sampled distribution, we can apply an appropriate matrix concentration inequality for matrices with independent, centered, isotropic, and subgaussian rows to the sample covariance formed by i.i.d. samples, produced using parallel chains.

F.3.1 SUBGAUSSIANITY OF SAMPLED DISTRIBUTION

As described in the preceding section, it will be to our advantage to first quantitatively characterize the subgaussianity of a distribution $\widehat{\mu}$ such that $\chi^2(\widehat{\mu}\|\mu) < \epsilon$, where μ is α-strongly log-concave and $(1/\alpha)$-smooth.

Lemma F.4. *Suppose that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is an α-strongly log-concave distribution and $\widehat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is such that $\chi^2(\widehat{\mu}\|\mu) < \frac{1}{2}$. Then, for all unit vectors $v \in \mathbb{R}^d$,*

$$\mathbb{P}_{\widehat{\mu}}(|\langle v, Y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t) \leq 2\exp\left(\frac{-\alpha t^2}{8}\right)$$

for all $t \geq 0$ so that $Y \sim \widehat{\mu}$ is $(\sqrt{8/\alpha})$-subgaussian.

Proof. We already showed above that if μ is such a distribution, then for all unit vectors $v \in \mathbb{R}^d$,

$$\mathbb{P}_{\mu}(|\langle v, X - \mathbb{E}_{\mu}[X]\rangle| \geq t) \leq 2\exp\left(-\frac{\alpha t^2}{2}\right),$$

which aligns with the definition of $(\sqrt{2/\alpha})$-subgaussianity for an $\mathbb{R}^d$-valued random vector provided in (Vershynin, 2018, Section 3.4). Now, take $Y \sim \widehat{\mu}$ such that $\chi^2(\widehat{\mu}\|\mu) < \epsilon$ for some $\epsilon \in (0, 1/2)$. Fixing a unit vector $v \in \mathbb{R}^d$, we compute

$$\begin{aligned}
\mathbb{P}_{\widehat{\mu}}(|\langle v, Y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t) &= \int_{|\langle v, y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t} d\widehat{\mu}(y) \\
&= \int_{|\langle v, y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t} \frac{d\widehat{\mu}}{d\mu}(y) d\mu(y) \\
&\leq \left(\int_{|\langle v, y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t} \left(\frac{d\widehat{\mu}}{d\mu}(y)\right)^2 d\mu(y)\right)^{1/2} \left(\int_{|\langle v, y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t} d\mu(y)\right)^{1/2} \\
&= \left(\int_{|\langle v, y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t} \left(\frac{d\widehat{\mu}}{d\mu}(y)\right)^2 d\mu(y)\right)^{1/2} \left(\mathbb{P}_{\mu}(|\langle v, X - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t)\right)^{1/2}.
\end{aligned}$$

We can now analyze both terms separately, relating the first term to the fact that $\chi^2(\widehat{\mu}\|\mu) < \epsilon$ and bounding the second term using the subgaussianity of μ. For the first term, we write

$$\int_{|\langle v, y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t} \left(\frac{d\widehat{\mu}}{d\mu}(y)\right)^2 d\mu(y) \leq \int_{\mathbb{R}^d} \left(\frac{d\widehat{\mu}}{d\mu}(y)\right)^2 d\mu(y) = \chi^2(\widehat{\mu}\|\mu) + 1 < \frac{3}{2},$$

as long as $\epsilon \in (0, 1/2)$. On the other hand, because

$$\begin{aligned}
|\langle v, X - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| &\leq |\langle v, X - \mathbb{E}_{\mu}[X]\rangle| + |\langle v, \mathbb{E}_{\mu}[X] - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \\
&\leq |\langle v, X - \mathbb{E}_{\mu}[X]\rangle| + \|\mathbb{E}_{\mu}[X] - \mathbb{E}_{\widehat{\mu}}[Y]\|_2
\end{aligned}$$

we have that

$$\begin{aligned}
\mathbb{P}_{\mu}(|\langle v, X - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t) &\leq \mathbb{P}_{\mu}(|\langle v, X - \mathbb{E}_{\mu}[X]\rangle| + \|\mathbb{E}_{\mu}[X] - \mathbb{E}_{\widehat{\mu}}[Y]\|_2 \geq t) \\
&= \mathbb{P}_{\mu}(|\langle v, X - \mathbb{E}_{\mu}[X]\rangle| \geq t - \|\mathbb{E}_{\mu}[X] - \mathbb{E}_{\widehat{\mu}}[Y]\|_2).
\end{aligned}$$

Because $\chi^2(\widehat{\mu}\|\mu) < \epsilon$, we have that

$$\|\mathbb{E}_{\mu}[X] - \mathbb{E}_{\widehat{\mu}}[Y]\|_2 \leq \sqrt{\epsilon \lambda_{\max}(\Sigma_{\mu})} < \frac{1}{2}\sqrt{2\lambda_{\max}(\Sigma_{\mu})} \leq \frac{\sqrt{2}}{2}\alpha^{-1/2},$$

taking the supremum over all unit vectors $v \in \mathbb{R}^d$ in Equation 33. We also apply the fact that μ is α-strongly log-concave so that $\lambda_{\max}(\Sigma_{\mu}) \leq 1/\alpha$. Thus, by subgaussianity of μ, for $t > \frac{\sqrt{2}}{2}\alpha^{-1/2}$, we have that

$$\begin{aligned}
\mathbb{P}_{\mu}(|\langle v, X - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t) &\leq \mathbb{P}_{\mu}\left(|\langle v, X - \mathbb{E}_{\mu}[X]\rangle| \geq t - \frac{\sqrt{2}}{2}\alpha^{-1/2}\right) \\
&\leq 2\exp\left(-\frac{\alpha(t - \frac{\sqrt{2}}{2}\alpha^{-1/2})^2}{2}\right).
\end{aligned}$$

Now, we apply the fact that $(t-c)^2 \geq \frac{1}{2}t^2 - c^2$ for all $t, c \in \mathbb{R}$ to obtain a bound of

$$2\exp\left(-\frac{\alpha(\frac{1}{2}t^2 - \frac{1}{2}\alpha^{-1})}{2}\right) = 2\exp(1/4)\exp\left(-\frac{\alpha t^2}{4}\right).$$

Moreover, if $t \leq \frac{\sqrt{2}}{2}\alpha^{-1/2}$, then we can simply bound

$$\mathbb{P}_\mu(|\langle v, X - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t) \leq 1 \leq 2\exp(1/4)\exp\left(-\frac{\alpha t^2}{4}\right),$$

so the above bound holds for all $t > 0$. Returning to the original expression that we sought to bound to establish the subgaussianity of $Y \sim \widehat{\mu}$, we have that

$$\mathbb{P}_{\widehat{\mu}}(|\langle v, Y - \mathbb{E}_{\widehat{\mu}}[Y]\rangle| \geq t) \leq \sqrt{3}\exp(1/8)\exp\left(\frac{-\alpha t^2}{8}\right) \leq 2\exp\left(\frac{-\alpha t^2}{8}\right)$$

for all $t \geq 0$. □

Crucially, we can now safely assume that the sampled distribution $\widehat{\mu}$ is subgaussian, paving the way for good concentration of the sample covariance.

F.3.2 Concentration of Sample Covariance

Consider taking n independent, identically distributed samples $X_1, \ldots, X_n \sim \widehat{\mu}$, where we use $\widehat{\mu}$ to indicate that we are drawing samples from parallel ULMC chains. From these samples, we form the empirical mean

$$\widehat{m}_n := \frac{1}{n}\sum_{i=1}^n X_i$$

and the empirical covariance

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{m}_n)(X_i - \widehat{m}_n)^\top.$$

With this normalization, $\widehat{\Sigma}_n$ provides a *biased* estimate of the population covariance

$$\widehat{\Sigma} = \mathbb{E}_{\widehat{\mu}}[(X - \mathbb{E}_{\widehat{\mu}}[X])(X - \mathbb{E}_{\widehat{\mu}}[X])^\top].$$

However, we only seek a guarantee that the sample covariance satisfies

$$(1+\delta)^{-1/2}\widehat{\Sigma} \preceq \widehat{\Sigma}_n \preceq (1+\delta)^{1/2}\widehat{\Sigma}, \tag{37}$$

and the biased estimate above eases the analysis slightly. If Equation 37 holds, and

$$(1+\delta)^{-1/2}\Sigma \preceq \widehat{\Sigma} \preceq (1+\delta)^{1/2}\Sigma$$

holds for the same constant $\delta > 0$, we then have that

$$(1+\delta)^{-1}\Sigma \preceq (1+\delta)^{-1/2}\widehat{\Sigma} \preceq \widehat{\Sigma}_n \preceq (1+\delta)^{1/2}\widehat{\Sigma} \preceq (1+\delta)\Sigma,$$

allowing us to use the sample covariance as a preconditioner at step k of Gaussian cooling. The biased estimate of the sample covariance turns out to suffice for these purposes, with cleaner analysis. Moreover, observe that Equation 37 is equivalent to showing that

$$[(1+\delta)^{-1/2} - 1]\widehat{\Sigma} \preceq \widehat{\Sigma}_n - \widehat{\Sigma} \preceq [(1+\delta)^{1/2} - 1]\widehat{\Sigma}.$$

Equivalently, we have that

$$[(1+\delta)^{-1/2} - 1]I \preceq \widehat{\Sigma}^{-1/2}\widehat{\Sigma}_n\widehat{\Sigma}^{-1/2} - I \preceq [(1+\delta)^{1/2} - 1]I.$$

In turn, this holds true if

$$\|\widehat{\Sigma}^{-1/2}\widehat{\Sigma}_n\widehat{\Sigma}^{-1/2} - I\|_2 \leq 1 - (1+\delta)^{-1/2}. \tag{38}$$

We now aim to quantify how many samples are needed in order for Equation 38 to hold with high probability, to be made precise below.

To quantify the sample complexity of obtaining Equation 38, we begin by decomposing $\widehat{\Sigma}_n$ as

$$\begin{aligned}\widehat{\Sigma}_n &= \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{m} + \widehat{m} - \widehat{m}_n)(X_i - \widehat{m} + \widehat{m} - \widehat{m}_n)^\top \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{m})(X_i - \widehat{m})^\top - (\widehat{m}_n - \widehat{m})(\widehat{m}_n - \widehat{m})^\top,\end{aligned}$$

where $\widehat{m} := \mathbb{E}_{\widehat{\mu}}[X]$. Hence, we have that

$$\widehat{\Sigma}^{-1/2}\widehat{\Sigma}_n\widehat{\Sigma}^{-1/2} = \frac{1}{n}\sum_{i=1}^n \widehat{\Sigma}^{-1/2}(X_i - \widehat{m})(X_i - \widehat{m})^\top\widehat{\Sigma}^{-1/2} - \widehat{\Sigma}^{-1/2}(\widehat{m}_n - \widehat{m})(\widehat{m}_n - \widehat{m})^\top\widehat{\Sigma}^{-1/2} \tag{39}$$

Define

$$Z_i := \widehat{\Sigma}^{-1/2}(X_i - \widehat{m}), \quad i \in [n].$$

Mirroring the proof of (Vershynin, 2018, 4.7.1), if we can show that the $Z_i \in \mathbb{R}^d$ are κ-subgaussian vectors for some $\kappa > 0$, then we will be able to show that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - I\right\| > \kappa^2 \max\{\delta(t), \delta(t)^2\}\right) \le 2\exp(-t^2),$$

where

$$\delta(t) := C\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$$

for some absolute constant $C > 0$.

To proceed in this manner, we observe that for any unit vector $v \in \mathbb{R}^d$ and $t \ge 0$,

$$\begin{aligned}\mathbb{P}(|\langle Z_i, v\rangle| \ge t) = \mathbb{P}\left(\left|\langle \widehat{\Sigma}^{-1/2}(X_i - \widehat{m}), v\rangle\right| \ge t\right) &= \mathbb{P}\left(\left|\langle X_i - \widehat{m}, \widehat{\Sigma}^{-1/2}v\rangle\right| \ge t\right) \\ &= \mathbb{P}\left(\left|\left\langle X_i - \widehat{m}, \frac{\widehat{\Sigma}^{-1/2}v}{\|\widehat{\Sigma}^{-1/2}v\|_2}\right\rangle\right| \ge \frac{t}{\|\widehat{\Sigma}^{-1/2}v\|_2}\right) \\ &\le 2\exp\left(-\frac{\alpha t^2}{8\|\widehat{\Sigma}^{-1/2}v\|_2^2}\right) \\ &\le 2\exp\left(-\frac{\alpha t^2}{8\lambda_{\max}(\widehat{\Sigma}^{-1})}\right) \\ &= 2\exp\left(-\frac{\alpha\lambda_{\min}(\widehat{\Sigma})t^2}{8}\right),\end{aligned}$$

applying Lemma F.4 to the distribution $\widehat{\mu}$ and the fact that $\|\widehat{\Sigma}^{-1/2}v\|_2^2 = v^\top\widehat{\Sigma}^{-1}v \le \lambda_{\max}(\widehat{\Sigma}^{-1})$ for any unit vector $v \in \mathbb{R}^d$. Returning to Lemma F.3, as long as $\varepsilon < (\sqrt{3} - \sqrt{2})^2$, we have that

$$\langle v, \widehat{\Sigma}v\rangle \ge (1 - \sqrt{8\varepsilon} - \varepsilon)\langle v, \Sigma v\rangle \ge (1 - \sqrt{8\varepsilon} - \varepsilon)\lambda_{\min}(\Sigma).$$

Hence, if μ is $(1/\alpha)$-smooth (as will be the case for Gaussian cooling), it follows that

$$\lambda_{\min}(\widehat{\Sigma}) \ge (1 - \sqrt{8\varepsilon} - \varepsilon)\alpha.$$

If we assume, say, that $\varepsilon < \frac{1}{16}$, then $1 - \sqrt{8\varepsilon} - \varepsilon > \frac{1}{5}$, and we may conclude that

$$\mathbb{P}(|\langle Z_i, v\rangle| \ge t) \le 2\exp\left(-\frac{\alpha^2 t^2}{40}\right),$$

showing that once we have placed the sample distribution in isotropic position, the resulting random vectors Z_i are $(\sqrt{40}/\alpha)$-subgaussian, as long as $\chi^2(\widehat{\mu}\|\mu) < \frac{1}{16}$. We may now apply the relevant subgaussian matrix concentration bound from Vershynin (2018) directly, which we include for the sake of completeness.

Theorem F.5 (Theorem 4.6.1 in Vershynin (2018)). *Let $A \in \mathbb{R}^{n\times d}$ be a random matrix with independent, mean-zero, κ-subgaussian, isotropic rows A_i. Then, for any $t \geq 0$, we have that*

$$\mathbb{P}\left(\left\|\frac{1}{n}A^\top A - I\right\| > \kappa^2 \max\{\delta(t), \delta(t)^2\}\right) \leq 2\exp(-t^2), \tag{40}$$

where

$$\delta(t) := C\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$$

for some absolute constant $C > 0$.

Indeed, taking A to the $n \times d$ matrix with rows Z_i, we have that

$$\frac{1}{n}A^\top A = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top,$$

and it therefore follows immediately from Theorem F.5 that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - I\right\| > \frac{40}{\alpha^2}\max\{\delta(t), \delta(t)^2\}\right) \leq 2\exp(-t^2).$$

In fact, with some simple algebraic manipulations (c.f. Remark 4.7.3 and Exercise 4.49 in Vershynin (2018)), one can show that Equation 40 implies that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - I\right\| > \frac{40}{\alpha^2}\left(\frac{d+u}{n} + \sqrt{\frac{d+u}{n}}\right)\right) \leq 2\exp(-u), \quad u \geq 0.$$

For notational simplicity, let $C(\delta) = 1 - (1+\delta)^{-1/2}$ as above. We aim to choose u and n so that

$$\frac{40}{\alpha^2}\left(\frac{d+u}{n} + \sqrt{\frac{d+u}{n}}\right) < \frac{1}{2}C(\delta),$$

while still maintaining a high probability guarantee. Now, if we consider the inequality

$$\frac{40}{\alpha^2}(s + \sqrt{s}) < \frac{1}{2}C(\delta),$$

we require that $\sqrt{s} < \frac{1}{80}\alpha^2 C(\delta)$, which implies that choosing $s = \Theta(\alpha^4\delta^2)$ will suffice. Moreover, we require that $\frac{d+u}{n} \leq s$, which holds if and only if $u \leq sn - d$.

Taking $u = \frac{s}{2}n$, and rearranging, we will have the desired bound as long as

$$n \geq \frac{2d}{s} = \Omega\left(\frac{d}{\alpha^4\delta^2}\right).$$

In particular, for this choice of n, we will have that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - I\right\| > \frac{1}{2}C(\delta)\right) \leq 2\exp\left(-\frac{s}{2}n\right) = 2\exp(-C_1\alpha^4\delta^2 n),$$

where $C_1 > 0$ is an absolute constant.

We now turn our attention to the second term, which is slightly easier to analyze. Take

$$R_n := \widehat{\Sigma}^{-1/2}(\widehat{m}_n - \widehat{m}) = \frac{1}{n}\sum_{i=1}^n \widehat{\Sigma}^{-1/2}(X_i - \widehat{m}) = \frac{1}{n}\sum_{i=1}^n Z_i,$$

so that

$$\|\widehat{\Sigma}^{-1/2}(\widehat{m}_n - \widehat{m})(\widehat{m}_n - \widehat{m})^\top\widehat{\Sigma}^{-1/2}\|_2 = \|R_n R_n^\top\|_2 = \|R_n\|_2^2.$$

Noting that the Z_i are independent, centered, and isotropic by construction, it follows that

$$Y_n := \sqrt{n}R_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$$

is centered and isotropic. Below, we take the Orlicz norm associated to the function $\Phi(u) = \exp(u^2) - 1$ to be $\|X\|_{\psi_2} := \inf\{\kappa > 0 : \mathbb{E}[\exp(X^2/\kappa^2)] \leq 2\}$, as in (Vershynin, 2018, Definition 2.6.4). Then, if X is κ-subgaussian, $\|X\|_{\psi_2} \leq c\kappa$ for some universal $c > 0$.

Now, because the Z_i are independent, centered, κ-subgaussian random variables, (Vershynin, 2018, Proposition 2.7.1) reveals that for some universal $C' > 0$,

$$\|Y_n\|_{\psi_2}^2 = \frac{1}{n}\left\|\sum_{i=1}^n Z_i\right\|_{\psi_2}^2 \leq \frac{C'}{n}\sum_{i=1}^n \|Z_i\|_{\psi_2}^2 = C'c^2\kappa^2$$

so that Y_n itself is $\Theta(\kappa)$-subgaussian. To proceed, we require a standard ϵ-net argument.

Lemma F.6. *With R_n and $C(\delta)$ defined as above, if $n = \Omega\left(\frac{d}{\alpha^4\delta^2}\right)$, then*

$$\mathbb{P}\left(\|R_n\|_2^2 > \frac{1}{2}C(\delta)\right) \leq \exp(-C_2\alpha^2\delta n).$$

Proof. Choose $\epsilon = 1/2$. By (Vershynin, 2018, Corollary 4.2.11), we can construct an ϵ-net $\mathcal{N}$ of $\mathbb{S}^{d-1}$ of size $|\mathcal{N}| \leq 5^d$. Choosing $v \in \mathbb{R}^d$ such that $\|R_n\|_2 = \langle R_n, v\rangle$, we may find $v' \in \mathcal{N}$ such that $\|v - v'\|_2 \leq \frac{1}{2}$. Thus,

$$|\langle R_n, v - v'\rangle| \leq \|R_n\|_2\|v - v'\|_2 \leq \frac{1}{2}\|R_n\|_2,$$

implying that

$$|\langle R_n, v'\rangle| \geq |\langle R_n, v\rangle| - |\langle R_n, v - v'\rangle| \geq \frac{1}{2}\|R_n\|_2.$$

In particular, $\|R_n\|_2 \leq 2\max_{v\in\mathcal{N}}|\langle R_n, v\rangle|$. Now, as R_n is $\Theta(\kappa/\sqrt{n})$-subgaussian, for any fixed $v \in \mathcal{N}$, we have that

$$\mathbb{P}(|\langle R_n, v\rangle| > t) \leq 2\exp\left(-\frac{c_0 nt^2}{\kappa^2}\right)$$

for some absolute constant $c_0 > 0$. Via a union bound over all $v \in \mathcal{N}$, we have that

$$\mathbb{P}(\|R_n\|_2 > t) \leq \mathbb{P}\left(\max_{v\in\mathcal{N}}|\langle R_n, v\rangle| > \frac{t}{2}\right) \leq 2\cdot 5^d\exp\left(-\frac{c_0 nt^2}{\kappa^2}\right) = 2\exp\left(-\frac{c_0 nt^2}{\kappa^2} + d\log 5\right)$$

taking c_0 smaller if necessary. Hence, for $t = \sqrt{C(\delta)/2}$, it follows that

$$\mathbb{P}(\|R_n\|_2^2 > C(\delta)) \leq 2\exp\left(-\frac{c_0 nC(\delta)}{\kappa^2} + d\log 5\right).$$

In particular, there exists universal $C_2 > 0$ such that for $n = \Omega\left(\frac{d}{\alpha^2\delta}\right)$, we have that

$$\mathbb{P}\left(\|R_n\|_2^2 > \frac{1}{2}C(\delta)\right) \leq 2\exp(-C_2\alpha^2\delta n),$$

using the fact that $\kappa = \Theta(1/\alpha)$ and $C(\delta) = \Theta(\delta)$ for δ sufficiently small. □

Combining our estimates of both terms in Equation 39, we have that for $n = \Omega\left(\frac{d}{\alpha^4\delta^2}\right)$,

$$\begin{aligned}\mathbb{P}\left(\|\widehat{\Sigma}^{-1/2}\widehat{\Sigma}_n\widehat{\Sigma}^{-1/2} - I\|_2 > C(\delta)\right) &\leq \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_iZ_i^\top - I\right\| > \frac{1}{2}C(\delta)\right) + \mathbb{P}\left(\|R_n\|_2^2 > \frac{1}{2}C(\delta)\right)\\ &\leq 2\exp(-C_1\alpha^4\delta^2 n) + 2\exp(-C_2\alpha^2\delta n)\\ &\leq 4\exp(-C_3\alpha^4\delta^2 n)\end{aligned}$$

for some universal $C_3 > 0$. This lemma is sufficient to conclude the main result of this section.

Theorem F.7. *Suppose that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is an α-strongly log-concave and $(1/\alpha)$-smooth measure on $\mathbb{R}^d$ with $\Sigma := \mathrm{Cov}(\mu)$ and that $\widehat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is such that for fixed $\delta \in (0,1)$,*

$$\chi^2(\widehat{\mu}\|\mu) < \min\left\{\left(\sqrt{3-(1+\delta)^{-1/2}} - \sqrt{2}\right)^2, \frac{1}{16}\right\}.$$

There exist universal constants $C_1, C_2 > 0$ such that if $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \widehat{\mu}$ and

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n \left(X_i - \frac{1}{n}\sum_{j=1}^n X_j\right)\left(X_i - \frac{1}{n}\sum_{j=1}^n X_j\right)^\top$$

is the corresponding (biased) sample covariance, formed from at least $n = \frac{C_1 d}{\alpha^4\delta^2}$ samples, then

$$(1+\delta)^{-1}\Sigma \preceq \widehat{\Sigma}_n \preceq (1+\delta)\Sigma$$

with probability at least $1 - 4\exp(-C_2\alpha^4\delta^2 n)$.

F.4 Concentration in the translation-invariant Case

As in the general case in the preceding section, we can reduce to well-established concentration inequalities for subgaussian random variables. However, we only need to establish that the d entries of the vector $\widehat{\sigma}$, defined in Equation 30, concentrate well. For completeness, recall that in the translation-invariant case, the target covariance Σ is diagonalized by a discrete Fourier transform in the sense that

$$\Sigma = \mathcal{F}\,\mathrm{diag}(\sigma)\mathcal{F}^*$$

for some $\sigma \in \mathbb{R}^d$. We then must establish that the entries of our estimate $\widehat{\sigma}$ satisfy

$$\left|\frac{\widehat{\sigma}_i}{\sigma_i} - 1\right| \le C(\delta), \quad i \in [d], \tag{41}$$

simultaneously, as seen in Equation 29. To this end, we may show that Equation 41 holds individually for each $i \in [d]$ and then perform a union bound over the d entries of $\widehat{\sigma}$. We begin by writing

$$\widehat{\sigma}_i = \frac{1}{n}\sum_{j=1}^n |\langle e_i, \mathcal{F}(x_j - \widehat{m}_n)\rangle|^2 = \frac{1}{n}\sum_{j=1}^n |\langle x_j - \widehat{m}_n, v_i\rangle|^2,$$

where $v_i = \mathcal{F}^* e_i \in \mathbb{C}^d$ (taking the physics convention, the inner product on $\mathbb{C}^d$ is linear in its second argument). Note that $\|v_i\| = 1$, as $\mathcal{F}^*$ is an isometry. On the other hand, if $X \sim \mu$ is drawn from the target distribution, we have that

$$\begin{aligned}\sigma_i = \langle e_i, \mathcal{F}^*\Sigma\mathcal{F}e_i\rangle &= \mathbb{E}[\langle e_i, \mathcal{F}^*(X - \mathbb{E}[X])(X-\mathbb{E}[X])^\top \mathcal{F}e_i\rangle] \\ &= \mathbb{E}[|\langle X - \mathbb{E}[X], \mathcal{F}^* e_i\rangle|^2] \\ &= \mathbb{E}[|\langle X - \mathbb{E}[X], v_i\rangle|^2].\end{aligned}$$

Splitting v_i into its real and imaginary components, we observe that if μ is an $(1/\alpha)$-smooth distribution, then

$$\mathbb{E}[|\langle X - \mathbb{E}[X], v_i\rangle|^2] = \mathbb{E}[|\langle X - \mathbb{E}[X], \Re(v_i)\rangle|^2] + \mathbb{E}[|\langle X - \mathbb{E}[X], \Im(v_i)\rangle|^2] \ge \alpha,$$

invoking Corollary A.1 to obtain a lower bound on the covariance of μ. Hence, the righthand side of Equation 41 is well-defined, and in fact, it suffices to show that

$$|\widehat{\sigma}_i - \sigma_i| \le \alpha C(\delta), \quad i \in [d].$$

Fixing $v \in \mathbb{R}^d$ to be a vector with $\|v\| \le 1$ (as is the case for both $\Re(v_i)$ and $\Im(v_i)$ above), it then suffices to obtain a sample complexity estimate such that

$$\left|\frac{1}{n}\sum_{j=1}^n |\langle x_j - \widehat{m}_n, v\rangle|^2 - \mathbb{E}[|\langle X - \widehat{m}, v\rangle|^2]\right| \le \frac{\alpha C(\delta)}{2}$$

with high probability, where $\widehat{m} = \mathbb{E}_{\widehat{\mu}}[X]$. The extra factor of two in the denominator arises because we will eventually need to apply such an inequality twice, to the real and imaginary components of the vector $v_i = \mathcal{F}^* e_i$. Now, for each $j \in [n]$, define

$$Z_j := \langle x_j - \widehat{m}, v \rangle. \tag{42}$$

We also take

$$\widehat{Z}_n := \frac{1}{n}\sum_{j=1}^n Z_j = \frac{1}{n}\sum_{j=1}^n \langle x_j - \widehat{m}, v\rangle = \langle \widehat{m}_n - \widehat{m}, v\rangle. \tag{43}$$

Now, it is easy to see that

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^n |\langle x_j - \widehat{m}_n, v\rangle|^2 &= \frac{1}{n}\sum_{j=1}^n |Z_j - \widehat{Z}_n|^2 \\ &= \frac{1}{n}\sum_{j=1}^n Z_j^2 - \widehat{Z}_n^2.\end{aligned}$$

If $Z \overset{d}{=} \langle X - \widehat{m}, v\rangle$, then

$$\mathbb{E}[|\langle X - \widehat{m}, v\rangle|^2] = \mathbb{E}[Z^2],$$

and have reduced to showing that for $Z_1, \dots, Z_n \overset{i.i.d.}{\sim} \mathrm{Law}(Z)$,

$$\left|\frac{1}{n}\sum_{j=1}^n Z_j^2 - \mathbb{E}[Z^2]\right| + \widehat{Z}_n^2 \le \frac{\alpha C(\delta)}{2} \tag{44}$$

via triangle inequality. In a very similar manner to the general case, the first term can be handled via an appropriate concentration inequality for sub-exponential random variables.

Lemma F.8. *Suppose that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is an α-strongly log-concave and $(1/\alpha)$-smooth measure on $\mathbb{R}^d$ and that $\widehat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is such that for fixed $\delta \in (0,1)$,*

$$\chi^2(\widehat{\mu}\|\mu) < \min\left\{\left(\sqrt{3 - (1+\delta)^{-1/2}} - \sqrt{2}\right)^2, \frac{1}{16}\right\}.$$

There exists a universal constant $C > 0$ such that if $X_1, \dots, X_n \overset{i.i.d.}{\sim} \widehat{\mu}$, then for $Z_1, \dots, Z_n$ as defined in Equation 42 and $Z \sim \mathrm{Law}(Z_1)$,

$$\left|\frac{1}{n}\sum_{j=1}^n Z_j^2 - \mathbb{E}[Z^2]\right| \le \frac{\alpha C(\delta)}{4}$$

with probability at least $1 - 2\exp(-Cn\alpha^4\delta^2)$.

Proof. We write

$$\left|\frac{1}{n}\sum_{j=1}^n Z_j^2 - \mathbb{E}[Z^2]\right| = \left|\sum_{j=1}^n \frac{1}{n}(Z_j^2 - \mathbb{E}[Z^2])\right|.$$

Now, the random variables $Y_j := Z_j^2 - \mathbb{E}[Z^2]$ are independent, mean-zero, and $O(1/\alpha)$-subexponential, assuming that $\widehat{\mu}$ is the same distribution discussed in Lemma F.4. Indeed, $Z_j = \langle X_j - \widehat{m}, v\rangle$ is shown to be $(\sqrt{8/\alpha})$-subgaussian in said lemma, and a combination of (Vershynin, 2018, Lemma 2.8.5 and Equation (2.26)) show that the centered variance $Z_j^2 - \mathbb{E}[Z^2]$ is then subexponential with

$$\|Z_j^2 - \mathbb{E}[Z^2]\|_{\psi_1} \le C\|Z_j^2\|_{\psi_1} = C\|Z_j\|_{\psi_2}^2 = \frac{8C}{\alpha},$$

where $C > 0$ is a universal constant and the (subexponential) Orlicz norm $\|\cdot\|_{\psi_1}$ is defined by $\|X\|_{\psi_1} := \inf\{K > 0 : \mathbb{E}[\exp(|X|/K)] \leq 2\}$. This observation allows us to apply the subexponential Bernstein inequality from (Vershynin, 2018, Corollary 2.9.2), which yields

$$\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{1}{n}(Z_j^2 - \mathbb{E}[Z^2])\right| \geq \frac{\alpha C(\delta)}{2}\right) \leq 2\exp\left[-Cn\min\left\{\frac{\alpha^2 C(\delta)^2}{\|Z_j^2 - \mathbb{E}[Z^2]\|_{\psi_1}^2}, \frac{\alpha C(\delta)}{\|Z_j^2 - \mathbb{E}[Z^2]\|_{\psi_1}}\right\}\right]$$
$$\leq 2\exp\left(-Cn\alpha^4 C(\delta)^2\right),$$

where $C > 0$ is some universal constant (taken larger if necessary in the second inequality). The conclusion then follows because $C(\delta) = \Theta(\delta)$. □

If $\varepsilon > 0$ is fixed, then as long as we take

$$n = \Omega\left(\frac{\log(d/\varepsilon)}{\alpha^4\delta^2}\right) \tag{45}$$

we can ensure that

$$\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{1}{n}(Z_j^2 - \mathbb{E}[Z^2])\right| \geq \frac{\alpha C(\delta)}{4}\right) \leq \frac{\varepsilon}{4d}.$$

Turning to the second term in Equation 44, we observe that

$$\mathbb{P}\left(\widehat{Z}_n^2 \geq \frac{\alpha C(\delta)}{4}\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n} Z_j\right| \geq \sqrt{\frac{\alpha C(\delta)}{4}}\right).$$

As noted in the proof Lemma F.8, the Z_j are $(\sqrt{8/\alpha})$-subgaussian. As a result, we can straightforwardly apply Hoeffding's inequality. From (Vershynin, 2018, Theorem 2.7.3), we know that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n} Z_j\right| \geq \sqrt{\frac{\alpha C(\delta)}{4}}\right) \leq 2\exp\left(-cn^2\frac{\alpha C(\delta)}{\sum_{j=1}^{n}\|Z_j\|_{\psi_2}^2}\right)$$
$$= 2\exp(-cn\alpha^2\delta),$$

for some universal $c > 0$, recalling that $\|Z_j\|_{\psi_2} = O(\alpha^{-1/2})$. Hence, taking n samples, where n is given as in Equation 45, suffices to ensure that

$$\widehat{Z}_n^2 \leq \frac{\alpha C(\delta)}{4}$$

with probability at least $1 - \frac{\varepsilon}{4d}$. We have in particular shown the following:

Lemma F.9. *Suppose that $\mu, \widehat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ are as in Lemma F.8. Then, for any fixed $v \in \mathbb{R}^d$ with $\|v\| \leq 1$ and $\varepsilon > 0$, as long as n is as in Equation 45, we have that*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}|\langle x_j - \widehat{m}_n, v\rangle|^2 - \mathbb{E}[|\langle X - \widehat{m}, v\rangle|^2]\right| \geq \frac{\alpha C(\delta)}{2}\right) \leq \frac{\varepsilon}{2d}.$$

We conclude, in the translation-invariant setting, by simply taking a union bound over the $2d$ vectors $\Re(v_i)$ and $\Im(v_i)$, where $v_i = \mathcal{F}^* e_i$ as identified above.

Corollary F.3. *Suppose that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is an α-strongly log-concave and $(1/\alpha)$-smooth measure on $\mathbb{R}^d$ with translation-invariant covariance $\Sigma = \mathcal{F}\,\mathrm{diag}(\sigma)\mathcal{F}^*$ and $\widehat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is such that*

$$\chi^2(\widehat{\mu}\|\mu) < \min\left\{\left(\sqrt{3 - (1+\delta)^{-1/2}} - \sqrt{2}\right)^2, \frac{1}{16}\right\}.$$

If $n = O\left(\frac{\varepsilon \log d}{\alpha^4\delta^2}\right)$ and $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \widehat{\mu}$, then $\widehat{\sigma}$ from Equation 30 satisfies

$$\|\widehat{\sigma} \odot \sigma^{-1} - \mathbf{1}\|_\infty \leq C(\delta).$$

with probability at least $1 - \varepsilon$.

Proof. As noted above, this claim will follow if we can show that $|\widehat{\sigma}_i - \sigma_i| \leq \alpha C(\delta)$ with probability at least $1 - \frac{\varepsilon}{d}$ for each $i \in [d]$, with the final claim then following directly from a union bound over all $i \in [d]$. If $v_i = \mathcal{F}^* e_i \in \mathbb{C}^d$, then

$$\begin{aligned}
|\sigma_i - \widehat{\sigma}_i| &= \left| \frac{1}{n} \sum_{j=1}^{n} |\langle x_j - \widehat{m}_n, v_i \rangle|^2 - \mathbb{E}[|\langle X - \widehat{m}, v_i \rangle|^2] \right| \\
&\leq \left| \frac{1}{n} \sum_{j=1}^{n} |\langle x_j - \widehat{m}_n, \Re(v_i) \rangle|^2 - \mathbb{E}[|\langle X - \widehat{m}, \Re(v_i) \rangle|^2] \right| \\
&\quad + \left| \frac{1}{n} \sum_{j=1}^{n} |\langle x_j - \widehat{m}_n, \Im(v_i) \rangle|^2 - \mathbb{E}[|\langle X - \widehat{m}, \Im(v_i) \rangle|^2] \right|.
\end{aligned}$$

By Lemma F.9, both summands on the righthand side of the preceding inequality are at most $\frac{\alpha C(\delta)}{2}$ with probability at least $1 - \frac{\varepsilon}{2d}$ given our choice of n. Hence, for each $i \in [d]$, we have that

$$\mathbb{P}(|\sigma_i - \widehat{\sigma}_i| \geq \alpha C(\delta)) \leq \frac{\varepsilon}{d}$$

by union bounding over the event that either summand exceeds the threshold $\alpha C(\delta)$. □

G Additional Lattice ϕ^4 Proofs and Experiments

In this appendix, we expand upon and provide proofs of several facts about lattice ϕ^4 distribution, as introduced in Section 4. Additionally, we determine the sensitivity of translation-invariant Gaussian cooling to the number of ULMC steps per chain N and the number of chains per stage n in Figure 7.

G.1 Additional Proofs

In full generality, the lattice ϕ^4 Hamiltonian takes the following form:

$$H(\phi) := \frac{1}{4} \sum_{x,y \in \Lambda} \beta_{xy} |\phi_x - \phi_y|^2 + h \sum_{x \in \Lambda} \phi_x + \sum_{x \in \Lambda} w(\phi_x) \tag{46}$$

Above, $\beta_{xy} = \beta_{yx} \geq 0$ is a symmetric coupling constant for each pair $(x, y) \in \Lambda \times \Lambda$, $h \in \mathbb{R}$ is an external magnetic field, and $w : \mathbb{R} \to \mathbb{R}$ is a scalar potential. For the lattice ϕ^4 model, we then take the scalar potential w to be

$$w(\phi_x) = \frac{\lambda}{4} \phi_x^4 + \frac{m}{2} \phi_x^2.$$

First, to understand why the lattice ϕ^4 model provides a useful testbed for Gaussian cooling, we compute the Hessian of the Hamiltonian:

$$\nabla^2 H(\phi) = (\Delta_\beta + mI) + 3\lambda \operatorname{diag}(\phi^2).$$

where ϕ^2 is interpreted elementwise. As long as $m > 0$ and the coupling constant β is also non-negative, having specialized to the case that Δ_β is a weighted graph Laplacian, it follows immediately that

$$\nabla^2 H(\phi) \succeq mI, \quad \forall \phi \in \mathbb{R}^\Lambda.$$

In other words, the H is m-strongly convex. However, it is *not* the case that H is L-smooth for any $L > 0$. Indeed, because the Hessian of the quartic term in the lattice ϕ^4 potential remains dependent on ϕ, taking $|\phi| \to \infty$ allows us to deduce that for any $L > 0$, there exists $\phi \in \mathbb{R}^\Lambda$ such that

$$\nabla^2 H(\phi) \succeq 3\lambda \operatorname{diag}(\phi^2) \succeq LI.$$

In order for this example to fit into our framework, we therefore need to truncate the scalar ϕ^4 potential to a quadratic potential past some threshold. More specifically, we replace the scalar potential w defined above with a truncated scalar potential of the form:

$$w_R(\phi_x) := \begin{cases} \frac{\lambda}{4}R^4 + \frac{m}{2}R^2 - (\lambda R^3 + mR)(\phi_x + R) + \frac{1}{2}(3\lambda R^2 + m)(\phi_x + R)^2 & \phi_x < -R, \\ w(\phi_x) & |\phi_x| \le R, \\ \frac{\lambda}{4}R^4 + \frac{m}{2}R^2 + (\lambda R^3 + mR)(\phi_x - R) + \frac{1}{2}(3\lambda R^2 + m)(\phi_x - R)^2 & \phi_x > R. \end{cases} \tag{47}$$

By construction, the truncated scalar potential w_R is twice continuously differentiable, and we have that

$$w_R''(\phi_x) = \begin{cases} w''(\phi_x) & |\phi_x| \le R, \\ 3\lambda R^2 + m & |\phi_x| > R. \end{cases}$$

Because $|w''(\phi_x)| \le 3\lambda R^2 + m$ for $|\phi_x| \le R$ and the contribution of the scalar potential to the Hessian of the Hamiltonian is diagonal, the truncated Hamiltonian $H_R(\phi)$ now satisfies, at the very least,

$$\Delta_\beta + mI \preceq \nabla^2 H_R(\phi) \preceq \Delta_\beta + (3\lambda R^2 + m)I.$$

for all $\phi \in \mathbb{R}^\Lambda$. If $\lambda_1 = \lambda_{\min}(\Delta_\beta)$ and $\lambda_n = \lambda_{\max}(\Delta_\beta)$, then we conclude that the measure $\pi_R(\phi) \propto \exp(-H_R(\phi))$ induced by the truncated Hamiltonian is indeed strongly log-concave, as

$$\nabla^2 H_R(\phi) \preceq (\lambda_n + 3\lambda R^2 + m)I.$$

We can guarantee that Assumptions A.4–A.5 are satisfied with the following parameters:

$$\begin{aligned} \alpha &= \left(\frac{3\lambda R^2}{\lambda_{\min}(\Delta_\beta) + m} + 1 \right)^{-1}, \\ \mu &= \lambda_{\min}(\Delta_\beta) + m, \\ L &= \lambda_{\max}(\Delta_\beta) + m + 3\lambda R^2. \end{aligned} \tag{48}$$

To show that the covariance is block circulant with circulant blocks, it suffices to show that for all $x, y \in \Lambda_d$

$$\mathbb{E}_{\phi\sim\pi}[\phi_x \phi_y] - \mathbb{E}_{\phi\sim\pi}[\phi_x]\mathbb{E}_{\phi\sim\pi}[\phi_y] = C([x-y]),$$

where $C : [d-1] \times [d-1] \to \mathbb{R}$ is a function of $[x-y] := (x_1 - y_1 \bmod d, x_2 - y_2 \bmod d)$. Indeed, if this holds for all $x, y \in \Lambda_d$, then we can partition $\Lambda_d \times \Lambda_d$ into $d \times d$ blocks by ordering the lattices sites (x, y) in ascending order as $(0,0), (0,1), \ldots, (0, d-1), \ldots, (d-1, 0,), \ldots, (d-1, d-1)$. It is then evident that the resulting covariance matrix $\Sigma = \mathrm{Cov}(\pi) \in \mathbb{R}^{d^2 \times d^2}$ is block circulant with $d \times d$ circulant blocks. Consequently, it is diagonalized by a discrete Fourier transform, exactly as in Section 3.1.

Lemma G.1. *The entries Σ_{xy} of the covariance $\Sigma := \mathrm{Cov}(\pi)$ of the (truncated) lattice ϕ^4 measure depend only on the increments $[x - y] := (x_1 - y_1 \bmod d, x_2 - y_2 \bmod d)$.*

Proof. Fix some $x, y \in \Lambda$. Define the linear shift operator $P_y : \mathbb{R}^\Lambda \to \mathbb{R}^\Lambda$ by $(P_y\phi)_x = \phi_{[x+y]}$. We observe that P_y, which simply permutes the entries of the periodic lattice Λ by constant shift of y, is unitary, with $(P_y^\dagger \phi)_x = \phi_{[x-y]}$. Now, let $\psi := P_y \phi$.

As a result, if we consider the original lattice ϕ^4 potential in Equation 11, we have that

$$\sum_{x\in\Lambda} w(\psi_x) = \sum_{x\in\Lambda} w(\phi_{[x+y]}) = \sum_{z\in\Lambda} w(\phi_z),$$

with the same holding true upon replacing w with the truncated potential w_R in Equation 47. We also compute

$$\langle \phi, \Delta_\beta \phi \rangle = \langle P_y^\dagger \psi, \Delta_\beta P_y^\dagger \psi \rangle = \langle \psi, P_y \Delta_\beta P_y^\dagger \psi \rangle.$$

Now, by definition of the Laplacian, we observe that for any $v, w \in \Lambda$,

$$\begin{aligned}\langle e_v, P_y \Delta_\beta P_y^\dagger e_w \rangle = \langle P_y^\dagger e_v, \Delta_\beta P_y^\dagger e_w \rangle &= \langle e_{[v+y]}, \Delta_\beta e_{[w+y]} \rangle \\ &= \begin{cases} \beta & [v+y] \sim [w+y], \\ 0 & \text{otherwise}, \end{cases} \\ &= (\Delta_\beta)_{vw}.\end{aligned}$$

Indeed, in the d-periodic two-dimensional lattice, $[v+y] \sim [w+y]$ if and only if $v \sim w$. Hence, $\langle \phi, \Delta_\beta \phi \rangle = \langle \psi, \Delta_\beta \psi \rangle$. With both of these facts together, we see that

$$H(\phi) = \frac{1}{2}\langle \phi, \Delta_\beta \phi \rangle + \sum_{x \in \Lambda} w(\phi_x) = \frac{1}{2}\langle \psi, \Delta_\beta \psi \rangle + \sum_{x \in \Lambda} w(\psi_x) = H(\psi)$$

is translation-invariant, with the same holding true for the truncated potential.

We can now directly compute each entry of the covariance to establish the claimed translation invariance. As noted above, this same computation goes through with the original potential and the truncated potential, so we only consider the former. Letting Z denote the constant normalization of $\pi(\phi) \propto \frac{1}{Z}\exp(-H(\phi))$ and μ^Λ denote the Lebesgue measure on $\mathbb{R}^\Lambda$, we first compute

$$\begin{aligned}\mathbb{E}_\pi[\phi_x \phi_y] &= \frac{1}{Z}\int_{\mathbb{R}^\Lambda} \phi_x \phi_y \exp(-H(\phi)) d\mu^\Lambda(\phi) \\ &= \frac{1}{Z}\int_{\mathbb{R}^\Lambda} (P_y^\dagger \psi)_x (P_y^\dagger \psi)_y \exp(-H(P_y^\dagger \psi)) d\mu^\Lambda(\psi) \\ &= \frac{1}{Z}\int_{\mathbb{R}^\Lambda} \psi_{[x-y]} \psi_0 \exp(-H(\psi)) d\mu^\Lambda(\psi) \\ &= \mathbb{E}_\pi[\phi_{[x-y]}\phi_0].\end{aligned}$$

Above, we change variables to $\psi = P_y \phi$, use the fact that $H(\phi) = H(P_y^\dagger \psi) = H(\psi)$, and note that P_y is a unitary and hence the Jacobian of the map $\phi \mapsto P_y \phi$ is one. An identical computation shows that $\mathbb{E}_\pi[\phi_x] = \mathbb{E}_\pi[\phi_{[x-y]}]$ and that $\mathbb{E}_\pi[\phi_y] = \mathbb{E}_\pi[\phi_0]$, applying the same change of variables in both cases. It follows that

$$\begin{aligned}\Sigma_{xy} = \mathbb{E}_\pi[\phi_x \phi_y] - \mathbb{E}_\pi[\phi_x]\mathbb{E}_\pi[\phi_y] &= \mathbb{E}_\pi[\phi_{[x-y]}\phi_0] - \mathbb{E}_\pi[\phi_{[x-y]}]\mathbb{E}_\pi[\phi_0] \\ &= \Sigma_{[x-y]0},\end{aligned}$$

which depends only on the increment $[x-y]$. □

We can now verify Equation 48 directly. Observe first that

$$\kappa(\nabla^2 H_R(\phi)) = \frac{\lambda_n + 3\lambda R^2 + m}{\lambda_1 + m}.$$

One may choose the entire quadratic part of the potential as the preconditioner, i.e., taking $M_\beta := \Delta_\beta + mI$. In this case, we obtain

$$I \preceq M_\beta^{-1/2} \nabla^2 H(M_\beta^{-1/2}\phi) M_\beta^{-1/2} \preceq I + 3\lambda R^2 M_\beta^{-1}.$$

In particular, the resulting condition number is

$$\kappa(M_\beta^{-1/2} \nabla^2 H(M_\beta^{-1/2}\phi) M_\beta^{-1/2}) = 1 + \frac{3\lambda R^2}{\lambda_1 + m},$$

which retains some undesirable but unavoidable dependence on R, the radius of truncation. To reconcile the example of lattice ϕ^4 theory with the condition in Equation 2, observe that because $M_\beta \succeq (\lambda_1 + m)I$, we have that

$$\frac{3\lambda R^2}{\lambda_1 + m} M_\beta \succeq 3\lambda R^2 I.$$

Hence, we have that

$$M_\beta \preceq \nabla^2 H(\phi) \preceq M_\beta + 3\lambda R^2 I \preceq \left(\frac{3\lambda R^2}{\lambda_1 + m} + 1\right) M_\beta.$$

If we define

$$P_\beta := \left(\frac{3\lambda R^2}{\lambda_1 + m} + 1\right) M_\beta,$$

then we further have that

$$\left(\frac{3\lambda R^2}{\lambda_1 + m} + 1\right)^{-1} P_\beta \preceq \nabla^2 H(\phi) \preceq P_\beta.$$

G.2 Additional Experiments

The heatmap in Figure 7 displays the sensitivity of the preconditioner estimate that Algorithm 1 outputs to both the number of chains per stage n and the number of ULMC steps per chain N for a lattice with sidelength $d = 100$. For all experiments, the number of stages of Gaussian cooling is limited to $K = 12$, with $\lambda = 0.5$, and $\beta = 2$. For this particular experiment, a truncation radius of $R = 4$ was used. The number of ULMC steps per chain appears to have a larger impact on the relative conditioning than the number of chains used.

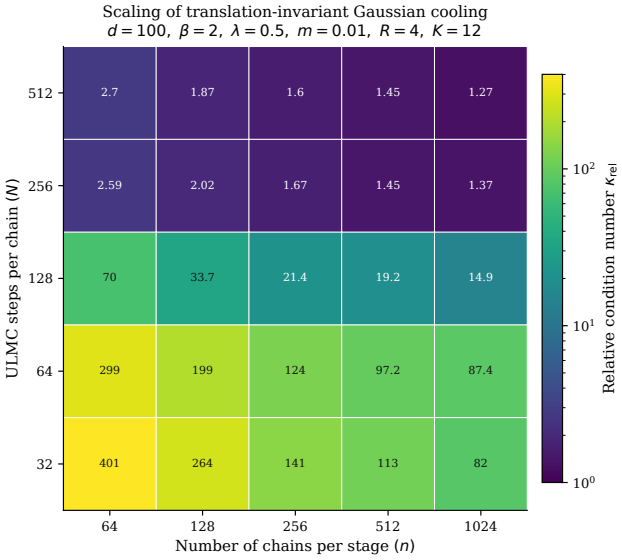


Figure 7: A comparison of translation-invariant Gaussian cooling performance across a range of N, the number of steps per chain, and n, the number of ULMC chains. For the lattice with sidelength $d = 100$ and mass $m = 0.01$, it is possible to reach nearly optimal preconditioning with relatively few chains and steps per chain.

H Experimental Details

All experiments were implemented fully in JAX and run on a single NVIDIA H100 SXM GPU with 80 GB VRAM. As stated in our AI disclosure, these experiments were implemented with the aid of generative AI tools, and users that wish to modify or build upon our code are encouraged to interface with the codebase through an LLM.

The supplementary material contains code for reproducing all experiments in Section 4. In particular, the supplementary code is structured as follows:

- `README.md` contains a high-level overview of the code, including descriptions of the files and directories listed below.
- `gaussian_cooling_algs.py` contains implementations of Algorithm 1 for both general and translation-invariant targets. The sampler for all experiments is ULMC, as described in Appendix D.
- `experiment_transformed_gaussian.py` contains the code for the experiments in Section 4.1, where we sample from a Gaussian that has been transformed by an ill-conditioned affine map. This file contains the code for reproducing Figures 1 and 2.
- `phi4` and the files therein contain the code for the lattice ϕ^4 experiments in Section 4.2. In particular, `PHI4_EXPERIMENTS.md` contains an in-depth description of all experiments used to produce the four figures in Section 4.2 and Appendix G.

To generate the ill-conditioned transformations in Section 4.1, we begin with a sequence of geometrically spaced eigenvalues in the interval $[1/\kappa(B), 1]$, where $\kappa(B)$ is the target condition number. Then, the eigenvectors of B are produced by a sampling a Haar random orthogonal matrix, at which point B can be recovered from its eigenvalues and eigenvectors. Table 4 contains the default parameters used in Section 4.1, all of which are also presented in the relevant figure subtitles.

Parameter	**Symbol**	**Default value**
Target dimension	d	100
Base covariance condition number	κ_0	10
Transformation condition numbers	$\kappa(B)$	$\{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}$
Chains per stage	n	256
ULMC steps per stage	N	96
Cooling stages	K	12
Cooling rate	γ	0.05
ULMC friction	$\widetilde{\gamma}$	1
ULMC step size	h	0.06
Repeats	—	3

Table 4: Default parameters for the experiments in Section 4.1.

Similarly, Table 5 contains all of default parameters for the experiments in Section 4.2 and Appendix G. For different experiments (e.g., for small mass $m = 0.001$), note that we needed to change the default parameters to obtain adequate performance. We also varied the cooling rate across experiments. Taking γ, the cooling rate, smaller pushes the performance of Gaussian cooling towards that of translation-invariant empirical preconditioning, as expected. For Figure 7, however, we took $\gamma = 0.35$ to verify that Gaussian cooling is still able to reduce the condition number of the target distribution to a constant near one, even if cooling is carried out slowly.

I Additional Algorithms

In this appendix, we elaborate on two variants of the algorithms presented in Section 2 above: (i) a precise description of the algorithm used for translation-invariant distributions in Algorithm 4 and (ii) a variant of Algorithm 1 that does not require a symmetric factorization of the covariance from the previous stage. In both cases, $\widehat{\Sigma}_n^{(-1)} = L^{-1}I$ so that $(\widehat{\Sigma}_n^{(-1)})^{1/2} = L^{-1/2}I$.

In Algorithm 4, note that both $(\widehat{\Sigma}_n^{(k)})^{1/2}$ and $(\widehat{\Sigma}_n^{(k)})^{-1/2}$ can be computed in time $\tilde{O}(d)$, as opposed to the $O(d^3)$ scaling of a full Cholesky factorization of the empirical covariance formed at each stage. Indeed, we may write

$$(\widehat{\Sigma}_n^{(k)})^{1/2} = \mathcal{F}\,\mathrm{diag}(\widehat{\sigma}_n^{(k)})^{1/2}\mathcal{F}^*,$$

and

$$(\widehat{\Sigma}_n^{(k)})^{-1/2} = \mathcal{F}\,\mathrm{diag}(\widehat{\sigma}_n^{(k)})^{-1/2}\mathcal{F}^*$$

with square roots and inverses taken elementwise.

Parameter	Figure 3	Figure 4	Figure 5	Figure 6	Figure 7
Lattice side d	512	100	100	100	100
Mass m	0.01	0.001	0.01	0.01	0.01
Truncation radius R	4	4	∞	2	4
Chains per stage n	512	2048	2048	512	2^6 through 2^{10}
Steps per stage N	256	2048	2048	256	2^5 through 2^9
Cooling stages K	12	12	12	12	12
Cooling rate γ	0.05	0.05	0.05	0.03	0.35
ULMC step size h	0.1	0.1	0.1	0.03	0.01
Repeats	1	3	3	3	5
Reference chains n_{ref}	512	2048	2048	512	4096
Reference steps N_{ref}	256	2048	2048	256	1634
Reference step size h_{ref}	0.001	0.001	0.001	0.03	0.03

Table 5: Parameters for the lattice ϕ^4 experiments in Section 4.2. All configurations use coupling $\beta = 2$, quartic coefficient $\lambda = 0.5$, and ULMC friction $\widetilde{\gamma} = 1$. Learning repeats denote independent preconditioner estimates. The IAT experiment selects one of its three learned preconditioners and runs one diagnostic trajectory with $10,000$ burn-in steps and $100,000$ retained samples. Reference covariances are estimated from endpoints of independent ULMC chains using the known quadratic preconditioner, mapped back to lattice coordinates and averaged over translations. To obtain the relative condition numbers presented in Figure 7, the reference number of steps was determined automatically as $N_{\mathrm{ref}} = \lceil T_{\mathrm{ref}}/h_{\mathrm{ref}} \rceil$, where $T_{\mathrm{ref}} = 5$ is an empirically selected target integration time.

Algorithm 4 Preconditioning via Gaussian Cooling, translation-invariant Case

Require: Preconditioning tolerance $\delta > 0$, smoothness parameter $L > 0$, global minimizer $x^* \in \mathbb{R}^d$, potential $U : \mathbb{R}^d \to \mathbb{R}$ with evaluation and gradient oracle, number of Gaussian cooling stages K, number of samples per stage n, number of ULMC steps per sample N

Ensure: A preconditioner $\widetilde{\Sigma}_n^{(K)}$ satisfying Equation 6 with high probability

1: **for** $k = 0, \dots, K$ **do**
2: Form $\widetilde{U}_k(x) = U_k((\widetilde{\Sigma}_n^{(k-1)})^{1/2} x)$
3: $x_1^{(k)}, \dots, x_n^{(k)} \leftarrow \mathrm{ULMC}\left(\widetilde{U}_k, (\widetilde{\Sigma}_n^{(k-1)})^{-1/2} x^*, \tilde{\gamma}, L, h, N, n\right)$
4: $\widehat{\sigma}_n^{(k)} \leftarrow \frac{1}{n} \sum_{i=1}^n \mathcal{F}(x_i^{(k)} - \widehat{m}_n(x_1^{(k)}, \dots, x_n^{(k)})) \odot \overline{\mathcal{F}(x_i^{(k)} - \widehat{m}_n(x_1^{(k)}, \dots, x_n^{(k)}))}$
5: $\widehat{\Sigma}_n^{(k)} \leftarrow \mathcal{F}\,\mathrm{diag}(\widehat{\sigma}_n^{(k)})\mathcal{F}^*$
6: $\widetilde{\Sigma}_n^{(k)} \leftarrow (\widetilde{\Sigma}_n^{(k-1)})^{1/2} \widehat{\Sigma}_n^{(k)} (\widetilde{\Sigma}_n^{(k-1)})^{1/2}$
7: **end for**
8: **return** $\widetilde{\Sigma}_n^{(K)}$

I.1 Factorization-Free Cooling

If dense Cholesky factorizations are undesirable, one may hope to obtain a version of Algorithm 1 that does not require symmetric factorizations of the empirical covariance at each stage, even when the target distribution is not translation-invariant. This is indeed possible, with the following algorithm motivated by ALDI (Garbuno-Inigo et al., 2020). In ALDI, it suffices to take the standard non-symmetric factorization of the empirical covariance:

$$C_n^{(k)} = \frac{1}{\sqrt{n}} \left[x_1^{(k)} - \widehat{m}_n(x_1^{(k)}, \dots, x_n^{(k)}) \quad \dots \quad x_n^{(k)} - \widehat{m}_n(x_1^{(k)}, \dots, x_n^{(k)}) \right] \in \mathbb{R}^{d \times n}.$$

In turn, we have that

$$C_n^{(k)} (C_n^{(k)})^\top = \frac{1}{n} \sum_{i=1}^n (x_i^{(k)} - \widehat{m}_n(x_1^{(k)}, \dots, x_n^{(k)}))(x_i^{(k)} - \widehat{m}_n(x_1^{(k)}, \dots, x_n^{(k)}))^\top$$

As is the case for the empirical covariance and mean, we can also view the non-symmetric factor above as a function of samples:

$$\widehat{C}_n(x_1, \dots, x_n) := \frac{1}{\sqrt{n}} \left[x_1 - \widehat{m}_n(x_1, \dots, x_n) \quad \dots \quad x_n - \widehat{m}_n(x_1, \dots, x_n) \right]. \tag{49}$$

As it turns out, with some modifications to the underdamped Langevin chain that Algorithm 1 utilizes, we can replace the symmetric square roots in the Gaussian cooling algorithm with their non-symmetric counterparts, greatly reducing the computational overhead of the algorithm by removing the need for full Cholesky factorizations.

Taking a step back, suppose that $A \in \mathbb{R}^{d \times d}$ is a matrix that admits two factorizations: $A = BB^\top$ and $A = CC^\top$, where $B \in \mathbb{R}^{d \times d}$ and $C \in \mathbb{R}^{d \times n}$ for some $n \geq d$. The underdamped Langevin diffusion in Equation 18, with potential $\tilde{V}(x) = V(Bx)$ and initial point $B^{-1}x^*$ then takes the form

$$\begin{cases} dX_t = P_t dt, \\ dP_t = -B^\top \nabla V(BX_t) dt - \tilde{\gamma} P_t dt + \sqrt{2\gamma} dB_t. \end{cases}$$

Now, suppose that we define processes $Y_t = BX_t$ and $Q_t = BP_t$, which satisfy

$$\begin{cases} dY_t = Q_t dt, \\ dQ_t = -A \nabla V(Y_t) dt - \tilde{\gamma} Q_t dt + \sqrt{2\gamma} B dB_t. \end{cases}$$

If W_t is an n-dimensional Brownian motion, it follows that the increments of BB_t and CW_t have the same law because $BB^\top = CC^\top$. Hence, we may instead consider the diffusion process

$$\begin{cases} dY_t = Q_t dt, \\ dQ_t = -A \nabla V(Y_t) dt - \tilde{\gamma} Q_t dt + \sqrt{2\gamma} C dW_t, \end{cases} \tag{50}$$

initialized at $x^* = BB^{-1}x^*$. To simulate this via the discretization in Equation 24, we perform the updates

$$\widehat{Y}_{(k+1)h} = \widehat{Y}_{kh} + \frac{1 - e^{-\tilde{\gamma}h}}{\tilde{\gamma}} \widehat{Q}_{kh} - \frac{\tilde{\gamma}h - 1 + e^{-\tilde{\gamma}h}}{\tilde{\gamma}^2} A \nabla V(\widehat{Y}_{kh}) + \eta_k$$

$$\widehat{Q}_{(k+1)h} = e^{-\tilde{\gamma}h} \widehat{Q}_{kh} - \frac{1 - e^{-\tilde{\gamma}h}}{\tilde{\gamma}} A \nabla V(\widehat{Y}_{kh}) + \xi_k,$$

where $(\eta_k, \xi_k) \sim \mathcal{N}(0, \widehat{\Sigma}_{\tilde{\gamma},h})$, with

$$\widehat{\Sigma}_{\tilde{\gamma},h} := \begin{bmatrix} \frac{2}{\tilde{\gamma}}[h - \frac{2}{\tilde{\gamma}}(1 - \exp(-\tilde{\gamma}h)) + \frac{1}{2\tilde{\gamma}}(1 - \exp(-2\tilde{\gamma}h))] & \frac{1}{\tilde{\gamma}}(1 - 2\exp(-\tilde{\gamma}h) + \exp(-2\tilde{\gamma}h)) \\ \frac{1}{\tilde{\gamma}}(1 - 2\exp(-\tilde{\gamma}h) + \exp(-2\tilde{\gamma}h)) & 1 - \exp(-2\tilde{\gamma}h) \end{bmatrix} \otimes A.$$

To sample the noise (η_k, ξ_k) in this formulation, factorize $\widehat{\Sigma}_{\tilde{\gamma},h} = \widehat{R}_{\tilde{\gamma},h} \widehat{R}_{\tilde{\gamma},h}^\top \otimes A$. Then, if $g \sim \mathcal{N}(0, I_{2n})$, we observe that

$$\mathrm{Cov}((\widehat{R}_{\tilde{\gamma},h} \otimes C)g) = \widehat{R}_{\tilde{\gamma},h} \widehat{R}_{\tilde{\gamma},h}^\top \otimes CC^\top = \widehat{R}_{\tilde{\gamma},h} \widehat{R}_{\tilde{\gamma},h}^\top \otimes A.$$

Hence, the noise can be sampled without ever forming a symmetric factorization of A. In summary, we can sample from the ULMC discretization of the process in Equation 50 using the following algorithm:

By construction, $\widehat{Y}_{Nh} \overset{d}{=} B\widehat{X}_{Nh}$ so that if $y_1^{(k)}, \dots, y_n^{(k)}$ are sampled from Algorithm 5, then $x_i^{(k)} := B^{-1} y_i^{(k)}$ are samples from Algorithm 3. In turn,

$$\widehat{S}_n(y_1^{(k)}, \dots, y_n^{(k)}) = \widehat{S}_n(Bx_1^{(k)}, \dots, Bx_n^{(k)}) = B\widehat{S}_n(x_1^{(k)}, \dots, x_n^{(k)})B^\top, \tag{51}$$

which is the desired output from Algorithm 1 at step K. Importantly, we can entirely sidestep any symmetric factorizations using this approach. Below, we introduce the shorthand

$$y_1, \dots, y_n = \mathrm{ULMC}_{\mathrm{FF}}(V, C, x^*, \tilde{\gamma}, L, h, N, n)$$

to call Algorithm 5 with the specified parameters. With this notation, we introduce Algorithm 6, which carries out Gaussian cooling without any symmetric factorizations of the empirical covariance.

Algorithm 5 Factorization Free Underdamped Langevin Monte Carlo

Require: Potential $V : \mathbb{R}^d \to \mathbb{R}$ with gradient oracle ∇V, non-symmetric factorization $C \in \mathbb{R}^{d\times n}$ with $A = CC^\top$ global minimizer x^*, friction parameter $\tilde{\gamma} > 0$, smoothness parameter $L > 0$, step size $h > 0$, Cholesky factor $\widehat{R}_{\tilde{\gamma},h} \in \mathbb{R}^{2\times 2}$, number of iterations N

Ensure: Sample $\widehat{Y}_{Nh}$

Draw $g_Y, g_Q \overset{i.i.d.}{\sim} \mathcal{N}(0, I_n)$
$(\widehat{Y}_0, \widehat{Q}_0) \leftarrow (x^* + L^{-1/2}Cg_Y, Cg_Q)$
for $m = 0, 1, \ldots, N-1$ **do**
  $\widehat{V}_{mh} \leftarrow \nabla V(\widehat{Y}_{mh})$
  Sample $g \sim \mathcal{N}(0, I_{2n})$
  $(\eta_m, \xi_m) \leftarrow (\widehat{R}_{\tilde{\gamma},h} \otimes C)g$
  $\widehat{Y}_{(m+1)h} \leftarrow \widehat{Y}_{mh} + \dfrac{1-e^{-\tilde{\gamma}h}}{\tilde{\gamma}}\,\widehat{Q}_{mh} - \dfrac{\tilde{\gamma}h - 1 + e^{-\tilde{\gamma}h}}{\tilde{\gamma}^2} CC^\top \widehat{V}_{mh} + \eta_m$
  $\widehat{Q}_{(m+1)h} \leftarrow e^{-\tilde{\gamma}h}\,\widehat{Q}_{mh} - \dfrac{1-e^{-\tilde{\gamma}h}}{\tilde{\gamma}} CC^\top \widehat{V}_{mh} + \xi_m$
end for
return $\widehat{Y}_{Nh}$

Algorithm 6 Factorization-Free Preconditioning via Gaussian Cooling

Require: Preconditioning tolerance $\delta > 0$, smoothness parameter $L > 0$, global minimizer $x^* \in \mathbb{R}^d$, potential $U : \mathbb{R}^d \to \mathbb{R}$ with evaluation and gradient oracle, number of Gaussian cooling stages K, number of samples per stage n, number of ULMC steps per sample N

Ensure: A preconditioner $\widehat{\Sigma}_n^{(K)}$ satisfying Equation 6 with high probability

for $k = 0, \ldots, K$ **do**
  $y_1^{(k)}, \ldots, y_n^{(k)} \leftarrow \mathrm{ULMC}_{\mathrm{FF}}\left(U_k, C_n^{(k-1)}, x^*, \tilde{\gamma}, L, h, N, n\right)$
  $C_n^{(k)} \leftarrow \widehat{C}_n(y_1^{(k)}, \ldots, y_n^{(k)})$
end for
$\widetilde{\Sigma}_n^{(K)} \leftarrow C_n^{(K)}(C_n^{(K)})^\top$
return $\widetilde{\Sigma}_n^{(K)}$

In the above algorithm, we take

$$C_n^{(-1)} := L^{-1/2} \begin{bmatrix} e_1 & e_2 & \cdots & e_d & \mathbf{0}_d & \cdots & \mathbf{0}_d \end{bmatrix} \in \mathbb{R}^{d \times n},$$

where $e_i \in \mathbb{R}^d$ is the ith standard basis vector and $\mathbf{0}_d \in \mathbb{R}^d$ is the vector of all zeros. Indeed, it then follows that $C_n^{(-1)}(C_n^{(-1)})^\top = L^{-1} I$, giving the desired non-symmetric factorization of $\widetilde{\Sigma}_n^{(-1)}$ from Algorithm 1.

J A CAUTIONARY (COUNTER)EXAMPLE

Although many distributions of interest will satisfy Equation 2, it is well-known in the HMC community that *nonlinear* correlations can stymie even the most effective mass matrix estimation procedures (Carpenter et al., 2017). This should come as no surprise, as Equation 2 imposes structure on the global conditioning of the potential, but it cannot capture local behavior of the Hessian. In turn, for any choice of $L > \mu > 0$, it is possible to construct potentials that are μ-strongly log-concave and L-smooth but, in order for Equation 2 to hold for some $B \succ 0$, one must choose $\alpha = O(1/\kappa)$. Unfortunately, for such potentials, Gaussian cooling appears to provide little practical benefit due to its query complexity's polynomial scaling in α, as presented in Theorem 3.2.

For the sake of concreteness, we construct a simple example of such a potential here, for an arbitrary dimension $d \in \mathbb{N}$. For fixed $L > \mu > 0$, define

$$U(x) := \sum_{i=1}^{d} \left(\frac{L+\mu}{4} x_i^2 + \frac{L-\mu}{8} \cos(2x_i) \right). \tag{52}$$

Because this potential is separable, its Hessian is diagonal, with

$$\nabla^2 U(x)_{ii} = \frac{L+\mu}{2} - \frac{L-\mu}{2} \cos(2x_i).$$

In particular, we have that

$$\mu \leq \nabla^2 U(x)_{ii} \leq L, \quad i \in [d]$$

so that U does indeed induce a μ-strongly log-concave and L-smooth measure with condition number $\kappa = \frac{L}{\mu}$.

Define $x^{(1)} = \frac{\pi}{2} e_1$ and $x^{(2)} = \frac{\pi}{2} e_2$, where $e_i \in \mathbb{R}^d$ denotes the ith standard basis vector. By construction, $\nabla^2 U(x^{(1)})_{11} = L$ and $\nabla^2 U(x^{(1)})_{22} = \mu$, while $\nabla^2 U(x^{(2)})_{11} = \mu$ and $\nabla^2 U(x^{(2)})_{22} = L$. If

$$\alpha B \preceq \nabla^2 U(x) \preceq B$$

is to hold for some $B \succ 0$ and $\alpha \in (0, 1)$ simultaneously for all $x \in \mathbb{R}^d$, then we would require that $\alpha B_{11} \leq \nabla^2 U(x)_{11} \leq B_{11}$ for all $x \in \mathbb{R}^d$. This condition implies that $L \leq B_{11}$ so that $\alpha L \leq \alpha B_{11} \leq \mu$. In other words, $\alpha \leq \frac{\mu}{L} = \frac{1}{\kappa}$.

As the ϕ^4 example in Section 4 indicates, the presence of a nonlinear term in the Hessian of the underlying potential does *not* necessarily mean that Equation 2 cannot hold for some $\alpha = \Theta(1)$.