The Manifold Joys of Sampling

Yin Tat Lee ✉ 🏠
University of Washington, Seattle, WA, USA
Microsoft Research, Seattle, WA, USA

Santosh S. Vempala ✉ 🏠
Georgia Tech, Atlanta, GA, USA

Abstract

We survey recent progress and many open questions in the field of sampling high-dimensional distributions, with specific focus on sampling with non-Euclidean metrics.

2012 ACM Subject Classification Theory of computation → Randomness, geometry and discrete structures

Keywords and phrases Sampling, Diffusion, Optimization, High Dimension

Digital Object Identifier 10.4230/LIPIcs.ICALP.2022.4

Category Invited Talk

Funding This work was supported in part by NSF awards DMS-1839116, DMS-1839323, CCF-1909756, CCF-2007443 and CCF-2134105.

Acknowledgements We thank Yunbum Kook and Andre Wibisono for helpful comments.

1 Introduction

Sampling high-dimensional distributions is a fundamental problem of growing importance in machine learning and related fields [7, 17, 2]. Progress on efficient algorithms for sampling has led to new mathematical connections and insights into a number of areas such as probability, convex geometry and analysis. The focus of this tutorial is to survey the state-of-the-art for the most general results along with open problems. There are many interesting results for special cases that are beyond the scope of this survey.

We begin by stating the most general version of the problem in Euclidean space.

▶ **Definition 1** (Sampling Problem)**.** Given oracle access to an integrable, real-valued function $f : \mathbb{R}^n \to \mathbb{R}_+$, an initial point $x_0 \in \mathbb{R}^n$ with $f(x_0) > \beta \int f(y)dy$ and an error parameter $\varepsilon > 0$, output a point x from a distribution that is within total variation distance ε from the distribution with density proportional to $f(x)$.

A major discovery in the theory of algorithms is that the above problem can be solved in randomized polynomial time for any *logconcave* function f. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}_+$ is logconcave if its logarithm is concave, i.e., for any $\lambda \in [0, 1]$ and any $x, y \in \mathbb{R}^n$,we have $f(\lambda x + (1-\lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}$. This class includes the important special cases of the uniform density over a convex body and a Gaussian restricted to a convex set. Note that we can alternatively think of a logconcave function as $e^{-f(x)}$ where now f is a convex function. As in optimization, the traditional frontier of polynomial-time algorithms for sampling has to do with convexity. In the past decade there has been progress on going beyond convexity, with appropriate weaker assumptions. Two approaches are (a) to show that isoperimetry of the target distribution suffices and (b) to use the convergence of the continuous time diffusion as a basis for proving the convergence of the discrete-time algorithm. We will illustrate these approaches later in this survey. Let us first introduce the main algorithms for the most general oracle setting.

49th International Colloquium on Automata, Languages, and Programming (ICALP 2022).
Editors: Mikołaj Bojańczyk, Emanuela Merelli, and David P. Woodruff;
Article No. 4; pp. 4:1–4:20
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Ball Walk

The ball walk with step-size parameter $\delta > 0$ is the following Markov Chain: at the current point x, pick a uniform random point y from the ball of radius δ centered at x; go to y with probability $\min\{1, f(y)/f(x)\}$. The starting point is chosen so that it satisfies $f(x_0) > 0$. When f is the indicator of a convex body K, this simply means that we start in K and at each step go to the proposed y only if y is also in K.

Hit-and-Run

The Hit-and-Run Markov chain is the following: at the current point x, pick a uniform random line ℓ through x, and go to a point y on the line with probability proportional to f restricted to ℓ. For a convex body, this means that we pick the next point uniformly from the chord in a random direction through the current point.

2 Mixing rates in the general oracle model

For a more detailed introduction to Markov chains in continuous state spaces, and these walks in particular, the reader is referred to [42, 36]. In this section by mixing rate we mean the rate to halve the χ^2-divergence between the current distribution and the target stationary distribution. We say that a distribution Q_0 is an M-warm start for a distribution Q if $\sup \frac{dQ_0(x)}{dQ(x)} \le M$. A weak M-warm start is when $\chi^2(Q_0, Q) \le M$. Recall that

$$\chi^2(P,Q) = \mathbb{E}_Q\left(\left(\frac{dP(x)}{dQ(x)} - 1\right)^2\right).$$

A key parameter in the analysis of a Markov chain is its conductance. For a Markov chain with state space K, transition kernel P and stationary distribution Q, the conductance of any measurable subset A of the state space is defined as

$$\phi(A) = \frac{\int_A P_x(K \setminus A) dQ(x)}{\min\{Q(A), Q(K \setminus A)\}}$$

and the conductance of the Markov chain itself is $\phi = \inf_A \phi(A)$. The conductance directly bounds the mixing rate.

▶ **Theorem 2** ([36]). *For a time-reversible Markov chain with conductance ϕ, the distribution after t steps satisfies*

$$\chi^2(Q_t, Q) \le \left(1 - \frac{\phi^2}{2}\right)^t \chi^2(Q_0, Q).$$

The paper [36] developed conductance-based analysis of the convergence of Markov chains in the setting of continuous state spaces, including an important extension to the setting when the conductance $\phi(A)$ can only be bounded from below for sets A of measure larger than some threshold.

2.1 Ball walk

The ball walk has the following convergence guarantee for a convex body. We use $\widetilde{O}(\cdot)$ to suppress logarithmic terms.

▶ **Theorem 3** (Ball Walk from Warm Start in Convex Body [22]). *Let $K \subseteq \mathbb{R}^n$ be a convex body containing the unit Euclidean ball with $R^2 = \mathbb{E}_K\left(\|x-\overline{x}\|^2\right)$ for $\overline{x} = \mathbb{E}_K x$. Then the mixing time of the ball walk in K with step size $\delta = 1/\sqrt{n}$ from an $O(1)$-warm start is bounded by $\widetilde{O}(n^2R^2)$.*

This guarantee generalizes cleanly to any logconcave density.

▶ **Theorem 4** (Ball Walk from Warm Start for Logconcave Density [38]). *The mixing time of the ball walk for a target logconcave density ν such that the level set of measure $1/8$ contains a unit ball and $R^2 = \mathbb{E}_\nu\left(\|x-\overline{x}\|^2\right)$, with step size $\delta = 1/\sqrt{n}$ from an $O(1)$-warm start is bounded by $\widetilde{O}(n^2R^2)$.*

The bound above is tight up to polylogarithmic factors; a cylinder with a unit ball cross section and height R shows a lower bound of $\Omega(n^2R^2)$. Can we reduce or eliminate the dependence on R? Classical results tell us that for any convex body there is an affine transformation (the John position) that ensures that the body is sandwiched between balls of radii 1 and at most n. Since all we need to bound is the average squared distance, there is a natural transformation of space that gives a bound of $R = \sqrt{n}$.

▶ **Definition 5.** A distribution with density ν in $\mathbb{R}^n$ is said to be in *isotropic position* if a random point X drawn from ν satisfies $\mathbb{E}_\nu(x) = 0$ and $\mathbb{E}_\nu(xx^\top) = I$.

Note that in isotropic position, $R^2 = \mathbb{E}(\|x\|^2) = n$ and this implies a bound of n^3 on the mixing rate for a logconcave distribution in isotropic position. As it turns out, it is possible to show a better convergence rate, where the dependence on diameter is essentially eliminated. The theorems below work also for *near*-isotropic position, by which we mean that the eigenvalues of the covariance matrix are bounded below and above by constants (rather than all being equal to 1).

▶ **Theorem 6** (Ball Walk from Warm Start with Isotropic Target). *The mixing time of the ball walk applied to a logconcave density in isotropic position with step size $\delta = 1/\sqrt{n}$ from an $O(1)$-warm start is bounded by $\widetilde{O}(n^2)$.*

The above theorem deserves some explanation and is the culmination of a quarter century of progress on geometric isoperimetric inequalities. First, we state the corresponding theorem which establishes the rate of convergence in terms of the *isoperimetry of the target distribution.* For this we define the KLS constant.

▶ **Definition 7.** For a density ν in $\mathbb{R}^n$, the KLS constant of ν is defined as

$$\frac{1}{\psi_\nu} = \inf_{S \subset \mathbb{R}^n} \frac{\nu(\partial S)}{\min\left\{\nu(S), 1-\nu(S)\right\}}$$

and the KLS constant for logconcave densities in $\mathbb{R}^n$ is $\psi_n = \sup\left\{\psi_\nu : \nu \text{ isotropic logconcave in } \mathbb{R}^n\right\}$.

We will discuss this constant and its significance presently. The following theorem [22] shows how it bounds the mixing rate of the ball walk.

▶ **Theorem 8** (Ball Walk and KLS [22]). *The mixing time of the ball walk applied to a logconcave density in isotropic position with step size $\delta = 1/\sqrt{n}$ from an $O(1)$-warm start is bounded by $\widetilde{O}(n^2\psi_\nu^2)$.*

Kannan, Lovász and Simonovits [21] conjectured that $\psi_n = O(1)$ for any logconcave isotropic density ν. Recently, Klartag and Lehec [25], following a line of work [21, 13, 16, 12, 32, 3], proved that $\psi_n = O(\log^5 n)$, which implies Theorem 6 above. Another way to state the underlying isoperimetric inequality is the following.

▶ **Theorem 9** (Euclidean Isoperimetry). *For any logconcave density ν in $\mathbb{R}^n$ whose covariance matrix A has largest eigenvalue λ_1 and any two disjoint subsets S_1, S_2 of $\mathbb{R}^n$ we have*

$$\nu(K \setminus S_1 \setminus S_2) \geq \frac{d(S_1, S_2)}{\psi_n \sqrt{\lambda_1}} \min\{\nu(S_1), \nu(S_2)\}$$

where $d(,)$ is Euclidean distance and ψ_n is known to be $O(\log^5 n)$ [25] and conjectured to be $O(1)$ [21].

An equivalent geometric way to state the KLS conjecture is that a hyperplane-induced subset achieves the minimum isoperimetric coefficient up to a universal constant factor. For a detailed discussion of the KLS conjecture, we refer the reader to [31]. Its full resolution remains an intriguing and fruitful open problem.

More recently, it has been shown that the complexity of isotropic transformation and volume computation can be reduced to the KLS constant.

▶ **Theorem 10** (Rounding and Volume Computation [8, 19]). *Given a convex body given in the membership oracle model, it can be brought into near-isotropic position in $\widetilde{O}(n^3\psi_n^2)$ queries and its volume can be computed to within relative error ε using $\widetilde{O}(n^3/\varepsilon^2)$ additional queries.*

We conclude this section with two questions about the analysis of the ball walk, which both seem to require the development of new tools. The first has to do with the convergence of the ball walk without a warm start. Although this does not converge quickly in general (e.g., when starting near a corner), it is conceivable that it does when started from a "nice" point. After all, the result for a warm start effectively states that "most" points are good starting points. But can we get our hands on one of them?

▶ **Question 11.** *Show that the ball walk started from the centroid of a convex body converges to its stationary distribution in polynomial time.*

The second question has to do with verifying that the ball walk has converged. This has value both theoretically and practically (to test convergence on an instance-by-instance basis rather than simply running up to the worst-case bound).

▶ **Question 12.** *Consider the ball walk in a convex body starting from a logconcave initial distribution (e.g., uniform in a ball contained inside the body). Show that the distance between the current distribution and the target can be bounded as a polynomial in the dimension and the distance to stationarity of a random one-dimensional marginal of the current distribution.*

A clean variant of this question, first proposed by Lovász, is whether the expected squared distance of a random point X_t of the ball walk from its starting point X_0 is a monotonic increasing function.

2.2 Hit-and-Run

One advantage of hit-and-run is that there is no step-size parameter. A more important property is that hit-and-run converges rapidly from *any* starting distribution (even a single point), unlike the ball walk. To see that the ball walk does not, consider starting the ball walk near a corner, e.g., near a vertex of a cube; then the probability of making a proper step (one that moves to a different point) can be arbitrarily small. Another way to see this is to note that the conductance of small sets for the ball walk Markov chain cannot, in general, be bounded from below. For hit-and-run, the picture is dramatically different – every subset can be shown to have large conductance! In particular, hit-and-run started from an arbitrary point in the interior (e.g., close to a corner) will quickly escape the corner.

▶ **Theorem 13** (Conductance of Hit-and-Run [37])**.** *The conductance of hit-and-run for any convex body in $\mathbb{R}^n$ containing a unit ball and contained in a ball of radius R is $\Omega(1/(nR))$.*

As a consequence it has the same general mixing rate as the ball walk, except that the dependence on the warm start parameter is logarithmic rather than polynomial, and in particular starting from any point x_0 in the interior of K, it is logarithmic in the inverse of the distance of x_0 to the boundary of K. It is conjectured (and observed in practice [17, 2]) that Hit-and-Run also mixes in n^2 steps for a body/logconcave distribution in isotropic position. However, it is an open problem to analyze this and show a bound better than n^3 as implied by the analysis in general position.

▶ **Question 14.** *Analyze the mixing rate of Hit-and-Run for a logconcave density or convex body in isotropic position.*

To understand the difficulty of answering this question, it is useful to see the isoperimetric inequality underlying the conductance bound for hit-and-run. This is our first departure from Euclidean distance, and an indication that the natural underlying geometry is different. For two points u, v in a convex body K, we define the *cross-ratio* distance as follows. Let p, q be the endpoints of the chord induced by u, v inside K, so that their order along the chord is p, u, v, q. Then,

$$d_K(u,v) = \frac{\|u-v\| \, \|p-q\|}{\|p-u\| \, \|v-q\|}.$$

While this distance is not a true distance in that it does not satisfy the triangle inequality, the closely related Hilbert distance, $d_H(u,v) = \ln(1 + d_K(u,v))$, is a metric. We have the following isoperimetric inequality.

▶ **Theorem 15** (Cross-ratio/Hilbert Isoperimetry [35, 38])**.** *For any logconcave density ν in $\mathbb{R}^n$ whose support is a convex body K and any two disjoint subsets S_1, S_2 of $\mathbb{R}^n$ we have*

$$\nu(K \setminus S_1 \setminus S_2) \geq d_K(S_1, S_2)\nu(S_1)\nu(S_2).$$

This isoperimetric inequality is affine-invariant and already tight. So it is unclear how to derive a more refined inequality that takes advantage of isotropic position.

▶ **Question 16.** *Is there a corresponding KLS conjecture that would imply a better mixing rate for hit-and-run?*

2.3 A bottleneck

For both the ball walk and hit-and-run, the mixing rate from a warm start for an isotropic target is $\Omega(n^2)$; this can be seen for a hypercube. The bottleneck for the ball walk is that the largest value that the step-size parameter δ can be set to is $O(1/\sqrt{n})$. This is because setting it to a larger value leads to most steps being rejected, i.e., WHP proposed steps are outside the body. In the case of hit-and-run, we have the same bottleneck. This is because the length of a random chord through a random point in a hypercube is still only $O(1/\sqrt{n})$, so this is the effective step-size on average.

3 In search of a larger step: the Dikin walk

For the walks discussed so far, the step-size is limited by points near the boundary. This suggests the idea of taking larger steps from points that are deeper inside, e.g., at the current point, use the largest ball that has a constant probability of a random step being accepted. So points deeper inside would take larger steps. However, this has the issue that the resulting stationary distribution is no longer uniform. To correct this, one can use a Metropolis filter based on the ratios of the volumes of the balls at the current point and the proposed point. Unfortunately, this means that this effective radius or step-size must change slowly. A smoother and more elegant approach is inspired by the seminal work of Dikin for convex optimization [10]. To optimize a linear function over a polytope, he proposed constructing a "maximal" ellipsoid around the current point that would be contained in the body, taking a large, objective-improving step within this ellipsoid and repeating. This was the first "interior-point" paradigm (now called Affine Scaling); it predates the central path interior-point method pioneered by Karmarkar [24] and generalized by Nesterov and Nemirovskii [41]. For a point x in the polytope defined by $\{Ax \geq b\}$, the Dikin ellipsoid is defined as follows:

$$E(x,r) = \left\{y : (y-x)^\top \mathbf{H}(x)(y-x) \leq r^2\right\} \qquad \text{where} \qquad \mathbf{H}(x) = \sum_{i=1}^{m} \frac{a_i a_i^\top}{(a_i^\top x - b_i)^2}$$

with a_i being the i'th row (normal vector) of the constraint matrix A for $i = 1, \ldots, m$. This ellipsoid adapts to the local geometry of the polytope, and $E(x,1)$ always fully contained in it. It suggests the following Dikin random walk: at the current point x, pick a random point y in the Dikin ellipsoid $E(x,r)$; if $x \in E(y,r)$, go to y with probability $\min\{1, \mathrm{vol}(E(x,r))/\mathrm{vol}(E(y,r))\}$. Kannan and Narayanan [23] analyzed its mixing rate with a suitable choice of the radius r.

▶ **Theorem 17** (Dikin in Polytope [23]). *The mixing rate of the Dikin walk from a warm start in a polytope in $\mathbb{R}^n$ given by m inequalities is bounded by $O(mn)$.*

Notably, the mixing rate of the walk *does not* depend on the affine position of the body – the process is affine-invariant, i.e., applying the walk to an affine transformation of the input body is the same as applying the transformation to the output of the process on the same body. This means in particular that one can assume for free that the polytope is in isotropic position. This avoids dependence on the outer ball radius parameter R in the previous section.

What should the step-size r be? It is chosen to be $O(1/\sqrt{n})$ and this turns out to be necessary to ensure that the rejection probability is not too high, i.e., that the volumes of Ellipsoids corresponding to nearby points are within a constant factor. The change in the ellipsoid can be bounded using the classical optimization property of *self-concordance*, i.e., when the matrix function $\mathbf{H}(x)$ is the Hessian of a convex function. Self-concordance is a key property in the analysis of the interior-point method for linear programming [41]. In fact, the ellipsoid matrix $\mathbf{H}(x)$ is the Hessian of the convex log barrier function at x,

$$\phi(x) = -\sum_{i=1}^{m} \ln(a_i^\top x - b_i),$$

i.e., $\mathbf{H}(x) = D^2\phi(x)$. So the log-barrier interior-point method for optimization corresponds to the Dikin walk for sampling!

To make this connection a bit more precise, let us define the norm of a vector v induced by a matrix function $\mathbf{H}$ as $\|v\|_x^2 = v^\top \mathbf{H}(x)v$.

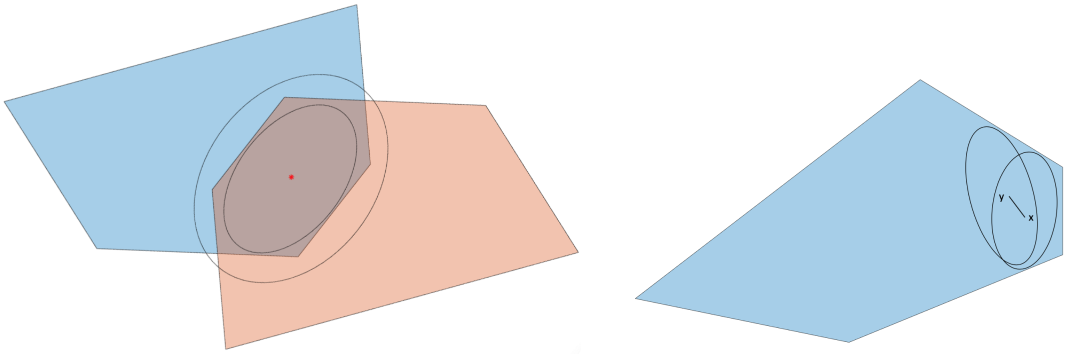

Figure 3.1 (a) $E_u(1) \subseteq K \cap (2u - K) \subseteq E_u(\sqrt{\nu})$. (b) (Strong) self-concordance measures the rate of change of Hessian.

▶ **Definition 18.** For a convex set $K \subset \mathbb{R}^n$, we say a matrix function $\mathbf{H} : K \to \mathbb{R}^{n \times n}$ is *self-concordant* if for any $x \in K$, we have

$$\left\|\mathbf{H}(x)^{-1/2} D\mathbf{H}(x)[h] \mathbf{H}(x)^{-1/2}\right\|_{\text{op}} \leq 2 \left\|h\right\|_x$$

where $D\mathbf{H}(x)[h]$ is the directional derivative of $\mathbf{H}$ at x in the direction h, i.e., $D\mathbf{H}(x)[h] = \frac{d}{dt}\mathbf{H}(x + th)$. We say $\mathbf{H}$ is symmetric ν-self-concordant if $\mathbf{H}$ is self-concordant and for any $x \in K$,

$$E(x, 1) \subseteq K \cap (2x - K) \subseteq E(x, \sqrt{\nu}).$$

Self-concordance relates the change in $\mathbf{H}$ to the change in x. Many natural self-concordant barriers, including the logarithmic barrier, satisfy a much stronger condition, replacing the operator norm above by the Frobenius norm. While this makes no difference for the worst-case bound for optimization, it turns out to be crucial for sampling.

▶ **Definition 19** (Strong Self-Concordance). For a convex set $K \subset \mathbb{R}^n$, we say a matrix function $\mathbf{H} : K \to \mathbb{R}^{n \times n}$ is *strongly self-concordant* if for any $x \in K$, we have

$$\left\|\mathbf{H}(x)^{-1/2} D\mathbf{H}(x)[h] \mathbf{H}(x)^{-1/2}\right\|_F \leq 2 \left\|h\right\|_x .$$

The canonical barrier [18] satisfies strong self-concordance. The situation with two other classical barriers, namely the universal and entropic barriers has a curious connection.

▶ **Lemma 20** ([27]). *Let $\mathbf{H}(x)$ be the Hessian of the universal or entropic barriers. Then,*

$$\left\|\mathbf{H}(x)^{-1/2} D\mathbf{H}(x)[h] \mathbf{H}(x)^{-1/2}\right\|_F = O(\psi_n) \left\|h\right\|_x .$$

In fact, up to a logarithmic factor, the strong self-concordance of these barriers is *equivalent* to the KLS conjecture.

▶ **Lemma 21.** *Given any strongly self-concordant matrix function $\mathbf{H}$ on $K \subset \mathbb{R}^n$. For any $x, y \in K$ with $\|x - y\|_x < 1$, we have*

$$\|\mathbf{H}(x)^{-\frac{1}{2}}(\mathbf{H}(y) - \mathbf{H}(x))\mathbf{H}(x)^{-\frac{1}{2}}\|_F \leq \frac{\|x - y\|_x}{(1 - \|x - y\|_x)^2}.$$

The following guarantee was shown for the convergence of the generalized Dikin walk in [27].

▶ **Theorem 22** (Convergence of general Dikin walk [27]). *The mixing rate of the Dikin walk for a symmetric, strongly self-concordant matrix function with convex log determinant is* $O(n\bar{\nu})$.

This implies faster mixing and sampling for polytopes using the LS barrier [28], which is strongly self-concordant, has a convex log determinant and has $\bar{\nu} = O(n \log^3 m)$.

▶ **Theorem 23** (Quadratic Convergence of Dikin [27]). *The mixing rate of the Dikin walk based on the LS barrier for any polytope in* $\mathbb{R}^n$ *is* $\tilde{O}(n^2)$ *and each step can be implemented in* $\tilde{O}(mn^{\omega-1})$ *arithmetic operations where* ω *is the matrix multiplication exponent.*

We note that the Dikin walk with the logarithmic barrier for a polytope $\{\mathbf{A}x \geq b\}$ can be implemented in time $O(\mathrm{nnz}(\mathbf{A}) + n^2)$ per step while maintaining the mixing rate of $O(mn)$.

The isoperimetry of the metric induced by the matrix function $\mathbf{H}$ follows by simply connecting it to the cross-ratio distance.

▶ **Lemma 24.** *For* $u, v \in K$, $d_K(u,v) \geq \frac{\|u-v\|_u}{\sqrt{\bar{\nu}}}$.

From this lemma and Theorem 15, we have that the isoperimetric coefficient for the Hessian norm distance is $\Omega(1/\sqrt{\bar{\nu}})$. This bound, as well as the bound of mn for the Dikin walk with the log barrier are tight as shown by a hypercube with one of its facets duplicated $m - n$ times. However, the situation is far from clear for the weighted Dikin walk, where the log barrier is replaced by one that weights each constraint, as in the LS barrier (and effectively eliminates this bad example).

▶ **Question 25.** *What is the isoperimetric coefficient of the Hessian distance induced by the LS barrier? Does the corresponding weighted Dikin walk mix in* $\tilde{O}(n)$ *steps?*

4 Large steps via non-Euclidean geometries

The Dikin walk and its weighted version are based on exploiting *local* geometry, where the norm is defined locally. As a direct consequence, this family of algorithms is affine-invariant. However, to ensure the correct stationary distribution and keep the probability of acceptance in the Metropolis filter reasonably large, the effective step-size is again $1/\sqrt{n}$, and the convergence rate is n^2. Is a larger step-size/smaller convergence rate possible?

To understand this, we delve further into the use of non-Euclidean geometry. So far, our random walks have only taken straight line steps in Euclidean space. The distribution of the direction of the next step depends on the current point in the case of the Dikin walk, but the step itself is still a straight line. A more drastic departure would be to use curves instead of straight lines. How?

First note that given a local metric such as the one induced by the Hessian of a convex function defines a manifold and using this metric we can define the length of a curve (a continuous path) and the distance between any two points (as the length of the shortest curve between them). Since the local metric varies, the shortest path between two points can be a curve (e.g., consider flight paths for the sphere metric). In the context of sampling polytopes, we consider the metric defined by the Hessian of the log barrier. The support of this manifold is the original polytope. Distances between points are magnified as the points get closer to the boundary. So geodesic paths tend to avoid the boundary!

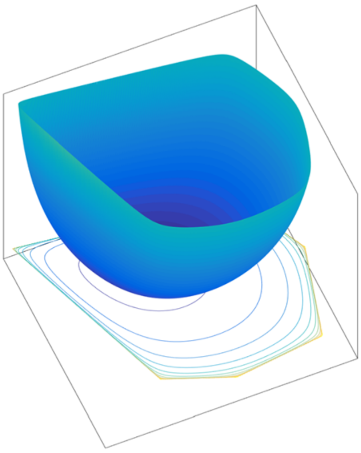

Figure 4.1 The Hessian manifold induced by the log barrier in a polytope.

How do we pick a random next step/direction? For this we can use the notion of tangent space attached to every point in a Riemannian manifold, and pick a random (Gaussian) vector in the tangent space. Then we go along the geodesic (locally shortest path) in the direction of the chosen vector along the manifold. This is the curved step. We make these notions more precise below. The main intuition for considering such a generalization is the possibility of taking larger steps unimpeded by the boundary. This family of walks have been called "geodesic" walks [31]. We will shortly see an even more natural variant. For some quick background on Riemannian geometry, we refer the reader to the appendix.

4.1 Geodesic walks

Consider an explicit polytope P and a manifold M with support P and metric induced by the Hessian of the log barrier in P. Moving to this manifold view allows us to avoid the constraint of small steps near the boundary, as there is no longer a hard boundary constraint. How large can we make the step size? This is limited by another factor, namely, when we take large steps, the filter acceptance probability (to maintain the desired target distribution) can become very small. Could we possibly avoid using a filter? To answer this, we first consider the continuous time limit of the corresponding "diffusion" process.

$$dx_t = \mu(x_t)dt + \left(2D^2\phi(x_t)\right)^{-1/2} dW_t.$$

The first term, called the *drift*, is given by

$$\mu_i(x_t) = \sum_{j=1}^{n} \frac{\partial}{\partial x_j} \left(\left(\nabla^2\phi(x_t)\right)^{-1}\right)_{ij} = D \cdot (D^2\phi(x_t)^{-1})_i$$

where $D\cdot$ is the divergence operator. The drift term is a deterministic vector field biasing the walk locally. Its purpose is to ensure that the process converges to the desired target uniform distribution. Notably, in the continuous setting, it replaces the Metropolis filter. The precise form of the drift can be derived using the Fokker-Planck equation.

▶ **Theorem 26** (Fokker-Planck equation). *For any stochastic differential equation (SDE) of the form*

$$dx_t = \mu(x_t)dt + \sqrt{H(x_t)}dW_t,$$

for a symmetric matrix function H, the probability density of the SDE is given by the diffusion equation

$$\frac{\partial}{\partial t}p(x) = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}[\mu_i(x)p(x)] + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_i x_j}[H_{ij}(x)p(x)].$$

In discrete time, we consider the following *geodesic walk*:

$$x^{(j+1)} = \exp_{x^{(j)}}(\sqrt{h}w + \frac{h}{2}\mu(x^{(j)})), \tag{4.1}$$

where $\exp_{x^{(j)}}$ is the exponential map from the tangent space at $x^{(j)}$, $T_{x^{(j)}}M$, back to the manifold, w is a random Gaussian vector in the tangent space $T_{x^{(j)}}M$, $\mu(x^{(j)}) \in T_{x^{(j)}}M$ is the drift term and h is the step size. The Gaussian vector w has mean 0 and variance 1 in the metric at x, i.e., for any u, $\mathbb{E}_w\langle w, u\rangle_x^2 = \|x\|_x^2$. We write it as $w \sim N_x(0, I)$. This discrete walk converges to continuous diffusion as $h \to 0$ and it converges at a rate faster than the walk suggested by Euclidean coordinates, namely, $x^{(j+1)} = x^{(j)} + \sqrt{h}w + h\mu(x^{(j)})$.

Implementing this discretization leads to substantial challenges. The stationary distribution of the geodesic walk is not uniform. To get around this issue, we use rejection sampling. For step-size h chosen in advance, let $p(x \xrightarrow{w} y)$ be the probability density of going from x to y using the local step w.

Algorithm 1 Geodesic Walk.

```
At current point $x$:
Pick a Gaussian random vector $w \sim N_x(0, I)$.
Compute $y = \exp_x(\sqrt{h}w + \frac{h}{2}\mu(x))$.
Let $p(x \xrightarrow{w} y)$ be the probability density of going from $x$ to $y$ using the above step $w$.
Compute a corresponding $w'$ s.t. $x = \exp_y(\sqrt{h}w' + \frac{h}{2}\mu(y))$.
With probability $\min\left(1, \frac{p(y \xrightarrow{w'} x)}{p(x \xrightarrow{w} y)}\right)$, go to $y$; otherwise, stay at $x$.
```

In this algorithm, computing the exponential map and computing the transition probability density are nontrivial steps that need the efficient solution of an ODE. Even though the walk uses a Metropolis filter in the end, the parameter h can be made as large as $n^{-3/4}$, and the overall mixing time is $O(m/h)$, leading to the first sub-quadratic mixing rate for sampling polytopes. Note that this is effectively a step-size of $\sqrt{h} = n^{-3/8} \gg n^{-1/2}$, the limitation of previous methods.

▶ **Theorem 27** (Convergence of Geodesic Walk [29]). *The geodesic walk in a polytope with the log barrier converges to the uniform density in the polytope in $O(mn^{3/4})$ steps and each step can be implemented in $\widetilde{O}(mn^{\omega-1})$ arithmetic operations.*

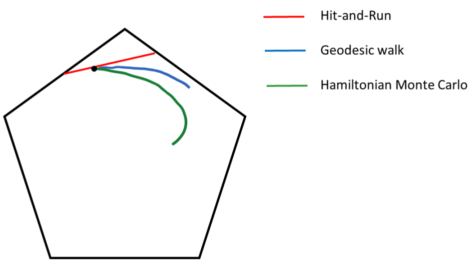

Figure 4.2 An illustration of Hit-and-run/Dikin, Geodesic walk and Hamiltonian Monte Carlo.

4.2 Riemannian Hamiltonian Monte Carlo

In the geodesic walk inspired by diffusion, in each step we choose a random direction and the drift along the entire step is determined by the (current) initial point. An even more natural discretization is to let the drift evolve along the trajectory in the chosen direction. The purpose of this modification would be to maintain the stationarity in spite of going to discrete time without introducing an explicit Metropolis filter. Can this be done? The answer is the method known as Hamiltonian Monte Carlo [40], and its manifold generalization [15]. To sample from a general distribution $e^{-H(x,y)}$. Hamiltonian Monte Carlo uses curves instead of straight lines in a time-reversible manner even if the target distribution is uniform.

▶ **Definition 28.** Given a continuous, twice-differentiable function $H : \mathcal{M} \times \mathbb{R}^n \subset \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ called the *Hamiltonian*, where $\mathcal{M}$ is the x domain of H, we say $(x(t), y(t))$ follows a *Hamiltonian curve* if it satisfies the *Hamiltonian equations*

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H(x,y)}{\partial y}, \\ \frac{dy}{dt} &= -\frac{\partial H(x,y)}{\partial x}. \end{aligned} \tag{4.2}$$

We define the map $T_\delta(x,y) \overset{\text{def}}{=} (x(\delta), y(\delta))$ where the $(x(t), y(t))$ follows the Hamiltonian curve with the initial condition $(x(0), y(0)) = (x, y)$.

Hamiltonian Monte Carlo is a sequence of randomly generated Hamiltonian curves.

Algorithm 2 Hamiltonian Monte Carlo.

```
Input: some initial point $x^{(1)} \in \mathcal{M}$.
for $k = 1, 2, \cdots, T$ do
    Sample $y^{(k+\frac{1}{2})}$ according to $e^{-H(x^{(k)},y)}/\pi(x^{(k)})$ where $\pi(x) = \int_{\mathbb{R}^n} e^{-H(x,y)} dy$.
    With probability $\frac{1}{2}$, set $(x^{(k+1)}, y^{(k+1)}) = T_\delta(x^{(k)}, y^{(k+\frac{1}{2})})$.
    Otherwise, $(x^{(k+1)}, y^{(k+1)}) = T_{-\delta}(x^{(k)}, y^{(k+\frac{1}{2})})$.
end
Output: $(x^{(T+1)}, y^{(T+1)})$.
```

▶ **Lemma 29.** *HMC has the following properties:*

1. **Energy conservation.** *For any Hamiltonian curve* $(x(t), y(t))$,

$$\frac{d}{dt} H(x(t), y(t)) = 0.$$

2. **Measure preservation.** *For any* $t \geq 0$,

$$\det\left(DT_t(x,y)\right) = 1$$

where $DT_t(x,y)$ *is the Jacobian of the map* T_t *at the point* (x,y).

3. **Time reversibility.** *For* $\pi(x) = \int_{\mathbb{R}^n} e^{-H(x,y)}dy$*, the probability density* $p_x(x')$ *of one step of HMC starting at* x *satisfies* $\pi(x)p_x(x') = \pi(x')p_{x'}(x)$.

Everything so far can be generalized to manifolds. To sample from the distribution $e^{-f(x)}$, we define

$$H(x,v) \stackrel{\text{def}}{=} f(x) + \frac{1}{2}\log((2\pi)^n \det g(x)) + \frac{1}{2}v^T g(x)^{-1}v \tag{4.3}$$

where g is the local metric of the manifold. In the Euclidean case, $g(x) = I$, and the Euclidean Hamiltonian is $H(x,v) = f(x) + \frac{1}{2}\|v\|^2$ and the corresponding dynamics becomes just $\frac{d^2x}{dt^2} = -\nabla f(x)$. One can view x as the position and v as the velocity. The convergence of HMC has been intensively studied in recent years, starting with [39], which established a mixing rate of $O(\kappa^2)$ for the case when f is strongly logconcave. For this setting it is now known that the optimal rate is $O(\kappa)$ [4].

We now restate the HMC equation for manifolds, as Riemannian HMC, noting that derivatives have to take the local metric into account.

▶ **Lemma 30.** *In Euclidean coordinates, the Hamiltonian equation for (4.3) can be rewritten as*

$$\begin{aligned} D_t\frac{dx}{dt} &= \mu(x), \\ \frac{dx}{dt}(0) &\sim N(0, g(x)^{-1}) \end{aligned}$$

where $\mu(x) = -g(x)^{-1}Df(x) - \frac{1}{2}g(x)^{-1}\mathrm{Tr}\left[g(x)^{-1}Dg(x)\right]$*, and* D_t *is the covariant derivative (or Levi-Civita connection) on the manifold* $\mathcal{M}$ *with metric* g.

With the above set-up, as long as we can compute the RHMC ODE accurately and efficiently, there is no need for a filtering step, leading to a significantly simpler implementation than that of the geodesic walk. In fact, it also gives a slightly improved mixing rate by allowing a larger time step in each iteration.

▶ **Theorem 31** (Convergence of RHMC [30])**.** *Riemannian Hamiltonian Monte Carlo in a polytope with the log barrier converges to the uniform density in the polytope in* $O(mn^{2/3})$ *steps.*

▶ **Question 32.** *Can the step-size of RHMC be improved so that the mixing rate is* $O(m\sqrt{n})$*? Can the mixing rate be improved to* $\widetilde{O}(n^{1.5})$ *with a weighted log barrier?*

RHMC appears to be quite practical, with recent implementations being able to sample from constrained distributions in dimension as high as 10^5 [26].

5 From diffusion to sampling: the manifold perspective

Langevin Diffusion (LD) is the following stochastic process[1]:

$$dX_t = -Df(X_t)dt + \sqrt{2}dB_t$$

[1] In this section, we use D for Euclidean derivative and reserve ∇ for manifold derivative.

where B_t is the standard Wiener process. The stationary distribution of this process is the density $\nu(x)$ proportional to $e^{-f(x)}$, a fact that can be verified using the Fokker-Planck equation for the time derivative of the density of X_t. To understand the rate of convergence, we recall two basic notions. First to measure the distance between distributions, we use the relative entropy (or KL divergence), defined as follows:

$$H_\nu(\rho) = \int \rho(x) \log \frac{\rho(x)}{\nu(x)}\, dx = \mathbb{E}_\rho \left(\log \frac{\rho}{\nu} \right).$$

In continuous time, the convergence will depend on what kind of isoperimetry is satisfied by the target density. So far we have seen Cheeger isoperimetry, which results in convergence in the χ^2-divergence. A stronger notion of isoperimetry is given by the Log-Sobolev Inequality. The relative Fisher information is defined as

$$J_\nu(\rho) = \int \rho \|\nabla \log(\frac{\rho}{\nu})\|^2 dx.$$

Note that this definition is for any Riemannian manifold.

▶ **Definition 33.** We say the distribution ν satisfies a Log-Sobolev inequality with constant α (called the log-Sobolev constant) if for every measure with density ρ we have

$$H_\nu(\rho) \le \frac{1}{2\alpha} J_\nu(\rho).$$

For logconcave densities in $\mathbb{R}^n$, an equivalent definition up to absolute constants is the following:

$$\sqrt{\alpha} \simeq \inf_{S \subseteq \mathbb{R}^n,\, \nu(S) \le \frac{1}{2}} \frac{\nu(\partial S)}{\nu(S)\sqrt{\log(1/\nu(S))}}$$

which looks like the Cheeger constant except for the additional log factor in the denominator – LSI requires the isoperimetric coefficient to get larger as the measure of the subset gets smaller.

A distribution that satisfies LSI has a strong convergence property in continuous time.

▶ **Theorem 34.** $H_\nu(\rho_t) \le e^{-2\alpha t} H_\nu(\rho_0)$.

This statement bears a remarkable resemblance to convergence of Gradient Descent to the optimal solution for strongly convex functions. In fact, as noted by Wibisono [45] (see also [20]), *Langevin Diffusion is Gradient Flow in the space of measures with the Wasserstein metric and the objective being the KL-divergence to the target density.*

We discuss this in a bit more detail below, starting from the simple optimization perspective.

▶ **Lemma 35.** *Let F be a function satisfying "Gradient Dominance":*

$$\|\nabla F(x)\|^2 \ge 2\alpha(F(x) - \min_x F(x)).$$

Then, the deterministic process $dx_t = -\nabla F(x_t)dt$ converges exponentially, i.e.,

$$F(x_t) - \min F \le e^{-2\alpha t}(F(x_0) - \min F).$$

.

Proof. We write

$$\begin{aligned}\frac{d}{dt}(F(x_t) - \min_x F(x)) &= \langle \nabla F(x_t), \frac{dx_t}{dt}\rangle \\ &= -\left\|\nabla F(x_t)\right\|^2 \\ &\leq -2\alpha \cdot (F(x_t) - \min_x F(x)).\end{aligned}$$

The conclusion follows. ◀

For $f : \mathbb{R}^n \to \mathbb{R}_+$, let ν be the density proportional to $e^{-f(x)}$ and $F(\rho) = H_\nu(\rho)$. We will apply the above lemma to this KL-divergence objective.

▶ **Lemma 36.** *For any two densities ρ, ν,*

$$\left\|\nabla_\rho H_\nu(\rho)\right\|_\rho^2 = \int \rho(x) \left\|D \log \frac{\rho(x)}{\nu(x)}\right\|^2 dx = J_\nu(\rho).$$

▶ **Theorem 37.** *Let f be a differentiable function with log-Sobolev constant α. Then the Langevin dynamics*

$$dx_t = -Df(x)dt + \sqrt{2}dW_t$$

converges exponentially in KL-divergence to the density $\nu(x) \propto e^{-f(x)}$ with mixing rate $O(1/\alpha)$, i.e., $H_\nu(\rho_t) \leq e^{-2\alpha t} H_\nu(\rho_0)$.

Proof. First, note that the Langevin SDE, by Fokker-Planck, corresponds to the following PDE:

$$\begin{aligned}\frac{d\rho(x)}{dt} &= D \cdot (\rho(x)Df) + \Delta\rho(x) \\ &= D \cdot \left(\rho(x) D \log \frac{\rho(x)}{\nu(x)}\right) \\ &= -\nabla_\rho H_\nu(\rho).\end{aligned}$$

Note that the last step above refers to the derivative with respect to the Wassterstein metric (see [44]). Next, we note that since ν satisfies the log-Sobolev inequality, $F(\rho) = H_\nu(\rho)$ satisfies the Gradient Dominance condition with parameter α. The theorem follows from Lemma 35. ◀

So far, we have only used the log-Sobolev inequality, not convexity or related properties. Will this suffice for an efficient algorithm? It was shown in [43] that this is indeed the case. The Unadjusted Langevin Algorithm (ULA) is the simple Euler discretization of Langevin dynamics, namely,

$$X_{k+1} = X_k - \delta \cdot Df(X_k) + \sqrt{2\delta} Z$$

where $Z \sim N(0, I)$ is standard Gaussian. The next theorem is for $\mathbb{R}^n$.

▶ **Theorem 38** (ULA converges under Isoperimetry [43])**.** *Assume $\nu = e^{-f}$ satisfies LSI with constant $\alpha > 0$ and Df is L-Lipschitz. Then ULA with step size $0 < \delta \leq \frac{\alpha}{4L^2}$ satisfies*

$$H_\nu(\rho_k) \leq e^{-\alpha\delta k} H_\nu(\rho_0) + \frac{8\delta n L^2}{\alpha}.$$

For any $0 < \varepsilon < 4n$, ULA with step size $\delta \leq \frac{\alpha\varepsilon}{16nL^2}$ reaches error $H_\nu(\rho_k) \leq \varepsilon$ after at most $k \geq \frac{1}{\alpha\delta} \log \frac{2H_\nu(\rho_0)}{\varepsilon}$ iterations.

Due to the simplicity and generality of the approach, Langevin Algorithms have been intensively studied in recent years for many special cases and under various additional conditions on the input [9, 5, 11, 34, 46, 6]. Here we consider an extension: what is the generalization of Langevin diffusion from Euclidean space to more general metrics?

▶ **Definition 39** (Riemannian Langevin Diffusion (RLD)). Let $\mathcal{M}$ be a manifold with metric g and measure $dv_g(x)$, whose density with respect to the Lebesgue measure is $\sqrt{|\det(g)|}$. For a distribution with density $\nu = e^{-F}$, Riemannian Langevin Diffusion is given by the following SDE whose stationary distribution is νdv_g:

$$dX_t = (\nabla \cdot (g^{-1}(X_t)) - \nabla F(X_t))dt + \sqrt{2g^{-1}(X_t)}dB_t.$$

Here $\nabla\cdot$ denotes the divergence with respect to the manifold and it is applied separately to each row of a matrix. When written in local Euclidean coordinates, this becomes

$$dX_t = (D \cdot (g^{-1}(X_t)) - g^{-1}(X_t)Df(X_t))dt + \sqrt{2g^{-1}(X_t)}dB_t$$

where $f(x) = F(x) - \frac{1}{2}\log\det g(x)$ and the derivative and divergence are Euclidean. When f is constant, then $Df = 0$ and we get the equation for Brownian motion on the manifold.

RLD has the same convergence guarantee in continuous time as stated by Theorem 34, with ν being the manifold stationary measure. While LD can only be applied to smooth target densities, RLD allows us to use the metric g to incorporate constraints and sample from e^{-f} subject to constraints. For example, by letting g be the log-barrier of a polytope we get to sample from densities restricted to the polytope, in continuous time. We next define the extension to a discrete time algorithm.

▶ **Definition 40** (Riemannian Langevin Algorithm (RLA)). To sample from the distribution $\nu dv_g(x)$ over the manifold $\mathcal{M}$, for a fixed step size ϵ, repeat:

$$\begin{aligned} y &= \exp_{x_k}(-\epsilon\nabla F(x_k)) \\ x_{k+1} &= \mathfrak{B}(y, \epsilon) \end{aligned}$$

where $\mathfrak{B}(x_0, t)$ samples from Brownian motion on the manifold, starting from x_0 after time t.

Recently, following work of [33, 1], [14] made progress on analyzing RLA in this general setting, showing that an extension of self-concordance of the metric suffices to guarantee convergence under the log-Sobolev inequality.

6 Discussion

The study of sampling has led to new tools, both for analysis and for algorithms, surprising connections with convex geometry, functional analysis and optimization, and practical algorithms in high dimension. We conclude by highlighting a few more open problems.

▶ **Question 41.** *Is there a natural KLS conjecture for Hessian manifolds?*

The close connection between diffusion and sampling raises an interesting sampling problem, namely efficiently simulating Brownian motion on manifolds.

▶ **Question 42.** *Given a metric g, an initial point x_0 and a time interval t, give an algorithm to sample from the distribution of Brownian motion on the manifold starting at x_0 for time t. For a manifold with metric g, Brownian motion is given by:*

$$dX_t = \nabla \cdot (g^{-1})dt + \sqrt{2g^{-1}}dB_t.$$

References

1 Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. *Advances in Neural Information Processing Systems*, 34, 2021.

2 Apostolos Chalkis and Vissarion Fisikopoulos. volEsti: Volume approximation and sampling for convex polytopes in R. *arXiv preprint*, 2020. `arXiv:2007.01578`.

3 Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.

4 Zongchen Chen and Santosh S. Vempala. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. *Theory of Computing*, 18(9):1–18, 2022. `doi:10.4086/toc.2022.v018a009`.

5 Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In *Conference on Learning Theory (COLT)*, pages 300–323. PMLR, 2018.

6 Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. *arXiv preprint*, 2021. `arXiv:2112.12662`.

7 Ben Cousins and Santosh Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016.

8 Ben Cousins and Santosh Vempala. Gaussian cooling and O*(nˆ3) algorithms for volume and gaussian volume. *SIAM Journal on Computing*, 47(3):1237–1273, 2018.

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

10 II Dikin. Iterative solution of problems of linear and quadratic programming. In *Doklady Akademii Nauk*, volume 174(4), pages 747–748. Russian Academy of Sciences, 1967.

11 Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

12 R. Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23:532–569, 2013.

13 B. Fleury. Concentration in a thin Euclidean shell for log-concave measures. *J. Funct. Anal.*, 259(4):832–841, 2010.

14 Khashayar Gatmiry and Santosh S Vempala. Convergence of the riemannian langevin algorithm. *arXiv preprint*, 2022. `arXiv:2204.10818`.

15 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, 2011.

16 Olivier Guedon and Emanuel Milman. Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures. *Geometric and Functional Analysis*, 21(5):1043–1068, 2011. `doi:10.1007/s00039-011-0136-5`.

17 Hulda S Haraldsdóttir, Ben Cousins, Ines Thiele, Ronan MT Fleming, and Santosh Vempala. Chrr: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743, 2017.

18 Roland Hildebrand. Canonical barriers on convex cones. *Mathematics of operations research*, 39(3):841–850, 2014.

19 He Jia, Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Reducing isotropy and volume to KLS: an $O^*(n^3\psi^2)$ volume algorithm. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 961–974, 2021.

20 Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

21 R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13:541–559, 1995.

22 R. Kannan, L. Lovász, and M. Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11:1–50, 1997.

23 Ravindran Kannan and Hariharan Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012.

24 N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–396, 1984.

25 Bo'az Klartag and Joseph Lehec. Bourgain's slicing problem and kls isoperimetry up to polylog. *arXiv preprint*, 2022. `arXiv:2203.15551`.

26 Yunbum Kook, Yin Tat Lee, Ruoqi Shen, and Santosh S. Vempala. Sampling with riemannian hamiltonian monte carlo in a constrained space, 2022. `doi:10.48550/ARXIV.2202.01908`.

27 Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Strong self-concordance and sampling. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1212–1222, 2020.

28 Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in sqrt(rank) iterations and faster algorithms for maximum flow. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 424–433, 2014.

29 Yin Tat Lee and Santosh Vempala. Geodesic walks in polytopes. *SIAM Journal on Computing*, 51(2):STOC17–400–STOC17–488, 2022. `doi:10.1137/17M1145999`.

30 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)*, pages 1115–1121, 2018.

31 Yin Tat Lee and Santosh S Vempala. The Kannan-Lovász-Simonovits conjecture. *arXiv preprint*, 2018. `arXiv:1807.03465`.

32 Yin Tat Lee and Santosh Srinivas Vempala. Eldan's stochastic localization and the KLS hyperplane conjecture: An improved lower bound for expansion. *CoRR*, abs/1612.01507, 2016. `arXiv:1612.01507`.

33 Mufan Bill Li and Murat A Erdogdu. Riemannian langevin algorithm for solving semidefinite programs. *arXiv preprint*, 2020. `arXiv:2010.11176`.

34 Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

35 L. Lovász. Hit-and-run mixes fast. *Math. Prog.*, 86:443–461, 1998.

36 L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. In *Random Structures and Alg.*, volume 4, pages 359–412, 1993.

37 L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM J. Computing*, 35:985–1005, 2006.

38 L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

39 Oren Mangoubi and Aaron Smith. Rapid mixing of hamiltonian monte carlo on strongly log-concave distributions. *arXiv preprint*, 2017. `arXiv:1708.07114`.

40 Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

41 Yurii Nesterov, Arkadii Nemirovskii, and Yinyu Ye. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

42 S. Vempala. Geometric random walks: A survey. *MSRI Combinatorial and Computational Geometry*, 52:573–612, 2005.

43 Santosh Vempala and Andre Wibisono. Rapid convergence of the Unadjusted Langevin Algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

44 Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

45 Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 2093–3027, 2018. URL: `http://proceedings.mlr.press/v75/wibisono18a.html`.

46 Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the metropolis-adjusted langevin algorithm for log-concave sampling. *arXiv preprint*, 2021. `arXiv:2109.13055`.

A Riemannian metrics

A manifold M can be viewed as an n-dimensional "surface" in $\mathbb{R}^k$ for some $k \geq n$.

1. Riemannian metric: For any $v, u \in T_pM$, the inner product (Riemannian metric) at p is given by $\langle v, u\rangle_p$ and this allows us to define the norm of a vector $\|v\|_p = \sqrt{\langle v, v\rangle_p}$. We call a manifold a Riemannian manifold if it is equipped with a Riemannian metric. When it is clear from context, we define $\langle v, u\rangle = \langle v, u\rangle_p$. In $\mathbb{R}^n$, $\langle v, u\rangle_p$ is the usual ℓ_2 inner product.
2. Tangent space T_pM: For any point p, the tangent space T_pM of M at point p is a linear subspace of $\mathbb{R}^k$ of dimension n. Intuitively, T_pM is the vector space of possible directions that are tangential to the manifold at x. Equivalently, it can be thought as the first-order linear approximation of the manifold M at p. For any curve c on M, the direction $\frac{d}{dt}c(t)$ is tangent to M and hence lies in $T_{c(t)}M$. For any open subset M of $\mathbb{R}^n$, we can identify T_pM with $\mathbb{R}^n$.

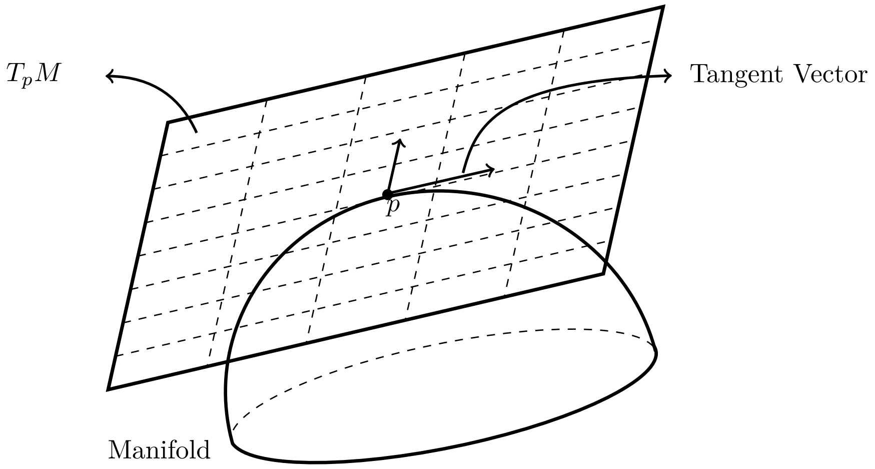

Figure A.1 Riemannian Manifold and Tangent Space.

3. Hessian manifold: We call M a Hessian manifold (induced by ϕ) if M is an open subset of $\mathbb{R}^n$ with the Riemannian metric at any point $p \in M$ defined by
$$\langle v, u\rangle_p = v^\top \nabla^2 \phi(p) u$$
where $v, u \in T_pM$ and ϕ is a smooth convex function on M.
4. Length: For any curve $c : [0, 1] \to M$, we define its length by
$$L(c) = \int_0^1 \left\| \frac{d}{dt} c(t) \right\|_{c(t)} dt.$$
5. Distance: For any $x, y \in M$, we define $d(x, y)$ be the infimum of the lengths of all paths connecting x and y. In $\mathbb{R}^n$, $d(x, y) = \|x - y\|_2$.

6. Geodesic: We call a curve $\gamma(t) : [a, b] \to M$ a geodesic if it satisfies both of the following conditions:

   a. The curve $\gamma(t)$ is parameterized with constant speed. Namely, $\left\|\frac{d}{dt}\gamma(t)\right\|_{\gamma(t)}$ is constant for $t \in [a, b]$.
   b. The curve is the locally shortest length curve between $\gamma(a)$ and $\gamma(b)$. Namely, for any family of curve $c(t, s)$ with $c(t, 0) = \gamma(t)$ and $c(a, s) = \gamma(a)$ and $c(b, s) = \gamma(b)$, we have that $\frac{d}{ds}\big|_{s=0} \int_a^b \left\|\frac{d}{dt}c(t, s)\right\|_{c(t,s)} dt = 0$.

   Note that, if $\gamma(t)$ is a geodesic, then $\gamma(\alpha t)$ is a geodesic for any α. Intuitively, geodesics are local shortest paths. In $\mathbb{R}^n$, geodesics are straight lines.
7. Exponential map: The map $\exp_p : T_pM \to M$ is defined as

   $$\exp_p(v) = \gamma_v(1)$$

   where γ_v is the unique geodesic starting at p with initial velocity $\gamma_v'(0)$ equal to v. The exponential map takes a straight line $tv \in T_pM$ to a geodesic $\gamma_{tv}(1) = \gamma_v(t) \in M$. Note that $\exp_p$ maps v and tv to points on the same geodesic. Intuitively, the exponential map can be thought as point-vector addition in a manifold. In $\mathbb{R}^n$, we have $\exp_p(v) = p + v$.
8. Parallel transport: Given any geodesic $c(t)$ and a vector v such that $\langle v, c'(0)\rangle_{c(0)} = 0$, we define the parallel transport of v along $c(t)$ by the following process: Take h to be infinitesimally small and $v_0 = v$. For $i = 1, 2, \cdots, 1/h$, we let v_{ih} be the vector orthogonal to $c'(ih)$ that minimizes the distance on the manifold between $\exp_{c(ih)}(hv_{ih})$ and $\exp_{c((i-1)h)}(hv_{(i-1)h})$. Intuitively, the parallel transport finds the vectors on the curve such that their end points are closest to the end points of v. For general vector $v \in T_{c'(0)}$, we write $v = \alpha c'(0) + w$ and we define the parallel transport of v along $c(t)$ is the sum of $\alpha c'(t)$ and the parallel transport of w along $c(t)$. For non-geodesic curve, see the definition in Fact 43.
9. Orthonormal frame: Given vector fields $v_1, v_2, \cdots, v_n$ on a subset of M, we call $\{v_i\}_{i=1}^n$ is an orthonormal frame if $\langle v_i, v_j\rangle_x = 1$ if $i = j$ and 0 otherwise. Given a curve $c(t)$ and an orthonormal frame at $c(0)$, we can extend it along the curve by parallel transport and it remains orthonormal on the whole curve.
10. Directional derivatives and the Levi-Civita connection: For a vector $v \in T_pM$ and a vector field u in a neighborhood of p, let γ_v be the unique geodesic starting at p with initial velocity $\gamma_v'(0) = v$. Define

    $$\nabla_v u = \lim_{h\to 0} \frac{u(h) - u(0)}{h}$$

    where $u(h) \in T_pM$ is the parallel transport of $u(\gamma_v(h))$ from $\gamma(h)$ to $\gamma(0)$. Intuitively, Levi-Civita connection is the directional derivative of u along direction v, *taking the metric into account.* In particular, for $\mathbb{R}^n$, we have $\nabla_v u(x) = \frac{d}{dt}u(x + tv)$. When u is defined on a curve c, we define $D_t u = \nabla_{c'(t)} u$. In $\mathbb{R}^n$, we have $D_t u(\gamma(t)) = \frac{d}{dt}u(\gamma(t))$. We reserve $\frac{d}{dt}$ for the usual derivative with Euclidean coordinates.

We list some basic facts about the definitions above.

▶ **Fact 43.** *Given a manifold M, a curve $c(t) \in M$, a vector v and vector fields u, w on M, we have the following:*

1. *(alternative definition of parallel transport) $v(t)$ is the parallel transport of v along $c(t)$ if and only if $\nabla_{c'(t)} v(t) = 0$.*

2. *(alternative definition of geodesic) c is a geodesic if and only if* $\nabla_{c'(t)} c'(t) = 0$.
3. *(linearity)* $\nabla_v (u + w) = \nabla_v u + \nabla_v w$.
4. *(product rule) For any scalar-valued function f,* $\nabla_v (f \cdot u) = \frac{\partial f}{\partial v} u + f \cdot \nabla_v u$.
5. *(metric preserving)* $\frac{d}{dt} \langle u, w \rangle_{c(t)} = \langle D_t u, w \rangle_{c(t)} + \langle u, D_t w \rangle_{c(t)}$.
6. *(torsion free-ness) For any map* $c(t, s)$ *from a subset of* $\mathbb{R}^2$ *to* M*, we have that* $D_s \frac{\partial c}{\partial t} = D_t \frac{\partial c}{\partial s}$ *where* $D_s = \nabla_{\frac{\partial c}{\partial s}}$ *and* $D_t = \nabla_{\frac{\partial c}{\partial t}}$.
7. *(alternative definition of Levi-Civita connection)* $\nabla_v u$ *is the unique linear mapping from the product of vector and vector field to vector field that satisfies (3), (4), (5) and (6).*