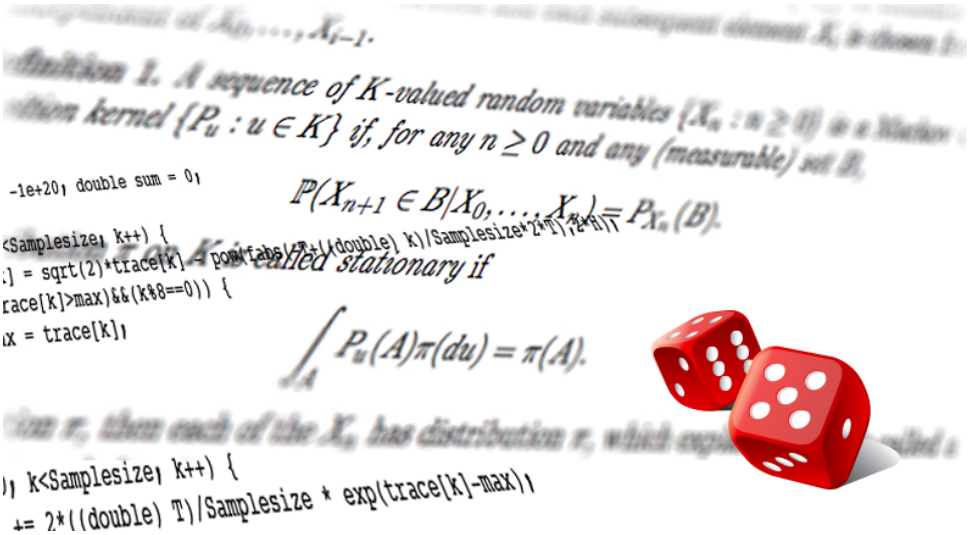
Stochastic processes and algorithms

Ton Dieker
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
ton.dieker@isye.gatech.edu

partially based on notes from Santosh Vempala



Lecture notes
University of Wrocław
November 19 – 23, 2012

List of Lectures

1 Introduction **1**

I Applications **4**

2 Optimization and volume **5**

3 Ranking and selection **7**

II Tools **10**

4 Markov chains **11**

5 Conductance and convergence **13**

6 Convexity and inequalities **15**

7 Isoperimetric inequalities **17**

III Algorithms and their analysis **20**

8 Geometric random walks **21**

9 Mixing for the ball walk **23**

10 Mixing for shake-and-bake **26**

1 Introduction

Sampling techniques play an increasingly important role in scientific and business applications. On the one hand, randomized algorithms been shown to remove (sometimes in an 'approximate' sense) some of the known computational barriers of deterministic algorithms, see the landmark papers [7, 16, 17, 24, 26]. On the other hand, techniques such as (Gibbs) sampling for missing data (e.g., [39, 22, 23]) are quickly gaining significance in the era of 'big data'. The past decades have witnessed great advances in this domain through the cross-fertilization of operations research, probability and statistics, computer science, and statistical physics.

In these lecture notes, we are interested in developing techniques for sampling from high-dimensional sets such as convex bodies and its applications, based on the Markov Chain Monte Carlo (MCMC) method. To assess the performance of these sampling schemes, we study the regime where the ambient dimension grows to infinity. Since the required computational effort is proportional to the mixing time (sometimes called burn-in or relaxation time), studying these algorithms requires tools for mixing times of Markov chains. This leads us to excursions into topics such as convex geometry and isoperimetric inequalities.

1.1 Sampling from high-dimensional objects

The problem of generating a sample of points according to a specified probability distribution over some set K is a fundamental problem in operations research with various uses [42]: (i) global optimization, where sampling forms the basis of simulated annealing and is used to generate candidate points for local searches [38, 43, 47], (ii) Monte Carlo integration, where one seeks to evaluate the integral $\int_K h(x)f(x)dx$ for some function h and some probability density f [8], (iii) identifying constraints that can be dropped in order to describe a polytope [3, 41], (iv) computing or approximating the volume of a convex body, which is a basic yet extremely difficult task in general [34]. The special case of sampling *uniformly* from the set K is the most fundamental setting in the sense that one can often bootstrap algorithms to handle a more general case, see for instance [4, 9, 36] and references therein.

Three different families of sets K are of particular interest: (1) K is convex, (2) K is discrete, and (3) K is the boundary of a convex set. The case of discrete K is fundamentally different since the geometry of the ambient space is often irrelevant, and this is not the focus of these lectures.

The global structure of the set K is rarely available in 'explicit' form. Therefore, we are interested in algorithms that only have access to a membership oracle, which can tell the algorithm whether any given point $x \in \mathbb{R}^n$ lies in K. We also briefly need so-called separation oracles, which additionally give a hyperplane separating x and K if x does not lie in K. Our main focus is not intrinsically dependent on the type of oracle, it is to find suitable algorithms.

It is instructive to see a simple example where an algorithm's performance can be assessed by letting the ambient dimension grow to infinity; this is the central feature of these notes. Suppose we want to sample from an n-dimensional ball B in $\mathbb{R}^n$ with radius 1. Also suppose we embed B within a hypercube with side length 2, and that we keep sampling from the hypercube until the realization lies in the ball. This is known as the *acceptance-rejection method*. The output

of this algorithm has the correct statistical properties regardless the value of the dimension n. However, the probability that a sample from the hypercube lies in the ball is

$$\frac{\pi^{n/2}}{\Gamma(n/2+1)} \approx \frac{1}{\sqrt{2}}\pi^{n/2-1}(n/2+1)^{-n/2-1/2}e^{n/2+1},$$

where Γ is the gamma function. Thus, as n grows, the ball only takes up a tiny fraction of the hypercube in terms of volume. Consequently this algorithm generates a large number of samples before finding a point inside the ball, and thus the algorithm performs poorly even for moderate n. In the following exercise, you are asked to devise an algorithm that *does* work efficiently in the setting of this example.

Exercise 1. Formulate a polynomial-time algorithm for sampling from a unit ball. ☺

On polynomial time algorithms

An algorithm is said to be a *polynomial time algorithm* if its running time is upper bounded by a polynomial in the size of the input for the algorithm. We think of any polynomial-time algorithm in n as a *fast*, *efficient*, or *tractable* algorithm. Problems for which a polynomial time algorithm exists belong to the complexity class P, which plays a central role in complexity theory.

Throughout these notes, we are interested in *fast* sampling algorithms in the sense that we study algorithms that turn out to be polynomial-time algorithms in the ambient dimension n. A central feature of the algorithms we discuss here is that they rely on *randomness*. To motivate the study of such randomized algorithms, consider the problem of estimating the volume of a convex body K in $\mathbb{R}^n$ which contains a unit ball. The next theorem essentially says that no deterministic algorithm can estimate the volume of K efficiently. For instance, it says that one needs exponential time to get a factor 2 approximation (say). The theorem is often attributed to Bárány and Füredi, but we refer to [6] for a more detailed history.

Theorem 1. *For every deterministic algorithm that runs in time $O(n^a)$ and outputs two numbers A and B such that $A \leq \mathrm{vol}(K) \leq B$ for any convex body K, there is some convex body for which the ratio B/A is at least*

$$\left(\frac{cn}{a\log n}\right)^n$$

where c is some absolute constant.

Given this negative result for deterministic algorithms, the following positive result from [17] is remarkable. Using randomness, we can overcome the preceding negative result!

Theorem 2. *For any convex body K and any $0 \leq \epsilon, \delta \leq 1$, there is a randomized algorithm which computes an estimate V such that with probability at least $1-\delta$, we have $(1-\epsilon)\mathrm{vol}(K) \leq V \leq (1+\epsilon)\mathrm{vol}(K)$, and the number of oracle calls is polynomial in n, $1/\epsilon$, and $\log(1/\delta)$.*

This algorithm and related ones are based on Markov Chain Monte Carlo methodology and the analysis of these algorithms boils down to establishing rapid mixing for Markov chains. Several elements are needed, and we will discuss the first two in much detail during this week: (1) Devising a Markov chain with a uniform stationary distribution. (2) Showing that one can sample efficiently from an 'almost' uniform distribution by showing that, after a polynomial number of steps of the Markov chain, the distribution of the chain is 'close' to the stationary distribution. This is called *rapid mixing*. (3) Using the sampling methodology to produce estimates for the volume.

1.2 About these notes

These notes consist of ten lectures, and they are organized into three parts. In Part I, we focus on applications of sampling from high-dimensional sets: convex optimization, volume computation, and ranking and selection problem. In Part II, we discuss the mathematical background and tools we need in this short course. In Part III, we introduce and analyze algorithms for sampling high-dimensional objects. Much of what I discuss is "well-known" in certain circles; general surveys of this material include [6, 45].

In these notes, I have aimed to emphasize *ideas*. As a result, I have had to make some sacrifices regarding mathematical precision. In that sense, the style in these notes is different from what you would see in a research paper or a semester-long specialized course, where details such as measurability are filled in carefully. I have chosen this approach in order to be able to summarize the full breadth of techniques in a short course of ten lectures.

I have used the following papers as the primary sources for the lectures, although much of the material has been presented in other papers as well. Lectures 2, 8, and 9 are adaptations of parts of [45], copied with permission. (But any errors, inaccuracies, and typos are mine.) Lectures 4 and 5 are based on [34]. Lecture 6 and part of Lecture 7 is based on [21], the other part is based on [27]. Lectures 3 and 10 are based on original work in progress; Lecture 3 with Seong-Hee Kim and Lecture 10 with Santosh Vempala.

1.3 Acknowledgments

Special thanks to Krzysztof Dębicki, Zbigniew Palmowski, Tomasz Rolski, and Ryszard Szekli for inviting me to deliver these lectures as a short course at the University of Wrocław. I would also like to thank Santosh Vempala for many discussions on stochastic processes and algorithms, which have heavily influenced these notes. Last but not least, I would like to thank the participants in the course for their good questions and comments, which have resulted in significant improvements to these notes.

Part I

Applications

2 | Optimization and volume

In this lecture, we discuss two applications of sampling from convex bodies.

2.1 Application I: Convex optimization

Let $S \subset \mathbb{R}^n$, and $f : S \to \mathbb{R}$ be a real-valued function. Optimization is the following basic problem: $\min f(x)$ s.t. $x \in S$, that is, find a point $x \in S$ which minimizes by f. We denote by x^* a solution for the problem. When the set S and the function f are convex, we obtain a class of problems which are solvable in polynomial time in n and $\log(1/\epsilon)$, where ϵ defines an optimality criterion. If x is the point found, then $|x - x^*| \leq \epsilon$.

The problem of minimizing a convex function over a convex set in $\mathbb{R}^n$ is a common generalization of well-known geometric optimization problems such as linear programming as well as a variety of combinatorial optimization problems including matchings, flows and matroid intersection, all of which have polynomial-time algorithms [25]. The so-called ellipsoid method solves these problems in polynomial time when K is given by a separation oracle. The algorithm discussed here is based on random sampling, see [4] for details. The idea is to minimize a convex function f by looking at the convex set $K_t = \{x \in \mathbb{R}^n : f(x) \leq t\}$ and search (in a binary fashion) for the smallest t for which this set is nonempty.

In the description below, we assume that the convex set K is contained in the axis-aligned cube of width R centered at the origin; further if K is non-empty then it contains a cube of width r. It is easy to show that any algorithm with this input specification needs to make at least $n \log(R/r)$ oracle queries. We assume a separation oracle is given to us; we later briefly discuss what happens when we only have a membership oracle. We next describe the algorithm; the parameter L is equal to $\log \frac{R}{r}$.

Input: A separation oracle for a convex set K and some number L.
Output: A point in K or a guarantee that K is empty.

1. Let P be the axis-aligned cube of side length R and center $z = 0$.
2. Check if z is in K. If so, report z and stop. If not, set
$$H = \{x : a^T x \leq a^T z\}.$$
where $a^T x \leq b$ is the halfspace containing K reported by the oracle.
3. Set $P = P \cap H$. Pick m random points $y^1, y^2, \ldots, y^m$ from P. Set z to be their average.
4. Repeat steps 2 and 3 at most $2nL$ times. Report K is empty.

The algorithm computes an approximate centroid in each iteration. Finding the exact centroid is #P-hard, i.e., computationally intractable. The idea behind the algorithm is that an

approximate centroid can be computed using $O(n)$ random points and the volume of P is likely to drop by a constant factor in each iteration with this choice of z, as formalized in the next lemma. This idea builds on Grünbaum's lemma, which says that the volume drops by a constant factor if z is the centroid of P.

Lemma 1. *Let z be the average of m points sampled from the uniform distribution π on a convex set P. If H is a halfspace containing z, then we have*

$$\mathbb{E}\left(\pi(H)\right) \geq \left(\frac{1}{e} - \sqrt{\frac{n}{m}}\right).$$

The guarantee on the algorithm follows immediately.

Theorem 3. *With high probability, the algorithm works correctly using at most $2nL$ oracle calls (and iterations).*

The algorithm can also be modified for optimization given a membership oracle only and a point in K. It has a similar flavor: sample random points from K; restrict K using the function value at the average of the random points; repeat. The oracle complexity turns out to be $O(n^5 L)$.

2.2 Application II: Volume computation

We next discuss perhaps the most important application and the principal motivation behind many developments in the theory of random walks: the problem of computing the volume of a convex body K. This has already briefly been discussed in Section 1.1. There are various ways to devise algorithms to compute the volume of a convex body, see for instance [6, 45].

Assume that the diameter D of K is polynomial in n, and that K contains a unit ball B_n. All known algorithms reduce volume computation to sampling from a convex body using the *multi-phase Monte Carlo* technique. They construct a sequence of convex bodies $K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K$, where $K_0 = B_n$ or some body whose volume is easily computed. The idea is to estimate the ratios $\mathrm{vol}(K_{i-1})/\mathrm{vol}(K_i)$ by generating sufficiently many independent (nearly) uniformly distributed random points in K_i and counting the fraction that lie in K_{i-1}. The product of these estimates is an estimate of $\mathrm{vol}(K_0)/\mathrm{vol}(K)$.

In order to get a sufficiently good estimate for the ratio $\mathrm{vol}(K_{i-1})/\mathrm{vol}(K_i)$, one needs about $m\mathrm{vol}(K_i)/\mathrm{vol}(K_{i-1})$ random points. So we would like to have the ratios $\mathrm{vol}(K_i)/\mathrm{vol}(K_{i-1})$ be small. But, the ratio of $\mathrm{vol}(K)$ and $\mathrm{vol}(K_0)$ could be $n^{\Omega(n)}$ (e.g., when K is a hypercube), and we need to keep $\mathrm{vol}(K_i)/\mathrm{vol}(K_{i-1})$ polynomial in order for the sampling scheme to work efficiently. A good choice is to keep these ratios bounded; this can for instance be achieved by setting $K_0 = B_n$ (unit ball) and $K_i = K \cap (2^{i/n} B_n)$ for $i = 1, 2, \ldots, m = O(n \log n)$.

3 | Ranking and selection

This lecture describes an application for sampling from high-dimensional objects I am currently working on.

In decision problems, one is often faced with the problem of selecting the best among several systems from noisy evaluations of their performance. For instance, one can think of investigating multiple possibilities to allocate service capacities in a stochastic network, where we want to select (say) the system with the lowest wait times [14, 15]. One could perform a simulation study to obtain performance estimates for each of the systems, but this results in noisy observations due to the use of simulation. It can be very costly to perform such a simulation study, especially if you perform a real-world simulation as opposed to a computer simulation. Another example are clinical trials, where different combinations of treatments are tested in order to find the combination that works best. In each of these cases, the sooner you know which systems underperform, the better; you can discard those systems from your simulations or trials. Therefore, we are interested in sequential procedures for judging which system is the best.

3.1 Background

The aforementioned selection-of-the-best problem has been studied actively in the context of simulation ranking and selection. Selection-of-the-best procedures are also found to be useful within optimization via simulation algorithms to clean up at the end of an optimization search or assist in selecting the most promising neighbor. Although many statistical procedures with statistical validity for correct selection have been presented, many of these use the Bonferroni inequality to reduce the many-system case to a (one-dimensional) two-system case [29, 30], and consequently the resulting statistical procedures are conservative even if the number of systems is moderately large. This tends to be a serious problem as the number of systems increase. It limits the number of simulated alternatives the procedures can handle and it limits their applications. We are hoping that the sampling algorithms from these lectures can break this barrier, but our current understanding is not good enough yet to answer in the affirmative. Regardless, I hope you find the proposed ideas interesting.

There exist at least three different approaches to the selection-of-the-best problem: the indifference-zone method, the Bayesian method, or the optimal computing budget allocation method. The indifference-zone method finds the best system with a guarantee on the probability of correct selection (PCS) while the other two methods maximize the PCS given a limited computational budget. A review of the Bayesian and optimal computing budget allocation methods can be found in [10], and a review of the indifference-zone method can be found in [31]; see also [11, 19, 29]. Some procedures are efficient in handling up to 1,000 simulated systems and have been successfully incorporated into commercial software packages such as Arena® from Rockwell Automation and Simio® from Simio LLC.

3.2 A sampling problem

It transpires from the above discussion that a critical issue with some existing ranking-and-selection methods is that one derives bounds for a high-dimensional problem in terms of one-dimensional problems. In doing so, one sacrifices accuracy. Therefore, we are interested in a procedure that directly works with high-dimensional objects.

As we shall see momentarily, this problem leads us to the question of sampling from an ellipsoid, which is the analog of an ellipse in arbitrary dimension. Suppose we seek a guarantee on the probability of selecting the correct (true best) system. One component of this probability is the probability that the best system does not get eliminated, and this is our main focus. Therefore, an approach based on stochastic processes in high dimensions holds promise for being able to identify the best system much faster. Some initial steps have been taken in joint work with Seong-Hee Kim [13, 28], where we make a first step away from one-dimensional frameworks by working with two-dimensional Brownian motions. Here we present it more generally (but we don't know yet if this can be pushed through completely).

We assume that there are n systems, $n \geq 2$. At each time epoch, we obtain an observation from each of the n systems. Motivated by the connection with simulation, we call these observations *replications*. We assume these replications are i.i.d. Let Y_{ij} be an observation from replication j of system i for $i = 1, \ldots, n$ and $j = 1, 2, \ldots$. The mean and variance of the outputs from system i are defined as $\mu_i = \mathbb{E}(Y_{ij})$ and $\sigma_i^2 = \mathbb{V}\mathrm{ar}(Y_{ij})$, respectively. Let $Y_j = (Y_{1j}, Y_{2j}, \ldots, Y_{kj})$ represent the observations from replication j for all n systems, and write $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$. Then we assume that the Y_j are i.i.d. with a normal distribution with mean μ and covariance matrix Σ. (In practice, variances are unknown and one instead resorts to estimates.) It is important to allow for correlations, since for instance the output of the simulation can be correlated due to the use of a simulation technique known as common random numbers.

In admittedly somewhat of a leap, we suppose that the observations of the system performance are generated in continuous time as opposed to discrete time, and we use the Brownian motion model to investigate how we can devise an algorithm with performance guarantees.

Performance guarantees

The goal is to give a guarantee that the probability of correct selection (PCS), which is the probability that the true best system remains after the elimination procedure, is sufficiently high:

$$\mathrm{PCS} = \mathbb{P}(\text{the best system is selected}) \geq 1 - \alpha.$$

We are thus let to deriving lower bounds on PCS. One particular component of the complementary probability is the probability that the best system is eliminated first, and we focus our discussion on this probability. Since it is only the relative order of the different Brownian motions that matters, we could study work in the subspace $\{x \in \mathbb{R}^n : x_1 + \cdots + x_n = 0\}$ as in [13, 28], but we do not do so here.

In an n-dimensional setting, this leads to the following concrete question of independent interest. Consider a Brownian motion in $\mathbb{R}^n$ with drift μ and covariance matrix Σ, started at the origin. When n is large, how can one calculate (or estimate) the probability that the location at which the Brownian motion first hits the ellipsoid $\partial K = \{x \in \mathbb{R}^n : x'\Sigma^{-1}x = 1\}$ satisfies $x_1 = \min_{i=1,\ldots,n} x_i$? The density of the exit location at $x \in \partial K$ is proportional to $\exp(\mu'\Sigma^{-1}x)$ [40], so we need to integrate this function over $\mathcal{C} \cap \partial K$, where $\mathcal{C} = \{x \in \mathbb{R}^n : x_1 = \min_{i=1}^n x_i\}$. A naive approach is to choose some mesh with small mesh size and approximate this integral with a (kind of) Riemann sum. However, this approach cannot be polynomial; imagine this approach on a unit hypercube instead of an ellipsoid, and note that one needs exponentially many small hypercubes of (some) width δ to cover the whole hypercube. Instead, one can use Monte Carlo

integration over the ellipsoid: writing $\sigma(dx)$ for the surface measure,

$$\int_{\mathcal{C}\cap\partial K} \exp(\mu'\Sigma^{-1}x)\sigma(dx) \approx \frac{1}{m}\sum_{i=1}^{m} \exp(\mu'\Sigma^{-1}X_i),$$

where the X_i are i.i.d. uniformly from $\mathcal{C}\cap\partial K$. Thus, to approximate this probability one needs to be able to sample (approximately) uniformly from the set $\mathcal{C}\cap\partial K$. This is the type of problem discussed in these notes.

Part II

Tools

4 | Markov chains

There is much to say about Markov chains, here we focus on those elements that are relevant from an algorithmic point of view. A good reference on Markov chains in the spirit of these lecture notes is [32].

Throughout, K is a compact subset of some Euclidean space. All of the sets K we are interested in are uncountable. Suppose $\{P_u : u \in K\}$ is a family of measures on K indexed by elements in K. (One actually needs some regularity in u, but we omit details as explained in Lecture 1.) Informally, a Markov chain on K is a sequence $\{X_t\}$ of elements of K, its first element has an arbitrary distribution and each subsequent element X_i is chosen from $P_{X_{i-1}}$ and independent of $X_0, \ldots, X_{i-1}$. We are only interested in discrete-time Markov chains in these notes, and we use t to denote time. When we write $t \geq 0$, it is implicit that t only takes integer values.

Definition 1. *A sequence of K-valued random variables $\{X_t : t \geq 0\}$ is a* Markov chain *with transition kernel $\{P_u : u \in K\}$ if, for any $t \geq 0$ and any (measurable) set B,*

$$\mathbb{P}(X_{t+1} \in B | X_0, \ldots, X_t) = P_{X_t}(B).$$

A distribution π on K is called *stationary* if

$$\int_K P_u(A)\pi(du) = \pi(A). \tag{4.1}$$

If X_0 has distribution π, then each of the X_t has distribution π, which explains why it is called a stationary distribution. A distribution π on K is called *atom-free* if $\pi(\{x\}) = 0$ for every $x \in K$.

Example 1. *Suppose $K = [-1, 1]^n$ and select some small $\delta > 0$. Let P_u be uniformly distributed on $[u-\epsilon, u+\epsilon]^n \cap K$. The resulting Markov chain can be thought of as a continuous-space analog of a nearest-neighbor random walk on $\{-m, \ldots, m\}^n$, and it is related to the* ball walks *to be discussed later.*

We will be interested in Markov chains with a unique stationary distribution, particularly in chains where the stationary distribution is uniform. Existence and uniqueness of π is a classical topic which is not covered here. One way to avoid thorny uniqueness issues is to require that the chain is *lazy*, meaning that we flip a fair coin to decide whether we stay in the current state or whether we make a move according to the transition kernel. From now on, we assume that our chain is lazy and that it has a unique stationary distribution π.

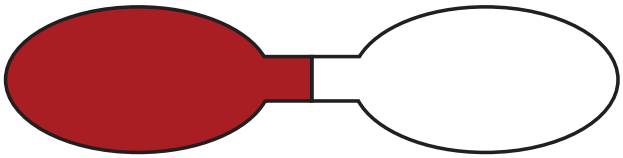
The *flow* of a subset A of K (with respect to π) is defined as

$$\Phi(A) = \int_A P_u(A^c)\pi(du),$$

where the complement A^c of a set A is defined through $A^c = K \backslash A$ for $A \subseteq K$. The *conductance* of a Markov chain is defined as

$$\phi = \inf_{A \subseteq K} \frac{\Phi(A)}{\min(\pi(A), \pi(A^c))}.$$

To get some intuition for these definitions, it is helpful to look at an example. Suppose K is a 'dumbbell' as depicted below.



In this example, the set A has been colored. The sets A and A^c have the same size. However, depending on the precise specification of the transition kernels $\{P_u\}$, the (probability) flow from A to A^c can be very small since the connecting piece acts as a 'bottleneck'. Therefore, the conductance can be very small in such a case.

Exercise 2. The kernel $\{P_u\}$ (or Markov chain) is said to be *reversible* with respect to π if $\pi(du)P_u(dx) = \pi(dx)P_x(du)$. Show that $\Phi(A) = \Phi(A^c)$ under reversibility. (All chains to be discussed will be reversible.) ☺

We need the following way to compare two probability measures P and Q on K. The *total variation distance* between P and Q is defined as

$$\|P - Q\|_{\text{TV}} = \sup_{A \subseteq K} |P(A) - Q(A)|.$$

The following lemma is useful when it is hard to manipulate *every* subset A as required in this definition, and we include it here because its proof is ingenious and it gives us some insight into how the supremum is attained.

Lemma 2. *If P and Q are absolutely continuous with densities f and g, respectively, then we have*

$$\|P - Q\|_{TV} = \frac{1}{2} \int_K |f(x) - g(x)| dx.$$

Proof. Let $B = \{x \in K : f(x) \geq g(x)\}$ and let $A \subseteq K$ be any event. Then

$$\begin{aligned} P(A) - Q(A) &= \int_{A \cap B} (f(x) - g(x))dx + \int_{A \cap B^c} (f(x) - g(x))dx \\ &\leq \int_{A \cap B} (f(x) - g(x))dx \leq \int_B (f(x) - g(x))dx = P(B) - Q(B). \end{aligned}$$

One can similarly show that $Q(A) - P(A) \leq Q(B^c) - P(B^c)$. Note that $Q(B^c) - P(B^c) = P(B) - Q(B)$, so we have shown that $|P(A) - Q(A)| \leq P(B) - Q(B) = |P(B) - Q(B)|$ and therefore $\|P - Q\|_{\text{TV}} = |P(B) - Q(B)|$. In particular,

$$\begin{aligned} \|P - Q\|_{\text{TV}} &= \frac{1}{2} \left[P(B) - Q(B) + Q(B^c) - P(B^c)\right] \\ &= \frac{1}{2} \left[\int_{\{x: f(x) \geq g(x)\}} (f(x) - g(x))dx + \int_{\{x: g(x) > f(x)\}} (g(x) - f(x))dx \right] \\ &= \frac{1}{2} \int_{x \in K} |f(x) - g(x)| dx, \end{aligned}$$

as claimed. □

5 | Conductance and convergence

In this lecture, we discuss some tools for studying how close the distribution at time t is to the stationary distribution. We write Q_t for the distribution at time t; this distribution critically depends on the initial distribution Q_0.

There is a wide variety of tools that are available to address this and similar questions, see for instance [32]. The conductance ϕ has emerged as a key quantity in this study; for instance, it is well-known that ϕ is connected to the spectral gap through the so-called Cheeger inequality. Our main tool is a proposition due to Lovász and Simonovits [34], which is less well-known yet powerful. We stress that it relies on our assumption that we work with a lazy Markov chain.

For each $x \in [0,1]$, we define

$$G_x = \left\{ g : K \to [0,1] \,\middle|\, \int_K g(u)\pi(du) = x \right\}$$

and

$$h_t(x) = \sup_{g \in G_x} \left(\int_K g(u) Q_t(du) - \int_K g(u)\pi(du) \right) = \sup_{g \in G_x} \int_K g(u) Q_t(du) - x.$$

To see the connection between h_t and total variation distance, we mention that, when π is atom-free,

$$h_t(x) = \sup_{A : \pi(A) = x} Q_t(A) - x, \tag{5.1}$$

and the supremum is attained. We omit a proof; see [34]. Note that $\|Q_t - \pi\|_{\mathrm{TV}} = \sup_{x \in [0,1]} h_t(x)$ in this case. The definition of h_t is useful because it exposes additional structural properties, such as in the following exercise. For the definition of concavity, refer to Lecture 6.

Exercise 3. Verify that $h_t(x)$ is concave in x for every integer $t \geq 0$. ☺

The next proposition is a result for atom-free π; see [18] for a similar result on finite state spaces.

Proposition 1. *Let π be atom-free. Let C_0 and C_1 be such that $h_0(x) \leq C_0 + C_1 \min(\sqrt{x}, \sqrt{1-x})$. Then we have, for $t \geq 0$,*

$$h_t(x) \leq C_0 + C_1 \min(\sqrt{x}, \sqrt{1-x}) \left(1 - \frac{\phi^2}{2}\right)^t.$$

Proof. For any $x \in [0,1]$, let $y = \min(x, 1-x)$. The proof relies on the inequality

$$h_t(x) \leq \frac{1}{2} h_{t-1}(x - 2\phi y) + \frac{1}{2} h_{t-1}(x + 2\phi y). \tag{5.2}$$

We delay the proof of this bound and give the remainder of the argument first, which is an induction argument. We assume that $x \leq 1/2$, so that $x = y$; the case $x > 1/2$ is similar. The

base of the induction has been assumed, so let us assume the claim with $t-1$ instead of t. Then by (5.2),

$$\begin{aligned} h_t(x) &\leq \frac{1}{2}h_{t-1}(x(1-2\phi)) + \frac{1}{2}h_{t-1}(x(1+2\phi)) \\ &\leq C_0 + \frac{1}{2}C_1\left(\sqrt{x(1-2\phi)} + \sqrt{x(1+2\phi)}\right)\left(1-\frac{\phi^2}{2}\right)^{t-1} \\ &\leq C_0 + C_1\sqrt{x}\left(1-\frac{\phi^2}{2}\right)^{t}, \end{aligned}$$

where the last inequality uses $\sqrt{1-2\phi}+\sqrt{1+2\phi} \leq 2(1-\phi^2/2)$.

It remains to verify (5.2). Let A satisfy (5.1). Then

$$h_t(x) = Q_t(A) - \pi(A) = \int_K P_u(A)(Q_{t-1}(du) - \pi(du)).$$

Set $g_1(u) = (2P_u(A)-1)\mathbf{1}_{u\in A}$ and $g_2(u) = 2P_u(A) - g_1(u)$, and note that $g_1 \in G_{x_1}$ and $g_2 \in G_{x_2}$ for $x_1 = \int_K g_1(u)\pi(du)$ and $x_2 = \int_K g_2(u)\pi(du)$. In view of this and the identity $2P_u(A) = g_1(u) + g_2(u)$, we deduce from the preceding display that

$$2h_t(x) \leq \int_K g_1(u)(Q_{t-1}(du) - \pi(du)) + \int_K g_2(u)(Q_{t-1}(du) - \pi(du)) \leq h_{t-1}(x_1) + h_{t-1}(x_2).$$

It is readily checked with (4.1) that $x_1 + x_2 = 2x$, and moreover that

$$x_1 = \int_K g_1(u)\pi(du) = \int_A \pi(du) - 2\int_A P_u(A^c)\pi(du) = x - 2\Phi(A) \leq x(1-2\phi),$$

so that $x_1 \leq x(1-2\phi) \leq x \leq x(1+2\phi) \leq x_2$. By concavity of h_{t-1}, the chord from x_1 to x_2 lies below the chord from $x(1-2\phi)$ to $x(1+2\phi)$, which implies the claim. □

The next corollary contains the key result from this lecture.

Corollary 1. *Let π be atom-free and set $M = \sup_A Q_0(A)/\pi(A)$. Then we have, for $t \geq 0$,*

$$\|Q_t - \pi\|_{TV} \leq \sqrt{M}\left(1-\frac{\phi^2}{2}\right)^{t}.$$

Proof. In view of (5.1), we have

$$h_0(x) = \sup_{A:\pi(A)=x} Q_0(A) - x \leq M \sup_{A:\pi(A)=x} \pi(A) = Mx.$$

We also have $h_0(x) \leq \sup_{A:\pi(A)=x}(1-\pi(A)) = 1-x$. Together these imply that $h_0(x) \leq \sqrt{M}\min(\sqrt{x}, \sqrt{1-x})$ for $x \in [0,1]$. The claim thus follows from Proposition 1. □

6 | Convexity and inequalities

A subset K of $\mathbb{R}^n$ is *convex* if, for any $x \in K$ and $y \in K$ and any $\alpha \in [0,1]$, the *convex combination* $\alpha x + (1-\alpha)y$ lies in K. A function $f : \mathbb{R}^n \to \infty$ is called *convex* if, for every x, y in its domain and $\alpha \in [0,1]$, we have $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$. If $-f$ is convex, then a function is *concave*.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *logconcave* if the function $\log(-f(\cdot))$ is convex, meaning that for $\alpha \in [0,1]$, $x, y \in \mathbb{R}^n$,

$$f(\alpha x + (1-\alpha)y) \geq f(x)^\alpha f(y)^{1-\alpha}. \tag{6.1}$$

Think of $e^{-\|x\|^2/2}$ as an example; many densities encountered in probability theory are logconcave. The product and the minimum of two logconcave functions are both logconcave; the sum is not in general. Momentarily we discuss several deeper properties of logconcave functions, as a corollary of a pivotal result known as the Brunn-Minkowski inequality.

To formulate the Brunn-Minkowski inequality, we need some more notation. Given a set A, we set $\alpha A = \{\alpha x : x \in A\}$. Note that, if A is n-dimensional, $\mathrm{vol}(\alpha A) = \alpha^n \mathrm{vol}(A)$ as is readily seen by calculating the Jacobian of the transformation $x \mapsto \alpha x$. Given two sets A_1 and A_2 in $\mathbb{R}^n$, we define their *Minkowski sum* as

$$A_1 + A_2 = \{y + z : y \in A_1, z \in A_2\}.$$

An example is when the set A_2 is a ball with radius ϵ. Writing B_n for the unit ball in $\mathbb{R}^n$ and assuming A is convex, the so-called Minkowski's theorem on mixed volumes implies that the quantity $\mathrm{vol}(A + \epsilon B_n)$ is a polynomial in ϵ; we use this in the next lecture.

The Brunn-Minkowski inequality says that for any bounded measurable A and B such that $\alpha A + (1-\alpha)B$ is measurable, we have

$$\mathrm{vol}(\alpha A + (1-\alpha)B)^{1/n} \geq \alpha \mathrm{vol}(A)^{1/n} + (1-\alpha)\mathrm{vol}(B)^{1/n}. \tag{6.2}$$

Note that it is natural for $\mathrm{vol}(A)^{1/n}$ to appear in an inequality since we have $\mathrm{vol}(\alpha A)^{1/n} = \alpha \mathrm{vol}(A)^{1/n}$, so the volume functional has a special homogeneity property when it is raised to the power $1/n$. By this homogeneity property, (6.2) is equivalent with the inequality

$$\mathrm{vol}(A + B)^{1/n} \geq \mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}. \tag{6.3}$$

Yet another equivalent statement is the inequality, for $\alpha \in [0,1]$ such that $\alpha A + (1-\alpha)B$ is measurable,

$$\mathrm{vol}(\alpha A + (1-\alpha)B) \geq \min\{\mathrm{vol}(A), \mathrm{vol}(B)\}. \tag{6.4}$$

For a beautiful survey on the Brunn-Minkowski inequality, see [21].

Exercise 4. Prove the equivalence of (6.2) and (6.4). ☺

We next mention a different version of the Brunn-Minkowski inequality, which states that

$$\mathrm{vol}(\alpha A + (1-\alpha)B) \geq \mathrm{vol}(A)^\alpha \mathrm{vol}(B)^{1-\alpha}. \tag{6.5}$$

cf. (6.1). Interestingly, this inequality is again equivalent to (6.2).

An important special case is when A and B are *parallel slices* of a convex set K: $A = K \cap \{x : x_1 = a\}$ and $B = K \cap \{x : x_1 = b\}$. Then $\alpha A + (1-\alpha)B = K \cap \{x : x_1 = \alpha a + (1-\alpha)b\}$, so (6.2) implies that $\mathrm{vol}(K \cap \{x : x_1 = y\})^{1/(n-1)}$ is concave and (6.5) implies that $\mathrm{vol}(K \cap \{x : x_1 = y\})$ is a logconcave function of y.

Using the Prékopa-Leindler inequality, which is stronger than the Brunn-Minkowski inequality, one can prove several remarkable properties of logconcave functions. Since we do not use these in the current notes, we do not go into detail; but I'm mentioning these properties here to pique your interest into the implications of these interesting geometric inequalities. There are many more, see [21].

1. Sections ('marginals') of a logconcave function are logconcave.

2. The convolution of two logconcave functions is logconcave.

If you liked the material in this lecture and you're interested in these inequalities, be sure to read about optimal transport [20, 46].

7 | Isoperimetric inequalities

The classical isoperimetric inequality states that, if K is a convex body, then

$$\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_n)}\right)^{1/n} \leq \left(\frac{S(K)}{S(B_n)}\right)^{1/(n-1)}, \tag{7.1}$$

where S denotes the surface area of the convex body K:

$$S(K) = \lim_{\epsilon \downarrow 0} \frac{\mathrm{vol}(K + \epsilon B_n) - \mathrm{vol}(K)}{\epsilon}. \tag{7.2}$$

(The limit exists as a consequence of Minkowski's mixed volume theorem, see the last lecture.)

The next exercise shows that, among all bodies with a given volume, a ball minimizes the surface area. (Some in the audience are probably more familiar with the corresponding inequality for Gaussian measures, where the analogs of balls are halfspaces and the isoperimetric inequality is related to the Borell inequality for the maximum of a Gaussian vector.)

Exercise 5. Show that (7.2) is the same as saying that, for any $x > 0$,

$$\inf_{K:\mathrm{vol}(K)=x} S(K) = S(rB_n), \tag{7.3}$$

where r satisfies $\mathrm{vol}(rB_n) = x$. ☺

The next exercise shows that the isoperimetric inequality (7.1) follows from the Brunn-Minkowski inequality.

Exercise 6. Use (7.2) and the Brunn-Minkowski inequality to deduce a lower bound on $S(K)$, and use this to prove (7.1). ☺

We need the related problem of finding a surface that partitions a convex body K into two parts, and whose measure is minimal relative to the volumes of the two parts. In general this is a very difficult problem and it is only partially understood. However, for our purposes it is sufficient to know that the surface area of S inside K is large compared to the volumes of S and $K \backslash S$: For any subset S of a convex body K with diameter D, we have

$$\mathrm{vol}_{n-1}(\partial S \cap K) \geq \frac{2}{D} \min(\mathrm{vol}(S), \mathrm{vol}(K \backslash S)).$$

The following theorem states an inequality of this form but for an analog (stronger) version where the sets S_1 and S_2 can be separated by a set with full measure. The bound is a bit worse than in the preceding display, but this claim is a bit easier to sketch the proof for. (Throughout, we are not worried about constants but we do worry about capturing the right dependencies on the dimension n and diameter D.)

Theorem 4. *Let S_1, S_2, and S_3 be a partition into measurable sets of a convex body K of diameter D. Then we have*

$$\mathrm{vol}(S_3) \geq \frac{d(S_1, S_2)}{4D} \min(\mathrm{vol}(S_1), \mathrm{vol}(S_2)).$$

where $d(S_1, S_2)$ denotes the Euclidean distance between S_1 and S_2:

$$d(S_1, S_2) = \min\{|u - v| : u \in S_1, v \in S_2\}.$$

There is also a related result for the *boundary* of a convex body. There are many subtleties when defining the 'volume' of an arbitrary set; this is the domain of geometric analysis, but here we simply assume that the boundary is a Riemannian manifold so that we can work with the so-called Riemannian volume form. This assumption also gives rise to a metric d, the *geodesic distance*. The following theorem is a weak version of a recent result by Milman [37].

Theorem 5. *Let S_1, S_2, and S_3 be a partition into measurable sets of the boundary ∂K of a convex body of diameter D. We then have, for some absolute constant $c > 0$,*

$$\mathrm{vol}(S_3) \geq \frac{c}{D}\, d(S_1, S_2) \min(\mathrm{vol}(S_1), \mathrm{vol}(S_2)),$$

where $d(S_1, S_2) = \inf_{x \in S_1, y \in S_2} d(x, y)$ denotes the geodesic distance between S_1 and S_2 on ∂K.

In the remainder of this lecture, we discuss some elements of the proof of Theorem 4. Throughout, fix K and S_1, S_2, and S_3. The key idea is to use *localization*, which reduces an n-dimensional inequality to a one-dimensional inequality.

Lemma 3. *Let $f : K \to \mathbb{R}$ be integrable, and $n \geq 2$. For any 2-dimensional affine subspace A of $\mathbb{R}^n$ and $z \in A$, there exists an $a \in A$ such that the halfspaces $H_+ = \{x \in \mathbb{R}^n : (x-z)'(a-z) \leq 0\}$ and $H_- = \{x \in \mathbb{R}^n : (x-z)'(a-z) \geq 0\}$ satisfy*

$$\int_{H_+ \cap K} f(x)dx = \int_{H_- \cap K} f(x)dx.$$

Proof. Given A and $z \in A$, consider the family of halfspaces $\mathcal{H} = \{H_b : b \in A\}$ given by $H_b = \{x \in \mathbb{R}^n : (x - z)'(b - z) \leq 0\}$. The functional $b \in A \mapsto \int_{H_b \cap K} f(x)dx - \int_{H_b^c \cap K} f(x)dx$ is continuous and changes sign under the transformation $b \mapsto 2z - b$ (antisymmetry about z). This yields the existence of some $a \in A$ for which the claim holds. □

Definition 2. *A partition of the convex set K is a finite collection $\mathcal{P} = \{P_1, \ldots, P_k\}$ of closed convex subsets of K such that*

1. $\bigcup_{i=1}^k P_i = K$,
2. $P_i \cap P_j = \partial P_i \cap \partial P_j, \quad 1 \leq i < j \leq k.$

We say that a partition $\mathcal{P}$ of K is ϵ-thin if each $P \in \mathcal{P}$ is contained in the Minkowski sum of some one-dimensional affine space and ϵB_n centered at the origin.

The following proposition is the key ingredient of the localization approach. I will sketch a 'picture proof' in class.

Proposition 2. *Let $f : K \to \mathbb{R}$ be an integrable function for which $\int_K f(x)dx = 0$. For any $\epsilon > 0$, there exists an ϵ-thin partition $\mathcal{P}$ of K for which $\int_P f(x)dx = 0$ for $P \in \mathcal{P}$.*

According to this proposition, the parts $P \in \mathcal{P}$ can be chosen as 'almost' line segments while maintaining the property that f integrates to zero. The function $f : K \to \mathbb{R}$ we apply this proposition to is given by $f(x) = \mathrm{vol}(S_2)1\{x \in S_1\} - \mathrm{vol}(S_1)1\{x \in S_2\}$. The equality $\int_P f(x)dx = 0$ translates to $\mathrm{vol}(P \cap S_1)/\mathrm{vol}(P \cap S_2) = \mathrm{vol}(S_1)/\mathrm{vol}(S_2)$. Suppose that $\mathrm{vol}(S_1) < \mathrm{vol}(S_2)$ without loss of generality.

Fix some part P, and let it be 'long' (i.e., not thin) in direction ℓ. Brunn-Minkowski implies that the volume on parallel slices orthogonal to ℓ yields a logconcave function. If $P \cap S_1$ and

$P \cap S_2$ are unions of such parallel slices intersected with P, then this can be used to derive the following 'one-dimensional isoperimetric inequality' for thin parts P,

$$\mathrm{vol}(S_3 \cap P) \geq \frac{2d(S_1,S_2)}{D}\mathrm{vol}(S_1 \cap P) = \frac{2d(S_1,S_2)}{D}\min(\mathrm{vol}(S_1 \cap P), \mathrm{vol}(S_2 \cap P)). \tag{7.4}$$

The arguments are slightly more subtle when S_1 and S_2 are not necessarily unions of parallel slices orthogonal to ℓ since one then needs to use this argument approximately and let ϵ converge to 0 appropriately, but the key idea remains the same.

The proof of Theorem 4 is completed as follows. First we classify the parts of $\mathcal{P}$ according to the relative size of S_3:

$$\mathcal{P}_+ = \left\{P \in \mathcal{P} : \mathrm{vol}(S_3 \cap P) \leq \frac{1}{2}\mathrm{vol}(P)\right\}, \quad \mathcal{P}_- = \left\{P \in \mathcal{P} : \mathrm{vol}(S_3 \cap P) > \frac{1}{2}\mathrm{vol}(P)\right\}.$$

Note that either $\sum_{P\in\mathcal{P}_-} \mathrm{vol}(P) \geq \frac{1}{2}\mathrm{vol}(K)$ (first case) or $\sum_{P\in\mathcal{P}_+} \mathrm{vol}(P) \geq \frac{1}{2}\mathrm{vol}(K)$ (second case). In the first case, we have

$$\mathrm{vol}(S_3) \geq \sum_{P\in\mathcal{P}_-} \mathrm{vol}(P \cap S_3) > \frac{1}{2}\sum_{P\in\mathcal{P}_-} \mathrm{vol}(P) \geq \frac{1}{4}\mathrm{vol}(K) \geq \frac{d(S_1,S_2)}{4D}\min(\mathrm{vol}(S_1), \mathrm{vol}(S_2)).$$

In the second case, we have by (7.4)

$$\mathrm{vol}(S_3) \geq \sum_{P\in\mathcal{P}_+} \mathrm{vol}(P \cap S_3) \geq \frac{2d(S_1,S_2)}{D}\sum_{P\in\mathcal{P}_+} \mathrm{vol}(S_1 \cap P),$$

where we note that (7.4) has only been justified 'approximately' so one needs to be quite a bit more careful to make this precise. From $\mathrm{vol}(S_1 \cap P)/(\mathrm{vol}(S_1 \cap P) + \mathrm{vol}(S_2 \cap P)) \leq \mathrm{vol}(P \cap S_1)/\mathrm{vol}(P)$, we deduce that

$$\begin{aligned}\sum_{P\in\mathcal{P}_+} \mathrm{vol}(S_1 \cap P) &\geq \sum_{P\in\mathcal{P}_+} \frac{\mathrm{vol}(S_1 \cap P)}{\mathrm{vol}(S_1 \cap P) + \mathrm{vol}(S_2 \cap P)}\mathrm{vol}(P) = \frac{\mathrm{vol}(S_1)}{\mathrm{vol}(S_1) + \mathrm{vol}(S_2)}\sum_{P\in\mathcal{P}_+} \mathrm{vol}(P) \\ &\geq \frac{\mathrm{vol}(S_1)}{2(\mathrm{vol}(S_1) + \mathrm{vol}(S_2))}\mathrm{vol}(K) \geq \frac{1}{2}\mathrm{vol}(S_1).\end{aligned}$$

Combining the latter two displays shows that the inequality in Theorem 4 also holds in the second case.

Part III

Algorithms and their analysis

8 | Geometric random walks

In this lecture we discuss several MCMC (approximate) sampling algorithms for sampling from convex sets or their boundary. Initially, all research in this direction was aimed at proving that the algorithm works in any dimension in the sense that the Markov chain underlying the MCMC algorithm converges to the correct target distribution. Interest then shifted to proving that the algorithms are efficient as the dimension increases, with key contributions by the computer science community. The running time of these MCMC algorithms is polynomial in the dimension n (sometimes under further assumptions on the geometry of the body), even though no deterministic algorithm can achieve polynomial running time as mentioned in the first lecture.

Throughout, let K be a convex set. Some of the walks defined here also 'work' for non-convex sets, but there are virtually no accompanying performance guarantees (except for [12]). Some of these walks are part of a more general family of similar walks; here we only discuss the simplest representative of the family.

The *ball walk* is a variation of the nearest-neighbor random walk on a grid. Its mixing properties were analyzed in [34], and we outline the proof in the next lecture.

Ball walk for sampling from K

- Pick a point y with a uniform distribution on the ball of radius δ centered at the current point $x \in K$.
- If y is in K, go to y; otherwise stay at x.

The *hit-and-run algorithm* of Smith [44] and its variations [1, 2] is inherently different from the ball walk in the sense that it could potentially take huge steps as opposed to being confined to a small neighborhood of the previous point. Its mixing properties were analyzed in [33, 35].

Hit-and-run for sampling from K

- Pick a line ℓ through the current point $x \in K$ according to a uniform distribution.
- The next point y has a uniform distribution on the chord $\ell \cap K$.

Shake-and-bake is a variation of hit-and-run which is suitable for sampling from the boundary of a set [5, 42]. Its mixing properties have not yet been analyzed, but Lecture 10 makes a first start.

Shake-and-bake for sampling from ∂K

- Pick a line ℓ through the current point $x \in \partial K$. according to cosine.
- The next point y is the intersection point of ℓ with ∂K different from x.

Exercise 7. Verify for each of these walks that the uniform distribution is a stationary distribution. ☺

These walks can be modified to sample much more general distributions. Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be a nonnegative integrable function. It defines a measure π_f (on measurable subsets of $\mathbb{R}^n$):

$$\pi_f(A) = \frac{\int_A f(x)\,dx}{\int_{\mathbb{R}^n} f(x)\,dx}.$$

The ball walk can be extended as follows.

Ball walk with Metropolis filter

- Pick a uniformly distributed random point y in the ball of radius δ centered at the current point x.
- Move to y with probability $\min\left\{1, \frac{f(y)}{f(x)}\right\}$; stay at x with the remaining probability.

Exercise 8. Prove that the ball walk with Metropolis filter has π_f for its stationary distribution. ☺

Hit-and-run can also be extended to sampling from such a general distribution π_f. For any line ℓ in $\mathbb{R}^n$, let $\pi_{\ell,f}$ be the *restriction* of π to ℓ, i.e.,

$$\pi_{\ell,f}(S) = \frac{\int_{p+tu\in S} f(p+tu)\,dt}{\int_\ell f(x)\,dx},$$

where p is any point on ℓ and u is a unit vector parallel to ℓ.

Hit-and-run for general functions

- Pick a uniform random line ℓ through the current point x.
- Go to a random point y along ℓ chosen from the distribution $\pi_{\ell,f}$.

Exercise 9. Prove that this version of hit-and-run has π_f for its stationary distribution. ☺

9 | Mixing for the ball walk

This lecture discusses the mixing properties of the ball walk. There are several reasons as to why it is natural to discuss this walk in more detail: (1) the proof structure is generic in the sense that most proofs of conductance follow the same rough outline, (2) the connection with isoperimetric inequalities is most apparent in this case, and (3) the proof is arguably the simplest of the geometric random walks presented in the previous lecture.

A geometric random walk is said to be *rapidly mixing* if its conductance is bounded from below by an inverse polynomial in the dimension. The ball walk is not rapidly mixing since it can be stuck in a 'corner'. For instance, if K is a hypercube in $\mathbb{R}^n$ with the current point being a vertex of the hypercube, then the chance of the ball walk making a move in the next step is 2^{-n}, so it takes an exponential amount of steps to get away from the vertex. You can make this arbitrarily bad. Below we discuss possible remedies. Throughout, we assume that K contains a unit ball.

Define the *local conductance* $\ell(x)$ at $x \in K$ as the probability of escaping from K, i.e., $\ell(u) = 1 - P_u(\{u\})$. We omit a proof of the next lemma.

Lemma 4. *Let $u, v \in K$. If $|u - v| \leq t\delta/\sqrt{n}$ and $\ell(u), \ell(v) \geq \ell$, then we have*

$$\|P_u - P_v\|_{TV} < 1 - \ell + t.$$

Theorem 6. *Let K be a convex body with diameter D containing a unit ball. Suppose that $\ell(u) \geq \ell$ for $u \in K$. Then the conductance ϕ of the ball walk satisfies*

$$\phi \geq \frac{\ell^2 \delta}{16\sqrt{n}D}.$$

Proof. Let $K = S_1 \cup S_2$ be a partition into measurable sets. We will prove that

$$\int_{S_1} P_x(S_2)\, dx \quad \geq \quad \frac{\ell^2 \delta}{16\sqrt{n}D} \min\{\mathrm{vol}(S_1), \mathrm{vol}(S_2)\}. \tag{9.1}$$

Note that since the uniform distribution is stationary,

$$\int_{S_1} P_x(S_2)\, dx = \int_{S_2} P_x(S_1)\, dx.$$

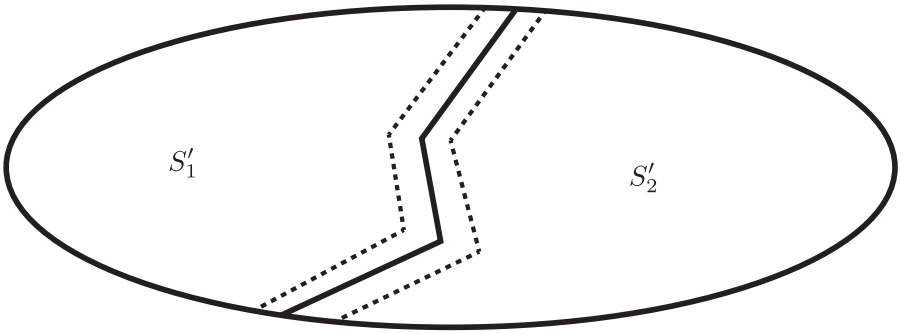
Consider the points that are "deep" inside these sets, i.e., points that are unlikely to jump out of the set:

$$S_1' = \left\{ x \in S_1 : P_x(S_2) < \frac{\ell}{4} \right\}$$

and

$$S_2' = \left\{ x \in S_2 : P_x(S_1) < \frac{\ell}{4} \right\}.$$

Let S_3' be the rest i.e., $S_3' = K \setminus S_1' \setminus S_2'$. A sketch of these definitions is given in the following diagram, where the boundary between S_1 and S_2 is represented by a solid curve, while the boundaries between S_1' and S_3' as well as between S_2' and S_3' are represented by a dashed curve.



Suppose $\mathrm{vol}(S_1') < \mathrm{vol}(S_1)/2$. Then

$$\int_{S_1} P_x(S_2)\,dx \geq \frac{\ell}{4}\mathrm{vol}(S_1 \setminus S_1') \geq \frac{\ell}{8}\mathrm{vol}(S_1)$$

which proves (9.1).

So we can assume that $\mathrm{vol}(S_1') \geq \mathrm{vol}(S_1)/2$ and similarly $\mathrm{vol}(S_2') \geq \mathrm{vol}(S_2)/2$. Now, for any $u \in S_1'$ and $v \in S_2'$,

$$||P_u - P_v||_{\mathrm{TV}} \geq 1 - P_u(S_2) - P_v(S_1) > 1 - \frac{\ell}{2}.$$

Applying Lemma 4 with $t = \ell/2$, we get that

$$|u - v| \geq \frac{\ell\delta}{2\sqrt{n}}.$$

Thus $d(S_1, S_2) \geq \ell\delta/2\sqrt{n}$. Applying Theorem 4 to the partition S_1', S_2', S_3', we have

$$\begin{aligned}\mathrm{vol}(S_3') &\geq \frac{\ell\delta}{\sqrt{n}D}\min\{\mathrm{vol}(S_1'), \mathrm{vol}(S_2')\} \\ &\geq \frac{\ell\delta}{2\sqrt{n}D}\min\{\mathrm{vol}(S_1), \mathrm{vol}(S_2)\}.\end{aligned}$$

We next note that

$$\begin{aligned}\int_{S_1} P_x(S_2)\,dx &= \frac{1}{2}\int_{S_1} P_x(S_2)\,dx + \frac{1}{2}\int_{S_2} P_x(S_1)\,dx \\ &\geq \frac{1}{2}\mathrm{vol}(S_3')\frac{\ell}{4} \\ &\geq \frac{\ell^2\delta}{16\sqrt{n}D}\min\{\mathrm{vol}(S_1), \mathrm{vol}(S_2)\},\end{aligned}$$

and this proves (9.1). □

The next corollary follows from the preceding theorem and Corollary 1. It shows that the mixing time is at most $nD^2/(\ell^4\delta^2)$.

Corollary 2. *Consider the setting of Theorem 6 and set $M = \sup_A Q_0(A)/\pi(A)$. Then there is a constant c such that, for $t \geq 0$,*

$$\|Q_t - \pi\|_{TV} \leq \sqrt{M}\left(1 - \frac{c\ell^4\delta^2}{nD^2}\right)^t.$$

Note that the presence of ℓ in this bound is related to the problem of being 'stuck' in a corner, as discussed already. There are several ways around this, which each have different implications on the choice of the parameter δ:

- Expand the body a bit, so that each point has a local conductance ℓ that is bounded away from 0. It is important not to increase the volume too much, because otherwise one cannot make a connection with sampling on the original body. In this case, one can choose $\delta = 1/(n\sqrt{n})$ and the mixing time will be $O(n^4 D^2)$.
- Change the body K a bit, so that ℓ is bounded away from 0. One can choose $\delta = 1/\sqrt{n}$ but there are conditions on the initial distribution to make this work (a so-called 'warm start').

The most important consequence of this is that (approximate) sampling can be done in time that is polynomial in D and n. Through a standard transformation, D can also be made polynomial in n, thus establishing Theorem 2.

10 | Mixing for shake-and-bake

This lecture discusses a result for sampling from the boundary of a convex body. There is much less work on sampling algorithms in this setting, with [5, 42] and related work as an exception. There are several reasons that warrant a detailed inquiry into such sampling algorithms: (1) Sampling from the boundary of a convex set generalizes sampling from a convex set K, since samples from K can be generated by sampling from the boundary of the set $K \times [0,1]$ in $\mathbb{R}^{n+1}$. (2) There are specific applications for sampling from the boundary of a convex set, see Lecture 3. (3) MCMC algorithms that exploit the boundary could prove to be faster, and the underlying ideas could lead to faster algorithms for sampling from convex sets as well. (4) New tools need to be developed, which promise to be useful in various other settings.

For $x, y \in \partial K$, define ϕ_{xy} to be the acute angle between $x - y$ and n_x, the inward normal at x. The one-step distribution is [5, 42]

$$P_u(A) = \frac{\pi^{(n-1)/2}}{|\partial K|\Gamma((n+1)/2)} \int_A \frac{\cos(\phi_{uv})\cos(\phi_{vu})}{\|u-v\|^{n-1}} dv, \tag{10.1}$$

where Γ is the gamma function. For $x \in \partial K$, define $F(x)$ through

$$P_x(\{y \in \partial K : d(x,y) \le F(x)\}) = \frac{3}{4}.$$

In other words, this is the 75'th percentile step length of a random point from x. For $1 \ge \gamma \ge 0$ and $x \in K$, we also define the *step size*

$$s_\gamma(x) = \sup\left\{ t \ge 0 \ : \ \frac{\mathrm{vol}(x + tB_n \cap K)}{\mathrm{vol}(tB_n)} \ge \gamma \right\}.$$

Lemma 5. *For any $x \in \partial K$ and $\gamma \le 1/2$,*

$$F(x) \ge \frac{s_\gamma(x)}{2}.$$

We say that K is a convex body with *curvature bounded from above by $\mathcal{C} < \infty$* if for each $x \in \partial K$, there is a ball B with radius $1/\mathcal{C}$ and center in K so that the tangent planes of K and B at x coincide and B lies in K.

Lemma 6. *Let $\gamma \in (0, 1/2)$. If K is a convex body with curvature bounded from above by $\mathcal{C} < \infty$, then*

$$s_\gamma(x) \ge \frac{c_\gamma}{\mathcal{C}\sqrt{n}},$$

where c_γ is a constant only depending on γ.

Proof sketch. Fix $x \in \partial K$ and let B be the aforementioned ball with radius $1/\mathcal{C}$. Let $\tilde{B}$ be a 'small' ball with radius r centered at x. Let x be the origin of a new coordinate system, in which the center of the larger ball is $(1/\mathcal{C}, 0, \ldots, 0)$.

We first characterize the points (in this new coordinate system) where the boundaries of the two balls intersect. All of these points have the same first coordinate, namely equal to t satisfying

$$\frac{1}{\mathcal{C}^2} - t^2 = r^2 - \left(\frac{1}{\mathcal{C}} - t\right)^2.$$

The solution is $t = 1/\mathcal{C} - \mathcal{C}r^2/2$.

We now use the property that, for a unit ball B, the volume of all points with first coordinate exceeding $1/\sqrt{n}$ takes up a constant fraction of the whole volume. Therefore, if $(1/\mathcal{C} - t)/r = 1/\sqrt{n}$, the volume fraction of the small ball inside the big ball is at least some constant. This shows that r should be at least of order $1/(\mathcal{C}\sqrt{n})$. □

Lemma 7. *Let $u, v \in \partial K$. If*

$$d(u, v) < \frac{1}{\sqrt{n}} \max(F(u), F(v)),$$

then we have

$$\|P_u - P_v\|_{TV} < 1 - \frac{1}{500}.$$

Theorem 7. *Let K be a convex body in $\mathbb{R}^n$ with diameter D containing a unit ball. Suppose the curvature is bounded from above by $\mathcal{C}$. Then the conductance ϕ of shake-and-bake satisfies*

$$\phi \geq \frac{c}{\mathcal{C}nD},$$

for some universal constant c.

Proof. Let $K = S_1 \cup S_2$ be a partition into measurable sets. We will prove that

$$\int_{S_1} P_x(S_2)\, dx \quad \geq \quad \frac{c}{\mathcal{C}nD} \min\{\mathrm{vol}(S_1), \mathrm{vol}(S_2)\}. \tag{10.2}$$

In this proof, the constant c can vary from line to line. Consider the points that are deep inside these sets, i.e., unlikely to jump out of the set:

$$S_1' = \left\{x \in S_1 : P_x(S_2) < \frac{1}{1000}\right\}, \qquad S_2' = \left\{x \in S_2 : P_x(S_1) < \frac{1}{1000}\right\}.$$

Set $S_3' = K \setminus S_1' \setminus S_2'$.

Suppose $\mathrm{vol}(S_1') < \mathrm{vol}(S_1)/2$. Then

$$\int_{S_1} P_x(S_2)\, dx \geq \frac{1}{1000}\mathrm{vol}(S_1 \setminus S_1') \geq \frac{1}{2000}\mathrm{vol}(S_1)$$

which proves (10.2).

So we can assume that $\mathrm{vol}(S_1') \geq \mathrm{vol}(S_1)/2$ and similarly $\mathrm{vol}(S_2') \geq \mathrm{vol}(S_2)/2$. For any $u \in S_1'$ and $v \in S_2'$,

$$\|P_u - P_v\|_{\mathrm{TV}} \geq 1 - P_u(S_2) - P_v(S_1) > 1 - \frac{1}{500}.$$

Thus, by Lemma 7, we must then have

$$d(u, v) \geq \frac{1}{\sqrt{n}} \max\{F(u), F(v)\}. \tag{10.3}$$

In particular, we have $d(S_1', S_2') \geq \inf_{x \in \partial K} F(x)/\sqrt{n}$.

We next apply Theorem 5 to obtain

$$\begin{aligned}\frac{\mathrm{vol}(S_3')}{\min\{\mathrm{vol}(S_1'),\mathrm{vol}(S_2')\}} &\geq \frac{c}{D\sqrt{n}}\inf_{x\in\partial K} F(x)\\ &\geq \frac{c}{2D\sqrt{n}}\inf_{x\in\partial K} s_\gamma(x),\end{aligned}$$

where the last inequality follows from Lemma 5. By Lemma 6, this is bounded from below by $c/(\mathcal{C}nD)$. Therefore,

$$\begin{aligned}\int_{S_1} P_x(S_2)\,dx &\geq \frac{1}{2}\cdot\frac{1}{1000}\mathrm{vol}(S_3')\\ &\geq \frac{c}{\mathcal{C}nD}\min\{\mathrm{vol}(S_1'),\mathrm{vol}(S_2')\}\\ &\geq \frac{c}{2\mathcal{C}nD}\min\{\mathrm{vol}(S_1),\mathrm{vol}(S_2)\}\end{aligned}$$

which again proves (10.2). □

The next corollary follows from the preceding theorem and Corollary 1. It shows that the mixing time is at most order $\mathcal{C}^2n^2D^2$.

Corollary 3. *Consider the setting of Theorem 7 and set* $M = \sup_A Q_0(A)/\pi(A)$. *Then there is a constant* c *such that, for* $t \geq 0$,

$$\|Q_t - \pi\|_{TV} \leq \sqrt{M}\left(1 - \frac{c}{\mathcal{C}^2n^2D^2}\right)^t.$$

Solutions to selected exercises

Solution to Exercise 1. (There are several correct answers.) Let U be uniformly distributed on $(0,1)$, and let Y be a standard Gaussian random vector in $\mathbb{R}^n$. Then let the algorithm output $U^{1/n}Y/\|Y\|$. It is readily seen this has the correct distribution; by symmetry it suffices to show that the norm $U^{1/n}$ has the correct distribution. Indeed, $P(U^{1/n} \leq x) = x^n$, which is the ratio of the volume of a ball with radius x and the volume of a unit ball.

Solution to Exercise 2. Under reversibility, we have

$$\begin{aligned}\Phi(A) &= \int_A P_u(A^c)\pi(du) = \int_A \int_{A^c} P_u(dx)\pi(du) \\ &= \int_A \int_{A^c} \pi(dx)P_x(du) = \int_{A^c} \int_A P_x(du)\pi(dx) = \Phi(A^c),\end{aligned}$$

where the second last equality follows from Fubini's theorem.

Solution to Exercise 3. Let $x_1, x_2 \in [0,1]$ and $\alpha \in [0,1]$. For $g_1 \in G_{x_1}$ and $g_2 \in G_{x_2}$, we have $\alpha g_1 + (1-\alpha)g_2 \in G_{\alpha x_1 + (1-\alpha)x_2}$. In particular, we have the following upper bound for $h_t(\alpha x_1 + (1-\alpha)x_2)$:

$$\sup_{g \in G_{\alpha x_1 + (1-\alpha)x_2}} \int_K g(u)Q_t(du) \leq \sup_{g_1 \in G_{x_1}} \sup_{g_2 \in G_{x_2}} \int_K (\alpha g_1(u) + (1-\alpha)g_2(u))Q_t(du) - x,$$

and the right-hand side equals $\alpha h_t(x_1) + (1-\alpha)h_t(x_2)$.

Solution to Exercise 4. For brevity we suppress the requirements that each set of which we take the volume is measurable. Clearly, (6.2) implies (6.4). To prove the converse, note that we may assume that $\mathrm{vol}(A) \neq 0$ and $\mathrm{vol}(B) \neq 0$. Replacing A and B in (6.4) by the unit-volume bodies $\mathrm{vol}(A)^{-1/n}A$ and $\mathrm{vol}(B)^{-1/n}B$, and setting $\alpha = \mathrm{vol}(A)^{1/n}/(\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n})$, (6.4) becomes

$$(\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n})^{-1}\mathrm{vol}(A+B) \geq 1,$$

which is (6.4).

Solution to Exercise 5. We must have $r = (x/\mathrm{vol}(B_n))^{1/n}$ to match the volume of rB_n with the volume of K, so the right-hand side becomes $S((x/\mathrm{vol}(B_n))^{1/n}B_n) = (x/\mathrm{vol}(B_n))^{(n-1)/n}S(B_n)$. Therefore, (7.3) is equivalent with

$$S(K) \geq \left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_n)}\right)^{(n-1)/n} S(B_n),$$

which gives us (7.1) upon rearranging.

Solution to Exercise 6. Note that

$$S(K) \geq \lim_{\epsilon \downarrow 0} \frac{(\mathrm{vol}(K)^{1/n} + \epsilon \mathrm{vol}(B_n)^{1/n})^n - \mathrm{vol}(K)}{\epsilon} = n\mathrm{vol}(K)^{(n-1)/n}\mathrm{vol}(B_n)^{1/n},$$

and that $S(B_n) = n\mathrm{vol}(B_n)$. The desired inequality follows upon rearranging.

Bibliography

[1] C. Bélisle, A. Boneh, and R. J. Caron, *Convergence properties of hit-and-run samplers*, Comm. Statist. Stochastic Models, 14 (1998), pp. 767–800.

[2] C. J. P. Bélisle, H. E. Romeijn, and R. L. Smith, *Hit-and-run algorithms for generating multivariate distributions*, Math. Oper. Res., 18 (1993), pp. 255–266.

[3] H. C. P. Berbee, C. G. E. Boender, A. H. G. Rinnooy Kan, C. L. Scheffer, R. L. Smith, and J. Telgen, *Hit-and-run algorithms for the identification of nonredundant linear inequalities*, Math. Programming, 37 (1987), pp. 184–207.

[4] D. Bertsimas and S. Vempala, *Solving convex programs by random walks*, J. ACM, 51 (2004), pp. 540–556.

[5] C. G. E. Boender, R. J. Caron, A. H. G. Rinnooy Kan, and et al., *Shake-and-bake algorithms for generating uniform points on the boundary of bounded polyhedra*, Oper. Res., 39 (1991), pp. 945–954.

[6] B. Bollobás, *Volume estimates and rapid mixing*, in Flavors of geometry, Cambridge Univ. Press, Cambridge, 1997, pp. 151–182.

[7] E. J. Candes and T. Tao, *Near-optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.

[8] M.-H. Chen and B. W. Schmeiser, *General hit-and-run Monte Carlo sampling for evaluating multidimensional integrals*, Oper. Res. Lett., 19 (1996), pp. 161–169.

[9] S. Chib and E. Greenberg, *Understanding the Metropolis-Hastings algorithm*, The American Statistician, 49 (1995), pp. 327–335.

[10] S. E. Chick, *Subjective probability and bayesian methodology*, in Handbooks in OR and MS: Simulation, B. Nelson and S. Henderson, eds., Elsevier Science, Oxford, 2007, pp. 225–258.

[11] S. E. Chick and Y. Wu, *Selection procedures with frequentist expected opportunity cost bounds*, Oper. Res., 53 (2005), pp. 867–878.

[12] D. Dadush, C. Chandresekaran, and S. Vempala, *Thin partitions: Isoperimetric inequalities and a sampling algorithm for star shaped bodies*, in ACM-SIAM Symposium on Discrete Algorithms (SODA), 2010.

[13] A. Dieker and S.-H. Kim, *Selecting the best by comparing simulated systems in a group of three when variances are known and unequal*, in Proceedings of the 2012 Winter Simulation Conference, 2012.

[14] A. B. Dieker and X. Gao, *Sensitivity analysis for diffusion processes constrained to an orthant.* arXiv:1107.2871, 2011.

[15] A. B. Dieker, S. Ghosh, and M. S. Squillante, *Capacity allocation for stochastic networks.* preprint, 2012.

[16] M. Dyer and A. Frieze, *Computing the volume of convex bodies: a case where randomness provably helps*, in Probabilistic combinatorics and its applications (San Francisco, CA, 1991), vol. 44, Amer. Math. Soc., 1991, pp. 123–169.

[17] M. Dyer, A. Frieze, and R. Kannan, *A random polynomial-time algorithm for approximating the volume of convex bodies*, J. Assoc. Comput. Mach., 38 (1991), pp. 1–17.

[18] J. A. Fill, *Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process*, Ann. Appl. Probab., 1 (1991), pp. 62–87.

[19] P. Frazier, *Indifference-zone ranking and selection for more than 15,000 alternatives.* Technical Report, Cornell University, Ithaca, NY.

[20] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.

[21] R. J. Gardner, *The Brunn-Minkowski inequality*, Bull. Amer. Math. Soc. (N.S.), 39 (2002), pp. 355–405.

[22] A. E. Gelfand and A. F. M. Smith, *Sampling-based approaches to calculating marginal densities*, J. Amer. Statist. Assoc., 85 (1990), pp. 398–409.

[23] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-6 (1984), pp. 721–741.

[24] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.

[25] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric algorithms and combinatorial optimization*, vol. 2, Springer-Verlag, Berlin, second ed., 1993.

[26] M. Jerrum, A. Sinclair, and E. Vigoda, *A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries*, J. ACM, 51 (2004), pp. 671–697.

[27] R. Kannan, L. Lovász, and M. Simonovits, *Isoperimetric problems for convex bodies and a localization lemma*, Discrete Comput. Geom., 13 (1995), pp. 541–559.

[28] S.-H. Kim and A. Dieker, *Selecting the best by comparing simulated systems in a group of three*, in Proceedings of the 2011 Winter Simulation Conference, 2011.

[29] S.-H. Kim and B. L. Nelson, *A fully sequential procedure for indifference-zone selection in simulations*, ACM Trans. Modeling Comp. Simulation, 11 (2001), pp. 251–273.

[30] ———, *On the asymptotic validity of fully sequential selection procedures for steady-state simulation*, Oper. Res., 54 (2006), pp. 475–488.

[31] ———, *Selecting the best: theory and method*, in Handbooks in OR and MS: Simulation, B. Nelson and S. Henderson, eds., Elsevier Science, Oxford, 2007, pp. 501–534.

[32] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*, American Mathematical Society, Providence, RI, 2009.

[33] L. Lovász, *Hit-and-run mixes fast*, Math. Program., 86 (1999), pp. 443–461.

[34] L. Lovász and M. Simonovits, *Random walks in a convex body and an improved volume algorithm*, Random Structures Algorithms, 4 (1993), pp. 359–412.

[35] L. Lovász and S. Vempala, *Hit-and-run from a corner*, SIAM J. Comput., 35 (2006), pp. 985–1005.

[36] ———, *The geometry of logconcave functions and sampling algorithms*, Random Structures Algorithms, 30 (2007), pp. 307–358.

[37] E. Milman, *Isoperimetric bounds on convex manifolds*, in Concentration, functional inequalities and isoperimetry, vol. 545 of Contemp. Math., Amer. Math. Soc., Providence, RI, 2011, pp. 195–208.

[38] D. J. Reaume, H. E. Romeijn, and R. L. Smith, *Implementing pure adaptive search for global optimization using Markov chain sampling*, J. Global Optim., 20 (2001), pp. 33–47.

[39] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, Springer-Verlag, New York, second ed., 2004.

[40] L. C. G. Rogers and J. W. Pitman, *Markov functions*, Ann. Probab., 9 (1981), pp. 573–582.

[41] H. E. Romeijn, *Shake-and-bake algorithms for the identification of nonredundant linear inequalities*, Statist. Neerlandica, 45 (1991), pp. 31–50.

[42] ———, *A general framework for approximate sampling with an application to generating points on the boundary of bounded convex regions*, Statist. Neerlandica, 52 (1998), pp. 42–59.

[43] H. E. Romeijn and R. L. Smith, *Simulated annealing for constrained global optimization*, J. Global Optim., 5 (1994), pp. 101–126.

[44] R. L. Smith, *Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions*, Oper. Res., 32 (1984), pp. 1296–1308.

[45] S. Vempala, *Geometric random walks: a survey*, in Combinatorial and computational geometry, Math. Sci. Res. Inst. Publ., Cambridge Univ. Press, Cambridge, 2005, pp. 577–616.

[46] C. Villani, *Optimal transport: Old and new*, Springer-Verlag, Berlin, 2009.

[47] Z. B. Zabinsky, *Stochastic adaptive search for global optimization*, Springer, 2003.