THE UNIVERSITY OF CHICAGO

DIFFUSION IN COMPUTER SCIENCE AND STATISTICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
HARIHARAN NARAYANAN

CHICAGO, ILLINOIS
AUGUST 2009

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I cannot overstate my debt to Partha Niyogi for being an excellent advisor. Not only did I gain much from his unique perspectives on geometry, learning theory and life in general, but I also, as a result of his constant encouragement, remained motivated for the past several years. I am grateful to Ketan Mulmuley for introducing me to computational geometry and numerous conversations about complexity theory. I thank Gregory Lawler for giving me a glimpse of the beautiful theory of Schramm-Loewner Evolution. I am grateful to Ravi Kannan and Steve Lalley for many illuminating discussions on stochastic processes. I thank Laci Babai for the numerous occasions on which he examined my preprints and gave critical advice. I thank Janos Simon, Sasha Razborov and other faculty at the University of Chicago who enriched my graduate experience. I am grateful to my collaborators Lawler, Raecke, Radhakrishnan, Mulmuley, Belkin, Niyogi, Harsha, Kannan, Mahoney and Hayes from whom I learnt much.

I thank my professors at IIT Bombay, ISI Calcutta and Tata institute who were the first to give me a taste of graduate school. I am particularly grateful to Vivek Borkar, B. V. Rao, Amartya Kumar Dutta, Kalyan Mukherjea and Somesh Bagchi.

My years at Chicago would have been very different without my friends. I would like to thank all of them for the memorable times we had together. I am especially grateful to Rahul, Jordan, Vikas, Deepankar, Garry, April, Tan, Dan, Harsh and Harish for many pleasant trips and get-togethers. I thank Gabri for more things than I could enumerate. I am grateful to Amma and Appa for their constant support and encouragement.

ABSTRACT

In this thesis, we investigate diffusion as an algorithmic and analytic tool in statistics and computer science.

We address a question arising from computational linguistics, where we wish to understand the behavior of a network of agents modeled as nodes of a graph that adaptively modify their lexicon using data from their neighbors. By introducing a model of memory and a family of coalescing random walks, we prove that they eventually reach a consensus with probability 1.

We study distributed averaging on graphs and devise a distributed algorithm that is based on a diffusion process having two time scales.

Addressing the question of routing in a network, we use steady-state diffusions corresponding to electrical flow in a network of resistors for oblivious routing and prove that this scheme performs well under a variety of performance measures.

Based on a microscopic view of diffusion as an ensemble of particles executing independent Brownian motions, we develop the fastest currently known algorithm for computing the area of the boundary of a convex set. A similar technique is used to produce samplers for the boundaries of convex sets and smooth hypersurfaces that are the boundaries of open sets in $\mathbb{R}^n$, assuming access to samplers for the interior. These algorithms are motivated by Goodness-of-Fit tests in statistics.

The halfplane capacity, a quantity often used to parameterize stochastic processes arising in statistical physics, known as Schramm-Loewner evolutions, is shown to be comparable to a more geometric notion.

We analyze a class of natural random walks on a Riemannian manifold, and give bounds on the mixing times in terms of the Cheeger constant and a notion of smoothness that relates the random walk to the metric underlying the manifold.

A Markov chain having a stationary distribution that is uniform on the interior of a polytope is developed. This is the first chain whose mixing time is strongly polynomial when initiated in the vicinity of the center of mass. This Markov chain can be interpreted as a random walk on a certain Riemannian manifold. The resulting algorithm for sampling polytopes outperforms known algorithms when the number of constraints is of the same order of magnitude as the dimension. We use a variant of this Markov chain to design a randomized version of Dikin's affine scaling algorithm for linear programming. We provide polynomial-time guarantees which do not exist for Dikin's algorithm.

Addressing a question from machine learning, under certain smoothness conditions, we prove that a form of weighted surface area is the limit of the weight of graph cuts in a family of random graphs arising in the context of clustering. This is done by relating both to the amount of diffusion across the surface in question.

Addressing a related issue on manifolds, we obtain an upper bound on the annealed entropy of the collection of open subsets of a manifold whose boundaries are well-conditioned. This result leads to an upper bound on the number of random samples needed before it is possible to accurately classify data lying on a manifold.

CHAPTER 1

INTRODUCTION

In this thesis, we investigate the use of diffusion as an algorithmic and analytic tool in several disciplines. The areas to which we apply the principle of diffusion include Statistics, Machine Learning, Convex Optimization, Statistical Physics and Computational Geometry.

1.1 History

The botanist Robert Brown was perhaps the first to identify, in a manuscript published in 1828, random movement of particles of matter, and note that this movement exists not only in particles from organic tissue but also from inorganic material. The latter issue had been a source of confusion for his predecessors. While many contributed to the understanding of random motion before 1900, it attracted widespread attention after a 1905 article of Albert Einstein titled "On the Motion of Small Particles Suspended in Liquids at Rest, Required by the Molecular-Kinetic Theory of Heat". In this article, Einstein tried to establish the existence and sizes of molecules and to compute Avogadro's number using the mean displacement of Brownian particles over a period of time. Louis Bachelier, who is credited with the first mathematical study of the Brownian process in his 1900 thesis, "The theory of speculation" did so in order to model stock options. In this work he introduced the model of a random walker. Many other scientists made key contributions to the theory of diffusion around this period, including Sutherland, Smoluchowski, Perrin and Langevin (see [24]).

Diffusion processes are intimately related to random walks, which in recent years have found numerous applications in Computer Science and Statistics. In the rest of this section, we will describe some basic random walks and delineate their relevance to the results contained in this thesis.

1.2 Random walks and diffusion

Consider the following random walk on $\mathbb{Z}$. Let $x_0 := 0$. Given x_i, toss a fair coin. If this lands *Heads*, set x_{i+1} to $x_i - 1$; otherwise set x_{i+1} to $x_i + 1$. This random process is canonically associated with a partial difference equation satisfied by the distributions of positions that the random walk occupies at successive time steps. Thus

$$\mathbb{P}[x_i = k] - \mathbb{P}[x_{i-1} = k] = \frac{\mathbb{P}[x_{i-1} = k-1] + \mathbb{P}[x_{i-1} = k+1] - 2\mathbb{P}[x_{i-1} = k]}{2}.$$

This is a discrete diffusion process. On the other hand, we have the heat equation or diffusion equation which causes suitable functions from $\mathbb{R}$ to $\mathbb{R}$ to evolve as a function of time (see Itô-McKean [35]).

$$\frac{\partial p_t(x)}{\partial t} = \frac{\partial^2 p_t(x)}{2(\partial x)^2}. \tag{1.1}$$

The operator $\frac{\partial^2}{2(\partial x)^2}$ is the infinitesimal generator of $1-$dimensional Brownian motion B_t, i. e. , for suitable functions f,

$$\lim_{t \downarrow 0} \frac{\mathbb{E}^{x_0}[f(B_t)] - f(x)}{t} = \frac{\partial^2 f}{2(\partial x)^2}\Big|_{x=x_0}.$$

Here the superscript x in $\mathbb{E}^x$ signifies that the Brownian motion was at x at time 0. If $p_0(x)$ is a density function, the distribution $p_t(x)$ obtained by solving the heat equation is also the probability density of Brownian motion witnessed at time t, if its position at time 0 was chosen according to the density $p_0(x)$. Thus, Brownian motions are related to continuous time diffusions.

1.3 Diffusion on graphs

Given an undirected graph (V, E), there is a natural random walk wherein the walker, when at a vertex v, picks a vertex uniformly at random from the neighbors of v and moves there. We will consider such random walks in Chapter 2, Chapter 3, and Chapter 4.

One particular question of interest has the following general form: *how might a group of linguistic agents arrive at a shared communication system purely through local patterns of interaction and without any global agency enforcing uniformity?* These agents could be artificial, for example robots or sensors, or natural such as animals or humans. In Chapter 2, we consider a model of the evolution of language among agents in a network and prove that in it, agents eventually arrive at a common language. This model generalizes a model studied by Liberman in [56] using simulations. Prior to our results, there was no theoretical analysis of either of these models. We assume that there is one concept and that each agent starts out with a probability distribution on words that may be used to describe the concept. At each time step, each agent produces a word that is heard by its neighbors and each agent modifies its probability distribution by taking a convex combination of the earlier distribution and the one supported equally on all the words it hears. We show that after a certain time period governed by the mixing time of a random walk on the network, with high probability, all agents produce the same word. In the analysis of this process, we trace the origin of words backwards in time. The trajectory of a word is a natural object to study since it leads us back to the word's source at the first time step, when the evolution was initiated. In our model, this path turns out to be a random walk, and by tracing multiple random walks backwards in time until they coalesce into a single source, we are able to give bounds on how quickly, all agents adopt the same word for a particular concept.

In Chapter 3, we consider the question of averaging on a graph that has one sparse cut separating two subgraphs that are internally well connected [70]. Such graphs could arise naturally in the context of sensor networks on uneven terrain. Consider a graph $G = (V, E)$, where i.i.d Poisson clocks with rate 1 are associated with each edge. Our algorithm uses non-convex combinations in an essential way. To the best of our knowledge, non-convex combinations have not appeared in past literature, in the context of distributed algorithms where each update is based only on current values.

We represent the "true" real valued time by T. Each node v_i holds a value $x_i(T)$ at time T. Let the average value held by the nodes be x_{av}. If the clock of an edge $e = (v_i, v_j)$ ticks at time T, it updates the values of vertices adjacent to it on the basis of $x_i(T), x_j(T)$ according to some algorithm $\mathcal{A}$. While there has been a large body of work devoted to algorithms for distributed averaging, nearly all algorithms involve only convex updates. In this chapter, we suggest that non-convex updates can lead to significant improvements. We do so by exhibiting a decentralized algorithm for graphs with one sparse cut that uses non-convex averages and has an averaging time that can be significantly smaller than the averaging time of known distributed algorithms, such as those of [8, 13]. Our algorithm makes non-convex updates as well as convex updates. The non-convex updates typically *increase* the variance of the values rather than decrease it, but in the process make a large transfer of mass across the sparse cut in the graph, which would not otherwise be possible. In order to obtain probabilistic bounds on the variance after t steps, we consider the logarithm of the variance as a function of time, and show that it is *stochastically dominated* by a random walk on the real line that possesses a negative drift.

In Chapter 4, we show that the asymptotic heat flow or the "electrical" flow is a good way of routing to minimize the sum of the p^{th} powers of edge loads in a graph [53]. We show that in graphs where the asymptotic heat flow from one vertex to another has a small ℓ_1 norm, this method of routing performs well under the aforementioned class of performance measures. We also show that this algorithm performs well on graphs on which a random walk mixes fast, such as expanders, which are of practical interest.

1.4 Diffusion and measures of sets

1.4.1 From random walks on $\mathbb{Z}$ to random walks on $\mathbb{R}$

It is possible to obtain a random walk on $\mathbb{R}$ by taking a scaling limit of the walk on $\mathbb{Z}$ mentioned in the beginning of Section 1.3. For any $t > 0$, $j \in \mathbb{N}$, we can define a random walk $\{z_i^j(t)\}_{i \geq 0}$ on $\left(\frac{1}{j}\right)\mathbb{Z}$, by setting

$$z_i^j(t) := \frac{x_{\lfloor 2t \cdot i \cdot j \rfloor}}{\sqrt{j}}.$$

As $j \to \infty$, for any fixed i, the distributions of $\{z_{i'}^j(t)\}_{1 \leq i' \leq i}$ converge in the Wasserstein metric defined below to a random walk on $\mathbb{R}$.

Definition 1.4.1. *Given two probability distributions μ_1, μ_2 supported on $\mathbb{R}^i$, we define the Wasserstein distance $W_2(\mu_1, \mu_2)$ by*

$$W_2(\mu_1, \mu_2) := \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \int_{\mathbb{R}^i \times \mathbb{R}^i} \|w - z\| d\gamma(w, z), \tag{1.2}$$

where $\Gamma(\mu_1, \mu_2)$ is the collection of all measures on $\mathbb{R}^i \times \mathbb{R}^i$ whose marginals on the first and second component are respectively μ_1 and μ_2.

This limit is, by the Central Limit Theorem [51], the same as the distribution of the first i steps of the random walk on $\mathbb{R}$ constructed by the following procedure.

Let

$$G_1^t(x, y) := \frac{1}{\sqrt{4\pi t}} e^{-\frac{\|x-y\|^2}{4t}}.$$

Random walk on $\mathbb{R}$:

1. Let $z_0(t) = 0$.

2. Given $z_i(t)$, let $z_{i+1}(t)$ be chosen from the distribution $G_1^t(z_i(t), \cdot)$ in a manner independent of $(z_1(t), \dots, z_i(t))$.

By putting together n independent copies of $\{z_i(t)\}$, we can construct a random walk on $\mathbb{R}^n$, whose transition kernel from a point x is an $n-$dimensional spherical Gaussian $G^t(x, \cdot)$ given by

$$G^t(x, y) := (4\pi t)^{-\frac{n}{2}} e^{-\frac{\|x-y\|^2}{4t}}. \tag{1.3}$$

Given an initial probability distribution μ_0 supported on $\mathbb{R}^n$, the probability distribution after one step can be expressed as a convolution of measures supported on $\mathbb{R}^n$,

$$\mu_1 \quad = \quad \mu_0 * G^t(0, \cdot).$$

Moreover μ_1 has a density ρ_1 with respect to the Lebesgue measure, that satisfies

$$\rho_1(x) \quad := \quad \int G^t(y, x) d\mu_0(y). \tag{1.4}$$

1.4.2 Measuring surface area

In Chapter 5, we consider a convex subset K of $\mathbb{R}^n$ respectively containing and contained in Euclidean balls of radius r and R. The convex set is specified by an oracle, which when presented with a point x in $\mathbb{R}^n$, returns "Yes" if $x \in K$ and "No" otherwise. We design a randomized algorithm [5] that computes the surface area of a convex set within a relative error ϵ, with a probability of failure less than δ, whose run-time, measured in terms of these quantities is

$$O\left(n^4 \log \frac{1}{\delta} \left(\frac{1}{\epsilon^2} \log^9 \frac{n}{\epsilon} + \log^8 n \log \frac{R}{r} + \frac{1}{\epsilon^3} \log^7 \left(\frac{n}{\epsilon}\right)\right)\right).$$

The task of developing a randomized algorithm for estimating the surface area of a convex set was mentioned as an open problem in the book “Geometric Algorithms in Combinatorial Optimization” by Grötschel, Lovász and Schrijver. Our algorithm is based on the fact that the amount of heat diffusing in a short period of time out of a uniformly heated body placed in a vacuum, is proportional to its surface area. If μ_0 is the uniform probability measure on K, the distribution of the point obtained by making one step from a random point in K is given by (1.4) above. We show that if $\sqrt{t} = O(\frac{\epsilon r_{in}}{n})$, where r_{in} is the radius of the largest Euclidean ball that can be placed in K, S is the surface area and V is the volume,

$$\mu_1(\mathbb{R}^n \setminus K) = \sqrt{\frac{t}{\pi}} \left(\frac{S}{V}\right) (1 + O(\epsilon)) \, . \tag{1.5}$$

The question of computing the volume of a convex body is a classical question in theoretical computer science and algorithms are known that perform this task, therefore S can be recovered from the ratio $\frac{S}{V}$ above.

Suppressing a polynomial dependence on $\frac{1}{\epsilon}, \ln\left(\frac{1}{\delta}\right)$ and $\ln\left(\frac{R}{r}\right)$, this algorithm makes $O^*(n^4)$ calls to a membership oracle compared to $O^*(n^{8.5})$ for the best previously known algorithm.

In Chapter 6, we use an extension of the above idea to develop an algorithm that samples the surface of an n-dimensional convex body. The underlying intuition is that if the initial distribution of particles is the uniform distribution on the interior of the set, then over a short period of time, Brownian particles diffuse out of the surface of a convex set almost uniformly. Therefore, if we approximately identify the points at which they exit the body, we obtain an approximately uniform sampler for the surface of the body [72].

In the same chapter, we also use this idea to sample a smooth hypersurface that is the boundary of a (not necessarily convex) subset of $\mathbb{R}^n$, if this subset can be sampled uniformly. These algorithms have applications to Goodness-of-Fit tests in statistics.

1.4.3 From random walks on $\mathbb{R}$ to Brownian motion

It is possible to "paste together" in a consistent manner, copies of the random walk on $\mathbb{R}$ mentioned above, corresponding to each $t \in \{2^{-i} | i = 1, 2, \ldots\}$ and obtain standard $1-$dimensional Brownian motion $\{B_t\}_{t\geq 0}$ in the limit. This is the basis of Lévy's construction of Brownian motion [91].

The standard $1-$dimensional Brownian motion, $\{B_t\}_{t\geq 0}$ is uniquely characterized by the following.

1. $B_0 = 0$.

2. For each $0 < s_1 < t_1 < \ldots < s_k < t_k$, $B_{t_1} - B_{s_1}, \ldots, B_{t_k} - B_{s_k}$ are independent Gaussians with mean 0 and respective variances $t_1 - s_1, \ldots, t_k - s_k$.

3. As a function of t, B_t is continuous with probability 1.

1.4.4 Schramm-Loewner Evolution and halfplane capacity

In Chapter 7, we consider a connected set A whose complement in the (complex) upper halfplane $\mathbb{H}$ is simply connected. We relate $\mathrm{hsiz}(A)$, which we define to be the 2-dimensional area of the union of all balls tangent to the real line and centered at points in A to a quantity known as halfplane capacity that is defined using diffusions.

Following Schramm's seminal paper [92] "Scaling limits of loop-erased random walks and uniform spanning trees," important progress was made towards understanding the conformal invariance of the scaling limits of several two dimensional lattice models in statistical physics by several researchers including Lawler, Schramm, Werner [54] and Smirnov [88]. These limits have been described using a new tool known as Schramm-Loewner Evolution (SLE).

The chordal Schramm-Loewner evolution with parameter $\kappa \geq 0$ is the random collection of conformal maps satisfying the following stochastic differential equation:

$$\dot{g}_t(z) = \frac{2}{g_t(z) - \sqrt{\kappa} B_t}, \; g_0(z) = z,$$

where z belongs to $\mathbb{H}$ and $\dot{g}_t$ represents the time derivative of g_t. Denoting the domain of g_t by H_t, we obtain a random collection of continuously growing hulls $K_t := \mathbb{H} \setminus H_t$. In this parametrization, the halfplane capacity (defined in Chapter 7) of K_t is equal to $2t$ [51]. Thus, halfplane capacity is a quantity arising naturally in the context of SLE. Our main theorem in Chapter 7 states that

$$\frac{1}{66} < \frac{\mathrm{hcap}(A)}{\mathrm{hsiz}(A)} \leq \frac{7}{2\pi}.$$

1.5 Random walks on manifolds and polytopes

1.5.1 From random walks on $\mathbb{R}^n$ to a ball walk

Suppose $t_n := \frac{t}{n}$. Then, as n tends to infinity, the Wasserstein distance between the $n-$dimensional spherical Gaussian distribution whose density is $G^{t_n}(0, \cdot)$ and the uniform distribution over the Euclidean ball of radius $\sqrt{2t}$ centered at the origin tends to 0. The latter distribution can be used to define a random walk on $\mathbb{R}^n$ in which a transition from x involves moving to a point chosen uniformly at random in a ball of radius $\sqrt{2t}$ centered at x. This random walk has a natural extension to manifolds given by an atlas, if the atlas possesses certain characteristics.

1.5.2 Random walks on manifolds

In statistics, one is interested in Goodness-of-Fit tests for a numerous of multivariate distributions. For example, testing for a Gamma distribution leads one to consider positive real valued random variables $X_1, \dots, X_n$ such that $\sum_i X_i = a$ and $\prod_j X_j = b$. The set of all $(X_1, \dots, X_n)$ under these constraints is a manifold of codimension two. The question of sampling this manifold has been raised by Diaconis et al (see [19]). We address this question and others of its kind in Chapter 6.

We take an intrinsic view of the question of sampling manifolds in Chapter 8. Let $\mathcal{M}$ be a manifold. Let $\mathcal{M}$ be specified by a family of smooth injective maps $\{U_x : B \to \mathcal{M}\}_{x \in \mathcal{M}}$ where B is the unit Euclidean ball centered at the origin and $x = U_x(0)$. Let $\rho(x)$ be a probability density function on $\mathcal{M}$, whose value at x is measured with respect to the push-forward of the Lebesgue measure via U_x at the point x. This family of maps corresponds to an atlas consisting of charts $\{U_x^{-1}\}_{x \in \mathcal{M}}$. We consider the Markov chain in which, for any $x \in \mathcal{M}$, the next point z is obtained as follows. Let Jac denote the Jacobian.

1. Toss a fair coin and if $Heads$, set z to x.
2. If $Tails$, do the following:
   (a) Pick a random point $w \in B$ from the uniform measure on the unit ball and let $z := U_x(w)$.
   (b) If $x \in U_z(B)$,
      i. with probability $\min\left(1, \frac{\rho(z) \det \mathrm{Jac}(U_z^{-1} U_x)(0)}{\rho(x)}\right)$ let z remain unchanged.
      ii. Else, set z to x.
   (c) If $x \notin U_z(B)$, set z to x.
3. Output z.

In Chapter 8, given a Riemannian metric, we relate the mixing time of this chain with the Cheeger constant of the weighted manifold, under some additional assumptions relating the family of maps $\{U_x\}_{x \in \mathcal{M}}$ to the metric.

1.5.3 Random walks on polytopes

In Chapter 9, we design a new Markov chain on the points of an $n-$dimensional polytope that mixes rapidly and whose stationary distribution is the uniform distribution [40]. The resulting algorithm for sampling polytopes outperforms existing algorithms when the number of constraints m is $O(n^{1.62})$. This random walk can be viewed as a random walk on a manifold in the framework of Chapter 8 with respect to a certain non-Euclidean metric. Algorithms for sampling polytopes have numerous applications. We mention one such application here. It was shown in [41] that if an $n-$dimensional polytope defined by m inequalities contains a ball of radius $\Omega(n\sqrt{\log m})$, then it is possible to sample the lattice points inside it in polynomial time by sampling the interior of the polytope and picking a nearby lattice point. Often, objects of interest such as contingency tables can be encoded as lattice points in a polytope, leading to algorithms for sampling them. Contingency tables are two-way tables that are used by statisticians to represent bivariate data. A solution proposed in [20] to the frequently encountered problem of testing the independence or dependence of two different attributes of empirical data involves sampling uniformly from the set of two-way tables having fixed row and column sums. It was shown in [67] that under some conditions, this can be achieved in polynomial time by quantizing random points from an associated polytope. The results of Chapter 9, together with the results of [67] lead to an algorithm that outperforms known algorithms for sampling contingency tables when the row and column sums are sufficiently large.

Let K be a polytope in $\mathbb{R}^n$ defined by m linear inequalities. The underlying Markov chain is the first to have a mixing time that is strongly polynomial when started from a “central” point x_0. If s is the supremum over all chords $\overline{pq}$ passing through x_0 of $\frac{|p-x_0|}{|q-x_0|}$ and ϵ is an upper bound on the desired total variation distance from the uniform distribution, it is sufficient to take $O\left(mn\left(n\log(sm)+\log\frac{1}{\epsilon}\right)\right)$ steps of the random walk. We use this result to design an affine interior point algorithm that does a *single* random walk to solve linear programs approximately. More precisely, suppose $Q = \{z \big| Bz \leq \mathbf{1}\}$ contains a point z such that $c^T z \geq d$ and $r := \sup_{z \in Q} \|Bz\| + 1$, where B is an $m \times n$ matrix, c is a $n-$dimensional vector and $d > 0$. Then, after $\tau = O\left(mn\left(n \ln\left(\frac{mr}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$ steps, the random walk is at a point x_τ for which $c^T x_\tau \geq d(1-\epsilon)$ with probability greater than $1-\delta$. The fact that this algorithm has a run-time that is provably polynomial is notable since the analogous deterministic affine algorithm analyzed by Dikin has no known polynomial guarantees.

1.6 Learning Theory

A standard assumption in learning theory is that data is generated by sampling independently at random from some unknown probability distribution. A class of algorithms termed graph-based methods first construct a weighted graph, whose vertices are the data points and whose edge weights are determined by the relative positions of the points and perform natural graph-theoretic operations on the resulting graph. Graph-based methods are extensively used for clustering and other machine learning tasks and are known to work very well. The reasons behind their excellent performance are not well understood. In Chapter 10, we prove that the weights of certain graph cuts corresponding to well-conditioned hypersurfaces tend to a weighted surface area of the associated hypersurfaces, thus proving that in the limit, these graph cuts tend to meaningful quantities [71]. We use Gaussian weights on edges . This was the first result of its kind, but subsequently, other schemes for producing graphs from data, such as k-Nearest Neighbors have been analyzed. For example, it was shown in [63] that as the number of data points tends to infinity, graph cuts corresponding to hyperplanes can have different limits depending on the scheme used to construct the graphs. One of the intuitions underlying many graph-based methods for clustering and semi-supervised learning, is that class or cluster boundaries pass through areas of low probability density. We provide formal analysis of this notion. We introduce a notion of weighted boundary area, which measures the area of the boundary weighted by the density of the underlying probability distribution. We show that the sizes of the cuts induced by a smooth hypersurface on commonly used adjacency graphs converge to the volume of the boundary weighted by the underlying density. The proof uses a relation between the amount of diffusion across a hypersurface and the expected weight of a graph cut induced by the hypersurface. The proof then relates the amount of diffusion across the hypersurface to the weighted surface area of the hypersurface.

While modern data sets are typically high dimensional, they can often be successfully modeled as lying on low dimensional manifolds embedded in high dimensional space because they are generated by a dynamical system that has relatively few degrees of freedom. In recent years, the hypothesis that data lie on a low dimensional manifold has gained acceptance, and the class of methods whose theoretical validity is contingent upon this hypothesis has formed a subfield of its own called manifold learning. Some of these methods belong to the family of graph-based methods discussed above. In Chapter 11, we study the following question central to manifold learning:

How many random samples of data are needed to learn a smooth cut on a manifold? [73]

Let $\mathcal{M}$ be a manifold that is a submanifold of $\mathbb{R}^m$. We consider *smooth cuts* where each cut corresponds to a submanifold (say $P \subset \mathcal{M}$) that divides $\mathcal{M}$ into two pieces. Since P is a submanifold of $\mathcal{M}$, and therefore $\mathbb{R}^m$, one can associate to it a measure of complexity given by its condition number $1/\tau$, where τ is the supremum over all r such that the tubular neighborhood of radius r does not self-intersect (see Definition 6.3.1). Corresponding to each cut whose condition number is $\leq \frac{1}{\tau}$, we may associate two indicator functions, one for either part. Let $\mathcal{C}_\tau$ be the class of functions that can be obtained in this manner. By letting τ vary, we obtain a structured family of classification functions. The number of samples needed to learn the elements of $\mathcal{C}_\tau$ is a natural quantity to study, and is a natural generalization of halfspace learning on the surface of a sphere, to arbitrary manifolds. In Chapter 11, we prove sample complexity bounds that depend on the maximum density ρ_{max}

of the distribution $\mathcal{P}$ from which samples are drawn, the curvatures of $\mathcal{M}$ and the class boundary, and the dimension of $\mathcal{M}$ (but not the ambient dimension). This is achieved by bounding the annealed entropy of $\mathcal{C}_\tau$ with respect to $\mathcal{P}$. We show that the dependence on the maximum density ρ_{max} of $\mathcal{P}$ is unavoidable by proving that for any fixed τ, there exist manifolds for which the VC-dimension of $\mathcal{C}_\tau$ is infinite.

CHAPTER 2

LANGUAGE EVOLUTION, COALESCENT PROCESSES AND THE CONSENSUS PROBLEM ON A SOCIAL NETWORK

In recent times, there has been an increased interest in theories of language evolution that have an applicability to the study of dialect formation, linguistic change, creolization, the origin of language, and animal and robot communication systems in general. (see [45, 76, 33] and references therein). One particular question of interest has the following general form: *how might a group of linguistic agents arrive at a shared communication system purely through local patterns of interaction and without any global agency enforcing uniformity?* The linguistic agents in question might be humans, animals, or machines in a multi-agent society. For an example of interesting simulations that suggest how a shared vocabulary might emerge in a population, (see Liberman [56]) (other simulations are also provided by [90, 10] among others). In this chapter, we consider a generalization of Liberman's model, prove several theoretical properties, and establish connections to related phenomena in population genetics through coalescent processes.

Our model is as follows. For simplicity, we consider how a common word for a particular concept might emerge through local interactions even though the agents had different initial beliefs about the word for this concept. For example agents might use the phonological forms "coconut","nariyal", "thengai" etc. to describe the concept of the fruit. Thus we imagine a situation where every time an event in the world occurs that requires the agents to use a word to describe this event, they may start out by using different words based on their initial belief about the word for this event or object. By observing the linguistic behavior of their neighbors agents might update their beliefs. The question is - will they eventually arrive at a common word and if so how fast.

2.0.1 Model

1. Let $\mathbf{W}$ be a set of words (phonological forms, codes, signals, etc.) that may be used to denote a certain concept (meaning or message).

2. Let each agent hold a belief that is a probability measure on $\mathbf{W}$. At time t, we denote the belief of agent i to be $\mathbf{b}_i^{(t)}$.

3. Agents are on a communication network which we model as a directed weighted graph where vertices correspond to agents. We further assume that the weight of each directed edge is positive and that there exists a directed path from any node to any other. An agent (say i) can only observe the linguistic actions of its out-neighbors, i. e. nodes to which a directed edge points from i. We denote weight of the edge from i to j by P_{ij}, where for any i, $\sum_j P_{ij} = 1$.

4. The update protocol for the $\mathbf{b}_i^{(t)}$ as a function of time is as follows:

   (a) At each time t, each agent i chooses a word $w = w_i^{(t)} \in \mathbf{W}$ (randomly from to its current belief $\mathbf{b}_i^{(t)}$) and produces it. Let $X_i^{(t)}$, denote the probability measure concentrated at $w_i^{(t)}$. Since $w_i^{(t)}$ is a random word, $X_i^{(t)}$ is correspondingly a random measure.

   (b) At every point in time, each agent can observe the words that their neighbors produce but they have no access to the private beliefs of these same neighbors.

   (c) Let P be the matrix whose ij^{th} entry satisfies

$$P_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n} A_{ij}}.$$

   At every time step, every agent updates its belief by a weighted combination of its current belief and the words it has just heard, i.e.,

$$\mathbf{b}_i^{(t+1)} = (1-\alpha)\mathbf{b}_i^{(t)} + \alpha \sum_{j=1}^{n} P_{ij} X_j^{(t)},$$

where α is a fixed real number in the interval $(0,1)$ that is not time dependent.

At a time t, the beliefs of the agents are represented by the vector

$$\mathbf{b}^{(t)} := (\mathbf{b}_1^{(t)}, \ldots, \mathbf{b}_n^{(t)})^T.$$

Similarly, let the point measures on words $X_i^{(t)}$ be organized into a vector

$$X^{(t)} := (X_1^{(t)}, \ldots, X_n^{(t)})^T.$$

Then the reassignment of beliefs can be expressed succinctly in matrix form where the entries in the vectors involved are measures rather than numbers as

$$\mathbf{b}^{(t+1)} = (1-\alpha)\mathbf{b}^{(t)} + \alpha P X^{(t)}. \tag{2.1}$$

2.0.2 Remarks:

1. If beliefs were directly observable and agents updated based on a weighted combination of their beliefs and that of their neighbors,

$$\mathbf{b}^{(t+1)} = (1-\alpha)\mathbf{b}^{(t)} + \alpha P \mathbf{b}^{(t)}, \tag{2.2}$$

the system has a simple linear dynamics, where all beliefs converge to a weighted average of the initial beliefs. Thus eventually, everyone has the same belief (see [8] for pioneering work and [36] for a recent elaboration in an economic context.)

2. Our focus in this chapter is on the situation where the beliefs are *not observable* but only the linguistic actions $X_i^{(t)}$ are (and only to the immediate neighbors). Therefore, the corresponding dynamics follows a Markov chain. The state space of this chain (defined by Equation 2.1) is the set of all n-tuples of belief vectors. Since this is continuous, the standard mixing results with finite state spaces do not apply directly.

3. Note that in our setting we have assumed that the communication matrix A_{ij} does not change with time. If this matrix changes with time the evolution is not Markovian in the usual sense but the arguments in this chapter when combined with results in [95] would lead to a proof of convergence under suitable conditions. We omit this analysis for ease of exposition.

2.0.3 Results:

Our main results are summarized below.

1. With probability 1 (w.p.1), as time tends to infinity, the belief of each agent converges in variation distance to one supported on a single word, common to all agents.

2. W.p.1, there is a finite time T such that for all times $t > T$, all agents produce the same fixed word.

3. The rate at which beliefs converge depends upon the mixing properties of the Markov chain whose transition matrix is P.

4. The rate of convergence is *independent* of the size of $\mathbf{W}$. One might think that a population where every agent has one of two words for the concept would arrive at a shared word faster than one in which every agent had a different word for the concept. This intuition turns out to be incorrect.

The proof of these results exposes a natural connection with coalescent processes and has a parallel in population genetics. Our analysis brings out two different interpretations of the behavior of a linguistic agent. In the most direct interpretation, the agent's linguistic knowledge of the word is internally encoded in terms of a belief vector. This belief vector is updated with experience. In a second interpretation an agent's representation of its linguistic knowledge is in terms of a memory stack in which it literally stores every single word it has heard weighted by how long ago it heard it and the importance of the person it heard it from. Such an interpretation is consistent with exemplar theory (see [14]). An external observer looking at this agent's linguistic actions will not be able to distinguish between these two different internal representations that the agent may have.

2.0.4 Connections to other fields

Linear update rules are often used in distributed systems, to achieve coherence among different agents or to share knowledge gathered individually. In a model that has been intensively studied, a number of sensors form a network, each of which measures a quantity such as temperature [8]. Neighbors communicate during each time step and make linear updates in a synchronous or asynchronous manner. The rate at which consensus is attained is studied. There is also a related body of work on Coordination and Distributed Control. A model of flocking has been considered in [17], where a group of birds, have a certain initial velocity, and the evolution of their velocities is governed by a differential equation wherein each bird modifies its velocity to bring it closer to that of its neighbors. The update rule involves a graph Laplacian. Some results are derived concerning the initial conditions that result in flocking behavior.

There are two connections to evolutionary theory that are worth mentioning. First, our proof of convergence exposes a natural coalescent process over words. Coalescent processes are, of course, widely used in modeling and making inferences about genetic evolution [34, 44].

Second, researchers have considered game-theoretic models of evolution [89] and more recent research in this tradition has addressed evolutionary games on graphs [79]. The question of how agents may learn an appropriate strategy for a coordination game on a graph has many high level similarities to the problem studied in this chapter.

Finally, there have been a large number of models on achieving coherence in a linguistic population. Many of these rely on simulations. Among mathematical studies, two strands are worth noting. The model of language evolution proposed in [18] has many similarities with languages of agents evolving on a graph. But it is worth noting that in that model, if at each time step, the number of linguistic examples (observations) collected by each agent is bounded from above by a constant (independent of time), the community fails to achieve a consensus language. A second strand is the collection of results obtained in [78, 45]. While there are many synergies with that body of work, there is nothing that is directly comparable.

2.1 Convergence to a Shared Belief: Quantitative results

Let $\tilde{P}$ be the transition matrix on the augmented agent space $\tilde{S} = S \cup \hat{S}$, where for $i, j \in S := \{1, \ldots, n\}$ and $\hat{S} = \{\hat{1}, \ldots, \hat{n}\}$.

$$\tilde{P}(i \to j) = \tilde{P}(\hat{i} \to j) = \alpha P_{ij},$$

$$\tilde{P}(i \to \hat{i}) = \tilde{P}(\hat{i} \to \hat{i}) = 1 - \alpha.$$

Definition 2.1.1. *Let $T_{mix}(\epsilon)$ denote the mixing time of $\tilde{P}$, defined as the smallest t for which, for each specific choice of $v, w \in \tilde{S}$,*

$$\sum_{u \in \tilde{S}} |\tilde{P}^{(t)}(v \to u) - \tilde{P}^{(t)}(w \to u)| < \epsilon.$$

Here $\tilde{P}^{(t)}(b \to c)$ denotes the probability that a Markov Chain governed by $\tilde{P}$ starting in b lands in c at the t^{th} time step.

The following is the main result of this chapter.

Theorem 1:

1. The probability that all agents produce the same word at times $T, T+1, \ldots$ tends to 1 as T tends to ∞. More precisely, if

$$\begin{aligned} \tau &= (4n/\alpha^2) T_{mix}(\frac{\alpha}{4}) \ln(4n/\alpha^2) \\ M &= e, \end{aligned}$$

   then

$$\mathbb{P}[\forall_{\substack{t \geq T \\ u \in S}} \; X_u^t = X_1^T] > 1 - \frac{MnTe^{-\frac{T}{\tau}}}{1 - e^{-\frac{T}{\tau}}}. \tag{2.3}$$

2. As time $t \to \infty$ all produced words converge (almost surely) to a word whose probability distribution is

$$\sum_{i=1}^{n} \pi_i \mathbf{b}_i^{(0)},$$

   where $(\pi_1, \ldots, \pi_n)$ is the stationary distribution of the Markov chain whose transition matrix is P.

2.1.1 A Model of Memory

The evolution of the $B^{(t)}$ is a Markov chain. It can be seen that its only absorbing states are of the form $(\mathbf{b}_1^{(t)}, \ldots, \mathbf{b}_n^{(t)})^T$, where $\forall i, \mathbf{b}_i^{(t)} = \delta_w$, and δ_w is the point measure concentrated on some word $w \in \mathbf{W}$. Formally, δ_w is the measure on $\mathbf{W}$, which assigns to a measurable set A the measure $\delta_w(A)$ according to the following rule.

$$\begin{aligned}\delta_w(A) &= 1 \quad \text{If } w \in A \\ &= 0 \quad \text{otherwise.}\end{aligned}$$

Therefore, if the Markov Chain were finite, a simple argument would suffice. However, the state space of our Markov Chain is uncountably infinite. Thus in principle, its dynamics could be hard to analyze. Our proof is based on coalescent processes, which have also been extensively used to study biological evolution [34, 44]. In analyzing the evolution of beliefs, we trace the origin of words backwards in time and find that all surviving words, are copies of a single word produced at some point in time sufficiently far in the past. Observe that if the process had begun at time 0, the beliefs at time $t+1$ would be

Observation 2.1.1.

$$B^{(t+1)} = \sum_{i=0}^{t} \alpha(1-\alpha)^i P X^{(t-i)} + (1-\alpha)^{t+1} B^{(0)}. \tag{2.4}$$

$X^{(t)} = (X_1^{(t)}, \ldots, X_n^{(t)})^T$ is a random vector whose entries are point measures, where $X_i^{(t)} = \delta(w_i^{(t)})$ and $w_i^{(t)}$ is chosen from the measure $\mathbf{b}_i^{(t)}$ on $\mathbf{W}$, independent of the choice of other coordinates of the vector $X^{(t)}$. This observation, motivates a model of memory that we define next.

Let each agent's memory be modeled as a stack. At the top level of the stack of agent i are all the words heard at time t. Below this are all words heard at time $t-1$ and so on tracing backwards in time until the first words heard at an initial time 1. At the lowest level, corresponding to time 0, is the initial belief $\mathbf{b}_i^{(0)}$ which is a probability distribution on the set of words. We may imagine this to be a form of vestigial memory.

Let agent j be adjacent to agent i. We shall describe the process by which agent j

produces word $w_j(t)$ and which induces $X_j(t)$, the point measure supported on $X_j(t)$. Let S_j be the stack held by agent j, and $S_j^{(t)}, \ldots, S_j^{(0)}$ be the levels in its stack from top to bottom. After j produces $X_j(t)$, i places $X_j(t)$, and all other $X_{j'}(t)$ produced by neighbors of i at time step t on the top of its stack. In order to describe the mechanism by which $X_j(t)$ is generated, let us introduce a geometric random variable Y where

$$\mathbb{P}[Y = i] = \alpha(1-\alpha)^i.$$

If $Y \leq t-1$, $X_j(t)$ is chosen to be the word produced by j' at time $t-1-Y$ (which is stored in S_{t-1-Y}) with probability $P_{jj'}$. If $Y \geq t$, $X_j(t)$ is chosen from the distribution in $\mathbf{b}_j^{(0)}$. This process has been illustrated in Figure 2.1.1. Note that in this model words are formal objects. While any two words present in the stack positions $S_j^{(t)}$ for $t = 1, 2, \ldots$ are considered distinct, there is a natural "parent-child" structure existing on the set of words. Under this scheme, let the probability distribution of $X_i^{(t)}$ be denoted $\tilde{\mathbf{b}}_i^{(t)}$. Denoting by $\tilde{B}^{(t)}$ the vector $(\tilde{\mathbf{b}}_1^{(t)}, \mathbf{b}_2^{(t)}, \ldots, \mathbf{b}_n^{(t)})$.

Observation 2.1.2. *A direct computation shows that in the model just described*

$$\tilde{B}^{(t+1)} = \sum_{i=0}^{t} \alpha(1-\alpha)^i P X^{(t-i)} + (1-\alpha)^{t+1} \tilde{B}^{(0)}. \tag{2.5}$$

This along with the fact that the randomness used in the generation of $X_j^{(t)}$ is independent of the randomness in the generation of all other words, tells us that the model of memory just described results in a system with the same dynamics as that introduced earlier. This particular model of memory may be viewed as an implementation of the ideas implicit in exemplar based accounts of linguistic behavior.

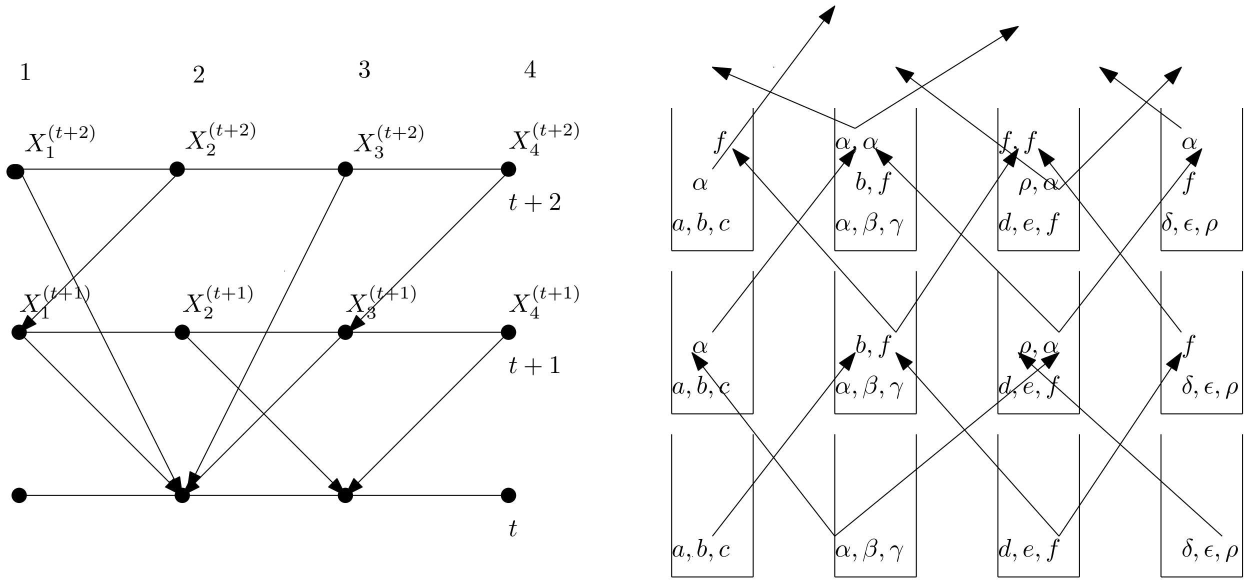

Figure 2.1: A coalescent process obtained by tracing the origin of words backwards in time, and the associated memory stacks of agents 1 to 4 for time steps t to $t+2$. Each agent produces α at time $t+2$ due to coalescence to a single word α produced by agent 2 at time t.

2.2 Proofs

By observations 2.1.1 and 2.1.2, in order to obtain an upper bound on $\mathbb{P}[X_i^{(t_1)} \neq X_j^{(t_2)}]$, it is sufficient to trace the ancestry of both words backwards in time and show that the probability that they do not have a common ancestor is small. Our results are best stated in terms of the coalescence time of a set of random walks. In Figure 2.2, we illustrate how the path tracing the origin of a word backwards in time can be encoded as a Markov chain on a state space $S \cup \hat{S} = \{1, \ldots, n, \hat{1}, \ldots, \hat{n}\}$. We use the states $\hat{1}, \ldots, \hat{n}$ as additional "memory" states. Since the random variable Y introduced in section 2.1.1 can be interpreted as the length of a run of heads in a biased coin (whose probability of coming heads is $1-\alpha$), we can account Y using additional memory states.

We define a variant of the meeting time between two Markov Chains as follows. Let $u, v \in S \cup \hat{S}$.

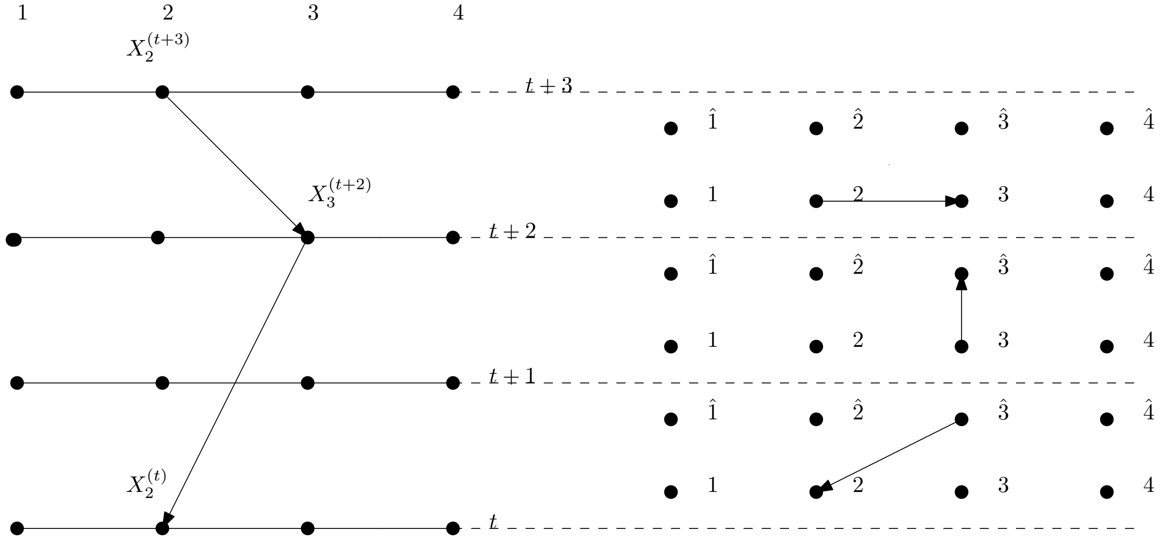

Figure 2.2: The ancestry of $X_2^{(t+3)}$ has been traced backwards in time to $X_2^{(t)}$. On the right,is an encoding of this path in terms of the transitions in a Markov Chain with "auxiliary states" $\hat{1}, \ldots, \hat{n}$. $\hat{3}$ is occupied at time step $t+1$ because the agent 3 produced a word at a time $t+2$ from past memory.

Definition 2.2.1. *For $t \geq 0$, let Y_t and Z_t be two independent random walks on $S \cup \hat{S}$ each of which has $\tilde{P}$ as its transition matrix and have initial states $Y_0 = u, Z_0 = v$. For $\Delta > 0$, let $M_{uv}(\Delta)$ be the smallest time $t > 0$ for which $Y_{t+\Delta} = Z_t \in S$.*

Theorem 2.2.1. *1. The probability that all agents produce the same word at times $T, T+1, \ldots$ tends to 1 as T tends to ∞. More precisely, if*

$$\begin{aligned} \tau &= (4n/\alpha^2) T_{mix}(\frac{\alpha}{4}) \ln(4n/\alpha^2) \\ M &= e, \end{aligned}$$

then

$$\mathbb{P}[\forall_{\substack{t\geq T\\ u\in S}}\ X_u^t = X_1^T] > 1 - \frac{MnTe^{-\frac{T}{\tau}}}{1-e^{-\frac{T}{\tau}}}. \tag{2.6}$$

2. *As time $t \to \infty$, all produced words converge (almost surely) to a random word chosen from the probability distribution*

$$\sum_{i=1}^{n} \pi_i \mathbf{b}_i^{(0)},$$

where $(\pi_1, \dots, \pi_n)$ is the stationary distribution of the Markov chain whose transition matrix is P.

Proof. To prove the first part, we observe that

$$\mathbb{P}\left[\neg\left(\forall_{\substack{t\geq T\\ u\in S}}\ X_u^t = X_1^T\right)\right]$$

$$\leq \sum_{j=1}^{\infty}\left(\mathbb{P}[X_1^{jT} \neq X_1^{(j+1)T}] + \sum_{k=0}^{T-1}\sum_{u=1}^{n}\mathbb{P}[X_u^{jT+k} \neq X_1^{jT}]\right)$$

by the union bound. The following application of Lemmas 2.2.1 and 2.2.2 completes the proof.

$$\mathbb{P}\left[\neg\left(\forall_{\substack{t\geq T\\ u\in S}} X_u^t = X_1^T\right)\right]$$

$$\leq \sum_{j=1}^{\infty}\left(\mathbb{P}[X_1^{jT} \neq X_1^{(j+1)T}] + \sum_{k=0}^{T-1}\sum_{u=1}^{n}\mathbb{P}[X_u^{jT+k} \neq X_1^{jT}]\right)$$

$$\leq \sum_{j=1}^{\infty}\left(\mathbb{P}[M_{11}(T) \geq jT] + \sum_{k=0}^{T-1}\sum_{u=1}^{n}\mathbb{P}[M_{u1}(k) \geq jT]\right)$$

$$\leq \frac{MnTe^{-\frac{T}{\tau}}}{1-e^{-\frac{T}{\tau}}},$$

where M and τ are the constants that appear in Lemma 2.2.2.

To prove the second part, we use the linearity of expectation to show that the expected value of the beliefs follows a simple rule. Namely

$$\begin{aligned}\mathbb{E}\mathbf{b}^{(t+1)} &= (1-\alpha)\mathbb{E}\mathbf{b}^{(t)} + \alpha P \mathbb{E}X^{(t)} \\ &= ((1-\alpha)I + \alpha P)\mathbb{E}\mathbf{b}^{(t)} \\ &= \dots \\ &= ((1-\alpha)I + \alpha P)^{t+1}\mathbb{E}\mathbf{b}^{(0)}.\end{aligned}$$

By well known results on Markov chains,

$$\lim_{t\to\infty}((1-\alpha)I + \alpha P)^t = (1,\dots,1)^T(\pi_1,\dots,\pi_n),$$

where π_i is the stationary probability of the state i under the chain P. Therefore, for each j,

$$\lim_{t\to\infty}\mathbb{E}\mathbf{b}_j^{(t)} = \sum_{i=1}^{n}\pi_i\mathbf{b}_i^{(0)},$$

By the first part of this theorem, as $t \to \infty$, $\mathbf{b}^{(t)}$ converges almost surely to a measure that is concentrated on a single common word w. Given a signed measure μ, let

$$|\mu| = \sup_{\|f\|_\infty \le 1}\int f d\mu.$$

Then,

$$\begin{aligned}\left|\mathbb{E}[\delta_w] - \mathbb{E}[X_i^T]\right| &\leq \mathbb{P}\left[\neg\left(\forall_{\substack{t\geq T\\u\in S}} X_u^t = X_1^T\right)\right]\\ &\leq \frac{MnTe^{-\frac{T}{\tau}}}{1-e^{-\frac{T}{\tau}}},\end{aligned}$$

It follows that this common word w must have the distribution $\sum_{i=1}^{n} \pi_i \mathbf{b}_i^{(0)}$. □

Lemma 2.2.1. *The probability that the word produced by agent u at time step t_1 is different from that produced by agent v at time step t_2 greater than t_1 can be bounded from above as follows.*

$$\mathbb{P}[X_u^{(t_1)} \neq X_v^{(t_2)}] \leq \mathbb{P}[M_{uv}(t_2 - t_1) \geq t_1].$$

Proof. In the model of memory introduced in section 2.1.1 we described a parent-child relationship between words, where a child word is identical to a parent word. The evolution of the Markov chain defined in this section corresponds to the geneology of a word. The event that the words $X_u^{(t_1)}$ and $X_v^{(t_2)}$ have a common ancestor produced at some time ≥ 0 is the event that $M_{uv}(t_2 - t_1) \leq t_1$. The lemma follows from the fact that two words that have a common ancestor are the same. □

Lemma 2.2.2. *The random variable $M_{uv}(\Delta)$ has an exponential tail bound uniform over u, v and Δ. More precisely, there exist constants $M, \tau > 0$ independent of u, v and Δ such that*

$$\mathbb{P}[M_{uv}(\Delta) \geq T] < Me^{-\frac{T}{\tau}}.$$

(In fact, this is satisfied for $\tau = \frac{4n}{\alpha^2} T_{mix}(\frac{\alpha}{4})$ and $M = e$.)

Proof. The stationary measure $\tilde{\mu}$ satisfies for each i, the identity $\alpha\tilde{\mu}(\hat{i}) = (1-\alpha)\tilde{\mu}(i)$.

Let $\tau_1 = T_{mix}(\frac{\alpha}{4})\ln(\frac{4n}{\alpha^2})$. Let us denote by $q_u(i)$ the probability $\mathbb{P}[Z_\tau = i \big| Z_0 = u]$. Then,

$$
\begin{aligned}
\sup_{u,v} \mathbb{P}[\neg(Y_{\tau+\Delta} = Z_\tau \in S) \big| Y_\Delta = u, Z_0 = v] \\
= 1 - \inf_{u,v} \sum_{i \in S} q_u(i) q_v(i) \\
\leq 1 - \inf_{u,v} \sum_{i \in S} \min(q_u(i), q_v(i))^2 \\
\leq 1 - \inf_{u,v} \frac{(\sum_{i \in S} \min(q_u(i), q_v(i)))^2}{n} \\
\leq 1 - \frac{\alpha^2}{4n}.
\end{aligned}
$$

Now, using the Markov property and conditioning repeatedly, we see that

$$
\begin{aligned}
\mathbb{P}[M_{uv}(\Delta) \geq T] \ \leq \ \mathbb{P}[\neg(Y_\Delta = Z_0 \in S)] \times \\
\prod_{i=1}^{\lfloor \frac{T}{\tau_1} \rfloor} \sup_{u,v} \mathbb{P}[\neg(Y_{\Delta + i\tau_1} = Z_{i\tau_1} \in S) \big| \\
(Y_{\Delta+(i-1)\tau_1}, Z_{(i-1)\tau_1}) = (u,v)] \\
\leq \mathbb{P}[\neg(Y_\Delta = Z_0 \in S)] \prod_{i=1}^{\lfloor \frac{T}{\tau_1} \rfloor} (1 - \frac{\alpha^2}{4n}) \\
\leq \left(1 - \frac{\alpha^2}{4n}\right)^{\frac{T}{\tau_1} - 1} \leq e^{1 - \frac{T}{\tau}}.
\end{aligned}
$$

where

$$
\tau = \frac{4n}{\alpha^2} T_{mix}(\frac{\alpha}{4}) \ln\left(\frac{4n}{\alpha^2}\right),
$$

which proves the Lemma. □

CHAPTER 3

DISTRIBUTED CONSENSUS ON A NETWORK HAVING A SPARSE CUT

3.1 Introduction

We consider the question of averaging on a graph that has one sparse cut separating two subgraphs that are internally well connected. While there has been a large body of work devoted to algorithms for distributed averaging, nearly all algorithms involve only *convex* updates. In this chapter, we suggest that *non-convex* updates can lead to significant improvements. We do so by exhibiting a decentralized algorithm for graphs with one sparse cut that uses non-convex averages and has an averaging time that can be significantly smaller than the averaging time of known distributed algorithms, such as those of [8, 13]. We use stochastic dominance to prove this result in a way that may be of independent interest.

Consider a Graph $G = (V, E)$, where i.i.d Poisson clocks with rate 1 are associated with each edge. We represent the "true" real valued time by T. Each node v_i holds a value $x_i(T)$ at time T. Let the average value held by the nodes be x_{av}. Every time an edge $e = (v, w)$ ticks, it updates the values of vertices adjacent to it on the basis of present and past values of v, w and their immediate neighbors according to some algorithm $\mathcal{A}$. There is an extensive body of work surrounding the subject of gossip algorithms in various contexts. Non-convex updates have been used in the context of a second order diffusion for load balancing [68] in a slightly different setting. The idea there was to take into account the value of the nodes during the previous two time steps rather than just the previous one, (in a synchronous setting), and set the future value of a node to a non-convex linear combination of the past values of some of its neighbors. There is also a line of research on averaging algorithms having two time scales, [12, 48] which is closely related to the present chapter.

In [69], the use of non-convex combinations for gossip on a geographic random graph on

n nodes was considered. It was shown that one can achieve averaging using $n^{1+o(1)}$ updates if one is willing to allow a certain amount of centralized control.The main technical difficulty in using non-convex updates is that they can skew the values held by nodes in the short term. We show that nonetheless, in the long term this leads to faster averaging. Let the values held by the nodes be $X(T) = (x_1(T), \dots, x_{|V|}(T))^T$. We study distributed averaging algorithms $\mathcal{A}$ which result in

$$\lim_{T \to \infty} X(T) = x_{av}\mathbf{1},$$

where x_{av} is invariant under the passage of time and show that in some cases there is an exponential speed-up in n if one allows the use of non-convex updates, as opposed to only convex ones.

Definition 3.1.1. *Let*

$$\operatorname{var} X(t) := \frac{\sum_{i=1}^{|V|}(x_i(t) - x_{av})^2}{|V|}.$$

Let T_{av} be the supremum over all $x \in \mathbb{R}^{|V|}$ of

$$\inf_t \mathbb{P}\left[\exists T > t, \frac{\operatorname{var} X(T)}{\operatorname{var} X(0)} > \frac{1}{e^2} \;\Big|\; X(0) = x\right] < \frac{1}{e}.$$

Notation 3.1.1. *Let a connected graph $G = (V, E)$ have a partition into connected graphs $G_1 = (V_1, E_1)$, and $G_2 = (V_2, E_2)$. Specifically, every vertex in V is either in V_1 or V_2, and every edge in E belongs to either E_1 or to E_2, or to the set of edges E_{12} that have one endpoint in V_1 and one in V_2. Let $|V_1| = n_1$, $|V_2| = n_2$ where without loss of generality, $n_1 \leq n_2$ and $|V| = n$. Let $T_{van}(G_1)$ and $T_{van}(G_2)$ be the averaging times of the "vanilla" algorithm that replaces at the clock tick of an edge e the values of the endpoints of e by the arithmetic mean of the two, applied to G_1 and G_2 respectively.*

Definition 3.1.2. *Let* $\mathbf{C}$ *denote the set of algorithms that use only convex updates of the form*

1. $x_i(t^+) = \alpha x_i(t^-) + \beta x_j(t^-)$.

2. $x_j(t^+) = \alpha x_j(t^-) + \beta x_i(t^-)$.

where $\alpha \in [0, 1]$ *and* $\alpha + \beta = 1$.

These updates have been extensively studied, see for example [8, 13].

Theorem 3.1.1. *The averaging time of any distributed algorithm in* $\mathbf{C}$ *is* $\Omega(\frac{\min(|V_1|,|V_2|)}{|E_{12}|})$

Theorem 3.1.2. *The averaging time of* $\mathcal{A}$ *is* $O(\log n(T_{van}(G_1) + T_{van}(G_2)))$.

Note that in the case where G_1 and G_2 are sufficiently well connected internally but poorly connected to each other, $\mathcal{A}$ outperforms any algorithm in $\mathbf{C}$. In fact for the graph G' obtained by joining two complete graphs G_1', G_2' each having $\frac{n}{2}$ vertices by a single edge, $\Omega(\frac{\min(|V_1'|,|V_2'|)}{|E_{12}'|}) = \Omega(n)$, while $O(\log n(T_{av}(G_1') + T_{av}(G_2'))) = O(\log n)$.

3.2 Limitations of convex combinations

Given a function $a(t)$, let its right limit at t be denoted by $a(t^+)$ and its left limit at t by $a(t^-)$. Consider an algorithm $\mathcal{C} \in \mathbf{C}$.

Proof of Theorem 3.1.1. Let us consider the initial condition where $X(0)$ is the vector that is 1 on vertices $v_1, \dots, v_{n_1}$ of G_1 and $-\frac{n_1}{n_2}$ on vertices $v_{n_1+1}, \dots, v_n$ of G_2. Let us denote $\frac{\sum_{i=1}^{n_1} x_i(t)}{n_1}$ by $y(t)$ and $\frac{\sum_{i=n_1+1}^{n} x_i(t)}{n_2}$ by $z(t)$. In the model we have considered, with probability 1, at no time does more than one clock tick.

In the course of the execution any algorithm in $\mathbf{C}$ $y(t)$ can change only during clock ticks of e_c and the same holds for $z(t)$. This is because during a clock tick of any other edge, both of whose end-vertices lie in G_1 or in G_2, $y(t)$ and $z(t)$ do not change. The vertices adjacent to e_c can change by at most 2 across these instants. Further, the values $x_n(t)$ and $x_{n+1}(t)$ are seen to lie in the interval

$$[\min_{i\in|V|} x_i(0), \max_{i\in|V|} x_i(0)] \subseteq [-1, 1].$$

If the clock of e_c ticks at time t, we therefore find that

$$|y(t^+) - y(t^-)| \leq \frac{2}{n_1}, \tag{3.1}$$

The number of clocks ticks of e_c until time t is a Poisson random variable whose mean is t.

A direct calculation tells us that

$$\operatorname{var}(X(t)) \quad \geq \quad \frac{n_1 y(t)^2}{n}. \tag{3.2}$$

To obtain a lower bound for $y(t)^2$, we note that the total number of times the clocks of edges belonging to E_{12} tick is a Poisson random variable ν_t with mean $t|E_{12}|$. It follows from Inequality (3.1) that $y(t) \geq 1 - \frac{2\nu_t}{n_1}$.

$$\begin{aligned}
|E_{12}|T_{av} &= \mathbb{E}[\nu_{T_{av}}] \\
&\geq \mathbb{P}\left[\nu_{T_{av}} \geq (1 - \frac{1}{e})\frac{n_1}{4}\right](1 - \frac{1}{e})\frac{n_1}{4}
\end{aligned}$$

However

$$\mathbb{P}\left[\nu_{T_{av}} \geq (1 - \frac{1}{e})\frac{n_1}{4}\right]$$

must be large, because otherwise $y(T_{av})$ would probably be large. More precisely,

$$\begin{aligned} \mathbb{P}\left[\nu_{T_{av}} \geq (1-\frac{1}{e})\frac{n_1}{4}\right] &\geq 1 - \mathbb{P}\left[\exists T > T_{av}, \mathrm{var}X(T) > \frac{1}{e^2}\right] \\ &> 1 - \frac{1}{e} \end{aligned}$$

Therefore,

$$\begin{aligned} T_{av} &\geq \mathbb{P}\left[\nu_{T_{av}} \geq (1-\frac{1}{e})\frac{n_1}{4|E_{12}|}\right](1-\frac{1}{e})\frac{n_1}{4} \\ &\geq \Omega(\frac{n_1}{|E_{12}|}) \end{aligned}$$

□

3.3 Using non-convex combinations

3.3.1 Algorithm $\mathcal{A}$

Let the vertices of G_1 be labeled by $[n_1]$ and those of G_2 by $[n]\backslash[n_1]$, where $[n] := \{1, \dots, n\}$. Let $e_c = (v_{n_1}, v_{n_1+1}$ be a fixed edge belonging to E_{12}. Let the time of the k^{th} clock tick of an edge e be t. Let $C >> 1$ be a sufficiently large absolute constant (independent of n.)

3.3.2 Analysis

Proof of Theorem 3.1.2. Since T_{av} is defined in terms of variance and algorithm $\mathcal{A}$ uses only linear updates, we may subtract out the mean from each $X_i(0)$ and it is sufficient to analyze the case when $x_{av} = 0$.

Let $V_1 = [n_1]$ and $V_2 = [n] \setminus [n_1]$. Let $\mu_1(t) = \frac{\sum_{i=1}^{n_1} x_i(t)}{n_1}$ and $\mu_2 = \frac{\sum_{i=n_1+1}^{n} x_i(t)}{n_2}$ and

- If the edge e is $e_c = (v_{n_1}, v_{n_1+1})$,
  1. If $k \equiv -1 \mod (\lceil C(T_{van}(G_1) + T_{van}(G_2)) \ln n \rceil)$
     (a) $x_{n_1}(t^+) = x_{n_1}(t^-) + n_1 \left\{x_{n_1+1}(t^-) - x_{n_1}(t^-)\right\}$
     (b) $x_{n_1+1}(t^+) = x_{n_1+1}(t^-) - n_1 \left\{x_{n_1+1}(t^-) - x_{n_1}(t^-)\right\}$
  2. If $k \not\equiv -1 \mod (\lceil C(T_{van}(G_1) + T_{van}(G_2)) \ln n \rceil)$ make no update.

- If the edge e is $(v_i, v_j) \notin E_{12}$
  1. $x_i(t^+) = \frac{x_i(t^-)+x_j(t^-)}{2}$.
  2. $x_j(t^+) = \frac{x_i(t^-)+x_j(t^-)}{2}$.

- If $e \in E_{12} \setminus \{e_c\}$ make no update.

$\mu(t) = |\mu_1(t)| + |\mu_2(t)|$. Let $\sigma(t)$ be

$$\sqrt{\frac{\sum_{i=1}^{n_1}(x_i(t) - \mu_1(t))^2 + \sum_{n_1+1}^{n}(x_i(t) - \mu_2(t))^2}{n}}.$$

We consider time instants $T_1, T_2, \ldots$ where T_i is the instant at which the clock of edge e ticks for the $\lceil iC(T_{van}(G_1) + T_{van}(G_2)) \ln n\rceil^{th}$ time. Observe that the value of $\mu(t)$ changes only across time instants $T_k, k = 1, 2, \ldots$.

The amount by which $x_{n_1}(t)$ and $x_{n_1+1}(t)$ deviate from $\mu_1(t)$ and $\mu_2(t)$ respectively, can be seen to be bounded above by $\sqrt{n}\sigma(t)$

$$\max\left\{|x_{n_1}(t) - \mu_1(t)|, |x_{n_1+1}(t) - \mu_2(t)|\right\} \tag{3.3}$$
$$\leq \sqrt{n}\sigma(t).$$

We now examine the evolution of $\sigma(T_k^+)$ and $\mu(T_k^+)$ as $k \to \infty$. The statements below are true if C is a sufficiently large universal constant (independent of n).

From T_k^+ to T_{k+1}^-, independent of x,

$$\mathbb{P}\left[\sigma(T_{k+1}^-) \geq \frac{\sigma(T_k^+)}{n^6} \,\middle|\, X(T_k^+) = x\right] \;\leq\; \frac{1}{4n} \tag{3.4}$$

$$\mu(T_{k+1}^-) \;=\; \mu(T_k^+) \tag{3.5}$$

Because of inequality (3.3), from T_k^+ to T_{k+1}^-

$$\sigma(T_{k+1}^+) \leq n(\sigma(T_{k+1}^-) + |\mu(T_{k+1}^-)|) \tag{3.6}$$

$$|\mu(T_{k+1}^+)| \leq n^{\frac{3}{2}}\sigma(T_{k+1}^-) \tag{3.7}$$

$$\text{var } X(t) \leq \mu(t)^2 + \sigma(t)^2.$$

We deduce from the above that

$$\mathbb{P}\left[\text{var } X(T_{k+1}^+) \geq \frac{\text{var } X(T_k^+)}{n^4}\right] \;\leq\; \frac{1}{4n} \tag{3.8}$$

Let A_k be the (random) operator obtained by composing the linear updates from time T_k^+ to T_{k+1}^+. Let $\|A\|$ denote the ℓ_2^n to ℓ_2^n norm of the linear map A.

$$\|A\| = \sup_{x \in \mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2}.$$

Lemma 3.3.1.

$$\mathbb{P}\left[\|A_k\|^2 \geq \frac{1}{n^3}\right] \;\leq\; \frac{1}{2} \tag{3.9}$$

To see this, let $v_1, \ldots, v_n$ be the canonical basis for $\mathbb{R}^n$. For any unit vector

$$x = \sum_{i=1}^{n} \lambda_i v_i$$

Then,

$$\begin{aligned} \|A_k(x)\| &\leq \sum_{i=1}^{n} |\lambda_i| \ \|A_k(v_i)\| \text{ (Triangle Ineq.)} \\ &\leq \sqrt{\sum_{i=1}^{n} \|A_k(v_i)\|^2} \text{ (Cauchy-Schwarz)} \end{aligned}$$

The Lemma now follows from Inequality (3.8) by an application of the Union Bound. □

Moreover, we observe by construction that the norm of A_k is less or equal to n,

$$\|A_k\| \leq n \tag{3.10}$$

Note that $\log(\operatorname{var} X(T_k^+))$ defines a random process (that is *not* Markov). The updates A_k from time T_k^+ to T_{k+1}^+ for successive k are i.i.d random operators acting on $\mathbb{R}^n$. Note that

$$\log(\operatorname{var} X(T_k^+)) - \log(\operatorname{var} X(0)) \leq \sum_{i=1}^{k} \log \|A_i\|$$

due to the presence of the supremum in the definition of operator norm.

$$W_k := \sum_{i=1}^{k} \log \|A_i\|$$

is a random walk on the real line for $k = 1, \ldots, \infty$.

The last and most important ingredient in this proof is *stochastic dominance.* It follows from Lemma 3.3.1 and Equation 3.10 that the random walk $\{W_k\}$ can be coupled with a random walk $\{\tilde{W}_k\}$ that is always to the right of it on the real line, i. e. for all k, $W_k \leq \tilde{W}_k$, where the increments

$$\begin{aligned}\tilde{W}_{k+1} - \tilde{W}_k &= \log n \ \ \text{(with probability } \frac{1}{2}\text{)} \\ &= -\frac{3}{2}\log n \ \ \text{(with probability } \frac{1}{2}.\text{)}\end{aligned}$$

By construction,

$$\log(\operatorname{var} X(T_k^+)) - \log(\operatorname{var} X(0)) \ \ \leq \ \ \tilde{W}_k \tag{3.11}$$

so it follows that T_{av} is upper bounded by any t_0 which satisfies

$$\mathbb{P}\Big[\ \forall T > t_0,\ \tilde{W}_T \leq -2\Big] > 1 - \frac{1}{e}.$$

Note that $\mathbb{E}[\tilde{W}_k] = -\frac{k \log n}{2}$ and $\mathbb{E}[\mathrm{var}\tilde{W}_k] = \frac{9k}{16}\log^2 n$.

In order to proceed, we shall need the following inequality (Proposition 2.1.2, [52]) about simple unbiased random walk $\{S_k\}_{k\geq 0}$ on $\mathbb{Z}$ starting at 0.

Proposition 3.3.1 (Proposition 2.1.2, [52])**.** *There exist constants c, β such that for any $n \in \mathbb{Z}$, $s > 0$*

$$\mathbb{P}[S_n \geq s\sqrt{n}] \leq ce^{-\beta s^2}.$$

Using this fact,

$$\begin{aligned}&\mathbb{P}[\forall T > t_0,\ \tilde{W}_T \leq -2] \\ = &\,\mathbb{P}[\forall T > t_0, (\log n)(S_T - \frac{T}{2}) \leq -2]\end{aligned}$$

For large n, this is the same as

$$\mathbb{P}[\forall T > t_0, S_T < \frac{T}{2}] \geq 1 - \sum_{T > t_0} ce^{-\beta T/4}.$$

Clearly, there is a constant t_0 independent of n such that $1 - \sum_{T>t_0} ce^{-\beta T/4} > 1 - \frac{1}{e}$. This completes the proof. □

CHAPTER 4

MIXING TIMES AND BOUNDS ON THE COMPETITIVE RATIO FOR OBLIVIOUS ROUTING

4.1 Introduction

Over the past three decades, there has been significant interest in the design and analysis of routing schemes in networks of various kinds. A network is typically modeled as a directed or undirected graph $G = (V, E)$, where E is a set of m edges representing links and V is a set of n vertices representing locations or servers. Each link is associated with a cost which is a function of the load that it carries. There is a set of demands, which has the form

$$\{(i,\, j,\, d_{ij})\big|(i,\, j) \in V \times V,\, d_{ij} \geq 0\}.$$

A routing scheme that routes a commodity from its source to its target independent of the demands at other source-target pairs is termed oblivious. The competitive ratio of an oblivious routing scheme Obl is the maximum taken over all demands, of cost that Obl incurs divided by the cost that the optimal adaptive scheme Opt incurs. Work in this area was initiated by Valiant and Brebner [98] who developed an oblivious routing protocol for parallel routing in the hypercube that routes any permutation in time that is only a logarithmic factor away from optimal. For the cost-measure of congestion (the maximum load of a network link) in a virtual circuit routing model, Räcke [81] proved the existence of an oblivious routing scheme with polylogarithmic competitive ratio for any undirected network. This result was subsequently made constructive by Harrelson, Hildrum and Rao [31] and improved to a competitive ratio of $O(\log^2 n \log\log n)$.

Oblivious routing has largely been studied in the context of minimizing the maximum congestion. A series of papers [81, 31, 82] has culminated recently in the development of an oblivious algorithm due to Räecke [82] whose competitive ratio with respect to congestion is $O(\log n)$. The algorithm Har that is studied in this chapter was introduced in [32] where it was shown to have a competitive ratio of $O(\sqrt{\log n})$ with respect to the ℓ_2 norm when demands route to a single common target. We study the task of uniformly minimizing all the ℓ_p norms of the vector of edge loads in an undirected graph while routing a multicommodity flow oblivious of the set of demands. As a matter of fact our results hold under any norm that transforms $\mathbb{R}^n$ into a Banach space symmetric and "unconditional" with respect to the canonical basis. These terms have been defined in section 4.3.

4.2 Our results

Let $G = (V, E)$ denote an undirected graph with a set V of n vertices and a set E of m edges. For any oblivious algorithm $\mathcal{A}$, let the competitive ratio of $\mathcal{A}$ in the norm $\|\cdot\|_p$ be denoted $\kappa_p(\mathcal{A})$. Let the performance index $\pi(\mathcal{A})$ of $\mathcal{A}$ be defined to be their supremum as $\|\cdot\|_p$ ranges over the set of all ℓ_p norms.

$$\pi(\mathcal{A}) := \sup_p \kappa_p(\mathcal{A}).$$

Let Har be the oblivious algorithm (formally defined in the following section,) which routes a flow from s to t in the (unique) way that minimizes the ℓ_2 norm of edge loads of that flow, assuming all other demands to be 0. The competitive ratio of this algorithm with respect to the ℓ_2 norm was shown in [32] to be $O(\sqrt{\log n})$ over demands having single common target. We show that Har has an index $\pi(\mathsf{Har})$ that is equal to its competitive ratio in the ℓ_1 norm, which is in turn bounded above by $\min(\sqrt{m}, O(T_{mix}))$ where T_{mix} is the mixing time of the canonical random walk, We obtain $O(\log n)$ upper bounds on $\pi(\mathsf{Har})$ for expanders. The constant in $O(\cdot)$ may depend on the family. Almost matching $\Omega(\frac{\log n}{\log \log n})$ lower bounds for expanders [30] and matching $\Omega(\log n)$ lower bounds for $2-$dimensional discrete tori [3] are known for the competitive ratio of an oblivious algorithm with respect to congestion or ℓ_∞ norm. In particular, for cost functions that are convex combinations of bounded powers of the various ℓ_p norms, such as $\sum_e g(\text{load}(e))$ where g is a polynomial with non-negative coefficients, Har has on these graphs a polylogarithmic competitive ratio. We show that there exist graphs for which no algorithm that is adaptive with respect to the demands but not p can simultaneously have a cost that is less than $\Omega(\sqrt{m})$ times the ℓ_p norm of the optimal adaptive algorithm that is permitted to vary with p, even if p can only take the values 1 and ∞. Lastly, we can handle a larger class of norms than ℓ_p, namely those norms that are invariant with respect to all permutations of the canonical basis vectors and reflections about any of them.

Theorem 4.2.1. *For any graph G, with n vertices, and m edges, on which the canonical random walk has a mixing time T_{mix}, $\pi(\mathsf{Har}) \leq \min(\sqrt{m}, O(T_{mix}))$.*

Hajiaghayi et al have shown in [30] that if G belongs to a family of expanders and $\mathcal{A}$ is any oblivious routing algorithm, the competitive ratio $\kappa_\infty(\mathcal{A})$ with respect to congestion is bounded from below by $\Omega(\frac{\log n}{\log\log n})$. Therefore Theorem 4.2.1 is tight up to an $O(\log\log n)$ factor for expanders.

Theorem 4.2.2. *For every m, there exists a graph G with m edges, such that for any oblivious algorithm $\mathcal{A}$, $\pi(\mathcal{A}) \geq \frac{\lfloor\sqrt{m-1}\rfloor}{2}$ on G.*

4.3 Definitions and Preliminaries

A network will be an undirected graph $G = (V, E)$, where V denotes a set of n vertices (or nodes) $\{1, \ldots, n\}$ and E a set of m edges. If a traffic vector $t = (t_1, \ldots, t_m)$ is transported across edges $e_1, \ldots, e_m$, we shall consider costs that are ℓ_p norms $\|t\|_p$. In our setting, the network is undirected and links are allowed to carry traffic in both directions simultaneously. For book keeping, it will be convenient to give each edge an orientation. For an edge e of the form $\{v, w\}$, we will write $e = (v, w)$ when we want to emphasize that the edge is oriented from v to w. The traffic on edge e will be a real number. If this number is positive, it will represent traffic along e from v to w; if it is negative, it will represent traffic along e from w to v. Let $\mathrm{In}(v)$ be the edges of G that are oriented into v and $\mathrm{Out}(v)$ be the edges of G that are oriented away from v. A potential ϕ on G is a function from V to $\mathbb{R}$. The gradient $\nabla\phi$ of a potential ϕ is a function from E to $\mathbb{R}$, whose value on an oriented edge $e := (u, v)$ is

$$\nabla\phi(e) := \phi(u) - \phi(v).$$

A flow f on G is a function from E to $\mathbb{R}$. The divergence $\operatorname{div} f$ of a flow f is a function from V to $\mathbb{R}$ whose value on a vertex v is given by

$$(\operatorname{div} f)(v) := \sum_{e \in \mathrm{Out}(v)} f(e) - \sum_{e \in \mathrm{In}(v)} f(e)$$

We shall denote by Δ the Laplacian operator that maps the space of real valued functions on V to itself as follows.

$$\Delta\phi := -\operatorname{div}(\nabla\phi).$$

We call such $f = \langle f_{ij} : i, j \in V(G)\rangle$ a multi-flow. We say that a multi-flow f meets the demand $\langle d_{ij} : i, j \in V\rangle$, if for all $i, j \in V$,

$$\operatorname{div} f_{ij} = d_{ij}\delta_i - d_{ij}\delta_j,$$

where $\delta_u(\cdot)$ is the Kronecker Delta function that takes a value 1 on u and 0 on all other vertices. If this is the case, we say "f routes D" and write $f \searrow D$. For a fixed i, j, we shall use $\|f_{ij}\|_1$ to denote $\sum_e |f_{ij}(e)|$. The traffic on the edge e under f is given by

$$t_f(e) = \sum_{i,j} |f_{ij}(e)|.$$

We shall call the vector $t_f := (t_f(e_1), \dots, t_f(e_m))$ the network traffic or network load, where $(e_1, \dots, e_m)$ is a list of the edges of G. If for every edge e, the total traffic on e under f is greater or equal to the total traffic on e under f', we shall say that $f' \lhd f$. i. e.

$$(\forall e) t_f(e) \geq t_{f'}(e) \Rightarrow f' \lhd f.$$

Definition 4.3.1. *An oblivious algorithm ($\mathcal{A}$), is a multi-flow $\{a_{ij}\}$ indexed by pairs of vertices i, j where each a_{ij} is a flow satisfying*

$$\operatorname{div} a_{ij} = \delta_i - \delta_j.$$

Given a demand D, $\mathcal{A}$ routes D using $D \cdot a := \langle d_{ij} a_{ij} : i, j \in V(G) \rangle$.

Definition 4.3.2. *For any oblivious algorithm $\mathcal{A}$, for every $p \in [1, \infty]$, we define its competitive ratio under the ℓ_p norm $\|\cdot\|_p$*

$$\kappa_p(\mathcal{A}) := \sup_D \sup_{f \searrow D} \frac{\|t_{D \cdot a}\|_p}{\|t_f\|_p}.$$

Let the performance index $\pi(\mathcal{A})$ of $\mathcal{A}$ be defined to be their supremum as $\|\cdot\|_p$ ranges over all possible ℓ_p norms,

$$\pi(\mathcal{A}) := \sup_{p \in [1, \infty]} \kappa_p(\mathcal{A}).$$

All results in this chapter hold without modification if the above definition of performance index, is altered to be the supremum over all norms that satisfy the symmetry and unconditionality conditions in Definition 4.3.6.

Definition 4.3.3. *We define* Har *to be the oblivious algorithm corresponding to the multi-flow* $h = \langle h_{ij} : i, j \in V(G)\rangle$, *where* h_{ij} *is the unique flow such that*

1. $\operatorname{div} h_{ij} = \delta_i - \delta_j$, *and*

2. *There exists a potential* ϕ_{ij} *such that* $\nabla \phi_{ij} = h_{ij}$.

These conditions uniquely determine $\{h_{ij}\}$ and determine the potential ϕ_{ij} up to an additive constant. The potential can be described in terms of random walk on the graph. Suppose $W_0, W_1, \ldots$ denotes simple random walk on the graph, and let $\tilde{\pi}(v) = \deg(v)/[2|E|)]$ denotes its stationary distribution.

Definition 4.3.4 (Hitting time)**.** *If* $S \subseteq V$, *let*

$$H_S = \min\{j \geq 1 : W_j \in S\}, \qquad \overline{H}_S = \min\{j \geq 0 : W_j \in S\}.$$

If $S = \{i, j\}$, *we write* $H_{ij}, \overline{H}_{ij}$.

Note that $H_S, \overline{H}_S$ agree if $W_0 \notin S$. The potential ϕ_{ij} with boundary condition $\phi_{ij}(j) = 0$ is given by

$$\phi_{ij}(v) = b_{ij}\, \mathbb{P}^v\{W(\overline{H}_{ij}) = i\},$$

where we write $\mathbb{P}^v$ to denote probabilities assuming $W_0 = v$ and the constant b_{ij} is given by

$$b_{ij}^{-1} = \mathbb{P}^j\{W(H_{ij}) = i\}.$$

Definition 4.3.5 (Mixing time)**.** *Let* $W_0, W_1, \ldots$ *be simple random walk on* G. *Let*

$$\rho_v^{(t)}(u) = \mathbb{P}^v\{W_t = u\}.$$

The mixing time as a function of ϵ is

$$T_{mix}(\epsilon) := \sup_{v \in V} \inf \left\{ t : \|\tilde{\pi} - \rho_v^{(t)}\|_1 \leq 2\epsilon \right\}.$$

Definition 4.3.6. *A Banach space X with a basis $\{e_1, \ldots, e_m\}$ is said to be symmetric and unconditional with respect to the basis if the following two conditions hold for any $x_1, \ldots, x_m \in \mathbb{R}$.*

S. For any permutation π, $\| \sum_{i=1}^m x_i e_i \|_X = \| \sum_{i=1}^m x_i e_{\pi(i)} \|_X$.

U. For any $\epsilon_1, \ldots, \epsilon_m \in \{-1, 1\}$, $\| \sum x_i e_i \|_X = \| \sum \epsilon_i x_i e_i \|_X$.

4.3.1 Interpolation Theorems

All ℓ_p norms satisfy the above conditions. Given a linear operator $A : \mathbb{R}^m \mapsto \mathbb{R}^m$, we shall define its $\ell_p \to \ell_p$ norm

$$\|A\|_{p \to p} = \sup_{\|x\|_p = 1} \frac{\|Ax\|_p}{\|x\|_p}.$$

More generally if $\mathbb{R}^m$ is endowed with a norm transforming it into a Banach space X, we shall denote its operator norm by $\|A\|_{X \to X}$. We will need the following special cases of the theorems of Riesz-Thorin [84, 94] and Mityagin [66].

Theorem 4.3.1 (Riesz-Thorin)**.** *For any $1 \leq p \leq r \leq q \leq \infty$,*

$$\|A\|_{r \to r} \leq \max(\|A\|_{p \to p}, \|A\|_{q \to q}).$$

The following theorem is due to B. Mityagin.

Theorem 4.3.2 (Mityagin)**.** *Let $\mathbb{R}^m$ be endowed with a norm transforming it into a Banach* *space X that is symmetric and unconditional with respect to the standard basis. Then,*

$$\|A\|_{X\to X} \leq \max(\|A\|_{1\to 1}, \|A\|_{\infty\to\infty}).$$

4.3.2 Some facts about harmonic functions and flows

$h_{ij} = \nabla\phi_{ij}$, where ϕ_{ij} is up to addition by a constant, the unique solution of the linear equation

$$\Delta\phi_{ij} = -\delta_i + \delta_j.$$

Therefore for every u, v and w, the following linear relation is true.

Fact 4.3.1. $h_{uv} + h_{vw} = h_{uw}$.

For $e = (u, v)$, let $h_e := h_{uv}$ and $\phi_e := \phi_{uv}$. More generally, we have the following.

Lemma 4.3.1. *Let g_{ij} be a flow such that* $\operatorname{div} g_{ij} = \delta_i - \delta_j$. *Then,*

$$\sum_e g_{ij}(e)h_e = h_{ij}.$$

Proof. By linearity,

$$\begin{aligned}\operatorname{div}\sum_e g_{ij}(e)h_e &= \sum_e g_{ij}(e)\operatorname{div} h_e\\ &= \sum_{e=(u,v)\in E} g_{ij}(e)(\delta_u - \delta_v),\end{aligned}$$

which is

$$\sum_{v\in V}\left(\sum_{e\in \mathrm{Out}(v)} g_{ij}(e) - \sum_{e\in \mathrm{In}(v)} g_{ij}(e)\right)\delta_v = \delta_i - \delta_j.$$

Secondly,

$$\nabla(\sum_e g_{ij}(e)\phi_e) = \sum_e g_{ij}(e)\nabla\phi_e = \sum_e g_{ij}(e)h_e.$$

According to the definition, h_{ij} is the unique flow that satisfies the above properties, so we are done. □

The following is a result from network theory [93].

Theorem 4.3.3 (Reciprocity Theorem)**.** *The flows comprising* Har *have the following symmetry property. For each* $i, j \in V$*, let* ϕ_{ij} *be a potential such that* $\nabla \phi_{ij} = h_{ij}$*. Then, for any* $u, v \in V$*,*

$$\phi_{ij}(u) - \phi_{ij}(v) = \phi_{uv}(i) - \phi_{uv}(j). \tag{4.1}$$

4.4 Using Interpolation to derive uniform bounds

Proof of Theorem 4.2.1. This Theorem follows from Proposition 4.4.3, Proposition 4.4.1 and Proposition 4.4.2.

Proposition 4.4.1. *For any graph G,*

$$\begin{aligned} \pi(\mathsf{Har}) &= \kappa_1(\mathsf{Har}) \\ &= \max_{e \in G} \|h_e\|_1, \end{aligned}$$

where the maximum is taken over all edges of G.

Proof of Proposition 4.4.1. Given a demand D, let D_p be constructed as follows. Let $\mathsf{Opt}\ _p(G, D)$ be an optimal multi-flow routing D with respect to ℓ_p, and $\mathrm{opt}_p(G, D)$ be the corresponding ℓ_p norm. For an edge $e = (u, w)$, let $(D_p)_{uw}$ be the total amount of traffic from u to v in $\mathsf{Opt}\ _p(G, D)$ that routes D. For any pair of vertices (u, w) that are not adjacent, let $(D_p)_{uw}$

be defined to be 0. Let $\|D\|_p$ be defined in the natural way to be

$$\|D\|_p := \left(\sum_{ij} d_{ij}^p \right)^{\frac{1}{p}} .$$

Lemma 4.4.1. $\mathrm{opt}_p(G, D) = \mathrm{opt}_p(G, D_p) = \|D_p\|_p$.

Proof of Lemma 4.4.1. Any multi-flow f that routes D_p can be converted to a multi-flow that routes D having the same total cost, since D_p was constructed from a multi-flow that routes D. Therefore $\mathrm{opt}_p(G, D) \leq \mathrm{opt}_p(G, D_p)$. By the definition of D_p, there exists an optimal solution to D that can be used to route D_p. This establishes that $\|D_p\|_p = \mathrm{opt}_p(G, D) \geq \mathrm{opt}_p(G, D_p)$, and proves the lemma. □

Let $\bar{\mathcal{D}}_p$ represent the set of all demands of the form D_p arising from some demand D by the above conversion procedure. By Lemma 4.4.1,

$$\forall_{D_p \in \bar{\mathcal{D}}_p} \sup_{f \searrow D} \frac{1}{\|t_f\|_p} = \frac{1}{\|D_p\|_p}.$$

Using $D \cdot h$ to denote the multi-flow $\langle d_{ij} h_{ij} : i, j \in V(G) \rangle$, it is sufficient to prove that for all $p \in [1, \infty]$.

$$\sup_{D_p \in \bar{\mathcal{D}}_p} \frac{\|t_{D_p \cdot h}\|_p}{\|D_p\|_p} \leq \sup_{D_1 \in \bar{\mathcal{D}}_1} \frac{\|t_{D_1 \cdot h}\|_1}{\|D_1\|_1}. \tag{4.2}$$

Let $e_1, \ldots, e_m$ be some enumeration of the edges of G and let R be the $m \times m$ matrix whose ij^{th} entry, where $e_i = (u, v)$ is given by $r_{ij} = |h_{uv}(e_j)|$. For $D_p \in \bar{\mathcal{D}}_p$, the "traffic vector" $t_{D_p \cdot h}$ can be obtained by applying the linear transformation R to d_p, for each $e_i = (u, v)$, $(d_p)_i = (D_p)_{uv}$ and $d_p = ((d_p)_1, \ldots, (d_p)_m)$.

$$\begin{pmatrix} t_{D_p \cdot h}(e_1) \\ t_{D_p \cdot h}(e_2) \\ \vdots \\ t_{D_p \cdot h}(e_m) \end{pmatrix} = \begin{pmatrix} r_{11} & \dots & r_{1m} \\ r_{21} & \dots & r_{2m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \dots & r_{mm} \end{pmatrix} \begin{pmatrix} (d_p)_1 \\ (d_p)_2 \\ \vdots \\ (d_p)_m \end{pmatrix}.$$

Given a linear operator $A : \mathbb{R}^m \mapsto \mathbb{R}^m$, we define its $\ell_p \to \ell_p$ norm

$$\|A\|_{p \to p} = \sup_{\|x\|_p = 1} \frac{\|Ax\|_p}{\|x\|_p}.$$

With this notation, for $p \in (1, \infty]$,

$$\sup_{D_p \in \bar{\mathcal{D}}_p} \frac{\|t_{D_p \cdot h}\|_p}{\|D_p\|_p} \quad \leq \quad \|R\|_{p \to p} \tag{4.3}$$

while for $p = 1$, because $\bar{\mathcal{D}}_1$ is equal to $\{\mathbb{R}^+\}^m$, we can make the stronger assertion that

$$\sup_{D_1 \in \bar{\mathcal{D}}_1} \frac{\|t_{D_1 \cdot h}\|_1}{\|D_1\|_1} \quad = \quad \|R\|_{1 \to 1}. \tag{4.4}$$

Lemma 4.4.2. $\|R\|_{1 \to 1} = \|R\|_{\infty \to \infty}$.

Proof of Lemma 4.4.2. Recall that R is an $m \times m$ matrix whose ij^{th} entry is $|h_{e_i}(e_j)|$. By the Reciprocity Theorem (Theorem 4.3.3), R is a symmetric matrix. Let $x_1 \in \mathbb{R}^m$ be a unit ℓ_1 normed vector that achieves the maximum dilation in the ℓ_1 norm when multiplied by R, i. e. $\|x_1\|_1 = 1$ and $\|Rx_1\|_1 = \|R\|_{1 \to 1}$. We may assume without loss of generality that all coordinates of x_1 are non-negative, because if we replace x_1 by the vector x_1' obtained by taking absolute values of the coordinates of x_1, $\|Rx_1'\| \geq \|Rx_1\|$. Let $u_1, \dots, u_m$ be the standard basis. We note that

$$\begin{aligned}
\|R\|_{1\to 1} &= \sum_{i=1}^{m} \|h_{e_i}\|_1 x_{1i} \\
&\leq \max_i \|h_{e_i}\|_1 \\
&= \max_i \|Ru_i\|_1 \\
&\leq \|R\|_{1\to 1}.
\end{aligned}$$

Therefore $\|R\|_{1\to 1} = \max_i \|h_{e_i}\|_1$. Since R is a symmetric matrix whose entries are non-negative,

$$\sup_{\|x\|_\infty = 1} \|Rx\|_\infty = \|R(u_1 + \ldots + u_m)\|_\infty.$$

Therefore,

$$\begin{aligned}
\|R\|_{\infty\to\infty} &= \|R(u_1 + \ldots + u_m)\|_\infty \\
&= \max_i \|h_{e_i}\|_1 \\
&= \|R\|_{1\to 1}.
\end{aligned}$$

□

By Lemma 4.4.2 and the Riesz-Thorin theorem (Theorem 4.3.1), we conclude that

$$\begin{aligned}
\sup_{D_p \in \bar{\mathcal{D}}_p} \frac{\|t_{D_p \cdot h}\|_p}{\|D_p\|_p} &\leq \|R\|_{p\to p} \\
&\leq \max(\|R\|_{1\to 1}, \|R\|_{\infty\to\infty}) \\
&= \|R\|_{1\to 1} (\text{By Lemma 4.4.2}) \\
&= \sup_{D_1 \in \bar{\mathcal{D}}_1} \frac{\|t_{D_1 \cdot h}\|_1}{\|D_1\|_1}
\end{aligned}$$

which establishes Inequality 4.2 and thereby completes the proof. □

Remark 4.4.1. *One can repeat the above argument using Mityagin's theorem (Theorem 4.3.2) instead of the Riesz-Thorin theorem, and prove the stronger statement that for any norm that transforms $\mathbb{R}^m$ into a symmetric unconditional Banach space X with respect to the standard basis,*

$$\kappa_X(\mathsf{Har}) \leq \kappa_1(\mathsf{Har}),$$

where $\kappa_X(\mathsf{Har})$ is the competitive ratio of Har with respect to the norm of X.

4.4.1 Bounding $\pi(\mathsf{Har})$ by hitting and mixing times

Proposition 4.4.2. *For any vertices $i, j \in V$,*

$$\|h_{ij}\|_1 \leq 8T_{mix}(\frac{1}{4}).$$

Proof of Proposition 4.4.2. It is always possible to find a potential ϕ_{ij} such that $\Delta\phi_{ij} = \delta_j - \delta_i$, and

$$\begin{aligned} \tilde{\pi}(\{u|\phi_{ij}(u) \leq 0\}) &\geq \frac{1}{2} \\ \tilde{\pi}(\{v|\phi_{ij}(v) \geq 0\}) &\geq \frac{1}{2}, \end{aligned}$$

because adding an arbitrary constant does not change the gradient of a potential. Let ϕ_{ij} satisfy the above conditions.

Recall that H_S be the (hitting) time taken for a random walk starting at v to hit the set S (Definition 4.3.4). Let $\mathbb{E}^v[\cdot]$ denote expectations assuming $W_0 = v$.

Lemma 4.4.3. *Suppose $\Delta\phi_{ij} = \delta_j - \delta_i$. Then,*

$$\|\nabla\phi_{ij}\|_1 \leq \sum_v \deg(v)|\phi_{ij}(v)|$$

Proof of Lemma 4.4.3.

$$\begin{aligned} \|\nabla\phi_{ij}\|_1 &= \sum_{(u,v)\in E} |\phi_{ij}(u) - \phi_{ij}(v)| \\ &\le \sum_{(u,v)\in E} (|\phi_{ij}(u)| + |\phi_{ij}(v)|) \\ &= \sum_{v} \deg(v)|\phi_{ij}(v)| \end{aligned}$$

□

Lemma 4.4.4. *Let* $\Delta\phi_{ij} = \delta_j - \delta_i$*,* $S_{\le} := \{v|\phi_{ij}(v) \le 0\}$ *and* $S_{\ge} := \{v|\phi_{ij}(v) \ge 0\}$*. Then,*

$$\sum_{v\in V} \deg(v)|\phi_{ij}(v)| \le \mathbb{E}H^i_{S_\le} + \mathbb{E}H^j_{S_\ge}.$$

Proof of Lemma 4.4.4. Let $W_0, W_1, \ldots$ be a random walk on G starting at i and ending the first time it hits $S_\le$. Given $v \in V$ and a subset S of V, let $N^i_S(v)$ be the number of times the walk *exits* v until hitting $S_\le$, and $\psi(v) := \frac{\mathbb{E}N^i_{S_\le}(v)}{\deg(v)}$. Note that $\psi(u) = 0$ for all $u \in S_\le$. We make the following claim.

Claim 4.4.1.

$$\begin{aligned} \Delta\psi(i) &= -1 \\ \Delta\psi(u) &= 0 \quad \textit{if} \;\; u \in V \setminus \{S_\le \cup \{i\}\}. \end{aligned}$$

Proof of Claim 4.4.1. To see why this is true, let $E(t, v)$ be the event that $W_t = v$. For any vertex u, let $\star(u)$ denote the set of vertices adjacent to u. We see that for $1 \le t \le H^t_{S_\le}$ and $u \in V \setminus \{S_\le \cup \{i\}\}$,

$$\mathbb{P}[E(t, v)] = \sum_{u\in\star(v)} \frac{\mathbb{P}[E(t-1, u)]}{\deg(u)}.$$

Summing up over time, this implies $\mathbb{E}[N^i_{S_\leq}(v)] = \sum_{u \in \star(v)} \frac{\mathbb{E}[N^i_{S_\leq}(u)]}{\deg(u)}$. This translates to $\psi(v) = \sum_{u \in \star(v)} \frac{\psi(u)}{\deg(v)}$. When $v = i$, a similar computation yields

$$\psi(i) = 1 + \sum_{u \in \star(i)} \frac{\psi(u)}{\deg(i)},$$

proving the claim. □

It follows that $\Delta(\psi - \phi)$ is 0 on *all* of $V \setminus S_\leq$. This implies that the maximum of $\phi - \psi$ cannot be achieved on $V \setminus S_\leq$ (Maximum principle for Harmonic functions). $\phi - \psi$ is ≤ 0 on $S_\leq$, therefore $\psi - \phi$ is a non-negative function. It follows that

$$\sum_{v \in S_\geq} \deg(v)\phi_{ij}(v) \leq \mathbb{E}^i \left[H_{S_\leq} \right].$$

An identical argument applied to $-\phi_{ij}$ instead of ϕ_{ij} gives us the following.

$$\sum_{v \in S_\leq} \deg(v)(-\phi_{ij}(v)) \leq \mathbb{E}^j \left[H_{S_\geq} \right].$$

Together, the last two inequalities complete the proof. □

Lemma 4.4.5. *Let S be a subset of V whose stationary measure is greater or equal to $\frac{1}{2}$. Let $v \in V \setminus S$. Then*

$$\mathbb{E}^v[H_S] \leq 4T_{mix}(1/4).$$

Proof of Lemma 4.4.5. Let $W_0, W_1, \ldots$ be a random walk on G starting at v. Recall that from Definition 4.3.5,

$$T_{mix}(\epsilon) = \sup_{v \in V} \inf \left\{ T \middle| \forall_{t \geq T} \|\tilde{\pi} - \rho_v^{(t)}\|_1 < \epsilon \right\}.$$

Let $\tau := T_{mix}(1/4)$ and $\mathbb{P} = \mathbb{P}^v$.

$$\begin{aligned}
\mathbb{P}[H_S \le \tau] &\ge \mathbb{P}[W_\tau \in S] \\
&= \tilde{\pi}(S) - (\tilde{\pi}(S) - \sum_{u\in S} \mathbb{P}[W_\tau = u]) \\
&\ge \frac{1}{2} - |\tilde{\pi}(S) - \sum_{u\in S} \mathbb{P}[W_\tau = u]| \\
&\ge \frac{1}{2} - |\sum_{u\in S} (\tilde{\pi}(u) - \mathbb{P}[W_\tau = u])| \ge \frac{1}{4}.
\end{aligned}$$

In order to get a bound on the expected hitting time from this bound on the hitting probability, we observe that the distribution of hitting times has an exponential tail. More precisely, using the Markovian property of the random walk,

$\mathbb{P}[H_S > k\tau]$ is less or equal to

$$\mathbb{P}\left[X_\tau \notin S\right] \prod_{i=1}^{k-1} \sup_{u \notin S} \mathbb{P}\left[X_{(i+1)\tau} \notin S \middle| X_{i\tau} = u\right] \le \frac{3^k}{4^k}.$$

Finally,

$$\mathbb{E}^v[H_S] \le \tau \left(\sum_{i=0}^{\infty} \mathbb{P}\left[H_S^v > i\tau\right] \right) \le 4\tau.$$

□

□

Proposition 4.4.3. *For any edge e, $\|h_e\|_1 \le \sqrt{m}$.*

Proof of Proposition 4.4.3. If $e = (u, v)$, h_e is by Thompson's principle ([23]) the minimizer of $\|f\|_2$ among all flows f for which $\operatorname{div} f = \delta_u - \delta_v$. Since u and v are adjacent, $\|f\|_2 = 1$ if f is the "shortest-path" flow defined by

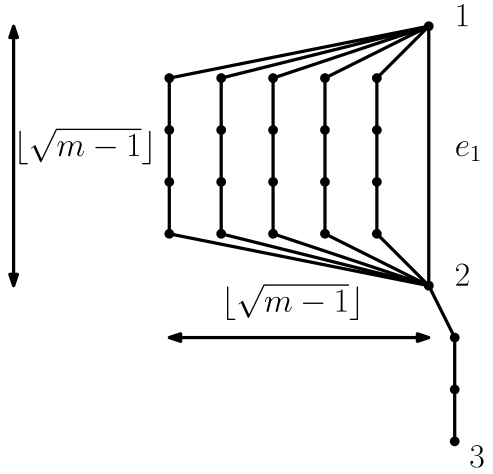

Figure 4.1: A graph on which the performance index $\pi(\mathcal{A})$ of any oblivious algorithm $\mathcal{A}$ is $\geq \frac{\lfloor\sqrt{m-1}\rfloor}{2}$

$$f(e') = \begin{cases} 1 & \text{if } e' = e \\ 0 & \text{otherwise.} \end{cases}$$

Therefore $\|h_e\|_2 \leq 1$. This implies that $\|h_e\|_1 \leq \sqrt{m}$, since for any vector $x \in \mathbb{R}^m$, $\|x\|_1 \leq \sqrt{m}\|x\|_2$. □

□

Proof of Theorem 4.2.2. Let G be a graph with $n = m - \lfloor\sqrt{m-1}\rfloor + 1$ vertices and m edges constructed as follows (see Figure 1). Let vertices labeled 1 and 2 be joined by an edge e_1, and also be connected by $r := \lfloor\sqrt{m-1}\rfloor$ vertex disjoint paths of length h. We make the remaining vertices and edges belong to a path from vertex 3 to 2, such that there is no path from any of these vertices to 1 which does not contain 2. We will fix a specific set of demands $D = \{d_{ij}\}$; namely a unit demand from 1 to 2, and 0 otherwise. Suppose that an algorithm $\mathcal{A}$ uses a flow denoted a_{12} to achieve this, where $a_{12}(e_1) = \alpha \in [0, 1]$. Then,

$$\|a_{ij}\|_1 \geq \alpha + (1 - \alpha)r,$$

and

$$\|a_{ij}\|_\infty \geq \alpha.$$

On the other hand a flow $\mathsf{Opt}_{\ 1}$ that uses only e_1 incurs an ℓ_1 norm of 1. A flow $\mathsf{Opt}_{\ \infty}$ that uses all the $r+1$ edge disjoint paths from 1 to 2 equally incurs an ℓ_∞ cost of $\frac{1}{r+1}$.

$$\frac{\|a_{12}\|_1}{\|\mathsf{Opt}_{\ 1}\|_1} \geq (1-\alpha)r + \alpha,$$

and

$$\frac{\|a_{12}\|_\infty}{\|\mathsf{Opt}_{\ \infty}\|_\infty} \geq \alpha(r+1).$$

Therefore,

$$\max(\kappa_1(\mathcal{A}), \kappa_\infty(\mathcal{A})) \geq \frac{\kappa_1(\mathcal{A}) + \kappa_\infty(\mathcal{A})}{2} \geq \frac{\lfloor\sqrt{m-1}\rfloor}{2}.$$

□

4.5 Random walks and $\pi(\textsf{Har})$

4.5.1 A Bound on the Spectral Gap using flows of Har

A well known method due to Diaconis-Stroock [21] and Sinclair [87], of bounding the spectral gap of a reversible Markov chain involves the construction of canonical paths from each node to every other. One may use flows instead of paths, and derive better mixing bounds, as was the case in the work of Morris-Sinclair [67] on sampling knapsack solutions. Sinclair suggested a natural way of constructing canonical flows using random walks in [87]. This scheme gives a bound of $O(\tau^2)$, if τ is the true mixing time. In this section, we observe that for a random walk on a graph, the flows of Har provide a certificate for rapid mixing as well, and give the same $O(\tau^2)$ bound on the mixing time. Let the stationary distribution be denoted $\tilde{\pi}$. $\tilde{\pi(i)} = \frac{\deg(i)}{2m}$. Let the *capacity* of an edge $e = (u, v)$, denoted $Q(e)$ be defined to be $\tilde{\pi}(u)p_{uv}$, where $p_{uv} = \frac{1}{\deg(u)}$ is the transition probability from u to v. Let the transition matrix be denoted P. In our setting of (unweighted) random walks, $\tilde{\pi(i)} = \frac{\deg(i)}{2m}$ and for each edge e, $Q(e)$ is $\frac{1}{2m}$.

Definition 4.5.1. *Let D be the demand $\langle d_{ij} : i, j \in V\rangle$, where $d_{ij} = \tilde{\pi}(i)\tilde{\pi}(j)$. Given a multi-flow f, where $f \searrow D$, let $\rho(f)$ denote the maximum load on any edge divided by its capacity i. e. $\rho(f) = \frac{\|t_f\|_\infty}{Q(e)}$.*

Theorem 4.5.1 (Sinclair)**.** *Let $f \searrow D$ as described above. Then, the second eigenvalue λ_1 of the transition matrix P satisfies*

$$\lambda_1 \leq 1 - \frac{1}{8\rho^2}.$$

Let us denote $T_{mix}(1/4)$ by τ. Let $h \cdot D$ be the multi-flow obtained by re-scaling h so that it meets D. Then, we have the following proposition.

Proposition 4.5.1.

$$\frac{1}{2\sqrt{2(1-\lambda_1)}} \leq \rho(h \cdot D) \leq 16\tau.$$

Proof of Proposition 4.5.1. The lower bound on $\rho(h \cdot D)$ follows from Theorem 4.5.1. We proceed to show the upper bound. In the case of a random walk on a graph with m edges, for every edge e, $Q(e) = \frac{1}{2m}$. Therefore

$$\begin{aligned} \rho(D \cdot h) &= \frac{\|t_{D \cdot h}\|_\infty}{Q(e)} \\ &= 2m\|t_{D \cdot h}\|_\infty. \end{aligned}$$

Given an edge $e = (u, w)$, let $\phi_e = \phi_{uw}$ be a potential such that $h_{uw} = \nabla\phi_{uw}$. For convenience, for any i, let h_{ii} be defined to be the flow that is zero on all edges. Let

$e = (u, w)$ be the edge that carries the maximum load. Then,

$$\begin{aligned}
\rho(D \cdot h) &= \sum_{i,j} \frac{|\deg(i)\deg(j) h_{ij}(e)|}{2m} \\
&= \sum_{i,j} \frac{\deg(i)\deg(j)}{2m} |\phi_{ij}(u) - \phi_{ij}(w)| \\
&= \sum_{i,j} \frac{\deg(i)\deg(j)}{2m} |\phi_e(i) - \phi_e(j)| \quad \text{(Thm 4.3.3)} \\
&\leq \sum_{i,j} \frac{\deg(i)\deg(j)}{2m} (|\phi_e(i)| + |\phi_e(j)|) \\
&= 2\sum_i \deg(i)|\phi_e(i)|.
\end{aligned}$$

Adding an appropriate constant to ϕ_e is necessary, we may assume that

$$\begin{aligned}
\tilde{\pi}(\{u|\phi_{ij}(u) \leq 0\}) &\geq \frac{1}{2} \\
\tilde{\pi}(\{v|\phi_{ij}(v) \geq 0\}) &\geq \frac{1}{2}.
\end{aligned}$$

Then, Lemma 4.4.4 and Lemma 4.4.5 together imply that

$$2\sum_i \deg(i)|\phi_e(i)| \leq 16\tau,$$

completing the proof. □

CHAPTER 5

COMPUTING THE SURFACE AREA OF A CONVEX SET

5.1 Introduction

An important class of algorithmic questions centers around estimating geometric invariants of convex bodies. Arguably, the most basic invariant is the volume. It can be shown [27], [2] that any deterministic algorithm to approximate the volume of a convex body within a constant factor in $\mathbb{R}^n$ needs time exponential in the dimension n. Remarkably, randomized algorithms turn out to be more powerful. In their pathbreaking paper [25] Dyer, Frieze and Kannan gave the first randomized polynomial time algorithm to approximate the volume of a convex body to arbitrary accuracy. Since then a considerable body of work has been devoted to improving the complexity of volume computation culminating with the recent best of $O^*(n^4)$ due to Lovász and Vempala [61].

Another fundamental geometric invariant associated with a convex body is surface area. Estimating the surface area was mentioned as an open problem by Grötschel, Lovász, and Schrijver in 1988 [29]. Dyer, Gritzmann and Hufnagel [26] showed in 1998 that it could be solved in randomized polynomial time. The primary focus of their paper was to establish that the computation of surface area and certain other mixed volumes was possible in randomized polynomial time, and they assumed access to oracles for δ-neighbourhoods of the convex body. They did not discuss the complexity of their algorithm given only a membership oracle for the convex body. Below, we indicate an $O^*(n^{8.5})$ analysis of their algorithm in terms of the more restricted queries.

In this chapter we develop a new technique for estimating volumes of boundaries based on ideas from heat propagation. The underlying intuition is that the amount of heat escaping from a heated object in a small interval of time is proportional to the surface area.

It turns out that this intuition lends itself to an efficient randomized algorithm for computing surface areas of convex bodies, given by a membership oracle. In this chapter we describe the algorithm and the analysis of the algorithm, proving a complexity bound of $O^*(n^4)$. The $O^*(\cdot)$ notation hides the polynomial dependence on the relative error ϵ, and poly-logarithmic factors in the parameters of the problem. Since, as will be shown below, surface area estimation is at least as hard as volume approximation, this bound is the best possible, given the current state-of-the-art in volume estimation.

We note that this bound cannot be obtained using methods previously proposed in [26] due to a bottleneck in their approach. The method in [26] exploited the fact that $\mathrm{vol}(K+B\delta)$ is a polynomial in δ, where $B\delta$ is a ball of radius δ and the Minkowski sum $K+B\delta$ corresponds to the set of points within a distance δ of K. The surface area is the coefficient of the linear term, which they then estimate by interpolation. However, in a natural setting, we only have access to a membership oracle for K, but not for $K + B\delta$. Therefore a membership oracle for $K + B\delta$ has to be constructed, which as far as we can see, requires solving a quadratic programming problem on a convex set. Given access only to a *membership oracle*, the best known algorithm to handle this task is due to Kalai and Vempala, and makes $O^*(n^{4.5})$ oracle calls ([38]), which gives a bound on the complexity of the algorithm in [26] that is $O^*(n^{8.5})$.

Even with a stronger *separation oracle* the complexity of the method in [26] is $O^*(n^5)$, since the associated quadratic programming problem requires $O^*(n)$ operations ([96], [9].) On the other hand, the complexity of our method is $O^*(n^4)$ using only a *membership oracle*, matching the complexity of the volume computation of Lovász and Vempala [61].

5.2 Overview of the algorithm

Notation. Throughout this chapter, B will denote the unit n-dimensional ball, K will denote an n-dimensional convex body such that $rB \subseteq K \subseteq RB$. $S = \mathrm{vol}(\partial K)$ will denote the surface area of K and $V = \mathrm{vol}(K)$, its volume.

We first observe that problem of estimating the surface area of a convex body is at least as hard as that of estimating the volume. This observation can be stated as

Proposition 5.2.1. *If the surface area of any n-dimensional convex body K can be approximated in $O(n^\beta \textsf{polylog}(\frac{nR}{\delta r})\textsf{poly}(\frac{1}{\epsilon}))$ time, the volume can be approximated in $O(n^\beta \textsf{polylog}(\frac{nR}{\delta r})\textsf{poly}(\frac{1}{\epsilon}))$ time, where δ is the probability that the relative error exceeds ϵ.*

The proof of the proposition relies on the fact that given a body K there is a simple relationship between the volume of K and the surface area of the cylinder $K \times [0, h]$. More specifically (see Fig. 2)

$$2\,\mathrm{vol}(K) = \mathrm{vol}(\partial(K \times [0, h])) - h\,\mathrm{vol}(\partial K)$$

Thus an efficient algorithm for surface area estimation would also lead to an almost equally efficient algorithm for estimating the volume.

Our approach provides an estimate for the isoperimetric ratio $\frac{S}{V}$. Using the fastest existing algorithm for volume approximation, we obtain a separate estimate for V. Multiplying these two estimates yields the surface area S.

The underlying intuition of our algorithm is that the heat diffuses from a heated body through its boundary. Therefore the amount of heat escaping in a short period of time is proportional to the surface area of the object. Recalling that a point source of heat diffuses at time t according to the Gaussian distribution $\frac{1}{(4\pi t)^{n/2}} e^{-\frac{\|x\|^2}{4t}}$ leads to the following informal description of the algorithm (see details in Section 3):

Step 1. Take $x_1, \ldots, x_N$ to be samples from the uniform distribution on K.

Step 2. For each x_i, let $y_i = x_i + v_i$, where v_i is sampled from the Gaussian distribution with density $\frac{1}{(4\pi t)^{n/2}} e^{-\frac{\|x\|^2}{4t}}$ for some appropriate value of t. Thus y_i is obtained from x_i by taking a random Gaussian step.

Step 3. Let $\hat{N}$ be the number of y's, which land outside of K. $\frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}$ is an estimate for $\frac{S}{V}$.

Step 4. Using an existing algorithm, produce an estimate $\hat{V}$ for the volume. Estimate the surface area as $\hat{V}\frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}$.

We will show that that each of the Steps 1,3,4 can be done using at most $O^*(n^4)$ calls to the membership oracle. It is customary to count the number of oracle calls rather than the number of arithmetic steps in the volume literature, while measuring the complexity. Step 2, of course, does not require any calls to the oracle at all.

The main technical result of this chapter is to show how to choose values of t and N, such that

$$(1-\epsilon)\left(\frac{S}{V}\sqrt{\frac{t}{\pi}}\right) < \frac{\hat{N}}{N} < (1+\epsilon)\left(\frac{S}{V}\sqrt{\frac{t}{\pi}}\right)$$

It is not known how to efficiently obtain independent random samples from the uniform distribution on K. We show how to relax this condition and use *almost independent* samples from a *nearly uniform* distribution instead, to derive these estimates.

We then apply certain results from [60] and [61], to generate $O\left(\frac{n}{\epsilon^3}\right)$ such samples making at most $O^*\left(\frac{n^4}{\epsilon^3}\right)$ oracle calls.

Putting these and some additional observations together, we obtain the following theorem which is the main result of this chapter:

Theorem 5.2.1. *The surface area of a convex body K, given by a membership oracle, and parameters r, R such that $rB \subseteq K \subseteq RB$ can be approximated to within a relative error of ϵ with probability $1-\delta$ using at most*

$$O\left(n^4 \log\frac{1}{\delta}\left(\frac{1}{\epsilon^2}\log^9\frac{n}{\epsilon} + \log^8 n \log\frac{R}{r} + \frac{1}{\epsilon^3}\log^7\left(\frac{n}{\epsilon}\right)\right)\right)$$

i. e. $O^*(n^4)$ oracle calls.

The number of arithmetic operations is $O^*(n^6)$, on numbers with a polylogarithmic number of digits. This is the same as that for volume computation in [61].

5.3 Algorithm to compute the surface area

5.3.1 Notation and Preliminaries

A body K is said to be in *t-isotropic position* if, for every unit vector u,

$$\frac{1}{t} \leq \int_K (u^T(x - \bar{x}))^2 dx \leq t,$$

where $\bar{x}$ is the center of mass of K. Let ρ be the uniform distribution on convex body K. We call a random point x ϵ-uniform if

$$\sup_{\text{measurable } A} P(x \in A) - \rho(A) \leq \frac{\epsilon}{2},$$

Two random variables will be called μ-independent if for any two Borel sets A and B in their ranges,

$$|P(X \in A, Y \in B) - P(X \in A)P(Y \in B)| \leq \mu.$$

A density ρ' is said to have $\mathcal{L}_2$ norm $\int_K \left(\frac{d\rho'}{d\rho}\right)^2 d\rho$ with respect to the uniform distribution on K.

A consequence of the results on page 4 of ([61]), and Theorem 7.2, [102] is, given a starting point that is ϵ-uniform, and comes from a distribution that has a bounded $\mathcal{L}_2$ norm it takes $O(n^3 \ln^7 \frac{n}{\epsilon\mu})$ oracle calls per point, to generate N points $x_1, \ldots, x_N$ that are ϵ-uniform, such that each pair is μ-independent from a convex body that is 2-isotropic. This fact plays a crucial role in allowing the surface area algorithm to have a complexity bounded by $O^*(n^4)$.

5.3.2 Algorithm

We present an algorithm below that outputs an ϵ-approximation to the surface area of a convex body K with probability $> 3/4$. Running it $\lceil 36 \ln \left(\frac{2}{\delta}\right) \rceil$ times and taking the median of the outputs gives the result with a confidence $> 1 - \delta$.

Input: Convex body K, given by a membership oracle, and parameters r, R such that $rB \subseteq K \subseteq RB$ and an error parameter $\epsilon < 1$.

Output: An estimate $\hat{S}$, that with probability $> 3/4$ has a relative error of less than ϵ with respect to S.

Set $\epsilon' := \frac{\epsilon}{8}$, $\mu := \frac{\epsilon'^4}{2^{18}n^2}$, $N := \lceil \frac{2^{13}n}{\epsilon'^3} \rceil$.

Step 1. Run a volume algorithm to obtain an estimate $\hat{V}$ of V that has a relative error ϵ' with probability $> \frac{15}{16}$.

Step 2 Generate a linear transformation T given by a symmetric positive-definite matrix such that TK is 2-isotropic with probability $> \frac{15}{16}$.

Step 3 Compute a lower bound r' to the smallest eigenvalue r_{opt} of $\frac{T^{-1}}{\sqrt{2}}$, that satisfies $\frac{2}{\sqrt{5}} r_{opt} < r' < r_{opt}$. Set $r_{in} := \max(r, r')$.

Step 4 Set $\sqrt{t} := \frac{\epsilon' r_{in}}{4n}$.

Step 5 Generate N random points $x_2, \ldots, x_N$ from K, such that with probability 15/16, they are $\frac{\epsilon'^2}{64n}$-uniform and each pair $\{x_i, x_j\}$ for $1 \leq i < j \leq N$ is μ-independent.

Step 6 Generate N independent random samples $v_1, \ldots, v_N$ from the spherically symmetric multivariate Gaussian distribution with mean $\vec{0}$ and variance $2nt$.

Step 7 Let $\hat{N} := |\{i | x_i + v_i \notin K\}|$ be the number of times $x_i + v_i$ lands outside of K.

Step 8 Output $\frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}\hat{V}$.

5.3.3 Analysis of the Run-time

Step 1 takes at most

$$O\left(\frac{n^4}{\epsilon^2}\log^9\frac{n}{\epsilon} + n^4\log^8 n \log\frac{R}{r}\right)$$

oracle calls, using the volume algorithm of Lovász and Vempala ([61].) The number of steps in the computation is $O^*(n^6)$.

Step 2 Such a transformation is obtained during the execution of the volume algorithm from [61] for no additional cost.

Step 3 takes $O(n^3)$ steps of computation ([80].)

Step 4 takes $O(1)$ steps.

Step 5 takes

$$O\left(\frac{n^4}{\epsilon^3}\log^7\left(\frac{n}{\epsilon}\right)\right)$$

steps of computation (including oracle calls) once a point x_1 is obtained that is $\frac{\epsilon'^2}{64n}$-uniform, and has an $\mathcal{L}^2$ norm that is bounded above by a constant. Such a point can be obtained from the algorithm in step 1, for no additional cost up to constants. The cost mentioned in this step is incurred because we are required to generate $O(\frac{n}{\epsilon^3})$ random points given the initial random point x_1 and the time per point is $O(n^3\ln^7\frac{n}{\epsilon})$. This last fact follows from the complexity per point mentioned on page 4 ([61]), and theorems 7.1 and 7.2 of ([102].)

Step 6 and **Step 7** take $O\left(\frac{n^2}{\epsilon^3}\mathsf{polylog}\frac{n}{\epsilon\delta}\right)$ steps each, assuming that a sample from univariate Gaussian distribution can be obtained upto $O(\mathsf{polylog}(\frac{n}{\epsilon\delta}))$ digits in $O(\mathsf{polylog}(\frac{n}{\epsilon\delta}))$ steps.

Step 8 takes $O(1)$ steps. Finally, to obtain the approximation with a confidence $> 1-\delta$, this algorithm must be run $O\left(\log\left(\frac{1}{\delta}\right)\right)$ times. Therefore the overall cost in terms of oracle calls is

$$O\left(n^4\log\frac{1}{\delta}\left(\frac{1}{\epsilon^2}\log^9\frac{n}{\epsilon}+\log^8 n\log\frac{R}{r}+\frac{1}{\epsilon^3}\log^7\left(\frac{n}{\epsilon}\right)\right)\right)$$

i. e. $O^*(n^4)$ oracle calls. The number of arithmetic operations is $O^*(n^6)$, on numbers with a polylogarithmic number of digits. This is the same as that for volume computation in [61].

5.4 Proving correctness of the algorithm

Definition 5.4.1. *Let*

$$G^t(x,y) := \frac{\mathrm{e}^{-\|x-y\|^2/4t}}{(4\pi t)^{n/2}}.$$

and

$$F_t := \sqrt{\frac{\pi}{t}} \int_K \int_{\mathbb{R}^n \setminus K} G^t(x,y) dy dx.$$

$\sqrt{\frac{t}{\pi}} \frac{F_t}{V}$ is the fraction of heat that would diffuse out of K in time t.

Our proof hinges on two main propositions. The first, Proposition 5.4.1, states that F_t is a good approximation for the surface area S. As in the surface area algorithm, let T be a linear transformation such that TK is 2-isotropic. Compute a lower bound r' to the smallest eigenvalue r_{opt} of $\frac{T^{-1}}{\sqrt{2}}$, that satisfies $\frac{2}{\sqrt{5}} r_{opt} < r' < r_{opt}$. Set $r_{in} := \max(r, r')$. Then, the following is true.

Proposition 5.4.1. *Let* $\sqrt{t} = \frac{\epsilon' r_{in}}{4n}$ *and* $\epsilon' < 1/2$. *Then,*

$$(1-\epsilon')S < F_t < (1+\epsilon')S.$$

Proposition 5.4.2 states that the empirical quantity $\hat{S}$ computed by the surface area algorithm is likely to be an ϵ-approximation of F_t with probability $> 3/4$. Let $x_1, x_2, \ldots, x_N$ from K, be ϵ'-uniform and each pair $\{x_i, x_j\}$ for $1 \le i < j \le N$ be μ-independent with probability $> 15/16$. Let $v_1, \ldots, v_N$ be N independent random samples from the spherically symmetric multivariate Gaussian distribution whose mean is $\vec{0}$ and variance is $2nt$. Let $\hat{N} := |\{i | x_i + v_i \notin K\}|$. Then,

Proposition 5.4.2. *Let* $\sqrt{t} = \frac{\epsilon' r_{in}}{4n}$ *and* $\epsilon' < 1/2$. *Then, with probability greater than* $\frac{3}{4}$,

$$(1-\epsilon')(1-2\epsilon')F_t < \frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}\hat{V} < (1+\epsilon')(1+2\epsilon')F_t.$$

These two Propositions together imply that with probability $> 3/4$,

$$(1-\epsilon)S < \frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}\hat{V} < (1+\epsilon)S.$$

The argument for boosting the confidence from $3/4$ to $1-\delta$ is along the lines of ([43],[37].) We devote the rest of this chapter to outlining the proofs of Proposition 5.4.1 and Proposition 5.4.2.

5.5 Relating surface area S to normalized heat flow F_t

In this section, we prove Proposition 5.4.1.

5.5.1 Notation and Preliminaries

The set of points within a distance δ of a convex body K (including K itself) shall be denoted K_δ. This is called the *outer parallel body* of K and is convex.
The set of points at a distance $\geq \delta$ to $\mathbb{R}^n - K$ shall be denoted $K_{-\delta}$. This is called the *inner parallel body* of K and again is convex. For any body K, we denote by ∂K, its boundary.

Given $x \in K$, let H_x be a closest halfspace to x not intersecting $K \setminus \partial K$. For $y \notin K$ define H_y to be the halfspace furthest from y containing K.

Observation 5.5.1. *If $x \in \partial K_{-\delta}$ then the distance between x and H_x is δ. If $y \in \partial K_\delta$ then the distance between y and H_y is δ.*

Definition 5.5.1. *Let*

$$e(t,\delta) = \frac{1 - \mathrm{Erf}\left(\frac{\delta}{2\sqrt{t}}\right)}{2}$$

where Erf *is the usual Gauss error function, defined by*

$$\mathrm{Erf}(z) := \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt.$$

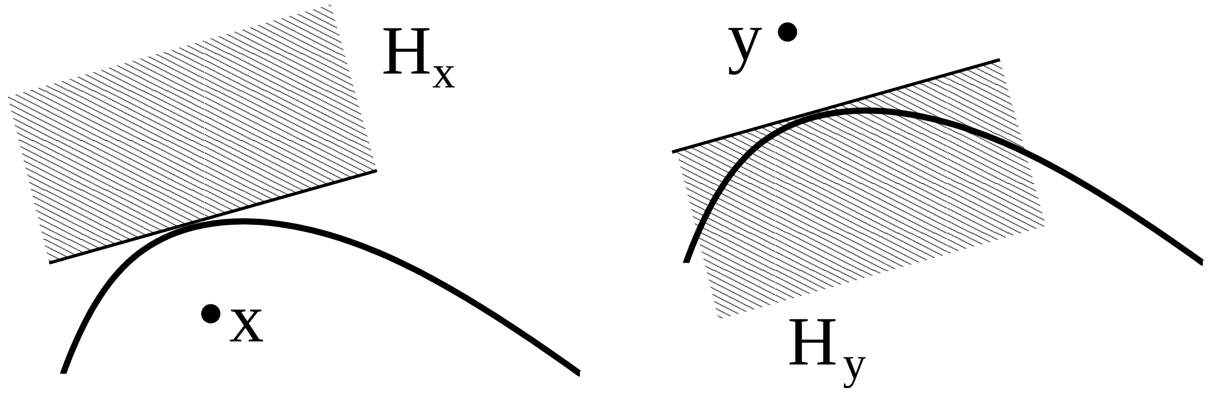

Figure 5.1: Points x and y and corresponding halfspaces H_x and H_y

Observation 5.5.2. *Let $x \in \partial K_{-\delta}$, and $y \in \partial K_\delta$. Then,*

$$\int_{H_y} G^t(z,y)dz = e(t,\delta)$$

and

$$\int_{H_x} G^t(z,x)dz = e(t,\delta)$$

The volume of K_δ is a polynomial in δ, given by the *Steiner formula* (see page 197, [86].)

$$\mathrm{vol}(K_\delta) = a_0 + \ldots + \binom{n}{i} a_i \delta^i + \ldots + a_n \delta^n.$$

The coefficients a_i satisfy the *Alexandrov-Fenchel inequalities* (see page 334, [86],) which state that the coefficients a_i are log-concave; i. e. $a_i^2 \geq a_{i-1}a_{i+1}$ for $1 \leq i \leq n-1$.

Definition 5.5.2. *The surface area* $\mathrm{vol}(\partial K)$ *of an arbitrary convex body K is defined as*

$$\lim_{\delta \to 0} \frac{\mathrm{vol} K_\delta - \mathrm{vol} K}{\delta}.$$

It follows from the *Steiner formula* that this limit exists and is finite. It is a consequence of Lemma 5.5.2 that the so defined surface area for an inner parallel body $\mathrm{vol}(\partial K_{-\delta})$ is a continuous function of δ. For an outer parallel body, the *Steiner formula* implies that $\mathrm{vol}(\partial K_\delta)$ is a polynomial in δ.

5.5.2 Proof of Proposition 5.4.1

Lemma 5.5.1 is the first step towards proving upper and lower bounds for the normalized heat flow F_t in terms of S. It bounds F_t above by a function of the $\mathrm{vol}(\partial K_\delta)$ and below by a function of $\mathrm{vol}(\partial K_{-\delta})$.

Lemma 5.5.1. *1.* $\sqrt{\frac{\pi}{t}}\int_{\delta\geq 0} \mathrm{vol}(\partial K_{-\delta})e(t,\delta)d\delta < F_t$

2. $F_t < \sqrt{\frac{\pi}{t}}\int_{\delta\geq 0} \mathrm{vol}(\partial K_\delta)e(t,\delta)d\delta.$

Proof

Note that for a fixed $x \in \partial K_{-\delta}$

$$\int_{\mathbb{R}^n\setminus K} G^t(x,y)dy > \int_{H_x} G^t(x,y)dy = e(t,\delta).$$

Therefore integrating over shells $\partial K_{-\delta}$, $F_t =$

$$\sqrt{\frac{\pi}{t}}\int_K\int_{\mathbb{R}^n\setminus K} G^t(x,y)dydx > \sqrt{\frac{\pi}{t}}\int_{\delta\geq 0} \mathrm{vol}(\partial K_{-\delta})e(t,\delta)d\delta.$$

By the same token for a fixed $y \in \partial K_\delta$

$$\int_K G^t(x,y)dx < \int_{H_y} G^t(x,y)dx = e(t,\delta)$$

and proceeding as before, we have the upper bound

$$F_t < \sqrt{\frac{\pi}{t}}\int_{\delta\geq 0} \mathrm{vol}(\partial K_\delta)e(t,\delta)d\delta.$$

□

The next step is to upper bound $\text{vol}(K_\delta)$ and lower bound $\text{vol}(\partial K_\delta)$, which is done in Lemmas 5.5.2 and 5.5.3 respectively.

Lemma 5.5.2.

$$\text{vol}(\partial K_{-\delta}) \geq \left(1 - n\frac{\delta}{r_{in}}\right) \text{vol}(\partial K)$$

Proof:

Let O be the center of the sphere of radius r_{in} contained inside K. We shall first prove that $K_{-\delta}$ contains $(1 - \frac{\delta}{r_{in}})K$ where this scaling is done from the origin O. Let A be a point on ∂K and let F be the image of A under this scaling. It suffices to prove that $F \in K_{-\delta}$.

We construct the smallest cone from A containing the sphere. Let B be a point where the cone touches the sphere. We have $OB = r_{in}$. Now consider the inscribed sphere centered at F. By similarity of triangles, we have

$$\frac{CF}{OB} = \frac{AF}{AO}$$

Noticing that $AF = \frac{\delta}{r_{in}} OA$, we obtain

$$CF = OB\frac{AF}{AO} = \delta$$

We thus see that the radius of the inscribed ball is δ and hence the δ-ball centered in F is contained in K. The fact that $F \in K_{-\delta}$ follows from the definition.

It is known that the surface area of a convex body is less or equal than the surface area of any convex body that contains it (page 284, [86]). Therefore

$$\text{vol}(\partial K_{-\delta}) \geq \text{vol}\left((1 - \frac{\delta}{r_{in}})\,\partial K\right)$$

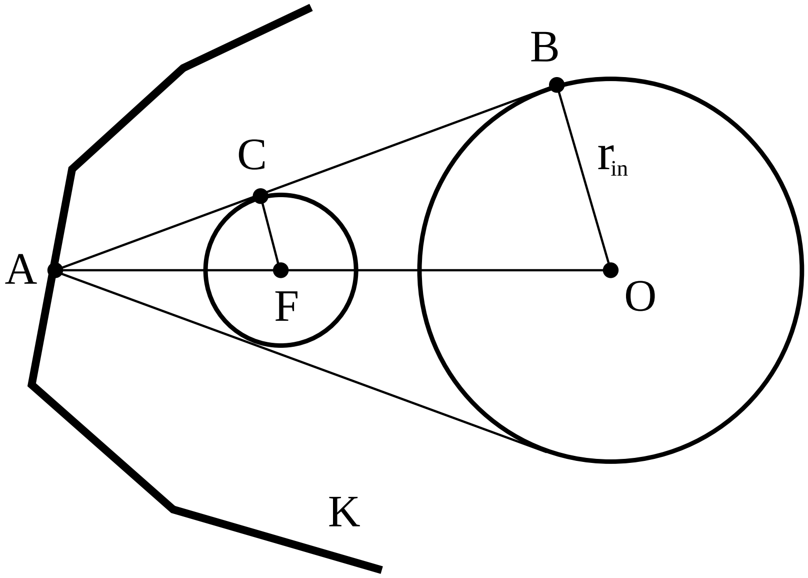

Figure 5.2: $K_{-\delta}$ contains $\left(1-\frac{\delta}{r_{in}}\right)K$

Since the volumes of $n-1$-dimensional objects scale as $n-1^{th}$ powers and observing that for $x<1$, $\max\{0,(1-x)^{n-1}\}>1-nx$, we arrive at the conclusion: $\mathrm{vol}\left((1-\frac{\delta}{r_{in}})\,\partial K\right)=$
$(1-\frac{\delta}{r_{in}})^{n-1}\ \mathrm{vol}(\partial K)$
$\geq (1-\frac{n\delta}{r_{in}})\,\mathrm{vol}(\partial K)$ □

□

Lemma 5.5.3.

$$\mathrm{vol}(K_\delta)\leq V\ \exp\left(\delta\frac{S}{V}\right).$$

Proof: The volume of K_δ is a polynomial in δ, given by the *Steiner formula* (see page 197, [86].)

$$\mathrm{vol}(K_\delta)=a_0+\ldots+\binom{n}{i}a_i\delta^i+\ldots+a_n\delta^n.$$

From the Alexandrov-Fenchel inequalities (see page 334, [86].) the coefficients a_i are log-concave; i. e.

$$a_i^2 \geq a_{i-1}a_{i+1}.$$

As a result

$$\frac{a_i}{a_0} \leq \left(\frac{a_1}{a_0}\right)^i.$$

a_0 is V, the volume of K while na_1 is the surface area S of K. Putting these inequalities together with the fact that $\binom{n}{i} \leq \frac{n^i}{i!}$, the lemma follows.

□

Although Lemma 5.5.3 is an upper bound on $\operatorname{vol}(K_\delta)$ rather than $\operatorname{vol}(\partial K_\delta)$, it can be applied after transforming the upper bound in Lemma 5.5.1 by integrating by parts. Lemmas 5.5.2, 5.5.3 and 5.5.1 together result in the following lemma.

Lemma 5.5.4. *Let* $\alpha = \left(\frac{S}{V}\right)^2 t$. *Then,*

$$S\left(1 - \frac{n\sqrt{\pi t}}{2r_{in}}\right) < F_t < S\left(\sqrt{\frac{\pi}{\alpha}}\frac{\exp(\alpha) - 1}{2} + \exp(\alpha)\right).$$

Finally, we prove the following bounds for the isoperimetric constant $\frac{V}{S}$ of K in terms of r_{in}.

Lemma 5.5.5.

$$\frac{r_{in}}{n} \leq \frac{V}{S} < 4\, r_{in}$$

Proof: It follows from Lemma 3.4 in [61] that a ball of radius $\frac{1}{\sqrt{2}}$ around the centroid of TK is entirely contained in TK. Therefore

Observation 5.5.3. *K contains a ball of radius r_{in}.*

Observation 5.5.4. *For any unit vector* u *that minimizes* $\frac{\|T^{-1}u\|}{\sqrt{2}}$, *if* x *is chosen uniformly at random from K,* $var(\mathrm{u} \cdot \mathrm{x}) \leq 5\mathrm{r}_{\mathrm{in}}^2$.

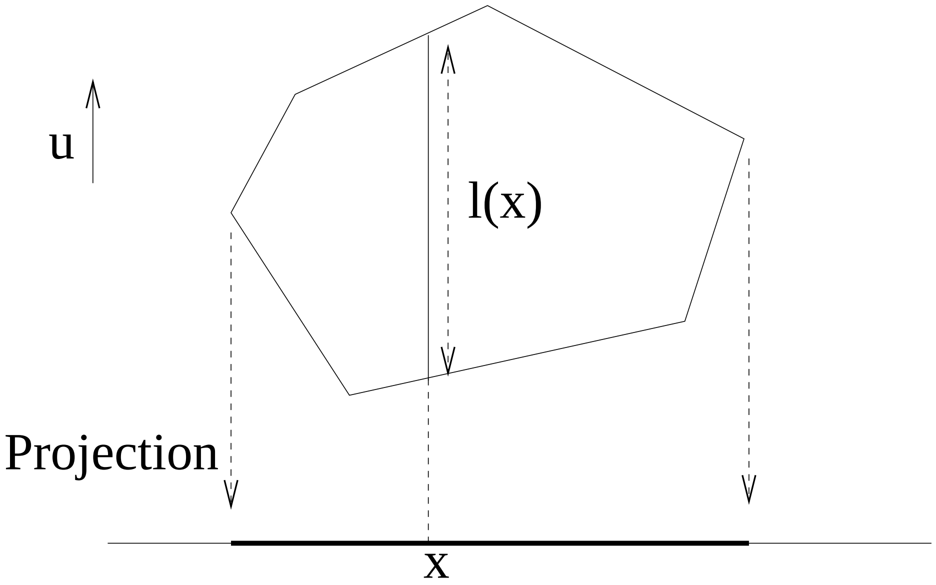

Figure 5.3: Projecting along a unit vector u minimizing $\|T^{-1}u\|$

We are now in a position to present the proof of Lemma 5.5.5. The first inequality $\frac{r_{in}}{n} \leq \frac{V}{S}$ can be obtained from Lemma 5.5.2 by integration. The only condition on r_{in} there, is that $r_{in}B \subseteq K$. This property is satisfied by r_{in} by Observation 5.5.3.

Fix a unit vector u such that for x chosen uniformly at random from K, $\mathrm{var}(\mathrm{u} \cdot \mathrm{x}) \leq 5\mathrm{r}_{\mathrm{in}}^2$. Observation 5.5.4 states that such a vector exists.

Definition 5.5.3. *Let π be an orthogonal projection of K onto a hyperplane perpendicular to* u. *Further, for a point $y \in \pi(K)$, let ℓ_y be the length of the preimage $\pi^{-1}(y)$.*

$$\mathrm{var}(\mathrm{u}^{\mathrm{T}}\mathrm{x}) \leq 5\mathrm{r}_{\mathrm{in}}^2.$$

The variance of $\mathrm{u} \cdot \mathrm{x}$ under the condition $\pi(x) = y$, is given by $\ell_y^2/12$, since this is the variance of a random variable that takes a value from an interval of length ℓ_y uniformly at random.

$$\begin{aligned} \mathrm{var}(\mathrm{u}\cdot\mathrm{x}) &\geq \frac{\int_{\pi(K)} \mathrm{var}(\mathrm{u}\cdot\mathrm{x}|\pi(\mathrm{x})=\mathrm{y})\ell_{\mathrm{y}}\mathrm{dy}}{V} \tag{5.1}\\ &= \frac{\int_{\pi(K)} \ell_y^3 dy}{12V}. \tag{5.2}\\ & \tag{5.3} \end{aligned}$$

$$\frac{\int_{\pi(K)} \ell_y^3 dy}{\mathrm{vol}(\pi(K))} \geq \left(\frac{\int_{\pi(K)} \ell_y dy}{\mathrm{vol}(\pi(K))}\right)^3 = \left(\frac{V}{\mathrm{vol}(\pi(K))}\right)^3.$$

since for any non-negative random variable X, $E[X^3] \geq E[X]^3$. Therefore,

$$\frac{\int_{\pi(K)} \ell_y^3 dy}{12V} \geq \left(\frac{V^2}{12\,\mathrm{vol}(\pi(K))^2}\right).$$

Further, $\mathrm{vol}(\pi(K)) \leq S/2$. Putting these facts together,

$$5r_{in}^2 \geq \left(\frac{V^2}{12\,\mathrm{vol}(\pi(K))^2}\right) \geq \left(\frac{V^2}{3S^2}\right),$$

and so $\frac{V}{S} < \sqrt{15}\, r_{in} < 4\, r_{in}$

□

Lemmas 5.5.4 and 5.5.5 together give the result of Proposition 5.4.1, as we show below. The lower bound on F_t is immediate for $\sqrt{t} = \frac{\epsilon r_{in}}{4n}$ using the lower bound in Lemma 5.5.4. To prove the upper bound, we observe that $\alpha = \left(\frac{S}{V}\right)^2 t \leq \left(\frac{n}{r_{in}}\right)^2 t$ from Lemma 5.5.1, which equals $\frac{\epsilon'^2}{16}$. Since $\epsilon < 0.5, \alpha < 1$. Therefore $\mathrm{e}^{\alpha} < 1 + 2\alpha$. It follows that

$$\begin{aligned} S\left(\sqrt{\frac{\pi}{\alpha}}\frac{\exp(\alpha)-1}{2} + \exp(\alpha)\right) &< S\left(\sqrt{\pi\alpha} + 1 + 2\alpha)\right)\\ &< S(1+4\sqrt{\alpha})\\ &< (1+\epsilon)S. \end{aligned}$$

□

5.6 Proof of Proposition 5.4.2

The proof of Proposition 5.4.2 is complicated by the large number of parameters involved. We mention the important steps below.

Lemma 5.6.1. *With probability greater than* 7/8*,*

$$(1-\epsilon')F_t < E\left[\frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}V\right] < (1+\epsilon')F_t,$$

and $(1-\epsilon')V < \hat{V} < (1+\epsilon')V$*.*

Lemma 5.6.2. *With probability greater than* 15/16*,*

$$var\left(\frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}V\right) < \frac{\epsilon'^2 F_t^2}{16}.$$

Using Chebycheff's inequality and Lemma 5.6.2,

$$P\left[\left|\frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}V - E\left[\frac{\hat{N}}{N}\sqrt{\frac{\pi}{t}}V\right]\right| > \epsilon' F_t\right] < \frac{1}{16}.$$

Putting this together with Lemma 5.6.1, we arrive at the desired result. □

CHAPTER 6

SAMPLING HYPERSURFACES

6.1 Introduction

Random sampling has numerous applications. They are ingredients in statistical goodness-of-fit tests and Monte-Carlo methods in numerical computation. In computer science, they have been used to obtain approximate solutions to problems that are otherwise intractable. A large fraction of known results in sampling that come with guarantees belong to the discrete setting. A notable exception is the question of sampling convex bodies in $\mathbb{R}^d$. A large body of work has been devoted to this question (in particular [25], [61]) spanning the past 15 years leading to important insights and algorithmic progress.

However, once one leaves the convex domain setting, much less is known. We are interested in the general setting in which we wish to sample a set that may be represented as a submanifold of Euclidean space. While continuous random processes on manifolds have been analyzed in several works, (such as those of P. Matthews [64],[65]), as far as we can see, these do not directly lead to algorithms with complexity guarantees.

6.1.1 Summary of Main Results for sampling Hypersurfaces

We develop algorithms for the following tasks.

Our basic setting is as follows. Consider an open set $A \subset \mathbb{R}^d$ specified through a membership oracle. Assume we have access to an efficient sampler for A and now consider the task of uniformly sampling the (hyper) surface ∂A. We consider two related but distinct problems in this setting.

(i) A is a convex body satisfying the usual constraint of $B_r \subset A \subset B_R$ where B_r and B_R are balls of radius r and R respectively. Then an efficient sampler for A is known to exist. However, no sampler is known for the surface of the convex body. It is worth noting that a number of intuitively plausible algorithms suggest themselves immediately. One idea may be draw a point x from A, shoot a ray in the direction from 0 to x and find its intersection with the boundary of the object. This will generate non-uniform samples from the surface (and it has been studied under the name Liouville measure.) A second idea may be to consider building a sampler for the set difference of a suitable expansion of the body from itself. This procedure has a complexity of at least $O^*(d^{8.5})$ oracle calls with the present technology because there is no method known to simulate each membership call to the expanded body using less than $O^*(d^{4.5})$ calls (see [9]).

Our main result here (Theorem 1) is to present an algorithm that will generate a sample from an approximately uniform distribution with $O^*(\frac{d^4}{\epsilon})$ calls to the membership oracle where ϵ is the desired variation distance to the target.

Beyond theoretical interest, the surface of the convex body setting has natural applications to many goodness of fit tests in statistics. The example of the gamma distribution discussed earlier requires one to sample from the set $\prod_i X_i = b$ embedded in the simplex (given by $\sum_j X_j = a$). This set corresponds to the boundary of a convex object.

(ii) A is a domain (not necessarily convex) such that its boundary ∂A has the structure of a smooth submanifold of Euclidean space of co-dimension one. A canonical example of such a setting is one in which the submanifold is the zeroset of a smooth function $f : \mathbb{R}^d \rightarrow \mathbb{R}$. A is therefore given by $A = \{x|f(x) < 0\}$. In machine learning applications, the function f may often be related to a classification or clustering function. In numerical computation and boundary value problems, one may wish to integrate a function subject to a constraint (given by $f(x) = 0$).

In this setting, we have access to a membership oracle for A (through f) and we assume a sampler for A exists. Alternatively, $A \subset K$ such that it has nontrivial fraction of a convex body K and one can construct a sampler for A sampling from K and using the membership oracle for rejection.

In this non-convex setting, not much is known and our main result (Theorem 2) is an algorithm that generates samples from ∂A that are approximately uniform with complexity $O^*(\frac{R}{\tau\sqrt{\epsilon}})$ where τ is a parameter related to the curvature of the manifold, R is the radius of a circumscribed ball and ϵ is an upper bound on the total variation distance of the output from uniform.

6.1.2 Notation

Let $\|.\|$ denote the Euclidean norm on $\mathbb{R}^d$. Let λ denote the Lebesgue measure on $\mathbb{R}^d$. The induced measure onto the surface of a manifold $\mathcal{M}$ shall be denoted $\lambda_{\mathcal{M}}$. Let

$$G^t(x, y) := \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|x-y\|^2}{4t}}.$$

be the d dimensional gaussian.

Definition 6.1.1. *Given two measures μ and ν over $\mathbb{R}^d$, let*

$$\|\mu - \nu\|_{TV} := \sup_{A \subseteq \mathbb{R}^d} |\mu(A) - \nu(A)|$$

denote the total variation distance between μ and ν.

Definition 6.1.2. *Given two measures μ and ν on $\mathbb{R}^d$, the transportation distance $d_{TR}(\mu, \nu)$ is defined to be the infimum*

$$\inf_{\gamma} \int \|x - y\| \, d\gamma(x, y).$$

taken over all measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that for measurable sets A and B, $\gamma(A \times \mathbb{R}^d) = \mu(A)$, $\gamma(\mathbb{R}^d \times B) = \nu(B)$.

Notation: *We say that $n = O^*(m)$, if $n = O(m \text{ polylog}(m))$. In the complexity analysis, we shall only consider the number of oracle calls made, as is customary in this literature.*

6.2 Sampling the Surface of a convex body

Let B be the unit ball in $\mathbb{R}^d$. Let B_α denote the ball of radius α centered at the origin. Consider a convex body K in $\mathbb{R}^d$ such that

$$B_r \subseteq K \subseteq B_R.$$

Let $\mathcal{B}$ be a source of random samples from K. Our main theorem is

Theorem 6.2.1. *Let K be a convex body whose boundary ∂K is a union of finitely many smooth Hypersurfaces.*

1. *The output of* Csample *has a distribution $\tilde{\mu}$, whose variation distance measured against the uniform distribution $\tilde{\lambda} = \tilde{\lambda}_{\partial K}$ is $O(\epsilon)$,*

$$\|\tilde{\mu} - \nu\|_{TV} \leq O(\epsilon).$$

2. *The expected number of oracles calls made by* Csample *(to $\mathcal{B}$ and the membership oracle of K) for each sample of* Csample *is $O^*(\frac{d}{\epsilon})$ (, giving a membership query complexity of $O^*(\frac{d^4}{\epsilon})$ for one random sample from ∂K.)*

6.2.1 Algorithm Csample

6.2.2 Correctness

In our calculations, $z \in \partial K$ will be be a generic point at which ∂K is smooth. In particular for all such z, there is a (unique) tangent hyperplane. Let $\lambda_{\partial K}$ denote the $n-1$-dimensional surface measure on ∂K. Let S and V denote the surface area and volume, respectively, of K. Let $\mu_{\partial K}$ denote the measure induced by the output of algorithm Csample . Let $|\mu|$ denote the total mass for any measure μ. We shall define a measure $\mu_{\partial K}$ on ∂K related to the

```
Csample

1. Estimate (see [80]) with confidence $> 1 - \epsilon$, the smallest eigenvalue $\kappa$ of the Inertia matrix $A(K) := \mathbb{E}[(x - \overline{x})(x - \overline{x})^T]$ where $x$ is random in uniformly $K$, to within relative error $1/2$ using $O(d\log^2(d)log\frac{1}{\epsilon})$ random samples (see Rudelson [85].)

2. Set
       $\sqrt{t} := \frac{\epsilon\sqrt{\kappa}}{32d}.$

3. (a) Set $p =$ Ctry $(t)$ .
   (b) If $p = \emptyset$, goto (3a). Else output $p$.
```

```
Ctry (t):

1. Use $\mathcal{B}$ to generate a random point $x$ from the uniform distribution on $K$.

2. Let $y :=$ Gaussian$(x, 2tI)$ be a random vector chosen from a spherical $d$-dimensional Gaussian distribution with covariance $2tI$ and mean $x$.

3. Let $\ell$ the segment whose endpoints are $x$ and $y$.

4. If $y \notin K$ output $\ell \cap \partial K$, else output $\emptyset$.
```

"local diffusion" out of small patches. Formally, if Δ a subset of ∂K, the measure assigned to it by $\mu_{\partial K}$ is

$$\mu_{\partial K}(\Delta) := \int_{x\in S}\int_{y\in\mathbb{R}^d\setminus S} G^t(x,y)\mathcal{I}\left[\overline{xy}\cap\Delta\neq\emptyset\right]d\lambda(x)d\lambda(y) \tag{6.1}$$

where $\mathcal{I}$ is the indicator function and $G^t(x,y)$ is the spherical Gaussian kernel with covariance matrix $2tI$. Note that

$$V\mathbb{P}[\mathsf{Ctry}\ (t)\in\Delta] = \mu_{\partial K}(\Delta).$$

Theorem 1 (part 1)

The output of $\mathsf{Csample}$ has a distribution $\tilde{\mu} = \frac{\mu_{\partial K}}{|\mu_{\partial K}|}$, whose variation distance measured against the uniform distribution $\tilde{\lambda}_{\partial K}$ is $O(\epsilon)$,

$$\|\tilde{\mu} - \tilde{\lambda}_{\partial K}\|_{TV} \leq O(\epsilon).$$

Proof: It follows from Lemma 6.3.1 to note that at generic points, *locally* the measure generated by one trial of $\mathsf{Ctry}\ (t)$ is always less than the value predicted by its small t asymptotics $\sqrt{\frac{t}{\pi}}\frac{S}{V}$, i. e.

$$\forall \text{ generic } z\in\partial K,\quad \frac{d\mu_{\partial K}}{d\lambda_{\partial k}} < \sqrt{\frac{t}{\pi}}S.$$

Thus we have a local upper bound on $\frac{d\mu_{\partial K}}{d\lambda_{\partial K}} \leq \sqrt{\frac{t}{\pi}}$ uniformly for all generic points $z\in\partial K$. It would now suffice to prove almost matching *global* lower bound on the total measure, of the form

$$|\mu_{\partial K}| > (1-O(\epsilon))\sqrt{\frac{t}{\pi}}S.$$

This is true by Proposition 4.1 in [5]. This proves that

$$\|\tilde{\mu} - \tilde{\lambda}_{\mathcal{M}}\|_{TV} \leq O(\epsilon).$$

□

6.2.3 Complexity

The number of random samples needed to estimate the Inertia matrix is $O^*(d)$ (so that the estimated eigenvalues are all within $(0.5, 1.5)$ of their true values with confidence $1-\epsilon$) from results of Rudelson ([85]). It is known that a convex body contains a ball of radius $\geq \sqrt{\Lambda_{min}(K)}$. Here $\Lambda_{min}(K)$ is the smallest eigenvalue of $A(K)$. Therefore, K contains a ball of radius r_{in}, where $r_{in}^2 = \frac{9}{10}\kappa$.

Theorem 1 (part 2):

The expected number of oracles calls made by Csample (to $\mathcal{B}$ and the membership oracle of K) for each sample of Csample is $O^*(\frac{d}{\epsilon})$ (, giving a total complexity of $O^*(\frac{d^4}{\epsilon})$ for one random sample from ∂K.)

Proof: The following two results will be used in this proof.

Lemma 6.2.1 ((Lemma 5.5 in [5]). *Suppose x has the distribution of a random vector (point) in K, define $A(K) := \mathbb{E}[(x-\overline{x})(x-\overline{x})^T]$. Let $\frac{5}{2}r_{in}^2$ be greater than the smallest eigenvalue of this (positive definite) matrix, as is the case in our setting. Then, $\frac{V}{S} < 4r_{in}$.*

Define $F_t := \sqrt{\frac{\pi}{t}}|\mu_{\partial K}|$.

Lemma 6.2.2 (Lemma 5.4 in [5]). *Suppose K contains a ball of radius r_{in}, (as is the case in our setting) then $S\left(1 - \frac{d\sqrt{\pi t}}{2r_{in}}\right) < F_t$.*

Applying Lemma 6.2.2, we see that

$$F_t > (1 - O(\epsilon))S.$$

The probability that Ctry succeeds in one trial is

$$\begin{aligned}
\mathbb{P}[\mathsf{Ctry}\ (t) \neq \emptyset] &= \sqrt{\frac{t}{\pi}}\frac{F_t}{V} && (6.2)\\
&> \sqrt{\frac{t}{\pi}}\frac{S}{V}(1 - O(\epsilon)) && (6.3)\\
&> \sqrt{\frac{t}{\pi}}\frac{1 - O(\epsilon)}{4r_{in}} \text{ (By Lemma 6.2.1)} && (6.4)\\
&> \Omega(\frac{\epsilon}{d}). && (6.5)
\end{aligned}$$

Therefore the expected number of calls to $\mathcal{B}$ and the membership oracle is $O^*(\frac{d}{\epsilon})$. By results of Lovász and Vempala ([60]) this number of random samples can be obtained using $O^*(\frac{d^4}{\epsilon})$ calls to the membership oracle. □

6.2.4 Extensions

S. Vempala [101] has remarked that these results can be extended more generally to sampling certain subsets of the surface ∂K of a convex body such as $\partial K \cap H$ for a halfspace H. In this case $K \cap H$ is convex too, and so Csample can be run on $K \cap H$. In order to obtain complexity guarantees, it is sufficient to bound from below, by a constant, the probability that Csample run on $H \cap K$ outputs a sample from $\partial K \cap H$ rather than $\partial H \cap K$. This follows from the fact that $\partial H \cap K$ is the unique minimal surface spanning $\partial K \cap \partial H$ and so has a surface area that is less than that of $\partial K \cap H$.

6.3 Sampling Well Conditioned Hypersurfaces

6.3.1 Preliminaries and Notation

Definition 6.3.1 (Condition Number)**.** *Let $\mathcal{M}$ be a smooth $d-$dimensional submanifold of $\mathbb{R}^m$. We define the condition number $c(\mathcal{M})$ to be $\frac{1}{\tau}$, where τ*

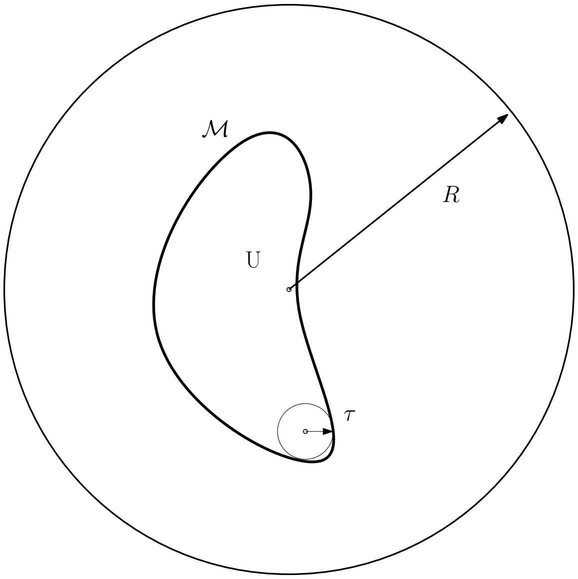

Figure 6.1: Condition number of a hypersurface

is the largest number to have the property that for any r < τ no two normals of length r that are incident on $\mathcal{M}$ at different points intersect.

In fact $\frac{1}{\tau}$ is an upper bound on the curvature of $\mathcal{M}$ ([77]). In this chapter, we shall restrict attention to a τ-conditioned manifold $\mathcal{M}$ that is also the boundary of a compact subset $U \in \mathbb{R}^d$.

Suppose we have access to a Black-Box $\mathcal{B}$ that produces i.i.d random points $x_1, x_2, \ldots$ from the uniform probability distribution on U. We shall describe a simple procedure to generate almost uniformly distributed points on $\mathcal{M}$.

6.3.2 Algorithm Msample

The input to Msample is an error parameter ϵ, a guarantee τ on the condition number of $\mathcal{M}$ and a Black-Box $\mathcal{B}$ that generates i.i.d random points from the uniform distribution on U as specified earlier. We are also provided with a membership oracle to U, of which $\mathcal{M}$ is the boundary. We shall assume that U is contained in a Euclidean ball of radius R, B_R. Msample , like Csample is a Las Vegas algorithm.

Let the probability measure of the output be $\tilde{\mu}_{out}$. The following is the main theorem of this section. Note that given perfectly random samples from U, the output probability density is close to the uniform in the $L_\infty-$norm, which is *stronger* than a total variation distance bound, and the number of calls to the Black box $\mathcal{B}$ is *independent* of dimension.

Theorem 6.3.1. *Let $\mathcal{M}$ be a τ-conditioned hypersurface that is the boundary of an open set contained in a ball of radius R. Let $\tilde{\mu}_{out}$ be the distribution of the output of Msample .*

1. *Let $\tilde{\lambda}_{\mathcal{M}}$ be the uniform probability measure on $\mathcal{M}$. Then, for any subset Δ of $\mathcal{M}$, the probability measure $\tilde{\mu}_{out}$ satisfies*

$$1 - O(\epsilon) < \frac{\tilde{\mu}_{out}(\Delta)}{\tilde{\lambda}_{\mathcal{M}}(\Delta)} < 1 + O(\epsilon).$$

2. *The total expected number of calls to $\mathcal{B}$ and the membership oracle of U is $O(\frac{R(1+\frac{2}{d}\ln\frac{1}{\epsilon})}{\tau\sqrt{\epsilon}})$.*

```
Msample
  1. Set $\sqrt{t} := \frac{\tau\sqrt{\epsilon}}{4(d+2\ln\frac{1}{\epsilon})}$.
  2. Set $p =$ Mtry $(t)$ .
  3. If $p = \emptyset$, goto (2). Else output $p$.
```

6.3.3 Correctness

Proof of part (1) of Theorem 6.3.1: We shall define a measure $\mu_{\mathcal{M}}$ on $\mathcal{M}$ related to the "local heat flow" out of small patches. Formally, if Δ a subset of $\mathcal{M}$, the measure assigned

Mtry (t)

1. Use $\mathcal{B}$ to generate a point x from U.
2. Generate a point $y := Gaussian(x, 2tI)$ from a spherical d-dimensional Gaussian of mean x and covariance matrix $2tI$.
3. If $y \in U$ output $\emptyset$.
   Else output an arbitrary element of $\overline{xy} \cap \mathcal{M}$ using binary search. (Unlike the convex case, $|\overline{xy} \cap \mathcal{M}|$ is no longer only 0 or 1.)

to it by $\mu_{\mathcal{M}}$ is

$$\mu_{\mathcal{M}}(\Delta) := \int_{x \in U} \int_{y \in \mathbb{R}^d \setminus U} G^t(x, y) \mathcal{I}\left[\overline{xy} \cap \Delta \neq \emptyset\right] d\lambda(x) d\lambda(y) \tag{6.6}$$

where $\mathcal{I}$ is the indicator function and $G^t(x, y)$ is the spherical Gaussian kernel with covariance matrix $2tI$. For comparison, we shall define μ_{out} by

$$\mu_{out} := V\tilde{\mu}_{out}\mathbb{P}[\mathsf{Mtry}\ (t) \neq \emptyset].$$

Since Msample outputs at most one point even when $|\overline{xy} \cap \mathcal{M}| > 1$, we see that for all $\Delta \subseteq \mathcal{M}$,

$$\mu_{out}(\Delta) \leq \mu_{\mathcal{M}}(\Delta).$$

The following Lemma provides a uniform *upper* bound on the Radon-Nikodym derivative of $\mu_{\mathcal{M}}$ with respect to the induced Lebesgue measure on $\mathcal{M}$.

Lemma 6.3.1. *Let $\lambda_{\mathcal{M}}$ be the measure induced on $\mathcal{M}$ by the Lebesgue measure λ on $\mathbb{R}^d$. Then*

$$\frac{d\mu_{\mathcal{M}}}{d\lambda_{\mathcal{M}}} < \sqrt{\frac{t}{\pi}}.$$

The Lemma below gives a uniform *lower* bound on $\frac{d\mu_{out}}{d\lambda_{\mathcal{M}}}$.

Lemma 6.3.2. *Let* $\sqrt{t} = \frac{\tau\sqrt{\epsilon}}{4(d+2\ln\frac{1}{\epsilon})}$. *Then*

$$\frac{d\mu_{out}}{d\lambda_{\mathcal{M}}} > \sqrt{\frac{t}{\pi}}(1 - O(\epsilon)).$$

Together the above Lemmas prove the first part of the Theorem. Their proofs have been provided below.

6.3.4 Complexity

Proof of part (2) of Theorem 6.3.1: Let S be the surface area of U (or the $d-1$-dimensional volume of $\mathcal{M}$.) Let V be the d-dimensional volume of U. We know that $U \subseteq B_R$. Since of all bodies of equal volume, the sphere minimizes the surface area, and $\frac{S}{V}$ decreases as the body is dilated,

$$\frac{S}{V} \geq \frac{d}{R}.$$

Lemma 6.3.2 implies that

$$\begin{aligned} \mathbb{P}[\mathsf{Mtry}\ (t) \neq \emptyset] &> \frac{S\sqrt{\frac{t}{\pi}}(1 - O(\epsilon))}{V} && (6.7)\\ &\geq \frac{d}{R}\frac{\tau\sqrt{\epsilon}(1 - O(\epsilon))}{8(d + 2\ln\frac{1}{\epsilon})} && (6.8)\\ &= \Omega(\frac{\tau\sqrt{\epsilon}}{R(1 + \frac{2}{d}\ln\frac{1}{\epsilon})}). && (6.9)\end{aligned}$$

This completes the proof. □

In our proofs of Lemma 6.3.1 and Lemma 6.3.2, we shall use the following Theorem of C. Borell.

Theorem 6.3.2 (Borell, [11])**.** *Let* $\mu_t = G^t(0, \cdot)$ *be the* d*-dimensional Gaussian measure with mean* 0 *and covariance matrix* $2It$*. Let* A *be any measurable set in* $\mathbb{R}^d$ *such that* $\mu(A) = \frac{1}{2}$.

Let A_ϵ be the set of points at a distance $\geq \epsilon$ from A. Then, $\mu_t(A_\epsilon) \geq 1 - e^{\frac{-\epsilon^2}{4t}}$.

Fact: With μ_t as above, and $B(R)$ the Euclidean ball of radius R centered at 0, $\frac{1}{2} < \mu_t(B(\sqrt{2dt}))$.

Proof of Lemma 6.3.1: Let H be a halfspace and ∂H be its hyperplane boundary. Halfspaces are invariant under translations that preserve their boundaries. Therefore for any halfspace H, $\mu_{\partial H}$ is uniform on ∂H. Noting that the image of a Gaussian under a linear transformation is a Gaussian, it is sufficient to consider the 1-dimensional case to compute the $d-1$-dimensional density $\frac{d\mu_{\partial H}}{d\lambda_{\partial H}}$.

$$\frac{d\mu_{\partial H}}{d\lambda_{\partial H}} = \int_{\mathbb{R}^-} \int_{\mathbb{R}^+} G^t(x, y) d\lambda(x) d\lambda(y), \tag{6.10}$$

which evaluates to $\sqrt{\frac{t}{\pi}}$ by a direct calculation. For any $z \in \mathcal{M}$, let H_z be the halfspace with the same outer normal as U such that ∂H_z is tangent to $\mathcal{M}$ at z. Let Δ be a small neighborhood of z in $\mathbb{R}^d$, and $|\Delta|$ denote its diameter.

$$\begin{aligned}
\frac{d\mu_{\mathcal{M}}}{d\lambda_{\mathcal{M}}}(z) &= \lim_{|\Delta| \to 0} \frac{\int_{x \in U} \int_{y \in \mathbb{R}^d \setminus U} G^t(x, y)\, \mathcal{I}\left[\overline{xy} \cap \Delta \neq \emptyset\right]\, d\lambda(x)\, d\lambda(y)}{\lambda_{\mathcal{M}}(\Delta)} \\
&= \lim_{|\Delta| \to 0} \frac{\int_{x \in \mathbb{R}^d} \int_{y \in \mathbb{R}^d} G^t(x, y)\, \mathcal{I}\left[\overline{xy} \cap \Delta \neq \emptyset\right]\, \mathcal{I}[x \in U \text{ and } y \in \mathbb{R}^d \setminus U]\, d\lambda(x)\, d\lambda(y)}{\lambda_{\mathcal{M}}(\Delta)} \\
&< \lim_{|\Delta| \to 0} \frac{\int_{x \in \mathbb{R}^d} \int_{y \in \mathbb{R}^d} G^t(x, y)\, \mathcal{I}\left[\overline{xy} \cap \Delta \neq \emptyset\right]\, d\lambda(x)\, d\lambda(y)}{2\lambda_{\mathcal{M}}(\Delta)} \\
&= \frac{d\mu_{\partial H_z}}{d\lambda_{\partial H_z}}(z) \\
&= \sqrt{\frac{t}{\pi}}.
\end{aligned}$$

The inequality in the above array of equations is strict because U is bounded. □

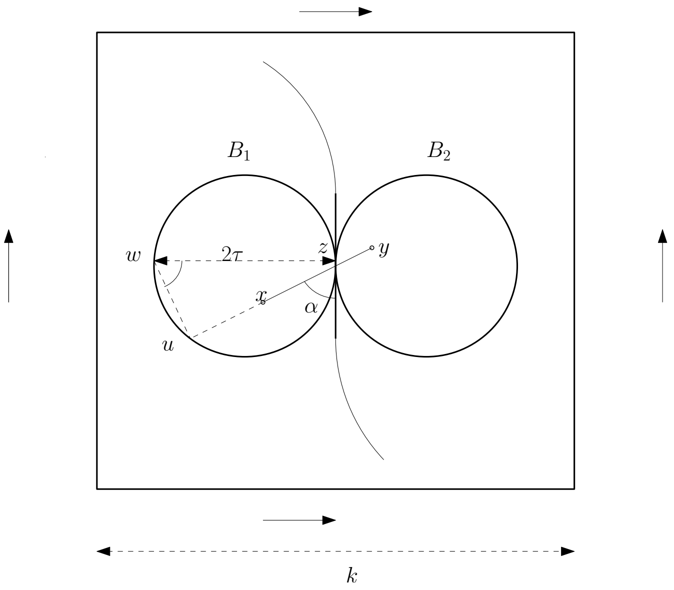

Figure 6.2: A transition $x \to y$ intersecting the hypersurface

Proof of Lemma 6.3.2: Let Δ be a small neighborhood of z in $\mathbb{R}^d$. Since $\mathcal{M}$ is a τ-conditioned manifold, for any $z \in \mathcal{M}$, there exist two balls $B_1 \subseteq U$ and $B_2 \subseteq \mathbb{R}^d \setminus U$ of radius τ that are tangent to $\mathcal{M}$ at z.

$$\frac{d\mu_{out}}{d\lambda_{\mathcal{M}}}(z) \quad > \quad \lim_{|\Delta| \to 0} \frac{\int_{x \in B_1} \int_{y \in B_2} G^t(x,y)\, \mathcal{I}\left[\overline{xy} \cap \Delta \neq \emptyset\right]\, d\lambda(x)\, d\lambda(y)}{\lambda_{\mathcal{M}}(\Delta)}.$$

The above is true because $|\overline{xy} \cap \mathcal{M}| = 1$ if $x \in B_1$ and $y \in B_2$. Let us define

$$\mathbb{P}_\tau := \lim_{|\Delta| \to 0} \frac{\int_{x \in B_1} \int_{y \in B_2} G^t(x,y)\, \mathcal{I}\left[\overline{xy} \cap \Delta \neq \emptyset\right]\, d\lambda(x)\, d\lambda(y)}{\int_{x \in H_z} \int_{y \in \mathbb{R}^d \setminus H_z} G^t(x,y)\, \mathcal{I}\left[\overline{xy} \cap \Delta \neq \emptyset\right]\, d\lambda(x)\, d\lambda(y)}. \tag{6.11}$$

Then

$$\mathbb{P}_\tau < \sqrt{\frac{\pi}{t}} \frac{d\mu_{out}}{d\lambda_{\mathcal{M}}}(z).$$

The proof now follows from

Lemma 6.3.3. *$\mathbb{P}_\tau > 1 - O(\epsilon)$.* □

Proof of Lemma 6.3.3: In order to obtain bounds on $\mathbb{P}_\tau$, we shall follow the strategy of mapping the picture onto a sufficiently large torus and doing the computations on this torus. This has the advantage that now averaging arguments can be used over the torus by virtue of its being compact (and a symmetric space.) These arguments do not transfer to $\mathbb{R}^d$ in particular because it is not possible to pick a point uniformly at random on $\mathbb{R}^d$. Consider the natural surjection

$$\phi_k : \mathbb{R}^d \to \mathbb{T}_k \tag{6.12}$$

onto a d dimensional torus of side k for $k >> \max(diam(U), \sqrt{t})$. For each point $p \in \mathbb{T}_k$, the fibre $\phi_k^{-1}(p)$ of this map is a translation of $k\mathbb{Z}^d$.

Let x be the origin in $\mathbb{R}^d$, and $e_1, \ldots, e_d$ be the canonical unit vectors. For a fixed k, let

$$\Xi_k := \phi_k(\kappa e_1 + span(e_2, \ldots, e_d)),$$

where κ is a random number distributed uniformly in $[0, k)$, be a random $d-1$-dimensional torus aligned parallel to $\phi_k(span(e_2, \ldots, e_k))$. Let $y := (y_1, \ldots, y_d)$ be chosen from a spherical d-dimensional Gaussian in $\mathbb{R}^d$ centered at 0 having covariance $2tI$.

Define $\mathbb{P}_\tau^{(k)}$ to be

$$\mathbb{P}_\tau^{(k)} := \mathbb{P}[y_2^2 + \ldots + y_d^2 < |y_1|\tau < \tau^2 \big| 1 = |\phi_k(\overline{xy}) \cap \Xi_k|] \tag{6.13}$$

It makes sense to define B_1 and B_2 on Ξ_k exactly as before i. e. tangent to Ξ_k at $\phi_k(\overline{xy}) \cap \Xi_k$ oriented so that B_1 is nearer to x than B_2 in geodesic distance. For geometric reasons, $\tilde{\mathbb{P}}_\tau^{(k)}$ is a lower bound on the probability that, even when the line segment $\overline{xy}$ in Figure 6.2 is slid along itself to the right until x occupies the position where z is now, y does not leave B_2. Figure 6.3 illustrates ball B_2 being slid, which is equivalent. In particular, this event would

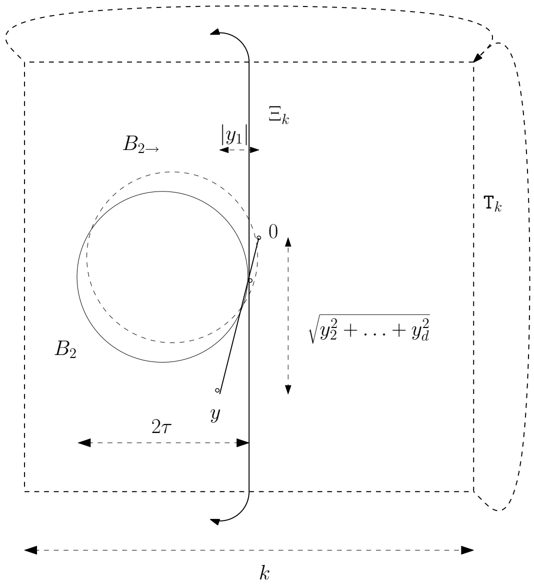

Figure 6.3: Sliding B_2

imply that $x \in B_1$ and $y \in B_2$.

$$\limsup_{k\to\infty} \mathbb{P}_\tau^{(k)} \le \mathbb{P}_\tau.$$

In the light of the above statement, it suffices to prove that for all sufficiently large k,

$$\mathbb{P}_\tau^{(k)} > 1 - O(\epsilon)$$

which will be done in Lemma 6.3.4. This completes the proof of this proposition. □

Lemma 6.3.4. *For all sufficiently large k,*

$$\mathbb{P}_\tau^{(k)} > 1 - O(\epsilon).$$

Proof: Recall that x is the origin and that $y := (y_1, \ldots, y_d)$ is Gaussian$(0, 2tI)$. Denote by E_k the event that

$$|\phi_k(\overline{xy}) \cap \Xi_k| = 1.$$

We note that

$$\mathbb{P}[E_k \mid y_1 = s] = \frac{|s|}{k}\mathcal{I}[|s| < k].$$

By Bayes' rule,

$$\rho[y_1 = s \,|\, E_k]\ \mathbb{P}[E_k] = \frac{|s|}{k}\left(\frac{e^{-s^2/4t}}{\sqrt{4\pi t}}\right)\mathcal{I}[|s| < k],$$

where $\mathcal{I}$ denotes the indicator function. In other words, there exists a constant $c_k := \frac{\mathbb{P}[E_k]^{-1}}{\sqrt{4\pi t}}$ such that

$$\rho[y_1 = s \,|\, |\Xi_k \cap \phi_k(\overline{xy})| = 1] \ = \ c_k \frac{|s|}{k}\, e^{-s^2/4t}\mathcal{I}[|s| < k].$$

A calculation tells us that

$$c_k \sim \frac{k}{4t}.$$

Let

$$\mathcal{I}_\tau := \mathcal{I}\left[\tau |\, y_1| > y_2^2 + \ldots + y_d^2\right]\mathcal{I}[|y_1| < \tau]\mathcal{I}[E_k].$$

By their definitions , $\mathbb{E}[\mathcal{I}_\tau | E_k] = \mathbb{P}_\tau^{(k)}$. Define

$$\mathcal{I}_{\parallel} := \mathcal{I}\left[|\, y_1| \not\in [\sqrt{\epsilon t}, \tau]\right]\mathcal{I}\left[E_k\right],$$

and

$$\mathcal{I}_\perp := \mathcal{I}\left[y_2^2 + \ldots y_d^2 > 4t(d + 2\ln\frac{1}{\epsilon})\right]\mathcal{I}\left[E_k\right].$$

A direct calculation tells us that $\mathbb{E}[\mathcal{I}_{\parallel} | E_k] = O(\epsilon)$. Similarly $\mathbb{E}[\mathcal{I}_\perp | E_k] = O(\epsilon)$ follows from Theorem 6.3.2 and the fact mentioned below it. This Lemma is implied by the following claim. □

Claim 6.3.1.

$$\mathcal{I}_\tau \geq \mathcal{I}[E_k] - \mathcal{I}_{\parallel} - \mathcal{I}_\perp.$$

Proof:

$$\begin{aligned}
\mathcal{I}_\perp &= \mathcal{I}\left[y_2^2 + \ldots + y_d^2 > 4t(d + 2\ln\frac{1}{\epsilon})\right]\mathcal{I}\left[E_k\right] \\
&= \mathcal{I}\left[y_2^2 + \ldots + y_d^2 > \tau\sqrt{\epsilon t}\right]\mathcal{I}\left[E_k\right]
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathcal{I}[E_k] - \mathcal{I}_{||} - \mathcal{I}_\perp &\leq \mathcal{I}[E_k]\,[y_2^2 + \ldots + y_d^2 < \tau\sqrt{\epsilon t} < \tau|y_1|\,]\,\mathcal{I}[|y_1| < \tau] \qquad (6.14) \\
&\leq \mathcal{I}_\tau \qquad (6.15)
\end{aligned}$$

□

CHAPTER 7

A GEOMETRIC INTERPRETATION OF HALFPLANE CAPACITY

Following Schramm's seminal paper [92] "Scaling limits of loop-erased random walks and uniform spanning trees," important progress has been made towards understanding the conformal invariance of the scaling limits of several two dimensional lattice models in statistical physics by several researchers including Lawler, Werner and Smirnov [54, 88]. These limits have been described using a new tool known as Schramm-Loewner Evolution (SLE). The chordal Schramm-Loewner evolution with parameter $\kappa \geq 0$ is the random collection of conformal maps satisfying the following stochastic differential equation:

$$\dot{g}_t(z) = \frac{2}{g_t(z) - \sqrt{\kappa} B_t}, \; g_0(z) = z,$$

where z belongs to the upper half plane $\mathbb{H}$. Denoting the domain of g_t by H_t, we obtain a random collection of continuously growing hulls $K_t := \mathbb{H} \setminus H_t$. In this parametrization, the halfplane capacity of K_t is equal to $2t$ [51]. Thus, halfplane capacity is a quantity arising naturally in the context of SLE.

Suppose A is a bounded, relatively closed subset of the upper half plane $\mathbb{H}$. We call A a compact $\mathbb{H}$-hull if A is bounded and $\mathbb{H} \setminus A$ is simply connected. The *halfplane capacity* of A, $\operatorname{hcap}(A)$, is defined in a number of equivalent ways (see [51], especially Chapter 3). If g_A denotes the unique conformal transformation of $\mathbb{H} \setminus A$ onto $\mathbb{H}$ with $g_A(z) = z + o(1)$ as $z \to \infty$, then g_A has the expansion

$$g_A(z) = z + \frac{\operatorname{hcap}(A)}{z} + O(|z|^{-2}), \quad z \to \infty.$$

Equivalently, if B_t is a standard complex Brownian motion and $\tau_A = \inf\{t \geq 0 : B_t \notin \mathbb{H}\backslash A\}$,

$$\text{hcap}(A) = \lim_{y\to\infty} y\, \mathbb{E}^{iy}\left[\Im(B_{\tau_A})\right].$$

Let $\Im[A] = \sup\{\Im(z) : z \in A\}$. Then if $y \geq \Im[A]$, we can also write

$$\text{hcap}(A) = \frac{1}{\pi}\int_{-\infty}^{\infty} \mathbb{E}^{x+iy}\left[\Im(B_{\tau_A})\right]\, dx.$$

These last two definitions do not require $\mathbb{H} \setminus A$ to be simply connected, and the latter definition does not require A to be bounded but only that $\Im[A] < \infty$.

For $\mathbb{H}$-hulls (that is, for A for which $\mathbb{H} \setminus A$ is simply connected), the halfplane capacity is comparable to a more geometric quantity that we define. This fact is not new (Gregory Lawler learned the fact from Oded Schramm in oral communication), but we do not know of a proof in the literature. In this note, we prove the fact giving (nonoptimal) bounds on the constant. We start with the definition of the geometric quantity.

Definition 7.0.2. *For an $\mathbb{H}$-hull A, let* hsiz(A) *be the 2-dimensional Lebesgue measure of the union of all balls centered at points in A that are tangent to the real line. In other words*

$$\text{hsiz}(A) = \text{area}\left[\bigcup_{x+iy\in A} \mathcal{B}(x+iy, y)\right],$$

where $\mathcal{B}(z,\epsilon)$ denotes the disk of radius ϵ about z.

In this chapter, we prove the following.

Theorem 7.0.3. *For every $\mathbb{H}$-hull A,*

$$\frac{1}{66}\,\text{hsiz}(A) < \text{hcap}(A) \leq \frac{7}{2\pi}\,\text{hsiz}(A).$$

We first comment that it suffices to prove the result for weakly bounded $\mathbb{H}$-hulls by which we mean $\mathbb{H}$-hulls A with $\Im(A) < \infty$ and such that for each $\epsilon > 0$, the set $\{x + iy : y > \epsilon\}$ is bounded. Indeed, for $\mathbb{H}$-hulls that are not weakly bounded, it is easy to verify that $\text{hsiz}(A) = \text{hcap}(A) = \infty$.

We start with a simple limit that is implied but not explicitly stated in [51].

Lemma 7.0.5. *If A is an $\mathbb{H}$-hull, then*

$$\text{hcap}(A) \geq \frac{\Im[A]^2}{2}. \tag{7.1}$$

Proof. It suffices to prove the result for hulls of the form $A = \eta(0, T]$ where η is a simple curve with $\eta(0+) \in \mathbb{R}$ parametrized so that $\text{hcap}[\eta(0, t]) = 2t$. In particular, $T = \text{hcap}(A)/2$. If $g_t = g_{\eta(0,t]}$, then g_t satisfies the Loewner equation

$$\partial_t g_t(z) = \frac{2}{g_t(z) - U_t}, \quad g_0(z) = z, \tag{7.2}$$

where $U : [0, T] \to \mathbb{R}$ is continuous. Suppose $\Im(z)^2 > 2\ \text{hcap}(A)$ and let $Y_t = \Im[g_t(z)]$. Then (7.2) gives

$$-\partial_t Y_t^2 \leq \frac{4Y_t}{|g_t(z) - U_t|^2} \leq 4,$$

which implies

$$Y_T^2 \geq Y_0^2 - 4T > 0.$$

This implies that $z \notin A$, and hence $\Im[A]^2 \leq 2\ \text{hcap}(A)$. □

The next lemma is a standard covering lemma. If $c > 0$ and $z = x + iy \in \mathbb{H}$, let

$$\mathcal{I}(z, c) = (x - cy, x + cy),$$

$$\mathcal{R}(z, c) = \mathcal{I}(z, c) \times (0, y] = \{x' + iy' : |x' - x| < cy, 0 < y' \leq y\}.$$

Lemma 7.0.6. *Suppose A is a weakly bounded $\mathbb{H}$-hull and $c > 0$. Then there exists a (finite or countable infinite) sequence of points $\{z_1 = x_i + iy_1, z_2 = x_2 + iy_2, , \ldots\} \subset A$ such that:*

- $y_1 \geq y_2 \geq y_3 \geq \cdots$;
- *the intervals $\mathcal{I}(x_1, c), \mathcal{I}(x_2, c), \ldots$ are disjoint;*
- $$A \subset \bigcup_{j=1}^{\infty} \mathcal{R}(z_j, 2c). \tag{7.3}$$

Proof. We define the points recursively. Let $A_0 = A$ and given $\{z_1, \ldots, z_j\}$, let

$$A_j = A \setminus \left[\bigcup_{k=1}^{j} \mathcal{R}(z_j, 2c) \right].$$

If $A_j = \emptyset$ we stop, and if $A_j \neq \emptyset$,we choose $z_{j+1} = x_{j+1}+iy_{j+1} \in A$ with $y_{j+1} = \Im[A_j]$. Note that if $k \leq j$, then $|x_{j+1} - x_k| \geq 2\,c\,y_k \geq c\,(y_k + y_{j+1})$ and hence $\mathcal{I}(z_{j+1}, c) \cap \mathcal{I}(z_k, c) = \emptyset$. Using the weak boundedness of A, we can see that $y_j \to 0$ and hence (7.3) holds. □

Lemma 7.0.7. *For every $c > 0$, let*

$$\rho_c := \frac{2\sqrt{2}}{\pi} \arctan\left(e^{-\theta}\right), \quad \theta = \theta_c = \frac{\pi}{4c}.$$

Then, the following is true. Suppose A is a weakly bounded $\mathbb{H}$-hull and $z = x_0 + iy_0 \in A$ with $y = \Im(A)$. Then

$$\operatorname{hcap}(A) \geq \rho_c^2\, y_0^2 + \operatorname{hcap}\left[A \setminus \mathcal{R}(z, 2c)\right].$$

Proof. By scaling and invariance under real translation, we may assume that $\Im[A] = y_0 = 1$ and $x_0 = 0$. Let $S = S_c$ be defined to be the set of all points z of the form $x + iuy$ where $x + iy \in A \setminus \mathcal{R}(i, 2c)$ and $0 < u \leq 1$.

Note that S is an $\mathbb{H}$-hull and that $S \cap A = A \setminus \mathcal{R}(i, 2c)$.

Using the capacity relation [51, (3.10)]

$$\operatorname{hcap}(A_1 \cup A_2) - \operatorname{hcap}(A_2) \leq \operatorname{hcap}(A_1) - \operatorname{hcap}(A_1 \cap A_2),$$

we see that

$$\operatorname{hcap}(S \cup A) - \operatorname{hcap}(S) \leq \operatorname{hcap}(A) - \operatorname{hcap}(S \cap A).$$

Hence, it suffices to show that

$$\operatorname{hcap}(S \cup A) - \operatorname{hcap}(S) \geq \rho_c^2.$$

Let f be the conformal map of $\mathbb{H} \setminus S$ onto $\mathbb{H}$ such that $z - f(z) = o(1)$ as $z \to \infty$. Let $S^* := S \cup A$. By properties of halfplane capacity [51, (3.8)] and (7.1),

$$\operatorname{hcap}(S^*) - \operatorname{hcap}(S) = \operatorname{hcap}[f(S^* \setminus S)] \geq \frac{\Im[f(i)]^2}{2}.$$

Hence, it suffices to prove that

$$\Im[f(i)] \geq \sqrt{2}\,\rho = \frac{4}{\pi} \arctan\left(e^{-\theta}\right). \tag{7.4}$$

By construction, $S \cap \mathcal{R}(z, 2c) = \emptyset$. Let $V = (-2c, 2c) \times (0, \infty) = \{x + iy : |x| < 2c, y > 0\}$ and let τ_V be the first time that a Brownian motion leaves the domain. Then [51, (3.5)],

$$\Im[f(i)] = 1 - \mathbb{E}^i\left[\Im(B_{\tau_S})\right] \geq \mathbb{P}\left\{B_{\tau_S} \in [-2c, 2c]\right\} \geq \mathbb{P}\left\{B_{\tau_V} \in [-2c, 2c]\right\}.$$

The map $\Phi(z) = \sin(\theta z)$ maps V onto $\mathbb{H}$ sending $[-2c, 2c]$ to $[-1, 1]$ and $\Phi(i) = i \sinh \theta$.

Using conformal invariance of Brownian motion and the Poisson kernel in $\mathbb{H}$, we see that

$$\mathbb{P}\left\{B_{\tau_V} \in [-2c, 2c]\right\} = \frac{2}{\pi} \arctan\left(\frac{1}{\sinh\theta}\right) = \frac{4}{\pi} \arctan\left(e^{-\theta}\right).$$

The second equality uses the double angle formula for the tangent. □

Lemma 7.0.8. *Suppose $c > 0$ and $x_1 + iy_1, x_2 + iy_2, \ldots$ are as in Lemma 7.0.6. Then*

$$\operatorname{hsiz}(A) \leq [\pi + 8c] \sum_{j=1}^{\infty} y_j^2. \tag{7.5}$$

If $c \geq 1$, then

$$\pi \sum_{j=1}^{\infty} y_j^2 \leq \operatorname{hsiz}(A). \tag{7.6}$$

Proof. A simple geometry exercise shows that

$$\text{area}\left[\bigcup_{x+iy \in \mathcal{R}(z_j, 2c)} \mathcal{B}(x+iy, y)\right] = [\pi + 8c]\, y_j^2.$$

Since

$$A \subset \bigcup_{j=1}^{\infty} \mathcal{R}(z_j, 2c),$$

the upper bound in (7.5) follows. Since $c \geq 1$, and the intervals $\mathcal{I}(z_j, c)$ are disjoint, so are the disks $\mathcal{B}(z_j, y_j)$. Hence,

$$\text{area}\left[\bigcup_{x+iy \in A} \mathcal{B}(x+iy, y)\right] \geq \text{area}\left[\bigcup_{j=1}^{\infty} \mathcal{B}(z_j, y_j)\right] = \pi \sum_{j=1}^{\infty} y_j^2.$$

□

Proof of Theorem 7.0.3. Let $V_j = A \cap \mathcal{R}(z_j, c)$. Lemma 7.0.7 tells us that

$$\text{hcap}\left[\bigcup_{k=j}^{\infty} V_j\right] \geq \rho_c^2\, y_j^2 + \text{hcap}\left[\bigcup_{k=j+1}^{\infty} V_j\right],$$

and hence

$$\text{hcap}(A) \geq \rho_c^2 \sum_{j=1}^{\infty} y_j^2.$$

Combining this with the upper bound in (7.5) with any $c > 0$ gives

$$\frac{\text{hcap}(A)}{\text{hsiz}(A)} \geq \frac{\rho_c^2}{\pi + 8c}.$$

Choosing $c = \frac{8}{5}$ gives us

$$\frac{\text{hcap}(A)}{\text{hsiz}(A)} > \frac{1}{66}.$$

For the upper bound, choose a covering as in Lemma 7.0.6 with $c = 1$. Subadditivity and scaling give

$$\text{hcap}(A) \leq \sum_{j=1}^{\infty} \text{hcap}\left[\mathcal{R}(z_j, 2y_j)\right] = \text{hcap}[\mathcal{R}(i,2)] \sum_{j=1}^{\infty} y_j^2.$$

Combining this with the lower bound in (7.5) gives

$$\frac{\text{hcap}(A)}{\text{hsiz}(A)} \leq \frac{\text{hcap}[\mathcal{R}(i,2)]}{\pi}.$$

Note that $\mathcal{R}(i,2)$ is the union of two real translates of $\mathcal{R}(i,1)$, $\text{hcap}[\mathcal{R}(i,2)] \leq 2\ \text{hcap}[\mathcal{R}(i,1)]$ whose intersection is the interval $(0, i]$. Using the capacity relation [51, (3.10)]

$$\text{hcap}(A_1 \cup A_2) \leq \text{hcap}(A_1) + \text{hcap}(A_2) - \text{hcap}(A_1 \cap A_2),$$

we see that

$$\text{hcap}(\mathcal{R}(i,2)) \leq 2\ \text{hcap}(\mathcal{R}(i,1)) -\ \text{hcap}((o,i]) = 2\ \text{hcap}(\mathcal{R}(i,1)) - \frac{1}{2}.$$

But $\mathcal{R}(i,1)$ is strictly contained in $A' := \{z \in \mathbb{H} : |z| \leq \sqrt{2}\}$, and hence

$$\text{hcap}[\mathcal{R}(i,1)] <\ \text{hcap}(A') = 2.$$

The last equality can be seen by considering $h(z) = z + 2z^{-1}$ which maps $\mathbb{H} \setminus A'$ onto $\mathbb{H}$. Therefore,

$$\text{hcap}[\mathcal{R}(i,2)] < \frac{7}{2},$$

and hence

$$\frac{\text{hcap}(A)}{\text{hsiz}(A)} \leq \frac{7}{2\pi}.$$

□

CHAPTER 8

RANDOM WALKS ON MANIFOLDS

8.0.5 Markov Schemes on metric spaces

In this section, we present a general setup in which bounds can be obtained for the mixing times of Markov chains on metric spaces. This setup has been used earlier in a number of settings. The division of the argument to bound mixing time into an isoperimetric inequality and a relation between geometric and probabilistic distance that is presented here appeared for a specific metric and measure in [60]. There is recent interest in the question of sampling manifolds, for example in [19]. The techniques used in this chapter do not lead to bounds for specific Markov Chains on manifolds, rather they give bounds that eventually hold on Markov chains that are part of a sequence parameterized by a "step-size" that tends to zero. Our application to sampling manifolds, on the other hand gives mixing bounds for specific step-sizes that we can state explicitly.

We begin with the definition of a Markov Scheme given by Lovász and Simonovits [58].

Definition 8.0.3. *Let $(\Omega, \mathcal{A})$ be a $\sigma-$algebra. For every $u \in \Omega$, let P_u be a probability measure on Ω and assume that for every $A \in \mathcal{A}$, the value $P_u(A)$ is measurable as a function of u. We assume M is lazy, i. e. $P_x(x) \geq \frac{1}{2}$ and reversible, i. e. for any measurable $S_1, S_2 \subseteq \Omega$,*

$$\int_{S_1} P_x(S_2) d\mu(x) = \int_{S_2} P_y(S_1) d\mu(y). \tag{8.1}$$

We call the triple $(\Omega, \mathcal{A}, \{P_u : u \in \Omega\})$ a Markov Scheme.

A Markov chain is a process governed by a Markov Scheme, started from some initial probability distribution. The *conductance* of the Markov Scheme is

$$\Phi = \inf_{0<\mu(A)\leq\frac{1}{2}} \frac{\int_A P_u(\Omega \setminus A)d\mu(A)}{\mu(A)}.$$

Let $(\mathcal{M}, d_{\mathcal{M}})$ be a metric space. For every point $x \in \mathcal{M}$, let P_x be a transition probability distribution on $\mathcal{M}$, defining a Markov Scheme M whose stationary distribution is μ.

For $x, y \in \mathcal{M}$, let $d_{\mathcal{M}}(x, y)$ be the distance between x and y and for any sets S_1 and S_2, let

$$d_{\mathcal{M}}(S_1, S_2) := \inf_{x\in S_1, y\in S_2} d_{\mathcal{M}}(S_1, S_2). \tag{8.2}$$

Suppose the following conditions hold.

1. For any $x, y \in \mathcal{M}$,

$$d_{\mathcal{M}}(x, y) < \delta_{\mathcal{M}} \Rightarrow d_{TV}(P_x, P_y) < 1 - \epsilon_{\mathcal{M}}. \tag{8.3}$$

2. For any partition of $\mathcal{M}$ into disjoint parts S_1, S_2 and S_3, if $d_{\mathcal{M}}(S_1, S_2) \geq \delta_{\mathcal{M}}$, then

$$\mu(S_3) \geq \alpha_{\mathcal{M}} \min(\mu(S_1), \mu(S_2)). \tag{8.4}$$

Theorem 8.0.4. *Let S_1 be a measurable subsets of $\mathcal{M}$, whose stationary measure $\mu(S_1)$ is less or equal to $\frac{1}{2}$. Then,*

$$\int_{S_1} P_x(\mathcal{M} \setminus S_1)d\mu(x) \geq \frac{\epsilon_{\mathcal{M}} \min(\alpha_{\mathcal{M}}, 1)\mu(S_1)}{4}.$$

Proof. Let $S_2 := \mathcal{M} \setminus S_1$.

Let $S_1' = S_1 \cap \{x \big| P_x(S_2) < \frac{\epsilon_{\mathcal{M}}}{2}\}$ and $S_2' = S_2 \cap \{y \big| P_y(S_1) < \frac{\epsilon_{\mathcal{M}}}{2}\}$.

If $x \in S_1'$ and $y \in S_2'$ then $d_{TV}(P_x, P_y) > 1 - \epsilon_{\mathcal{M}}$. By (8.3), $d_{\mathcal{M}}(x, y) \geq \delta_{\mathcal{M}}$. However, by the isoperimetric inequality (8.4), we then have

$$\mu(\mathcal{M} \setminus (S_1' \cup S_2')) \geq \alpha_{\mathcal{M}} \, \min(\mu(S_1'), \mu(S_2')). \tag{8.5}$$

$\mathcal{M} \setminus (S_1' \cup S_2)$ consists, of those points in S_1, from which the probability of moving to the other part in one step of the Markov chain is greater or equal to $\frac{\epsilon_{\mathcal{M}}}{2}$. Similarly, $\mathcal{M} \setminus (S_1 \cup S_2')$ consists, precisely of those points in S_2, from which the probability of moving to the other part in one step of the Markov chain is greater than or equal to $\frac{\epsilon_{\mathcal{M}}}{2}$. By the reversibility of the Markov chain,

$$\int_{S_1} P_x(S_2) d\mu(x) = \int_{S_2} P_y(S_1) d\mu(y).$$

Therefore,

$$\begin{aligned} \int_{S_1} P_x(\mathcal{M} \setminus S_1) d\mu(x) &= \frac{\int_{S_1} P_x(\mathcal{M} \setminus S_1) d\mu(x) + \int_{\mathcal{M} \setminus S_1} P_y(S_1) d\mu(y)}{2} && (8.6) \\ &\geq \frac{\epsilon_{\mathcal{M}} \, \mu(\mathcal{M} \setminus (S_1' \cup S_2'))}{4}. && (8.7) \end{aligned}$$

We consider two cases separately. First, suppose $\mu(S_1') \geq \frac{\mu(S_1)}{2}$. Then, by (8.5),

$$\mu(\mathcal{M} \setminus (S_1' \cup S_2')) \geq \alpha_{\mathcal{M}} \, \min\left(\frac{\mu(S_1)}{2}, \mu(S_1) - \mu(\mathcal{M} \setminus (S_1' \cup S_2'))\right).$$

It follows that

$$\mu(\mathcal{M} \setminus (S_1' \cup S_2')) \geq \frac{\alpha_{\mathcal{M}} \, \mu(S_1)}{2}, \tag{8.8}$$

and therefore,

$$\int_{S_1} P_x(\mathcal{M} \setminus S_1) d\mu(x) \geq \frac{\epsilon_{\mathcal{M}} \alpha_{\mathcal{M}} \mu(S_1)}{4}. \tag{8.9}$$

Next, suppose $\mu(S_1') \leq \frac{\mu(S_1)}{2}$. Then,

$$\mu(\mathcal{M} \setminus (S_1' \cup S_2')) \geq \frac{\mu(S_1)}{2},$$

implying that

$$\int_{S_1} P_x(\mathcal{M} \setminus S_1) d\mu(x) \geq \frac{\epsilon_{\mathcal{M}} \mu(S_1)}{4}.$$

□

Now applying Theorem 9.3.5 due to Lovász and Simonovits, we have the following theorem.

Theorem 8.0.5. *Let μ_0 be the initial distribution for a lazy reversible ergodic Markov chain on a metric space satisfying the above conditions and μ_k be the distribution of the k^{th} step. Let $s := \sup_S \frac{\mu_0(S)}{\mu(S)}$ where the supremum is over all measurable subsets S of K. Then, for all such S,*

$$|\mu_k(S) - \mu(S)| \leq \sqrt{s}\left(1 - \frac{(\epsilon_{\mathcal{M}} \min(\alpha_{\mathcal{M}}, 1))^2}{32}\right)^k.$$

8.1 The case of manifolds

In this section, we will consider the special case of a manifold with density, specified by a collection of smoothly varying charts indexed by the points of the manifold. Let B be the Euclidean unit ball. Let $\mathcal{M}$ be a Riemannian manifold. For $x, y \in \mathcal{M}$, let $d_{\mathcal{M}}(x, y)$ be the geodesic distance between x and y. Let $\mathcal{M}$ be specified by an a family of injective maps

$\{U_x : B \to \mathcal{M}\}_{x \in \mathcal{M}}$ where $x \in U_x(B)$. Let $\rho(x) = e^{-V(x)}$ be a probability density function on $\mathcal{M}$, whose value at x is measured with respect to the push-forward of the Lebesgue measure via U_x at the point x. Let Jac denote the Jacobian of a map. An atlas can be constructed from this family of maps, namely $\{U_x^{-1}\}_{x \in \mathcal{M}}$. We consider the Markov Scheme in which, for any $x \in \mathcal{M}$, P_x is the distribution of the random point z obtained as follows.

1. Toss a fair coin and if $Heads$, set z to x.
2. If $Tails$, do the following:
   (a) Pick a random point $w \in B$ from the uniform measure on the unit ball and let $z := U_x(w)$.
   (b) If $x \in U_z(B)$,
      i. with probability $\min\left(1, \frac{\rho(z) \det \mathrm{Jac}(U_z^{-1} U_x)(0)}{\rho(x)}\right)$ let z remain unchanged.
      ii. Else, set z to x.
   (c) If $x \notin U_z(B)$, set z to x.
3. Output z.

8.1.1 A good atlas

Let the atlas $\{U_x^{-1}\}_{x \in \mathcal{M}}$ and V, the logarithm of the density function, satisfy the following. There exists $r_m > 0$ such that for any $\alpha \in (0, 1)$,

1.

$$d_{\mathcal{M}}(x, y) \leq \alpha e^{-\frac{1}{n}} r_{\mathcal{M}} \Rightarrow y \in U_x(\alpha B), \tag{8.10}$$

and

$$y' \in \alpha e^{-\frac{1}{n}} B \Rightarrow d_{\mathcal{M}}(x, U_x(y')) \leq \alpha r_{\mathcal{M}}. \tag{8.11}$$

2. On its domain of definition, $U_z^{-1}U_x$ has continuous partial derivatives upto order 3, and for all $z \in U_x(B)$,

$$\left|V(z) - V(x) + \ln \det \mathrm{Jac}(U_z^{-1}U_x)(0)\right| < 1. \tag{8.12}$$

For a manifold $\mathcal{M}$ equipped with a measure μ, let the Minkowski outer measure of a (measurable) set A be defined as

$$\mu^+(\partial A) := \lim_{\epsilon \to 0^+} \frac{\mu(A_\epsilon) - \mu(A)}{\epsilon}, \tag{8.13}$$

where $A_\epsilon := \{x | d_{\mathcal{M}}(x, A) < \epsilon\}$.

Definition 8.1.1. *The Cheeger constant of the weighted manifold $(\mathcal{M}, \mu)$ is*

$$\beta_{\mathcal{M}} = \inf_{A \subset \mathcal{M}, \mu(A) \leq \frac{1}{2}} \frac{\mu^+(\partial A)}{\mu(A)}, \tag{8.14}$$

where the infimum is taken over measurable subsets.

With the above terminology, we have the following theorem.

Theorem 8.1.1. *Let μ_0 be the initial distribution for the Markov chain whose transitions are made as above. Let μ_k be the distribution of the k^{th} step. Let $s := \sup_S \frac{\mu_0(S)}{\mu(S)}$ where the supremum is taken over all measurable subsets S of K. Then, for all such S,*

$$|\mu_k(S) - \mu(S)| \leq \sqrt{s}\left(1 - \frac{(\epsilon_{\mathcal{M}} \min(\frac{r_{\mathcal{M}}\beta_{\mathcal{M}}}{n}, 1))^2}{128e^{12}}\right)^k.$$

We will need two lemmas for the proof of this theorem. These lemmas appear below.

Lemma 8.1.1. *Let $C, D \subseteq \mathcal{M}$ and $d_{\mathcal{M}}(C, D) \geq \delta_{\mathcal{M}}$. Then,*

$$\mu(\mathcal{M} \setminus \{C \cup D\}) \geq 2\min(\mu(C), \mu(D))(e^{\frac{\beta_{\mathcal{M}}\delta_{\mathcal{M}}}{2}} - 1). \tag{8.15}$$

Proof. We will consider two cases.

First, suppose that $\max(\mu(C), \mu(D)) > \frac{1}{2}$. Without loss of generality, we assume that $\mu(C) \leq \mu(B)$. Then, let

$$\delta_1 := \sup_{\mu(C_\delta) < \frac{1}{2}} \delta.$$

We proceed by contradiction. Suppose for some $\beta < \beta_{\mathcal{M}}$,

$$\exists \delta \in [0, \delta_1), \mu(C_\delta) < e^{\beta\delta}\mu(C). \tag{8.16}$$

Let δ' be the infimum of such δ. Note that since $\mu(C_\delta)$ is a monotonically increasing function of δ,

$$\mu(C_{\delta'}) = e^{\beta\delta'}\mu(C).$$

However, we know that

$$\mu^+(\partial C_{\delta'}) := \lim_{\epsilon \to 0^+} \frac{\mu(C_\epsilon) - \mu(C)}{\epsilon} \geq \beta_{\mathcal{M}}, \tag{8.17}$$

which contradicts the fact that in any right neighborhood of δ', there is a δ for which (8.16) holds. This proves that for all $\delta \in [0, \delta_1)$, $\mu(C_\delta) \geq e^{\delta\beta_{\mathcal{M}}}\mu(C)$. We note that $C_{\delta_1} \cap D_{\delta_{\mathcal{M}} - \delta_1} = \emptyset$, therefore $\mu(\delta_{\mathcal{M}} - \delta_1) \leq \frac{1}{2}$. So the same argument tells us that

$$\mu(D_{\delta_{\mathcal{M}} - \delta_1}) \geq e^{\beta_{\mathcal{M}}(\delta_{\mathcal{M}} - \delta_1)}\mu(D). \tag{8.18}$$

Thus, $\mu(\mathcal{M} \setminus \{C \cup D\}) \geq \mu(C)(e^{\beta_{\mathcal{M}}(\delta_{\mathcal{M}} - \delta_1)} + e^{\beta_{\mathcal{M}}\delta_1} - 2)$. This implies that

$$\mu(\mathcal{M} \setminus \{C \cup D\}) \geq 2\mu(C)\left(e^{\frac{\beta_{\mathcal{M}}\delta_{\mathcal{M}}}{2}} - 1\right).$$

Next, suppose $\mu(D) \geq \frac{1}{2}$. We then set $\delta_1 := \delta_{\mathcal{M}}$, and see that the arguments from (8.16) to (8.18) carry through verbatim. Thus, in this case, $\mu(\mathcal{M}\backslash\{C\cup D\}) \geq \min(\mu(C),\mu(D))(e^{\beta_{\mathcal{M}}\delta_{\mathcal{M}}}-1)$.

□

Lemma 8.1.2. *If* $d_{\mathcal{M}}(x,y) < \frac{e^{-\frac{1}{n}}r_{\mathcal{M}}}{n}$, *then*

$$d_{TV}(P_x, P_y) \leq 1 - \frac{1}{2e^5}. \tag{8.19}$$

Proof. For any $x \in \mathcal{M}$, let $D_x := U_x(B)$. Let us fix the convention that $\frac{dP_y}{dP_x}(x) := 0$ and $\frac{dP_y}{dP_x}(y) := +\infty$. Suppose $x \to w$ is one step of the chain. Then,

$$d(P_x, P_y) = 1 - \mathbb{E}_w\left[\min\left(1, \frac{dP_y}{dP_x}(w)\right)\right].$$

By a direct computation,

$$\mathbb{E}_w\left[\min\left(1, \frac{dP_y}{dP_x}(w)\right)\right] \geq \min\left(1, \frac{\rho(x)\det \mathrm{Jac}(U_x^{-1}U_y)(0)}{\rho(y)}\right)\mathbb{P}\left[(y \in D_w)\wedge(w \in D_y\setminus\{x\})\right].$$

As a consequence of (8.10),

$$\mathbb{P}\left[(y \in D_w)\wedge(w \in D_y\setminus\{x\})\right] \geq \mathbb{P}\left[(y \in D_w)\wedge(U_y^{-1}(w) \in e^{-\frac{1}{n}}B\setminus\{0\})\right].$$

By the triangle inequality, and (8.11),

$$\begin{aligned} d_{\mathcal{M}}(w,y) &\leq d_{\mathcal{M}}(x,y) + d_{\mathcal{M}}(x,w) \\ &\leq d_{\mathcal{M}}(x,y) + e^{\frac{1}{n}}r_{\mathcal{M}}\|U_x^{-1}(w)\|. \end{aligned}$$

Let E_1 denote the event that $U_x^{-1}(w) \in e^{-\frac{3}{n}} B \setminus \{0\}$. Then, if $d_{\mathcal{M}}(x, y) \leq \frac{e^{-\frac{1}{n}} r_{\mathcal{M}}}{n}$,

$$\begin{aligned}
\mathbb{P}\left[d_{\mathcal{M}}(w, y) \leq r_{\mathcal{M}}(e^{-\frac{1}{n}}) \middle| E_1\right] &\geq \mathbb{P}\left[d_{\mathcal{M}}(x, y) + e^{\frac{1}{n}} r_{\mathcal{M}} \|U_x^{-1}(w)\| \leq r_{\mathcal{M}}(e^{-\frac{1}{n}}) \middle| E_1\right] \\
&\geq \mathbb{P}\left[e^{\frac{1}{n}} r_{\mathcal{M}} \|U_x^{-1}(w)\| \leq r_{\mathcal{M}}(e^{-\frac{2}{n}}) \middle| E_1\right] \\
&= 1.
\end{aligned}$$

Analyzing the transition probabilities of the chain, we see that $\mathbb{P}[E_1] \geq \frac{1}{2e^4}$. Therefore,

$$\begin{aligned}
\mathbb{P}\left[(y \in D_w) \wedge (E_1)\right] &= \mathbb{P}\left[(y \in D_w) \middle| E_1\right] \mathbb{P}[E_1] \\
&= \frac{1}{2e^4}.
\end{aligned}$$

As a consequence of (8.12),

$$\min\left(1, \frac{\rho(x) \det \mathrm{Jac}(U_x^{-1} U_y)(0)}{\rho(y)}\right) \geq \frac{1}{e}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_w\left[\min\left(1, \frac{dP_y}{dP_x}(w)\right)\right] &\geq \min\left(1, \frac{\rho(x) \det \mathrm{Jac}(U_x^{-1} U_y)(0)}{\rho(y)}\right) \mathbb{P}\left[(y \in D_w) \wedge (w \in D_y \setminus \{x\})\right] \\
&\geq \frac{1}{2e^5}.
\end{aligned}$$

□

Proof of Theorem 8.1.1. Theorem 8.0.5 gives mixing bounds for general metric spaces in terms of $\epsilon_{\mathcal{M}}$, $\alpha_{\mathcal{M}}$ and implicitly, $\delta_{\mathcal{M}}$. By Lemma 8.1.1 and Lemma 8.1.2, we see that in the present context we can set $\delta_{\mathcal{M}}$ to $\frac{r_{\mathcal{M}}}{ne^{\frac{1}{n}}}$, $\epsilon_{\mathcal{M}}$ to $\frac{1}{2e^5}$ and $\alpha_{\mathcal{M}}$ to $\frac{r_{\mathcal{M}} \beta_{\mathcal{M}}}{ne^{\frac{1}{n}}}$. The theorem follows by substituting these values into Theorem 8.0.5. □

CHAPTER 9

RANDOM WALKS ON POLYTOPES AND AN AFFINE INTERIOR POINT METHOD FOR LINEAR PROGRAMMING

9.1 Introduction

In this chapter, we use ideas from interior point algorithms to define a random walk on a polytope. We call this walk *Dikin walk* after I. I. Dikin, because it relies on ellipsoids introduced by Dikin [22] in 1967, which he used to develop an affine-scaling interior point algorithm for linear programming.

There is a large body of literature addressing algorithms for sampling convex bodies in $\mathbb{R}^n$. Previous sampling algorithms were applicable to convex sets specified in the following way. The input consists of an n-dimensional convex set K circumscribed around and inscribed in balls of radius r and R respectively. The algorithm has access to an oracle which, when supplied with a point in $\mathbb{R}^n$ answers “yes” if the point is in K and “no” otherwise.

The first polynomial time algorithm for sampling convex sets appeared in [25]. It did a random walk on a sufficiently dense grid. The dependence of its mixing time on the dimension was $O^*(n^{23})$. It resulted in the first randomized polynomial time algorithm to approximate the volume of a convex set.

Another random walk that has been analyzed for sampling convex sets is known as the ball walk, which does the following. Suppose the current point is x_i. y is chosen uniformly at random from a ball of radius δ centered at x_i. If $y \in K$, x_{i+1} is set to K; otherwise $x_{i+1} = x_i$. After many successive improvements over several papers, it was shown in [39] that a ball walk mixes in $O^*(n\frac{R^2}{\delta^2})$ steps from a warm start if $\delta < \frac{r}{\sqrt{n}}$. A ball walk has not been proved to mix rapidly from any single point. A third random walk analyzed recently is known as Hit-and-Run [57, 60]. This walk mixes in $O\left(n^3(\frac{R}{r})^2 \ln \frac{R}{d\epsilon}\right)$ steps from a point at a distance d from the boundary [60], where ϵ is the desired variation distance to stationarity.

9.2 Results

The Markov Chain defining Dikin walk is invariant under affine transformations of the polytope. Consequently, the complex interleaving of rounding and sampling present in previous sampling algorithms for convex sets (see [25, 39, 62]) is unnecessary.

Some features of Dikin walk are the following.

1. The measures defined by the transition probabilities of Dikin walk are affine invariants, so there is no dependence on R/r (where R is the radius of the smallest ball containing the polytope K and r is the radius of the largest ball contained in K).

2. If K is an n-dimensional polytope defined by m linear constraints, the mixing time of the Dikin walk is $O(nm)$ from a warm start (i. e. if the starting distribution has a density bounded above by a constant).

3. If the walk is started at the "analytic center" (which can be found efficiently by interior point methods [83, 96]), it achieves a variation distance of ϵ in $O\left(mn\left(n\log m + \log\frac{1}{\epsilon}\right)\right)$ steps. This is strongly polynomial in the description of the polytope.

Dikin walk is similar to ball walk except that Dikin ellipsoids (defined later) are used instead of balls. Dikin walk is the first walk to mix in *strongly polynomial* time from a central point such as the center of mass (for which s, as defined below, is $O(n)$) and the analytic center (for which $s = O(m)$). Our main result related to the Dikin walk is the following.

Theorem 9.2.1. *Let n be greater than some universal constant. Let K be an n-dimensional polytope defined by m linear constraints and $x_0 \in K$ be a point such that s is the supremum over all chords $\overline{pq}$ passing through x_0 of $\frac{|p-x_0|}{|q-x_0|}$ and $\epsilon > 0$ be the desired variation distance to the uniform distribution. Let $\tau > 7 \times 10^8 \times mn\left(n\ln\left(20\, s\sqrt{m}\right) + \ln\left(\frac{32}{\epsilon}\right)\right)$ and $x_0, x_1, \ldots$*

be a Dikin walk. Then, for any measurable set $S \subseteq K$, the distribution of $x_{\mathcal{T}}$ satisfies $\left|\mathbb{P}[x_{\mathcal{T}} \in S] - \frac{\mathrm{vol}(S)}{\mathrm{vol}(K)}\right| < \epsilon$.

Running times

The mixing time for Hit-and-Run from a warm start is $O\left(\frac{n^2R^2}{r^2}\right)$, while for Dikin walk this is $O(mn)$. Hit-and-Run takes more random walk steps to provably mix on any class of polytopes where $m = o\left(\frac{nR^2}{r^2}\right)$. For polytopes with polynomially many faces, R/r cannot be $O\left(n^{\frac{1}{2}-\epsilon}\right)$ (but can be arbitrarily larger). Thus, $m = o(n\left(\frac{R}{r}\right)^2)$ holds true for some important classes of polytopes, such as those arising from the question of sampling contingency tables with fixed row and column sums (where $m = O(n)$). Each step of Dikin walk can be implemented using $O(mn^{\gamma-1})$ arithmetic operations, $\gamma < 2.376$ being the exponent of matrix multiplication (see 9.3.1). One step of Hit-and-Run implemented naively would need $O(mn)$ arithmetic operations. Evaluating costs in this manner, Hit-and-Run takes more random walk steps to provably mix on any class of polytopes where $m^\gamma = o\left(\frac{n^2R^2}{r^2}\right)$. A sufficient condition for $m = o\left(\frac{n^{3-\gamma}R^2}{r^2}\right)$ to hold is $m = o(n^{4-\gamma})$.

9.2.1 Applications

Sampling lattice points in polytopes

While polytopes form a restricted subclass of the set of all convex bodies, algorithms for sampling polytopes have numerous applications. It was shown in [41] that if an n dimensional polytope defined by m inequalities contains a ball of radius $\Omega(n\sqrt{\log m})$, then it is possible to sample the lattice points inside it in polynomial time by sampling the interior of the polytope and picking a nearby lattice point. Often, combinatorial structures can be encoded as lattice points in a polytope, leading in this way to algorithms for sampling them. Contingency tables are two-way tables that are used by statisticians to represent bivariate data. A solution proposed in [20] to the frequently encountered problem of testing the independence of two characteristics of empirical data involves sampling uniformly from the set of two-way tables having fixed row and column sums. It was shown in [67] that under some conditions, this can be achieved in polynomial time by quantizing random points from an associated polytope.

Linear Programming

We use this result to design an affine interior point algorithm that does a *single* random walk to solve linear programs approximately. In this respect, our algorithm differs from existing randomized algorithms for linear programming such as that of Lovász and Vempala [59], which solves more general convex programs. While optimizing over a polytope specified as in the previous subsection, if $m = O(n^{2-\epsilon})$, the number of random steps taken by our algorithm is less than that of [59]. Given a polytope Q containing the origin and a linear objective c, our aim is to find with probability $> 1-\delta$, a point $y \in Q$ such that $c^T y \geq 1-\epsilon$ if there exists a point $z \in Q$ such that $c^T z \geq 1$. We first truncate Q using a hyperplane $c^T y = 1 - \hat{\epsilon}$, for $\hat{\epsilon} << \epsilon$ and obtain $Q_{\hat{\epsilon}} = Q \cap \{y \big| c^T y \leq 1 - \hat{\epsilon}\}$. We then projectively transform $Q_{\hat{\epsilon}}$ to "stretch" it into a new polytope $\gamma(Q_{\hat{\epsilon}})$ where $\gamma : y \mapsto \frac{y}{1-c^T y}$. Finally, we do a simplified Dikin walk (without the Metropolis filter) on $\gamma(Q_{\hat{\epsilon}})$ which approaches close to the optimum in polynomial time. This algorithm is purely affine after one preliminary projective transformation, in the sense that Dikin ellipsoids are used that are affine invariants but not projective invariants. This is an important distinction in the theory of interior point methods and the fact that our algorithm is polynomial time is notable since the corresponding deterministic affine algorithm analyzed by Dikin [22, 99] has no known polynomial guarantees on its run-time. Its projective counterpart, the algorithm of Karmarkar however does [42]. In related work [7], Belloni and Freund have explored the use of randomization for preconditioning. While there is no "local" potential function that is improved upon in each step, our analysis may be interpreted as using the $\mathcal{L}_{2,\mu}$ norm (μ being the appropriate stationary measure) of the probability density of the k^{th} point as a potential, and showing that this reduces at each step by a multiplicative factor of $(1 - \frac{\Phi^2}{2})$ where Φ is the conductance of the walk on the transformed polytope. We use the $\mathcal{L}_{2,\mu}$ norm rather than variation distance because this allows us to give guarantees of exiting the region where the objective function is low before the relevant Markov Chain has reached approximate stationarity. The main result related to algorithm (Dikin) is the following.

Theorem 9.2.2. *Let n be larger than some universal constant. Given a system of inequalities $By \leq \mathbf{1}$, a linear objective c such that the polytope*

$$Q := \{y : By \leq \mathbf{1} \textit{ and } |c^T y| \leq 1\}$$

is bounded, and $\epsilon, \delta > 0$, the following is true. If $\exists\, z$ such that $Bz \leq \mathbf{1}$ and $c^T z \geq 1$, then y, the output of Dikin , satisfies

$$By \leq \mathbf{1}$$

$$c^T y \geq 1 - \epsilon$$

with probability greater than $1 - \delta$.

Strong Polynomiality

Let us call a point x central if $\ln s$, where s is the function of x defined in Theorem 9.2.1, is polynomial in m. The mixing time of Dikin walk both from a warm start, and from a starting point that is central, is strongly polynomial in that the number of arithmetic operations depends only on m and n. Previous Markov Chains for sampling convex sets (and hence polytopes) do not possess either of these characteristics. In the setting of approximate Linear

Programming that we have considered, the numbers of iterations taken by known interior point methods such as those of Karmarkar [42], Renegar [83], Vaidya [96] etc are strongly polynomial when started from a point that is central in the above sense. The algorithm Dikin presented here is no different in this respect. The fact that Dikin walk has a mixing time that is strongly polynomial from a central point such as the center of mass, is related to two properties of Dikin ellipsoids listed below.

Dikin ellipsoids and their virtues

Let K be a polytope in $n-$dimensional Euclidean space given as the intersection of m halfspaces $a_i^T x \leq 1$, $1 \leq i \leq m$. Defining A to be the $m \times n$ matrix whose i^{th} row is a_i^T, the polytope can be specified by $Ax \leq \mathbf{1}$. Let $x_0 \in int(K)$ belong to the interior of K. Let

$$H(x) = \sum_{1 \leq i \leq m} \frac{a_i a_i^T}{(1 - a_i^T x)^2}$$

and $\|z - x\|_x^2 := (z - x)^T H(x)(z - x)$. The *Dikin* ellipsoid D_x^r of radius r for $x \in K$ is the ellipsoid containing all points z such that

$$\|z - x\|_x \leq r.$$

Fact 9.2.1. *(1) Dikin ellipsoids are affine invariants in that if T is an affine transformation and $x \in K$, the Dikin ellipsoid of radius r centered at the point Tx for the polytope $T(K)$ is $T(D_x^r)$. This is easy to verify from their definition.*

(2) For any interior point x, the Dikin ellipsoid centered at x, having radius 1, is contained in K. This has been shown in Theorem 2.1.1 of [74]. Also, the Dikin ellipsoid at x having radius $\sqrt{m}$ contains $Sym_x(K) := K \cap \{y \big| 2x - y \in K\}$. This can be derived by an argument along the lines of Theorem 9.3.2.

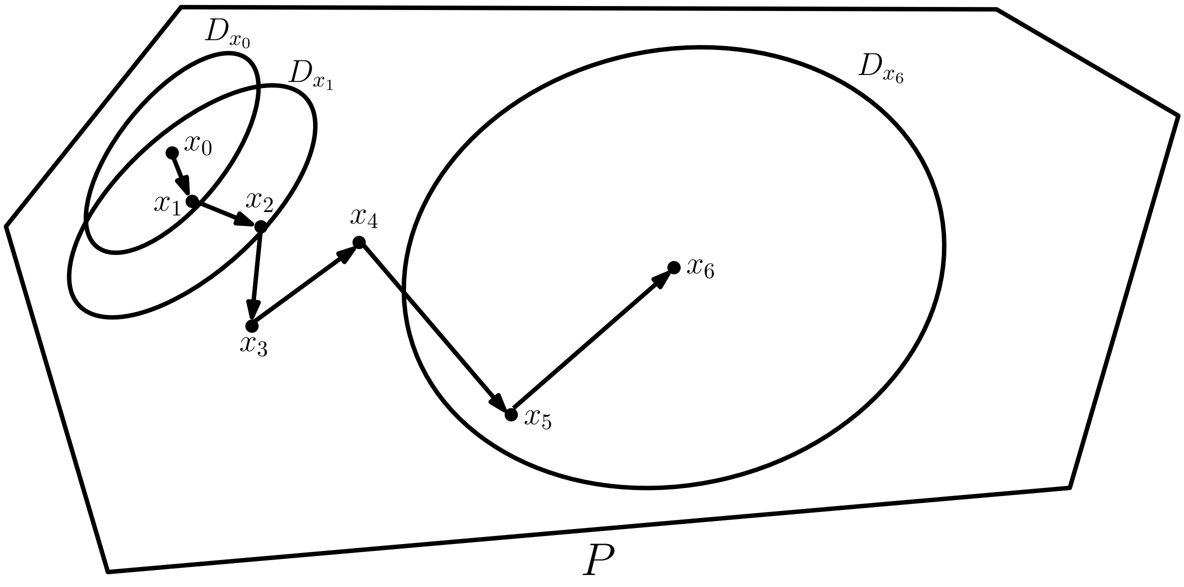

Figure 9.1: A realization of Dikin walk. Dikin ellipsoids D_{x_0}, D_{x_1} and D_{x_6} have been depicted.

9.3 Randomly Sampling Polytopes

9.3.1 Preliminaries

For two vectors v_1, v_2, let $\langle v_1, v_2 \rangle_x = v_1^T H(x) v_2$. For $x \in K$, we denote by D_x, the Dikin ellipsoid of radius $\frac{3}{40}$ centered at x. Dikin ellipsoids have been studied in the context of optimization [22] and have recently been used in online learning [1]. The second property mentioned in the subsection below implies that the Dikin walk does not leave K.

The "Dikin walk" is a "Metropolis" type walk which picks a move and then decides whether to "accept" the move and go there or "reject" and stay. The transition probabilities of the Dikin walk are listed below. When at x, one step of the walk is made as follows.

Therefore,

$$\mathbb{P}[x \to y] = \begin{cases} \min\left(\frac{1}{2\operatorname{vol}(D_x)}, \frac{1}{2\operatorname{vol}(D_y)}\right), \\ \text{if } y \in D_x \text{ and } x \in D_y; \\ 0, \text{otherwise}. \end{cases}$$

1. Flip an unbiased coin. If `Heads`, stay at x.
2. If `Tails` pick a random point y from D_x.
3. If $x \notin D_y$, then reject y (stay at x);
   if $x \in D_y$, then accept y with probability
   $\min\left(1, \frac{\mathrm{vol}(D_x)}{\mathrm{vol}(D_y)}\right) = \min\left(1, \sqrt{\frac{\det H(y)}{\det H(x)}}\right)$.

and $\mathbb{P}[x \to x] = 1 - \int_y d\mathbb{P}[x \to y]$.

Implementation of a Dikin step

Let K be the set of points satisfying the system of inequalities $Ax \leq \mathbf{1}$. $H(x) = A^T D(x)^2 A$ where $D(x)$ is the diagonal matrix whose i^{th} diagonal entry $d_{ii}(x) = \frac{1}{1-a_i^T x}$.

We can generate a Gaussian vector v such that $\mathbb{E}[vv^T] = (A^T D^2 A)^{-1}$ by the following procedure. Let u be a random m-vector from a Gaussian distribution whose covariance matrix is Id. Find v that satisfies the linear equations:

$$DAv = z$$

$$A^T D(z - u) = 0,$$

or equivalently,

$$A^T D^2 Av = A^T Du.$$

Allowing $(DA)^\dagger$ to be the Moore-Penrose pseudo-inverse of DA,

$$(DA)^\dagger (z - u) = 0 \Leftrightarrow (z - u) \perp \mathrm{column\,span}(DA)$$

$$\Leftrightarrow A^T D(z - u) = 0.$$

Thus, $\mathbb{E}vv^T = (DA)^\dagger \mathbb{E}zz^T (DA)^{\dagger T}$. z is the orthogonal projection of u onto the column span of DA,

therefore $(DA)^\dagger \mathbb{E}zz^T (DA)^{\dagger T} = H(x)^{-1}$. We can now generate a random point from the Dikin ellipsoid by scaling $v/\|v\|_x$ appropriately. The probability of accepting a Dikin step, is either 0 or the minimum of 1 and ratio of two determinants. Two matrix-vector products suffice to test whether the original point lies in the Dikin ellipsoid of the new one. By results of Baur and Strassen [4], the complexity of solving linear equations and of computing the determinant of an $n \times n$ matrix is $O(n^\gamma)$. The most expensive step, the computation of $A^T D(x)^2 A$ can be acheived using $mn^{\gamma-1}$, by partitioning a padded extension of $A^T D$ into $\leq \frac{m+n-1}{n}$ square matrices. Thus, all the operations needed for one step of Dikin walk can be computed using $O(mn^{\gamma-1})$ arithmetic operations where $\gamma < 2.377$ is the exponent for matrix multiplication.

9.3.2 Isoperimetric inequality

Given interior points x, y in a polytope K, suppose p, q are the ends of the chord in K containing x, y and p, x, y, q lie in that order. Then we denote $\frac{|x-y||p-q|}{|p-x||q-y|}$ by $\sigma(x, y)$. $\ln(1 + \sigma(x, y))$ is a metric known as the Hilbert metric, and given four collinear points a, b, c, d, $(a : b : c : d) = \frac{(a-c)\cdot(b-d)}{(a-d)\cdot(b-c)}$ is known as the cross ratio.

The theorem below was proved by Lovász in [57].

Theorem 9.3.1 (Lovász). *Let S_1 and S_2 be measurable subsets of K. Then,*

$$\mathrm{vol}(K \setminus S_1 \setminus S_2)\,\mathrm{vol}(K) \geq \sigma(S_1, S_2)\,\mathrm{vol}(S_1)\,\mathrm{vol}(S_2).$$

9.3.3 Dikin norm and Hilbert metric

Theorem 9.3.2 relates the Dikin norm to the Hilbert metric. The Dikin norms can be used to define a Riemannian manifold by using the associated bilinear form $< \cdot, \cdot >_x$ to construct a metric tensor. Dikin walk is a random walk on such a manifold in the spirit of the random walks discussed in Section 8.1.

Observation 9.3.1. *The isoperimetric properties of this manifold can be deduced from those of the Hilbert metric, and in fact, Theorem 9.3.1 and Theorem 9.3.2 together imply that the weighted Cheeger constant (see Definition 8.1.1) of this manifold is bounded below by* $\frac{1}{2\sqrt{m}}$.

Theorem 9.3.2. *Let x, y be interior points of K. Then,*

$$\sigma(x, y) \geq \frac{\|x-y\|_x}{\sqrt{m}}.$$

Proof. It is easy to see that we can restrict attention to the line ℓ containing x, y. We may also assume that $x = 0$ after translation. So now $b_i \geq 0$. Let c_i be the component of a_i along ℓ; we may view c_i, y as real numbers with ℓ as the real line now. $K \cap \ell = \{y : c_i y \leq b_i\}$ (where b_i had been taken to be 1). Dividing constraint i by $|c_i|$, we may assume that $|c_i| = 1$. After renumbering constraints so that $b_1 = \min\{b_i \big| c_i = -1\}$ and $b_2 = \min\{b_i \big| c_i = 1\}$, we have $K \cap \ell = [-b_1, b_2]$. Also

$$\|x-y\|_x^2 = y^2 \sum_i \frac{1}{b_i^2}.$$

Without loss of generality, assume that $y \geq 0$. [The proof is symmetric for $y \leq 0$.] Then, $\sigma(x, y) = \frac{y(b_1+b_2)}{b_1(b_2-y)}$, which is $\geq y \max_i(1/|b_i|)$. This is in turn $\geq \frac{\|x-y\|_x}{\sqrt{m}}$. □

9.3.4 Geometric and probabilistic distance

Let the Lebesgue measure be denoted λ. The total variation distance between two distributions π_1 and π_2 is $d(\pi_1, \pi_2) := \sup_S |\pi_1(S) - \pi_2(S)|$ where S ranges over all measurable sets.

Let the marginal distributions of transition probabilities starting from a point u be denoted P_u. Let us fix $r := 3/40$ for the remainder of this chapter. The main lemma of this section is stated below.

Lemma 9.3.1. *Let x, y be points such that $\sigma(x, y) \leq \frac{3}{400\sqrt{mn}}$. Then, the total variation distance between P_x and P_y is less than $1 - \frac{13}{200} + o(1)$.*

Proof. Let us fix the convention that $\frac{dP_y}{dP_x}(x) := 0$ and $\frac{dP_y}{dP_x}(y) := +\infty$. If $x \to w$ is one step of the Dikin walk,

$$d(P_x, P_y) = 1 - \mathbb{E}_w \left[\min\left(1, \frac{dP_y}{dP_x}(w)\right)\right].$$

It follows from Lemma 9.3.2 that

$$\mathbb{E}_w \left[\min\left(1, \frac{dP_y}{dP_x}(w)\right)\right] \geq \min\left(1, \frac{\mathrm{vol} D_x}{\mathrm{vol}(D_y)}\right) \mathbb{P}\left[(y \in D_w) \wedge (w \in D_y \setminus \{x\})\right].$$

It follows from Lemma 9.3.4 that

$$\min\left(1, \frac{\mathrm{vol}(D_x)}{\mathrm{vol} D_y}\right) \mathbb{P}\left[(y \in D_w) \wedge (w \in D_y \setminus \{x\})\right] \geq \tag{9.1}$$

$$e^{-\frac{r}{5}} \mathbb{P}\left[(y \in D_w) \wedge (w \in D_y \setminus \{x\})\right]. \tag{9.2}$$

Let E_x denote the event that

$$0 < \max\left(\|x - w\|_w^2, \|x - w\|_x^2\right) \leq r^2 \left(1 - \frac{1}{n}\right),$$

E_y denote the event that $\max\left(\|y - w\|_w, \|y - w\|_y\right) \leq r$ and E_{vol} denote the event that $\mathrm{vol}(D_w) \geq e^{4r} \mathrm{vol}(D_x)$. The complement of an event E shall be denoted $\overline{E}$.

The probability of E_y when $x \to w$ is a transition of Dikin walk can be bounded from below by $\left(\frac{e^{-4r}}{2}\right) \mathbb{P}\left[E_y \wedge E_x \wedge \overline{E_{vol}}\right]$ where w is chosen uniformly at random from D_x. It thus suffices to find a lower bound for $\mathbb{P}\left[E_y \wedge E_x \wedge \overline{E_{vol}}\right]$ where w is chosen uniformly at random from D_x, which we proceed to do. Let $\operatorname{erf}(x)$ denote the well known error function $\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ and $\operatorname{erfc}(x) := 1 - \operatorname{erf}(x)$.

$$\mathbb{P}\left[E_y \wedge E_x \wedge \overline{E_{vol}}\right] \geq \tag{9.3}$$

$$\mathbb{P}\left[E_y \wedge E_x\right] - \mathbb{P}\left[E_{vol}\right]. \tag{9.4}$$

Lemma 9.3.3 implies that $\mathbb{P}\left[E_{vol}\right] \leq \frac{\operatorname{erfc}(2)}{2} + o(1)$. Let E_x^1 be the event that

$$\|x - w\|_x^2 \leq r^2\left(1 - \frac{1}{n}\right).$$

As a consequence of Lemma 9.3.5,

$$\begin{aligned}\mathbb{P}\left[E_x\right] + o(1) &\geq \left(\frac{1 - 3\sqrt{2}r}{2}\right)\mathbb{P}\left[E_x^1\right] \\ &\geq \left(\frac{1 - 3\sqrt{2}r}{2\sqrt{e}}\right) - o(1). \qquad (9.5)\end{aligned}$$

Lemma 9.3.6 and Lemma 9.3.7 together tell us that

$$\begin{aligned}\mathbb{P}\left[E_y \middle| E_x\right] &\geq 1 - \left(\frac{4r^2 + \operatorname{erfc}(2) + o(1)}{1 - 3\sqrt{2}r}\right) - \left(\frac{4r^2 + \operatorname{erfc}(3/2) + o(1)}{1 - 3\sqrt{2}r}\right) \qquad (9.6) \\ &= 1 - \left(\frac{8r^2 + \operatorname{erfc}(2) + \operatorname{erfc}(\frac{3}{2}) + o(1)}{1 - 3\sqrt{2}r}\right). \qquad (9.7)\end{aligned}$$

Putting (9.5) and (9.7) together gives us that

$$\begin{aligned}\mathbb{P}\left[E_y \wedge E_x\right] &= \mathbb{P}\left[E_y \middle| E_x\right]\mathbb{P}\left[E_x\right] \qquad (9.8) \\ &\geq \frac{1 - 3\sqrt{2}r}{2\sqrt{e}} - \left(\frac{8r^2 + \operatorname{erfc}(2) + \operatorname{erfc}(\frac{3}{2})}{2\sqrt{e}}\right) - o(1). \qquad (9.9)\end{aligned}$$

Putting together (9.2), (9.4) and (9.9), we see that if $x \to w$ is a transition of the Dikin walk,

$$\mathbb{E}_w\left[\min\left(1, \frac{dP_y}{dP_x}(w)\right)\right] \geq \frac{e^{-\frac{21r}{5}}}{4\sqrt{e}}\left(1 - (3\sqrt{2}r + 8r^2 + \operatorname{erfc}(2)(1+\sqrt{e}) + \operatorname{erfc}(\frac{3}{2}))\right) - o(1).$$

For our choice of $r = 3/40$, this evaluates to more than $\frac{13}{200} - o(1)$. □

Since Dikin ellipsoids are affine-invariant, we shall assume without loss of generality that x is the origin and the Dikin ellipsoid at x is the Euclidean unit ball of radius r. This also means that in system of coordinates, the local norm $\|\cdot\|_x = \|\cdot\|_o$ is the Euclidean norm $\|\cdot\|$ and the local inner product $\langle\cdot,\cdot\rangle_x = \langle\cdot,\cdot\rangle_o$ is the usual inner product $\langle\cdot,\cdot\rangle$. On occasion we have used $a \cdot b$ to signify $\langle a, b\rangle$.

Lemma 9.3.2. *Let $w \in \operatorname{supp}(P_x) \setminus \{x, y\}$ and $y \in D_w$ and $w \in D_y$. Then,*

$$\frac{dP_y}{dP_x}(w) \quad \geq \quad \min\left(1, \frac{\operatorname{vol}(D_x)}{\operatorname{vol}(D_y)}\right).$$

Proof. Under the hypothesis of the lemma,

$$\begin{aligned}\frac{dP_y}{dP_x}(w) \quad &= \quad \frac{\min\left(\frac{1}{\operatorname{vol}(D_y)}, \frac{1}{\operatorname{vol}D_w}\right)}{\min\left(\frac{1}{\operatorname{vol}(D_x)}, \frac{1}{\operatorname{vol}D_w}\right)} \\ &= \quad \frac{\min\left(\frac{\operatorname{vol}(D_w)}{\operatorname{vol}(D_y)}, 1\right)}{\min\left(\frac{\operatorname{vol}(D_w)}{\operatorname{vol}(D_x)}, 1\right)}.\end{aligned}$$

The above expression can be further simplified by considering two cases.

1. Suppose $\min\left(\frac{\operatorname{vol}(D_w)}{\operatorname{vol}(D_y)}, 1\right) = 1$, then

$$\frac{\min\left(\frac{\operatorname{vol}D_w}{\operatorname{vol}(D_y)}, 1\right)}{\min\left(\frac{\operatorname{vol}(D_w)}{\operatorname{vol}(D_x)}, 1\right)} \quad \geq \quad 1.$$

2. Suppose $\min\left(\frac{\text{vol}(D_w)}{\text{vol}(D_y)}, 1\right) = \frac{\text{vol}(D_w)}{\text{vol}(D_y)}$, then

$$\frac{\min\left(\frac{\text{vol}(D_w)}{\text{vol}(D_y)}, 1\right)}{\min\left(\frac{\text{vol}(D_w)}{\text{vol}(D_x)}, 1\right)} \quad \geq \quad \frac{\text{vol}(D_x)}{\text{vol}(D_y)}.$$

Therefore,

$$\frac{dP_y}{dP_x}(w) \quad \geq \quad \min\left(1, \frac{\text{vol}(D_x)}{\text{vol}D_y}\right).$$

□

Lemma 9.3.3. *Let w be chosen uniformly at random from D_x. The probability that $\text{vol}(D_x) \leq e^{2rc}\,\text{vol}(D_w)$ is greater or equal to $1 - \frac{\text{erfc}(c)}{2} - o(1)$, i. e.*

$$\mathbb{P}\left[\frac{\text{vol}(D_w)}{\text{vol}(D_x)} \leq e^{2rc}\right] \quad \geq \quad 1 - \frac{\text{erfc}\,(c)}{2} - o(1).$$

Proof. By Lemma 9.3.13, $\ln(\frac{1}{\text{vol}(D_x)})$ is a convex function. Therefore,

$$\ln\,\text{vol}(D_w) - \ln\,\text{vol}(D_x) \quad \leq \quad \nabla\ln(\frac{1}{\text{vol}(D_x)}) \cdot (w - x).$$

By Lemma 9.3.12, $\|\nabla\ln(\frac{1}{\text{vol}(D_x)})\| \leq 2\sqrt{n}$. Therefore,

$$\nabla\ln(\frac{1}{\text{vol}(D_x)}) \cdot (w-x) \leq 2r\left(\frac{\sqrt{n}\,\nabla\ln(\frac{1}{\text{vol}(D_x)}) \cdot (w-x)}{\|\nabla\ln(\frac{1}{\text{vol}(D_x)})\|\|w-x\|}\right)$$

As stated in Theorem 9.3.3, when the dimension $n \to \infty$,

$$\frac{\sqrt{n}\,\nabla\ln(\frac{1}{\text{vol}(D_x)}) \cdot (w-x)}{\|\nabla\ln(\frac{1}{\text{vol}(D_x)})\|\|w-x\|}$$

converges in distribution to a standard Gaussian random variable whose mean is 0 and variance is 1. Therefore,

$$\mathbb{P}\left[\frac{\sqrt{n}\,\nabla\ln(\frac{1}{\mathrm{vol}(D_x)})\cdot(w-x)}{\|\nabla\ln(\frac{1}{\mathrm{vol}D_x})\|\|w-x\|}\le c\right] \quad\ge\quad \frac{1+\,\mathrm{erf}(c)}{2}-o(1).$$

This implies that

$$\begin{aligned}\mathbb{P}\left[\frac{\mathrm{vol}(D_w)}{\mathrm{vol}(D_x)}\le e^c\right] &\ge \mathbb{P}\left[\nabla\ln(\frac{1}{\mathrm{vol}(D_x)})\cdot(w-x)\le c\right]\\ &\ge \left(\frac{1+\,\mathrm{erf}\left(\frac{c}{2r}\right)}{2}\right)-o(1).\end{aligned}$$

□

Lemma 9.3.4.

$$\ln\left(\frac{\mathrm{vol}(D_y)}{\mathrm{vol}(D_x)}\right)\le n\sigma(x,y).$$

Proof. Suppose $\overline{pq}$ is a chord and p,x,y,q appear in that order. By Theorem 9.6.1,

$$\begin{aligned}\ln\left(\frac{\mathrm{vol}(D_y)}{\mathrm{vol}(D_x)}\right) &\le \ln\left(\frac{|p-y|^n}{|p-x|^n}\right)\\ &\le n\sigma(x,y).\end{aligned}$$

□

Lemma 9.3.5. *Let w be chosen uniformly at random from D_x. Then,*

$$\begin{aligned}&\mathbb{P}\left[\|x-w\|_w^2\le r^2\left(1-\frac{1}{n}\right)\Big|\|x-w\|_x^2\le r^2\left(1-\frac{1}{n}\right)\right]\\ &\qquad\qquad\ge\frac{1-3\sqrt{2}r}{2}-o(1).\end{aligned}$$

Proof. Let E_x^1 be the event that

$$\|x - w\|_x^2 \leq r^2 \left(1 - \frac{1}{n}\right).$$

We set c to $3\sqrt{2}r$ in Lemma 9.3.8 and see that

$$\mathbb{P}\left[\|x - w\|_w^2 + \|x - w\|_{2x-w}^2 \geq 2r^2 \left(1 - \frac{1}{n}\right) \Big| E_x^1\right] \\ \leq 3\sqrt{2}r + o(1).$$

If $\|x - w\|_w^2 + \|x - w\|_{2x-w}^2 \leq 2r^2 \left(1 - \frac{1}{n}\right)$, then either $\|x - w\|_w^2$ or $\|x - w\|_{2x-w}^2$ must be less or equal to $r^2 \left(1 - \frac{1}{n}\right)$. □

Lemma 9.3.6. *Let $\sigma(x, y) \leq \frac{3}{400\sqrt{mn}}$. Then, if w is chosen uniformly at random from D_x,*

$$\mathbb{P}\left[\|y - w\|_y \geq r \Big| \max\left(\|x - w\|_x^2, \|x - w\|_w^2\right) \leq r^2 \left(1 - \frac{1}{n}\right)\right] \\ \leq \frac{4r^2 + \operatorname{erfc}(2) + o(1)}{1 - 3\sqrt{2}r}.$$

Proof. It follows from Lemma 9.3.10, after substituting 1 for η and 2 for η_1 that

$$\mathbb{P}\left[\|y - w\|_y \geq r \Big| \|x - w\|_x^2 \leq r^2 \left(1 - \frac{1}{n}\right)\right] \\ \leq 2r^2 + \frac{\operatorname{erfc}(2)}{2} + o(1).$$

This lemma follows using the upper bound from Lemma 9.3.5 for

$$\mathbb{P}\left[\|x - w\|_w^2 \leq r^2 \left(1 - \frac{1}{n}\right) \Big| \|x - w\|_x^2 \leq r^2 \left(1 - \frac{1}{n}\right)\right].$$

An application of Theorem 9.3.2 completes the proof. □

Lemma 9.3.7. *Suppose $\sigma(x, y) \leq \frac{3}{400\sqrt{mn}}$. Let w be chosen uniformly at random from D_x. Then,*

$$\mathbb{P}\left[\|y - w\|_w \geq r \middle| \max(\|x - w\|_w^2, \|x - w\|_x^2) \leq r^2 \left(1 - \frac{1}{n}\right)\right]$$
$$\leq \frac{4r^2 + \operatorname{erfc}(3/2) + o(1)}{1 - 3\sqrt{2}r}.$$

Proof. Substituting $c = 1$ in Lemma 9.3.9, we see that

$$\mathbb{P}\left[\|y - w\|_w^2 - \|x - w\|_w^2 \geq \psi_1 \middle| \|x - w\|_x^2 \leq r^2(1 - \frac{c}{n})\right]$$
$$\leq 2r^2 + \frac{\operatorname{erfc}(3/2)}{2} + o(1).$$

This implies that

$$\mathbb{P}\left[\|y - w\|_w^2 - \|x - w\|_w^2 \geq \frac{r}{n} \middle| \|x - w\|_x^2 \leq r^2 \left(1 - \frac{1}{n}\right)\right]$$
$$\leq 2r^2 + \frac{\operatorname{erfc}(3/2)}{2} + o(1).$$

This lemma follows using the lower bound from Lemma 9.3.5 for

$$\mathbb{P}\left[\|x - w\|_w^2 \leq r^2 \left(1 - \frac{1}{n}\right) \middle| \|x - w\|_x^2 \leq r^2 \left(1 - \frac{1}{n}\right)\right].$$

□

The following theorem has the geometric interpretation that the probability distribution obtained by orthogonally projecting a random vector v_n from an n-dimensional ball of radius $\sqrt{n}$ onto a line converges in distribution to the standard mean zero, variance 1, normal distribution $N[0, 1]$. This was known to Poincaré, and is a fact often mentioned in the context of measure concentration phenomena, see for example [55].

Theorem 9.3.3 (Poincaré)**.** *Let v_n be any n-dimensional vector and h_n be a random vector chosen uniformly from the n-dimensional unit Euclidean ball. Then, as $n \to \infty$, $\frac{\sqrt{n}<v_n,h_n>}{\|v_n\|\|h_n\|}$ converges in distribution to a zero-mean Gaussian whose variance is* 1*, i. e.* $N[0, 1]$.

Let

$$\psi_1 := \frac{\|y-x\|_x^2}{(1-r)^2} + \frac{(3+2\sqrt{6})r\|y-x\|_x}{\sqrt{n}}.$$

Lemma 9.3.8. *Let v be chosen uniformly at random from D_x and c be a positive constant. Then,*

$$\mathbb{P}\left[\|x-v\|_v^2 + \|x-v\|_{2x-v}^2 \geq 2r^2\left(1 - \frac{(c - \frac{18r^2}{c})}{n}\right)\right] \\ \leq c + o(1).$$

Proof. Let the i^{th} constraint be $a_i^T x \leq 1$ for all $i \in \{1, \dots, m\}$. Let $x - v$ be denoted h. In the present frame, for any vector v, $\|v\|_x = \|v\|$.

$$\|x-v\|_v^2 + \|x-v\|_{2x-v}^2 = \sum_i \frac{(a_i^T h)^2}{(1-a_i^T h)^2} + \sum_i \frac{(a_i^T h)^2}{(1+a_i^T h)^2} \tag{9.10}$$

In the present coordinate frame $\sum_i a_i a_i^T = I$. Consequently for each i,

$$\mathbb{E}[(a_i^T h)^2] = \frac{\|a_i\|^2 \mathbb{E}[\|h\|^2]}{n} \tag{9.11}$$

$$\leq \frac{r^2}{n}. \tag{9.12}$$

$$\sum_i \left(\frac{(a_i^T h)^2}{2(1 - a_i^T h)^2} + \frac{(a_i^T h)^2}{2(1 + a_i^T h)^2} \right) = \sum_i (a_i^T h)^2 \left(\frac{1 + (a_i^T h)^2}{(1 - (a_i^T h)^2)^2} \right) \tag{9.13}$$

$$= \sum_i \left((a_i^T h)^2 + \frac{3(a_i^T h)^4 - (a_i^T h)^6}{(1 - (a_i^T h)^2)^2} \right)$$

$$= \|h\|_x^2 + \sum_i \frac{3(a_i^T h)^4 - (a_i^T h)^6}{(1 - (a_i^T h)^2)^2}. \tag{9.14}$$

In the present coordinate frame $\sum_i a_i a_i^T = I$. Consequently for each i,

$$\mathbb{E}\left[\frac{(a_i^T h)^2}{\|a_i\|^2 \|h\|^2} \right] = \frac{1}{n}. \tag{9.15}$$

By Theorem 9.3.3, the probability that $|a_i^T h| \geq n^{-\frac{1}{4}}$ is $O(e^{-\sqrt{n}/2})$. $|a_i^T h|$ is $\leq \|a_i^T\| r$, which is less than $\frac{1}{2}$. This allows us to write

$$\mathbb{E}\left[\frac{3(a_i^T h)^4 - (a_i^T h)^6}{(1 - (a_i^T h)^2)^2} \right] = 3\mathbb{E}[(a_i^T h)^4](1 + o(1)), \tag{9.16}$$

and so

$$\mathbb{E}\left[\sum_i \frac{3(a_i^T h)^4 - (a_i^T h)^6}{(1 - (a_i^T h)^2)^2} \right] = \sum_i 3\mathbb{E}[(a_i^T h)^4](1 + o(1)). \tag{9.17}$$

Next, we shall find an upper bound on $\mathbb{E}[\sum_i (a_i^T h)^4]$. The length of h and its direction are independent, therefore

$$\mathbb{E}\left[\sum_i (a_i^T h)^4 \right] = \sum_i \|a_i\|^4 \mathbb{E}[\|h\|^4] \mathbb{E}\left[\frac{(a_i^T h)^4}{\|a_i\|^4 \|h\|^4} \right]. \tag{9.18}$$

A direct integration by parts tells us that if the distribution of X is $N[0, 1]$, then $\mathbb{E}[X^4] = 3$.

Therefore,

$$\mathbb{E}\left[\frac{(a_i^T h)^4}{\|a_i\|^4\|h\|^4}\right] \;=\; \frac{3+o(1)}{n^2}. \tag{9.19}$$

$\mathbb{E}[\|h\|^4]$ is equal to $r^4(1+o(1))$ and so

$$\mathbb{E}\left[\sum_i (a_i^T h)^4\right] \;=\; \sum_i \left(\frac{3+o(1)}{n^2}\right)\|a_i\|^4 r^4. \tag{9.20}$$

This implies that

$$\begin{aligned}
\mathbb{E}\left[\sum_i \frac{3(a_i^T h)^4}{(1-(a_i^T h)^2)^2}\right] &= \frac{9+o(1)}{n^2}\sum_i \|a_i\|^4 r^4 && (9.21)\\
&\le \frac{9+o(1)}{n^2}\sum_i \|a_i\|^2 r^4 && (9.22)\\
&= \frac{(9+o(1))r^4}{n}. && (9.23)
\end{aligned}$$

In (9.22), we used the fact that $\sum_i a_i a_i^T = I$ and so $\|a_i\|^2 \le 1$ for each i. Together, Markov's inequality and (9.23) yield the following.

$$\begin{aligned}
\mathbb{P}\left[\sum_i \frac{3(a_i^T h)^4-(a_i^T h)^6}{(1-(a_i^T h)^2)^2} \ge \frac{c_2 r^4}{n}\right] &\le \mathbb{P}\left[\sum_i \frac{3(a_i^T h)^4}{(1-(a_i^T h)^2)^2} \ge \frac{c_2 r^4}{n}\right] && (9.24)\\
&\le \frac{9+o(1)}{c_2}. && (9.25)
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{P}[\|h\|_x^2 \ge r^2(1-\frac{c_1}{n})] &= \mathbb{P}[\|h\|_x^n \ge r^n(1-\frac{c_1}{n})^{n/2}] && (9.26)\\
&\le 1-e^{-\frac{c_1}{2}}+o(1). && (9.27)
\end{aligned}$$

We infer from (9.25) and (9.27) that

$$\begin{aligned}\mathbb{P}\left[\|h\|_x^2 + \sum_i \frac{3(a_i^T h)^4 - (a_i^T h)^6}{(1-(a_i^T h)^2)^2} \geq r^2(1 - \frac{c_1 - c_2 r^2}{n})\right] &\leq 1 - e^{-\frac{c_1}{2}} + \frac{9}{c_2} + o(1) \\ &\leq \frac{c_1}{2} + \frac{9}{c_2} + o(1). \qquad (9.28)\end{aligned}$$

Setting c_1 to c and c_2 to $\frac{18}{c}$ proves the lemma. □

Let E_x^c be the event that $\|x-w\|_x^2 \leq r^2(1 - \frac{c}{n})$.

Lemma 9.3.9. *Let w be a point chosen uniformly at random from D_x. Then, for any positive constant c, independent of n,*

$$\mathbb{P}\left[\|y-w\|_w^2 - \|x-w\|_w^2 \geq \psi_1 \middle| E_x^c\right]$$

$$\leq 2r^2 + \frac{\operatorname{erfc}(3/2)}{2} + o(1).$$

Proof. $\|y\|_w^2$ can be bounded above in terms of $\|y\|_o$ as follows.

$$\begin{aligned}\|y\|_w^2 &\leq y^T\left(\sum_i \frac{a_i a_i^T}{(1-a_i^T w)^2}\right) y \qquad (9.29) \\ &\leq \left(\sup_i \frac{1}{(1-a_i^T w)^2}\right) \sum_i y^T a_i a_i^T y. \qquad (9.30)\end{aligned}$$

For each i, $\|a_i\| \leq 1$, therefore

$$\left(\sup_i \frac{1}{(1-a_i^T w)^2}\right) \sum_i y^T a_i a_i^T y \quad \leq \quad \frac{\|y\|_o^2}{(1-r)^2}. \qquad (9.31)$$

Let E_w^c be the event that $\|w\|_o^2 \leq 1 - \frac{c}{n}$.

By Theorem 9.3.3,

$$\mathbb{P}\left[(-2\langle y,w\rangle_o) \geq \frac{2r\eta_1\|y\|_o}{\sqrt{n}}\Big|E_w^c\right] \;\leq\; \frac{1-\operatorname{erf}(\eta_1)}{2}+o(1). \tag{9.32}$$

$(\langle y,w\rangle_o - \langle y,w\rangle_w)^2$ can be bounded above using the Cauchy-Schwarz inequality as follows.

$$\begin{aligned}
(\langle y,w\rangle_o - \langle y,w\rangle_w)^2 &= \left(w^T\left(1-\sum_i \frac{a_ia_i^T}{(1-a_i^Tw)^2}\right)y\right)^2\\
&= \left(\sum_i \frac{w^Ta_i((1-a_i^Tw)^2-1)a_i^Ty}{(1-a_i^Tw)^2}\right)^2\\
&\leq \left(\sum_i \frac{\left(w^Ta_i((1-a_i^Tw)^2-1)\right)^2}{(1-a_i^Tw)^4}\right)\left(\sum_i (a_i^Ty)^2\right).
\end{aligned}$$

Let κ be a standard one-dimensional Gaussian random variable whose variance is 1 and mean is 0 (i. e. having distribution $N[0,1]$). Since $r<\frac{1}{2}$ and each $\|a_i\| = \|a_i\|_o$ is less or equal to 1, it follows from Theorem 9.3.3 that conditional on E_w^c,

$$\frac{\left(nw^Ta_i((1-a_i^Tw)^2-1)\right)^2}{4r^2\|a_i\|^2(1-a_i^Tw)^4}$$

converges in distribution to the distribution of κ^4, whose expectation can be shown using integration by parts to be 3. So,

$$\begin{aligned}
\mathbb{E}\left[\sum_i \frac{\left(w^Ta_i((1-a_i^Tw)^2-1)\right)^2}{(1-a_i^Tw)^4}\Big|E_w^c\right] &\leq \sum_i \left(\frac{4}{n^2}\right)\|a_i\|_o^4 r^4(3+o(1))\\
&\leq \left(\frac{12+o(1)}{n^2}\right) r^4 \sum_i \|a_i\|_o^2\\
&= \frac{(12+o(1))r^4}{n}.
\end{aligned}$$

Thus by Markov's inequality,

$$\mathbb{P}\left[\sum_i \frac{\left(w^T a_i((1-a_i^T w)^2-1)\right)^2}{(1-a_i^T w)^4} \geq \frac{12\eta_2 r^4}{n}\Big|E_w^c\right] \leq \frac{1+o(1)}{\eta_2}. \tag{9.33}$$

$\sum_i (a_i^T y)^2$ is equal to $\|y\|_o^2$. Therefore (9.33) implies that

$$\mathbb{P}\left[(\langle y,w\rangle_o - \langle y,w\rangle_w)^2 \geq \frac{12\eta_2 r^4\|y\|_o^2}{n}\right] \leq \frac{1+o(1)}{\eta_2}. \tag{9.34}$$

Putting (9.32) and (9.34) together, we see that

$$\mathbb{P}\left[-2\langle y,w\rangle_w \geq \frac{2r\eta_1\|y\|_o}{\sqrt{n}} + 2\sqrt{\frac{12\eta_2 r^4\|y\|_o^2}{n}}\,\Bigg|E_w^c\right] \leq \frac{1-\ \mathrm{erf}(\eta_1)}{2} + \frac{1+o(1)}{\eta_2} \tag{9.35}$$

Conditional on E_w^c, $\|w\|_w^2$ is less or equal to $r(1-\frac{c}{n})$.

Therefore, using erfc(x) to denote $1-\ \mathrm{erf}(x)$,

$$\mathbb{P}\left[\|y-w\|_w^2 - \|w\|_w^2 \geq \frac{\|y\|_o^2}{(1-r)^2} + \frac{2r\|y\|_o}{\sqrt{n}}\left(\eta_1 + r\sqrt{12\eta_2}\right)\Big|E_w^c\right] \leq \eta_2^{-1} + \frac{\mathrm{erfc}(\eta_1)}{2} + o(1).$$

Setting $\eta_1 = 3/2$ and $\eta_2 = \frac{1}{2r^2}$, gives

$$\mathbb{P}\left[\|y-w\|_w^2 - \|w\|_w^2 \geq \Big|E_w^c\right] \leq 2r^2 + \frac{\mathrm{erfc}(3/2)}{2} + o(1). \tag{9.36}$$

□

Lemma 9.3.10. *Let c be a positive constant. Let*

$$\psi_2 := \|y-x\|_y^2 + \frac{2r\eta_1\|y-x\|_x}{\sqrt{n}} + \frac{2\eta\|y-x\|_x}{\sqrt{n}}\left(\sqrt{3}r + \|y-x\|_x\right).$$

If w is a point chosen uniformly at random from D_x, for any positive constants η and η_1, Then,

$$\mathbb{P}\left[\|y-w\|_y^2 - \|x-w\|_x^2 \geq \psi_2 \middle| E_w^c\right]$$

$$\leq \frac{2r^2}{\eta^2} + \frac{\operatorname{erfc}(\eta_1)}{2} + o(1).$$

Proof.

$$\begin{aligned}
\|y-w\|_y^2 &= \|y\|_y^2 + \|w\|_y^2 - 2\langle w, y\rangle_y && (9.37)\\
&\leq \|y\|_y^2 + \|w\|_o^2 && (9.38)\\
&+ \sqrt{(\|w\|_y^2 - \|w\|_o^2)^2} - 2\langle w, y\rangle_o + 2\sqrt{(\langle w, y\rangle_o - \langle w, y\rangle_y)^2}. && (9.39)
\end{aligned}$$

We shall obtain probabilistic upper bounds on each term in (9.39).

$$\begin{aligned}
(\|w\|_y^2 - \|w\|_o^2)^2 &= \left(w^T \left(\sum_i a_i a_i^T \left(\frac{1-(1-a_i^T y)^2}{(1-a_i^T y)^2}\right)\right) w\right)^2 && (9.40)\\
&\leq \left(\sum_i (w^T a_i)^4\right)\left(\sum_i \left(\frac{1-(1-a_i^T y)^2}{(1-a_i^T y)^2}\right)^2\right) && (9.41)\\
&= \left(\sum_i (w^T a_i)^4\right)\left(\sum_i 4\left(a_i^T y\right)^2 (1+o(1))\right) && (9.42)\\
&= (4+o(1))\, \|y\|_o^2 \sum_i (w^T a_i)^4. && (9.43)
\end{aligned}$$

In inferring (9.42) from (9.41) we have used the fact that $\|y\|_o$ is $O(\frac{1}{\sqrt{n}})$ which is $o(1)$. As was stated in (9.19) in slightly different terms,

$$\mathbb{E}\left[(w^T a_i)^4\right] = \frac{\|a_i\|^4 r^4 (3+o(1))}{n^2}.$$

Therefore by Markov's inequality, for any constant c,

$$\begin{aligned}\mathbb{E}\left[\sum_i (w^T a_i)^4 \Big| \|w\|_o^2 \leq r^2(1-\frac{c}{n})\right] &= \sum_i \frac{\|a_i\|^4 r^4(3+o(1))}{n^2} \\ &\leq \frac{r^4(3+o(1))}{n^2}\sum_i \|a_i\|^2 \\ &= \frac{r^4(3+o(1))}{n}.\end{aligned}$$

Therefore,

$$\mathbb{P}\left[(\|w\|_y^2 - \|w\|_o^2)^2 \geq \eta^2 \frac{12\|y\|_o^2 r^4}{n}\right] \leq \frac{1+o(1)}{\eta^2}. \tag{9.44}$$

By Theorem 9.3.3, as $n \to \infty$, the distribution of $\frac{\sqrt{n}\langle w,y\rangle_o}{r\|y\|_o}$ converges in distribution to $N[0,1]$. Therefore

$$\mathbb{P}\left[(-2\langle w,y\rangle_o) \geq \frac{2\eta_1 r\|y\|_o}{\sqrt{n}}\Big| \|w\|_o^2 \leq r^2(1-\frac{c}{n})\right] \leq \frac{\text{erfc}(\eta_1)}{2} + o(1). \tag{9.45}$$

Finally, we need similar tail bounds for $(\langle w,y\rangle_o - \langle w,y\rangle_y)^2$. Note that

$$\begin{aligned}(\langle w,y\rangle_o - \langle w,y\rangle_y)^2 &= \left(w^T\left(\sum_i a_i a_i^T\left(\frac{1-(1-a_i^T y)^2}{(1-a_i^T y)^2}\right)\right)y\right)^2 &(9.46)\\ &\leq \left(\sum_i (w^T a_i a_i^T y)^2\right)\left(\sum_i\left(\frac{1-(1-a_i^T y)^2}{(1-a_i^T y)^2}\right)^2\right) &(9.47)\\ &= \left(\sum_i (w^T a_i a_i^T y)^2\right)\left(\sum_i (4+o(1))(a_i^T y)^2\right) &(9.48)\\ &= (4+o(1))\left(\sum_i (w^T a_i a_i^T y)^2\right)\|y\|_o^2. &(9.49)\end{aligned}$$

It suffices now to obtain a tail bound on $\sum_i (w^T a_i a_i^T y)^2$. By Theorem 9.3.3,

$$\begin{aligned}
\mathbb{E}\left[\sum_i (w^T a_i a_i^T y)^2 \Big| \|w\|_o^2 \leq r^2(1 - \frac{c}{n})\right] &\leq \left(\sum_i \|a_i a_i^T y\|^2\right) \frac{r^2(1 + o(1))}{n} \\
&\leq \left(\sum_i (a_i^T y)^2\right) \frac{r^2(1 + o(1))}{n} \\
&\leq \frac{\|y\|_o^2 r^2 (1 + o(1))}{n}.
\end{aligned}$$

Therefore,

$$\mathbb{P}\left[(\langle w, y\rangle_o - \langle w, y\rangle_y)^2 \leq \frac{4\eta^2 \|y\|_o^4 r^2}{n}\right] \leq \frac{1 + o(1)}{\eta^2}. \tag{9.50}$$

Putting together (9.44), (9.45) and (9.50), we see that

$$\mathbb{P}\left[\|y - w\|_y^2 - \|w\|_o^2 \geq \|y\|_y^2 + \frac{2\eta \|y\|_o}{\sqrt{n}}\left(\sqrt{3}r + \frac{r\eta_1}{\eta} + \|y\|_o\right) \Big| E_w^c\right] \leq \frac{2r^2}{\eta^2} + \frac{\mathrm{erfc}(\eta_1)}{2} + o(1).$$

□

The following is a generalization of the Cauchy-Schwarz inequality that takes values in a cone of semidefinite matrices where inequality is replaced by dominance in the semidefinite cone. It will be used to prove Lemma 9.3.12 and may be of independent interest.

Lemma 9.3.11 (Semidefinite Cauchy-Schwartz). *Let $\alpha_1, \dots, \alpha_m$ be reals and $A_1, \dots, A_m$ be $r \times n$ matrices. Let $B \preccurlyeq C$ signify that B is dominated by C in the semidefinite cone. Then*

$$\left(\sum_{i=1}^m \alpha_i A_i\right)\left(\sum_{i=1}^m \alpha_i A_i\right)^T \preccurlyeq \left(\sum_{i=1}^m \alpha_i^2\right)\left(\sum_{i=1}^m A_i A_i^T\right). \tag{9.51}$$

Proof. For each i and j,

$$0 \preccurlyeq \left(\alpha_j A_i - \alpha_i A_j\right)\left(\alpha_j A_i - \alpha_i A_j\right)^T$$

Therefore,

$$\begin{aligned} 0 &\preccurlyeq \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\alpha_j A_i - \alpha_i A_j\right)\left(\alpha_j A_i - \alpha_i A_j\right)^T \\ &= \left(\sum_{i=1}^{m}\alpha_i^2\right)\left(\sum_{i=1}^{m}A_i A_i^T\right) - \left(\sum_{i=1}^{m}\alpha_i A_i\right)\left(\sum_{i=1}^{m}\alpha_i A_i\right)^T \end{aligned}$$

□

We shall obtain an upper bound of $2\sqrt{n}$ on

$$\|\nabla \ln(\frac{1}{\mathrm{vol} D_x})\|\Big|_{x=o} = \|\nabla \ln \det H\|\Big|_o.$$

Lemma 9.3.12. $\|\nabla \ln \det H|_x\|_x \leq 2\sqrt{n}$.

Proof. In our frame,

$$\sum a_i a_i^T = I, \tag{9.52}$$

where I is the $n \times n$ identity matrix, and for any vector v,

$$\|v\|_o = \|v\|. \tag{9.53}$$

If X is a matrix whose $\ell_2 \to \ell_2$ norm is less than 1, $\log(I+X)$ can be assigned a unique value by equating it with the power series

$$\sum_{i=1}^{\infty}(-1)^{i-1}\frac{X^i}{i}.$$

Using this formalism when y is in a small neighborhood of the identity.

$$\ln\det H(y) \quad = \quad \operatorname{trace}\ln H(y). \tag{9.54}$$

In order to obtain an upper bound on $\|\nabla \ln\det H\|$ at o, it suffices to uniformly bound $|\frac{\partial \ln\det H}{\partial h}|$ along all unit vectors h, since

$$\|\nabla \ln\det H\| = \sup_{\|h\|=1} \big|\frac{\partial}{\partial h}\operatorname{trace}\ln H\big|. \tag{9.55}$$

$$\left[\frac{\partial}{\partial h}\operatorname{trace}\ln H\right]\Bigg|_o$$

$$\begin{aligned}
&= \lim_{\delta\to 0}\frac{\left(\operatorname{trace}\ln\left(\sum\frac{a_ia_i^T}{(1-\delta a_i^T h)^2}\right)-\ln I\right)}{\delta} && (9.56)\\
&= \sum_i 2(a_i^T h)(\operatorname{trace} a_ia_i^T) && (9.57)\\
&= 2\sum_i \|a_i\|^2 a_i^T h. && (9.58)
\end{aligned}$$

The Semidefinite Cauchy-Schwarz inequality from Lemma 9.3.11 gives us the following.

$$(\sum_i \|a_i\|^2 a_i)(\sum_i \|a_i\|^2 a_i^T) \preccurlyeq (\sum_i \|a_i\|^4)(\sum_i a_ia_i^T) \tag{9.59}$$

$\sum_i a_i a_i^T = I$, so the magnitude of each vector a_i must be less or equal to 1, and $\sum_i \|a_i\|^2$ must equal n.

Therefore

$$\begin{aligned}
(\sum_i \|a_i\|^4)(\sum_i a_i a_i^T) &= (\sum_i \|a_i\|^4)I & (9.60)\\
&\preccurlyeq (\sum_i \|a_i\|^2)I & (9.61)\\
&= nI & (9.62)
\end{aligned}$$

(9.59) and (9.62) imply that

$$(\sum_i \|a_i\|^2 a_i)(\sum_i \|a_i\|^2 a_i^T) \preccurlyeq nI. \tag{9.63}$$

(9.55), (9.58) and (9.63) together imply that

$$\|\nabla \ln \det H\| \quad \leq \quad 2\sqrt{n}. \tag{9.64}$$

□

The following is due to P. Vaidya [97].

Lemma 9.3.13. $\ln \det H$ *is a convex function.*

Proof. Let $\frac{\partial}{\partial h}$ denote partial differentiation along a unit vector h. Recall that $\sum_i a_i a_i^T = I$.

$$\frac{\partial^2 \ln \det H}{(\partial h)^2}\Big|_o$$

$$\begin{aligned}
&= \lim_{\delta\to 0} \frac{1}{\delta^2}\mathrm{trace}\ln\left(\left(\sum \frac{a_i a_i^T}{(1-\delta a_i^T h)^2}\right)\left(\sum \frac{a_i a_i^T}{(1+\delta a_i^T h)^2}\right)\right) \\
&= \lim_{\delta\to 0} \frac{\mathrm{trace}\left(\ln\left(\sum_i a_i a_i^T (\sum_{j\geq 0}(j+1)(\delta a_i^T h)^j)\right)\right)}{\delta^2} \\
&+ \frac{\mathrm{trace}\left(\ln\left(\sum_i a_i a_i^T (\sum_{j\geq 0}(j+1)(-\delta a_i^T h)^j)\right)\right)}{\delta^2} \\
&= \lim_{\delta\to 0} \frac{\mathrm{trace}\sum_{k\geq 1}\frac{(-1)^{k-1}}{k}\left(\sum_i a_i a_i^T (\sum_{j\geq 1}(j+1)(\delta a_i^T h)^j)\right)^k}{\delta^2} \\
&+ \frac{\mathrm{trace}\sum_{k\geq 1}\frac{(-1)^{k-1}}{k}\left(\sum_i a_i a_i^T (\sum_{j\geq 1}(j+1)(-\delta a_i^T h)^j)\right)^k}{\delta^2}.
\end{aligned}$$

The only terms in the numerators of the above limit that matter are those involving δ^2. So this simplifies to

$$\begin{aligned}
2\sum_i \mathrm{trace}\, a_i a_i^T (a_i^T h)^2 &= 2\sum_i \|a_i\|^2 (a_i^T h)^2 \\
&\geq 2\sum_i (a_i^T h)^4 \\
&\geq \frac{2\left(\sum_i (a_i^T h)^2\right)^2}{m} \\
&= \frac{2}{m}.
\end{aligned}$$

This proves the lemma. □

9.3.5 Conductance and mixing time

The proof of the following theorem is along the lines of Theorem 11 in [57].

Theorem 9.3.4. *Let n be greater than some universal constant. Let S_1 and $S_2 := K \setminus S_1$*

be measurable subsets of K. Then,

$$\int_{S_1} P_x(S_2)d\lambda(x) \geq \frac{6}{10^5\sqrt{mn}} \min\left(\operatorname{vol}(S_1), \operatorname{vol}(S_2)\right).$$

Proof. Let ρ be the density of the uniform distribution on K. We shall use ρ in some places where it is seemingly unnecessary because, then, most of this proof transfers verbatim to a proof of Theorem 9.6.4 as well. For any $x \neq y \in K$,

$$\rho(y)\frac{dP_y}{d\lambda}(x) \quad = \quad \rho(x)\frac{dP_x}{d\lambda}(y),$$

therefore ρ is the stationary density of the Markov chain. Let $\delta = \frac{3}{400\sqrt{mn}}$ and $\epsilon = \frac{13}{200}$. Let $S_1' = S_1 \cap \{x \big| \rho(x)P_x(S_2) \leq \frac{\epsilon}{2\operatorname{vol}(K)}\}$ and $S_2' = S_2 \cap \{y \big| \rho(y)P_y(S_1) \leq \frac{\epsilon}{2\operatorname{vol}(K)}\}$. By the reversibility of the chain, which is easily checked,

$$\int_{S_1} \rho(x)P_x(S_2)d\lambda(x) = \int_{S_2} \rho(y)P_y(S_1)d\lambda(y).$$

If $x \in S_1'$ and $y \in S_2'$ then

$$\int_K \min\left(\rho(x)\frac{dP_x}{d\lambda}(w), \rho(y)\frac{dP_y}{d\lambda}(w)\right) d\lambda(w) < \frac{\epsilon}{\operatorname{vol}(K)}.$$

For sufficiently large n, Lemma 9.3.1 implies that $\sigma(S_1', S_2') \geq \delta$. Therefore Theorem 9.3.1 implies that

$$\pi(K \setminus S_1' \setminus S_2') \geq \delta\pi(S_1')\pi(S_2').$$

First suppose $\pi(S_1') \geq (1-\delta)\pi(S_1)$ and $\pi(S_2') \geq (1-\delta)\pi(S_2)$. Then,

$$\begin{aligned}\int_{S_1} P_x(S_2)d\rho(x) &\geq \frac{\epsilon\pi(K \setminus S_1' \setminus S_2')}{2} \\ &\geq \frac{\epsilon\delta\pi(S_1')\pi(S_2')}{2} \\ &\geq \left(\frac{(1-\delta)^2\epsilon\delta}{8}\right)\min(\pi(S_1), \pi(S_2))\end{aligned}$$

and we are done. Otherwise, without loss of generality, suppose $\pi(S_1') \leq (1-\delta)\pi(S_1)$. Then

$$\int_{S_1} P_x(S_2)d\rho(x) \geq \frac{\epsilon\delta}{2}\pi(S_1)$$

and we are done. □

The following theorem was proved in [58].

Theorem 9.3.5 (Lovász-Simonovits). *Let μ_0 be the initial distribution for a lazy reversible ergodic Markov chain whose conductance is Φ and stationary measure is μ, and μ_k be the distribution of the k^{th} step. Let $M := \sup_S \frac{\mu_0(S)}{\mu(S)}$ where the supremum is over all measurable subsets S of K. Then, for all such S,*

$$|\mu_k(S) - \mu(S)| \leq \sqrt{M}\left(1 - \frac{\Phi^2}{2}\right)^k.$$

We now in a position to prove the main theorem regarding Dikin walk, Theorem 9.2.1.

Proof of Theorem 9.2.1. Let t be the time when the first proper move is made. $\mathbb{P}[t \geq t' \big| t \geq t'-1] \leq 1 - \frac{13}{200} + o(1)$ by Lemma 9.3.1 applied when $x = x_0$ and y approaches x_0. Therefore when n is sufficiently large,

$$\mathbb{P}\left[t < \frac{\ln(\frac{\epsilon}{2})}{\ln(1-\frac{6}{100})}\right] \geq 1 - \frac{\epsilon}{2}.$$

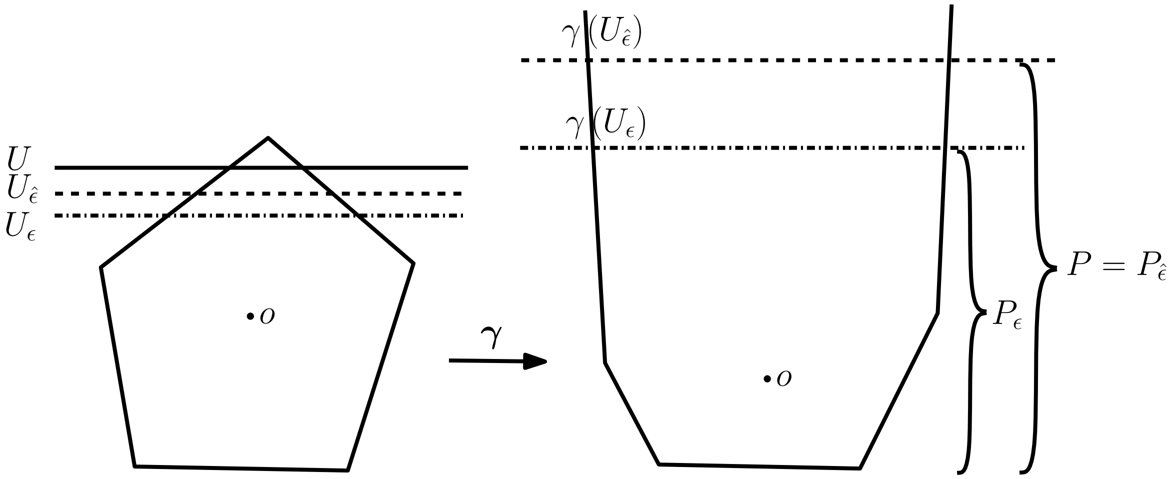

Figure 9.2: The effect of the projective transformation γ.

Let μ_k be the distribution of x_k and μ be the stationary distribution, which is uniform. Let ρ_k and ρ likewise be the density of μ_k and $\rho = \frac{1}{\mathrm{vol}(K)}$ the density of the uniform distribution. We shall now find an upper bound for $\frac{\rho_{k+t}}{\rho}$. For any $x \in K$, $\rho_t(x) \geq \frac{100}{6\,\mathrm{vol}(D_x)}$ by Lemma 9.3.1, applied when $x = x_0$ and y approaches x_0. By (2) in Fact 9.2.1 $\frac{\mathrm{vol}(D_x)}{\mathrm{vol}(K)} \geq \left(\frac{r}{\sqrt{2m}\,s}\right)^n$, which implies that

$$\sup_{S \subseteq K} \frac{\mu_t(S)}{\mu(S)} = \sup_{x \in K} \frac{\rho_t(x)}{\rho} \tag{9.65}$$

$$\leq \left(\frac{\sqrt{2m}\,s}{r}\right)^n \left(\frac{100}{6}\right). \tag{9.66}$$

The theorem follows by plugging in Equation 9.66 and the lower bound on the conductance of Dikin walk given by Theorem 9.3.4 into Theorem 9.3.5. □

9.4 Affine algorithm for linear programming

We shall consider problems of the following form. Given a system of inequalities $By \leq \mathbf{1}$, a linear objective c such that the polytope

$$Q := \{y : By \leq \mathbf{1} \text{ and } |c^T y| \leq 1\}$$

is bounded, and $\epsilon, \delta > 0$ the algorithm is required to do the following.

- If $\exists\, y$ such that $By \leq \mathbf{1}$ and $c^T y \geq 1$, output y such that $By \leq \mathbf{1}$ and $c^T y \geq 1 - \epsilon$ with probability greater than $1 - \delta$.

Any linear program can be converted to such a form, either by the sliding objective method or by combining the primal and dual problems and using the duality gap added to an appropriate slack variable as the new objective (see [47] and references therein). Before the iterative stage of the algorithm which is purely affine, we need to transform the problem using a projective transformation. Let $s \geq \sup_{y \in Q} \|By\| + 1$, and

$$\tau := \left\lceil 4 \times 10^8 \times mn \left(n \ln \left(\frac{4ms^2}{\epsilon^2} \right) + 2 \ln \left(\frac{2}{\delta} \right) \right) \right\rceil. \tag{9.67}$$

Let γ be the projective transformation $\gamma : y \mapsto \frac{y}{1 - c^T y}$, and γ^{-1} the inverse map, $\gamma^{-1} : x \mapsto \frac{x}{1 + c^T x}$. For any $\epsilon' > 0$, let $Q_{\epsilon'} := Q \cap \{y | c^T y \leq 1 - \epsilon'\}$ and $U_{\epsilon'}$ be the hyperplane $\{y | c^T y = 1 - \epsilon'\}$. Let $\hat{\epsilon} = \frac{\epsilon\delta}{4n}$ and $K_\epsilon := \gamma(Q_\epsilon)$. Let $K := K_{\hat{\epsilon}} = \gamma\left(Q_{\hat{\epsilon}}\right)$. For $x \in K$, let D_x denote the Dikin ellipsoid (with respect to K) of radius $r := \frac{3}{40}$, centered at x.

9.5 Algorithm

```
1. Choose $x_0$ uniformly at random from $r^{-1}D_o$, where $o$ is the origin.
2. While $i < \tau$ and $c^T\gamma^{-1}(x_i) < 1 - \epsilon$, choose $x_{i+1}$ using the rule below.
   (a) Flip an unbiased coin. If Heads, set $x_{i+1}$ to $x_i$.
   (b) If Tails pick a random point $y$ from $D_{x_i}$.
   (c) If $x_i \notin D_y$, then reject $y$ and set $x_{i+1}$ to $x_i$; if $x_i \in D_y$, then set $x_{i+1}$ to $y$.
3. If $c^T\gamma^{-1}(x_\tau) \geq 1 - \epsilon$ output $\gamma^{-1}(x_\tau)$, otherwise declare that there is no $y$ such that $By \leq \mathbf{1}$ and $c^T y \geq 1$.
```

9.6 Analysis

For any bounded $f : K \to \mathbb{R}$, we define

$$\|f\|_2 := \sqrt{\int_K f(x)^2 \rho(x) d\lambda(x)}$$

where $\rho(x) = \frac{\text{vol}(D_x)}{\int_K \text{vol}(D_x)d\lambda(x)}$. The following lemma shows that cross ratio is a projective invariant.

Lemma 9.6.1. *Let $\gamma : \mathbb{R}^n \to \mathbb{R}^n$ be a projective transformation. Then, for any 4 collinear points a, b, c and d, $(a : b : c : d) = (\gamma(a) : \gamma(b) : \gamma(c) : \gamma(d))$.*

Proof. Let $\{e_1, \dots, e_n\}$ be a basis for $\mathbb{R}^n$. Without loss of generality, suppose that $a, b, c, d \in \mathbb{R}e_1$. γ can be factorized as $\gamma = \gamma_2 \circ \gamma_1$ where $\gamma_1 : \mathbb{R}^n \to \mathbb{R}^n$ is a projective transformation and maps $\mathbb{R}e_1$ to $\mathbb{R}e_1$ and $\gamma_2 : \mathbb{R}^n \to \mathbb{R}^n$ is an affine transformation. Affine transformations clearly preserve the cross ratio, so the problem reduces to showing that $(a : b : c : d) = (\gamma_1(a) : \gamma_1(b) : \gamma_1(c) : \gamma_1(d))$, which is a 1-dimensional question. In 1-dimension, the group of projective transformations is generated by translations $(x \mapsto x + \beta)$, scalar multiplication

$(x \mapsto \alpha x)$ and inversion $(x \mapsto x^{-1})$, where $\alpha, \beta \in \mathbb{R} \setminus \{0\}$. In each of these cases the equality is easily checked. □

The following was proved in a more general context by Nesterov and Todd in Theorem 4.1, [75].

Theorem 9.6.1 (Nesterov-Todd)**.** *Let $\overline{pq}$ be a chord of K and x, y be interior points on it so that p, x, y, q are in order. Then $z \in D_y$ implies that $p + \frac{|p-x|}{|p-y|}(z - p) \in D_x$.*

The following theorem is from [58].

Theorem 9.6.2 (Lovász-Simonovits)**.** *Let M be a lazy reversible ergodic Markov chain on $K \subseteq \mathbb{R}^n$ with conductance Φ, whose stationary distribution is μ. For every bounded f, let $\|f\|_{2,\mu}$ denote $\sqrt{\int_K f(x)^2 d\mu(x)}$. For any fixed f, let Mf be the function that takes x to $\int_K f(y) dP_x(y)$. Then if $\int_K f(x) d\mu(x) = 0$,*

$$\|M^k f\|_{2,\mu} \leq \left(1 - \frac{\Phi^2}{2}\right)^k \|f\|_{2,\mu}.$$

We shall now prove the main theorem regarding Algorithm Dikin, Theorem 9.2.2.

Proof of Theorem 9.2.2. Let $\overline{pq}$ be a chord of the polytope K_ϵ containing the origin o such that

$c^T(\gamma^{-1}(p)) \geq c^T(\gamma^{-1}(q))$. Let $p' = \gamma^{-1}(p)$, $q' = \gamma^{-1}(q)$ and r' be the intersection of the chord $\overline{p'q'}$ with the hyperplane $U := \{y \big| c^T y = 1\}$. Then, $\frac{|q-o|}{|p-o|} \leq \frac{|q'-o|}{|p'-o|} \leq s$. $\frac{|p-o|}{|q-o|}$ is equal to $|(\infty : o : q : p)|$. By Lemma 9.6.1, the cross ratio is a projective invariant. Therefore,

$$\frac{|p-o|}{|q-o|} = \left(\frac{|p'-o|}{|p'-r'|}\right)\left(\frac{|r'-q'|}{|q'-o|}\right) \tag{9.68}$$

$$\leq \left(\frac{1}{\epsilon}\right)(s). \tag{9.69}$$

Therefore, for any chord $\overline{pq}$ of K_ϵ through o, $\frac{|p|}{|q|} \leq \frac{s}{\epsilon}$.

Let $D = \int_K \mathrm{vol}(D_y) d\lambda(y)$. Let

$$\rho_o(x) = \begin{cases} \frac{1}{\mathrm{vol}(D_o)}, & x \in D_o; \\ 0, & \text{otherwise}, \end{cases}$$

be the density of x_o and likewise $\rho_\mathcal{T}$ be the density of the distribution of $x_\mathcal{T}$. Let $f_0(x) = \frac{\rho_0(x)}{\rho(x)}$ and $f_\mathcal{T}(x) = \frac{\rho_\mathcal{T}(x)}{\rho(x)}$.

$$\begin{aligned} \|f_0\|_2^2 &= \int_{D_o} \left(\frac{\rho_0(x)}{\rho(x)}\right)^2 \rho(x) d\lambda(x) \\ &\leq \frac{D}{\mathrm{vol}(D_o) \inf\limits_{x \in D_o} \mathrm{vol}(D_x)} \end{aligned}$$

By Fact 9.2.1 and the fact that the Dikin ellipsoid of radius r with respect to K_ϵ is contained in the Dikin ellipsoid of the same radius with respect to K, $\sqrt{2m}D_o \supseteq Sym_o(K_\epsilon)$. (9.69) implies that $Sym_o(K_\epsilon) \supseteq \left(\frac{\epsilon}{s}\right) K_\epsilon$. We see from Theorem 9.6.1 that $\inf\limits_{x \in D_o} \mathrm{vol}(D_x) \geq \mathrm{vol}((1-r)D_o)$. Therefore,

$$\begin{aligned} \|f_0\|_2^2 &\leq \frac{D}{\mathrm{vol}(D_o) \inf\limits_{x \in D_o} \mathrm{vol}(D_x)} \\ &\leq \left(\frac{2m(\frac{s}{\epsilon})^2}{1-r}\right)^n \left(\frac{D}{\int_{K_\epsilon} \mathrm{vol}(D_y) d\lambda(y)}\right) \\ &= \left(\frac{2m(\frac{s}{\epsilon})^2}{1-r}\right)^n \left(\frac{1}{\pi(K_\epsilon)}\right), \end{aligned} \tag{9.70}$$

where π is the stationary distribution. For a line $\ell \perp U$, let π_ℓ and ρ_ℓ be interpreted as the induced measure and density respectively. Let ℓ intersect the facet of K that belongs to $U_{\hat{\epsilon}}$ at u. Then by Theorem 9.6.1, for any $x, y \in \ell \cap K$ such that $|x - u| > |y - u|$, $\frac{\rho_\ell(x)}{|u-x|^n} \leq \frac{\rho_\ell(y)}{|u-y|^n}$.

By integrating over such 1-dimensional fibres ℓ perpendicular to U, we see that

$$\begin{aligned}
\pi(K_\epsilon) &= \frac{\int_{\ell \perp U} \pi_\ell(\ell \cap K_\epsilon) du}{\int_{\ell \perp U} \pi_\ell(\ell) du} \\
&\leq \sup_{\ell \perp U} \frac{\pi_\ell(\ell \cap K_\epsilon)}{\pi_\ell(\ell)} \\
&\leq \left(\frac{(1 - 1/\hat{\epsilon})^{n+1} - (1/\epsilon - 1/\hat{\epsilon})^{n+1}}{(1/\epsilon - 1/\hat{\epsilon})^{n+1}} \right) \\
&\lesssim \exp(\frac{\delta}{4}) - 1 \quad \text{as } n \to \infty. \qquad (9.71)
\end{aligned}$$

The relationship between conductance Φ and decay of the $\mathcal{L}_2$ norm from Theorem 9.6.2 tells us that

$$\begin{aligned}
\|f_\tau - \mathbb{E}_\rho f_\tau\|_2^2 &\leq \|f_0 - \mathbb{E}_\rho f_0\|_2^2 \, e^{-\tau \Phi^2} \\
&= \left(\|f_0\|_2^2 - \|(\mathbb{E}_\rho f_0)\mathbf{1}\|_2^2 \right) e^{-\tau \Phi^2} \\
&\leq \left(\frac{2m(\frac{s}{\epsilon})^2}{1 - r} \right)^n \left(\frac{e^{-\tau \Phi^2}}{\pi(K_\epsilon)} \right) \text{ (from (9.70))}
\end{aligned}$$

which is less than $\frac{\delta^2}{4\pi(K_\epsilon)}$, when we substitute Φ from Theorem 9.6.4 and the value of τ from (9.67).

$$\begin{aligned}
\frac{\delta^2}{4\pi(K_\epsilon)} &\geq \int_{K_\epsilon} (f_\tau(x) - \mathbb{E}_\rho f_\tau)^2 \rho(x) d\lambda(x) \\
&\geq \frac{\left(\int_{K_\epsilon} (f_\tau(x) - \mathbb{E}_\rho f_\tau) \rho(x) d\lambda(x) \right)^2}{\int_{K_\epsilon} \rho(x) d\lambda(x)} \\
&= \frac{(\mathbb{P}[x_\tau \in K_\epsilon] - \pi(K_\epsilon))^2}{\pi(K_\epsilon)}.
\end{aligned}$$

which together with (9.71) implies that $\mathbb{P}[x_\tau \in K_\epsilon] \lesssim \delta$ and completes the proof. □

The following generalization of Theorem 9.3.1 was proved in [62].

Theorem 9.6.3 (Lovász-Vempala). *Let S_1 and S_2 be measurable subsets of K and μ a measure supported on K that possesses a density whose logarithm is concave. Then,*

$$\mu(K \setminus S_1 \setminus S_2)\mu(K) \geq \sigma(S_1, S_2)\mu(S_1)\mu(S_2).$$

The proof of the following lemma is along the lines of Lemma 9.3.1 and is provided below.

Lemma 9.6.2. *Let x, y be points such that $\sigma(x, y) \leq \frac{3}{400\sqrt{mn}}$. Then, the overlap*

$$\int_{\mathbb{R}^n} \min\big(\operatorname{vol}(D_x)P_x,\, \operatorname{vol}(D_y)P_y\big)\, d\lambda(x)$$

between $\operatorname{vol}(D_x)P_x$ and $\operatorname{vol}(D_y)P_y$ in algorithm Dikin *is greater than $(\frac{9}{100} - o(1))\operatorname{vol}(D_x)$.*

Proof. If $x \to w$ is one step of Dikin ,

$$\int_{\mathbb{R}^n} \min\big(\operatorname{vol}(D_x)P_x,\, \operatorname{vol}(D_y)P_y\big)\, d\lambda(x) =$$
$$\mathbb{E}_w\left[\min\left(\operatorname{vol}(D_x),\, \operatorname{vol}(D_y)\frac{dP_y}{dP_x}(w)\right)\right].$$

$$\mathbb{E}_w\left[\min\left(\operatorname{vol}(D_x),\, \operatorname{vol}(D_y)\frac{dP_y}{dP_x}(w)\right)\right] =$$
$$\operatorname{vol}(D_x)\mathbb{P}\,\big[(y \in D_w) \wedge (w \in D_y \setminus \{x\})\big].$$

Let E_x denote the event that
$0 < \max\left(\|x - w\|_w^2, \|x - w\|_x^2\right) \leq r^2\left(1 - \frac{1}{n}\right)$ and
E_y denote the event that $\max\left(\|y - w\|_w, \|y - w\|_y\right) \leq r$. The probability of E_y when $x \to w$ is a transition of Dikin is greater or equal to $\frac{\mathbb{P}[E_y \wedge E_x]}{2}$ when w is chosen uniformly at random

from D_x. Thus, using Lemmas 9.3.5, 9.3.6 and 9.3.7,

$$\int_{\mathbb{R}^n} \min\big(\mathrm{vol}(D_x)P_x, \mathrm{vol}(D_y)P_y\big)\, d\lambda(x) \geq$$
$$\mathrm{vol}(D_x)\frac{\mathbb{P}\left[E_y\middle|E_x\right]\mathbb{P}\left[E_x\right]}{2} \geq$$
$$\frac{\mathrm{vol}(D_x)(1 - 3\sqrt{2}r - 8r^2 - \mathrm{erfc}(2) - \mathrm{erfc}(\frac{3}{2}) - o(1))}{4\sqrt{e}}.$$

When $r = 3/40$, this evaluates to more than
$\mathrm{vol}(D_x)(\frac{9}{100} - o(1))$. □

The proof of the following theorem closely follows that of Theorem 9.3.2.

Theorem 9.6.4. *If K is a bounded polytope, the conductance of the Markov chain in Algorithm* Dikin *is bounded below by* $\frac{8}{10^5\sqrt{mn}}$.

Proof. For any $x \neq y \in K$, $\mathrm{vol}(D_y)\frac{dP_y}{d\lambda}(x) =$
$\mathrm{vol}(D_x)\frac{dP_x}{d\lambda}(y)$, and therefore

$$\rho(x) := \frac{\mathrm{vol}(D_x)}{\int_K \mathrm{vol}(D_x)d\lambda(x)}$$

is the stationary density. Let $\delta = \frac{3}{400\sqrt{mn}}$ and $\epsilon = \frac{9}{100}$. Theorem 9.6.3 is applicable in our situation because by Lemma 9.3.13, the stationary density ρ is log-concave. The proof of Theorem 9.3.2 now applies verbatim apart from using Lemma 9.6.2 instead of Lemma 9.3.1, and Theorem 9.6.3 instead of Theorem 9.3.1. This gives us

$$\int_{S_1} P_x(S_2)d\rho(x) \;\;\geq\;\; \left(\frac{(1-\delta)^2\epsilon\delta}{8}\right)\min(\pi(S_1), \pi(S_2)).$$

Thus we are done. □

CHAPTER 10

LOW DENSITY SEPARATION, SPECTRAL CLUSTERING AND GRAPH CUTS

10.1 Introduction

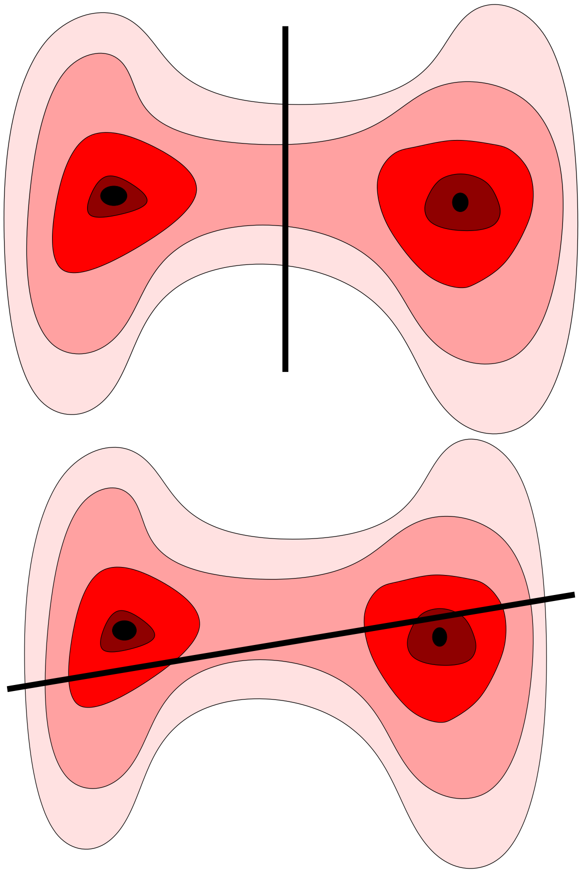

Figure 10.1: A likely cut and a less likely cut.

In this chapter we propose a formal measure on the complexity of the boundary, which intuitively corresponds to the Low Density Separation assumption. We will show that given a class boundary, this measure can be computed from a finite sample from the probability distribution p. The number of samples used is exponential in the dimension. We do not provide theorems in this setting, but if the distribution has support on a low-dimensional submanifold the number of samples would be exponential in the intrinsic dimension rather than the ambient dimension. Moreover, we show that it is closely related to a cut of a certain standard adjacency graph, defined on that sample. We will also point out some interesting connections to spectral clustering. We propose the *weighted area* of the boundary, represented by the contour integral along the boundary $\int_{\text{cut}} p(s)ds$ to measure the complexity of a cut. It is clear that the boundary in the left panel has a considerably lower weighted length than the boundary in the right panel of our Fig. 10.1.

To formalize this notion further consider a (marginal) probability distribution with density $p(x)$ supported on some domain or manifold M. This domain is partitioned in two disjoint clusters/parts. Assuming that the boundary S is a smooth hypersurface we define the *weighted volume of the cut* to be

$$\int_S p(s)ds, \tag{10.1}$$

where ds ranges over all $d-1-$dimensional infinitesimal volume elements tangent to the hypersurface. Note that just as in the example above, the

We will show how this quantity can be approximated given empirical data and establish connections with some popular graph-based methods.

10.2 Connections and related work

10.2.1 Spectral Clustering

Over the last two decades there has been considerable interest in various spectral clustering techniques . The idea of spectral clustering can be expressed very simply. Given a graph we would like to construct a balanced partitioning of the vertices, such that this partitioning minimizes the number (or total weight) of edges across the cut. This is generally an NP-hard optimization problem. However a simple real valued relaxation can be used to reduce it to standard linear algebra, typically to finding eigenvectors of a certain *graph Laplacian*. We note that the quality of partition is typically measured in terms of the corresponding *cut size*.

10.2.2 Graph-based semi-supervised learning

Similarly to spectral clustering, graph-based semi-supervised learning constructs a graph from the data. In contrast to clustering, however, some of the data is labeled. The problem is typically to either label the unlabeled points (transduction) or, more generally, to build a classifier defined on the whole space. This may be done trying to find the minimum cut, which respects the labeled data, or, using graph Laplacian as a penalty functional.

One of the important intuitions of semi-supervised learning is the *cluster assumption* or, more specifically, the *low density separation assumption* suggested in [15], which states that the class boundary passes through a low density region. We will modify that intuition slightly by suggesting that cutting through a high density region may be acceptable as long as the length of the cut is very short. For example imagine two high-density round clusters connected by a very thin high-density thread. Cutting the thread is appropriate as long as the width of the thread is much smaller than the radius of the clusters.

The goal of this chapter is to take a step toward making a theoretical connection between cuts of data adjacency graphs and the underlying probability distributions. We will show that as more and more data is sampled from a probability distribution the (weighted) size of a fixed cut for the data adjacency graph converges to the (weighted) volume of the boundary for the partition of the underlying distribution.

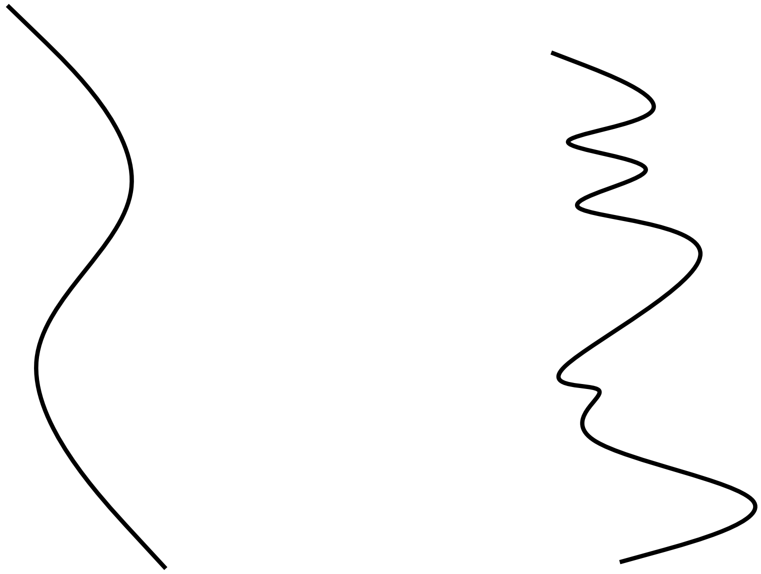

Figure 10.2: Curves of small and high condition number respectively

10.3 Summary of Main Results

Let ρ be a probability density function on a domain $M \subseteq \mathbb{R}^d$.

Let S be a smooth hypersurface that separates M into two parts, S_1 and S_2. The smoothness of S will be quantified by a *condition number* $1/\tau$. A formal definition of the condition number appears in Definition 6.3.1.

Definition 10.3.1. *Let $G^t(x, y)$ be the heat kernel in $\mathbb{R}^d$ given by*

$$G^t(x, y) := \frac{1}{(4\pi t)^{d/2}} \; \mathrm{e}^{-\|x-y\|^2/4t}.$$

Let $M_t := G^t(x, x) = \frac{1}{(4\pi t)^{d/2}}$.

Let $X := \{x_1, \ldots, x_N\}$ be a set of N points chosen independently at random from ρ. Consider the complete graph whose vertices are associated with the points in X, and where the weight of the edge between x_i and x_j is given by

$$W_{ij} = \begin{cases} G^t(x_i, x_j) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

Let W be the matrix $\{W_{ij}\}_{i,j}$. Let $X_1 = X \cap S_1$ and $X_2 = X \cap S_2$. Let D be the (diagonal) matrix whose entries are given by

$$D_{ij} = \begin{cases} \sum\limits_k w_{ik} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Let d_i be the weighted degree of the vertex corresponding to x_i. The normalized Laplacian is the random matrix $\tilde{L}(t, X) := I - D^{-1/2} W D^{-1/2}$. Let f be the vector $(f_1, \dots, f_N)$ where

$$f_i = \begin{cases} 1 & \text{if } x_i \in X_1 \\ 0 & \text{otherwise} \end{cases}$$

10.3.1 Regularity conditions on ρ and S

We make the following assumptions about ρ:

1. ρ can be extended to a function ρ' that is $L-$Lipshitz and which is bounded above by ρ_{max}.

2. For $0 < t < t_0$,

   $$\min\left(\rho(x), \int G^t(x, y)\rho(y)dy\right) \geq \rho_{min}.$$

   Note that this is a property of both of the boundary ∂M and ρ.

We note that since ρ' is $L-$Lipshitz over $\mathbb{R}^d$, so is $\int_M G^t(x, z)\rho'(z)dz$.

We assume that S has condition number $1/\tau$. We also make the following assumption about S:-

The volume of the set of points whose distance to both S and ∂M is $\leq R$, is $O(R^2)$ as $R \to 0$. This is reasonable, and is true if $S \cap \partial M$ is a manifold of codimension 2.

Under these conditions, our main theorem is

Theorem 10.3.1. *Let the number of points, $|X| = N$ tend to infinity and $\{t_N\}_0^\infty$, be a sequence of values of t that tend to zero such that $t_N > \frac{1}{N^{\frac{1}{2d+2}}}$. Then with probability* 1,

$$\frac{\sqrt{\pi}}{N\sqrt{t}} f^T \tilde{L}(t_N, X) f \to \int_S \rho(s) ds.$$

Further, for any $\delta \in (0,1)$ and any $\epsilon \in (0, 1/2)$, there exists a positive constant C and an integer N_0(depending on ρ, ϵ, δ and S) such that with probability $1 - \delta$,

$$(\forall N > N_0), \left| \frac{\sqrt{\pi}}{N\sqrt{t}} f^T \tilde{L}(t_N, X) f - \int_S \rho(s) ds \right| < C t_N^\epsilon.$$

This theorem is proved by first relating the empirical quantity $\frac{\sqrt{\pi}}{N\sqrt{t}} f^T \tilde{L}(t_N, X) f$ to a heat flow across the relevant cut (on the continuous domain), and then relating the heat flow to the measure of the cut. In order to state these results, we need the following notation.

Definition 10.3.2. *Let*

$$\psi_t(x) = \frac{\rho(x)}{\sqrt{\int_M G^t(x,z)\rho(z)dz}}.$$

Let

$$\beta(t, X) := \frac{\sqrt{\pi}}{N\sqrt{t}} f^T \tilde{L}(t, X) f$$

and

$$\alpha(t) := \sqrt{\frac{\pi}{t}} \int_{S_1} \int_{S_2} G^t(x,y) \psi_t(x) \psi_t(y) dx dy.$$

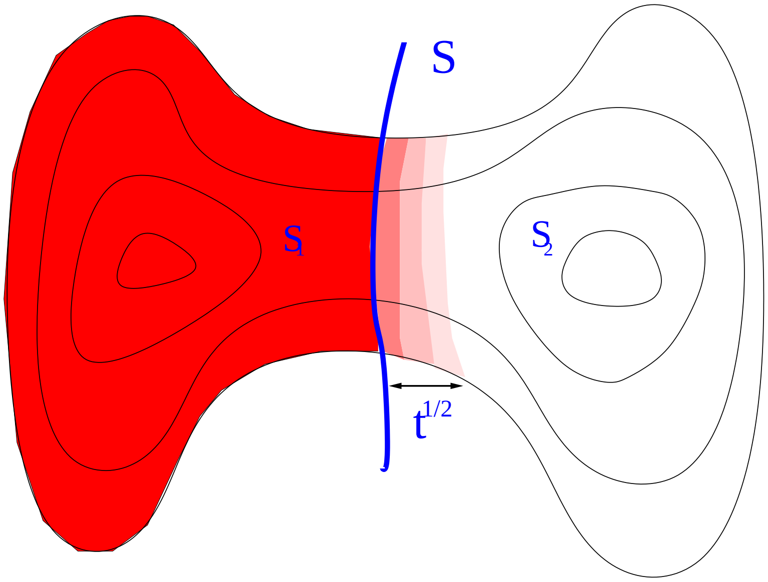

Figure 10.3: Heat flow α tends to $\int_S \rho(s)ds$

Where t and X are clear from context, we shall abbreviate $\beta(t, X)$ to β and $\alpha(t)$ to α. In theorem 10.3.2 we show that as the number of points $|X| = N$ tends to infinity, with probability 1, β tends to α.

In theorem 10.3.3 we show that $\alpha(t)$ can be made arbitrarily close to the weighted volume of the boundary by making t tend to 0.

Theorem 10.3.2. *Let $0 < \mu < 1$. Let $u := 1/\sqrt{t^{2d+1}N^{1-\mu}}$. Then, there exist positive constants C_1, C_2 depending only on ρ and S such that with probability greater than $1 - \exp\left(-C_1 N^{\mu}\right)$*

$$|\beta(t, X) - \alpha(t)| < C_2 \left(1 + t^{\frac{d+1}{2}}\right) u\alpha(t). \tag{10.2}$$

Theorem 10.3.3. *For any $\epsilon \in (0, \frac{1}{2})$, there exists a constant C such that for all t such that*

$0 < t < \tau(2d)^{-\frac{\mathrm{e}}{\mathrm{e}-1}}$,

$$\left|\sqrt{\frac{\pi}{t}}\int_{S_2}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdy - \int_S \rho(s)ds\right| < C\,t^{\epsilon}. \tag{10.3}$$

By letting $N \to \infty$ and $t_N \to 0$ at suitable rates and putting together theorems 10.3.2 and 10.3.3, we obtain the following theorem:

Theorem 10.3.4. *Let the number of random data points $N \to \infty$, and $t_N \to 0$, at rates so that $u := 1/\sqrt{t^{2d+1}N^{1-\mu}} \to 0$. Then, for any $\epsilon \in (0, 1/2)$, there exist positive constants C_1, C_2, such that for any $N > 1$ with probability greater than $1 - \exp(-C_1(N^{\mu}))$,*

$$\left|\beta\,(t_N, X) - \int_S \rho(s)ds\right| < C_2\,(t^{\epsilon} + u) \tag{10.4}$$

10.4 Outline of Proofs

Theorem 10.3.1 is a corollary of Theorem 10.3.4, obtained by setting u to be t^{ϵ}, and making μ as close to 0 as necessary. Theorem 10.3.4 is a direct consequence of Theorem 10.3.2 and Theorem 10.3.3.

Proof of Theorem 10.3.2:

We prove theorem 10.3.2 using a generalization of McDiarmid's inequality from [49, 50]. McDiarmid's inequality asserts that a function of a large number of independent random variables, that is not very influenced by the value of any one of these, takes a value close to its mean. In the generalization that we use, it is permitted that over a bad set that has a small probability mass, the function is highly influenced by some of the random variables. In our setting, it can be shown that our measure of a cut, $f^T\tilde{L}f$ is such a function of the independent random points in X, and so the result is applicable. There is another step involved, since the mean of $f^T\tilde{L}f$ is not α, the quantity to which we wish to prove convergence. Therefore we need to prove that the mean $E[\frac{\sqrt{\pi}}{N\sqrt{t}}f^T\tilde{L}(t,X)f]$ tends to $\alpha(t)$ as

N tends to infinity. Now,

$$\frac{\sqrt{\pi}}{N\sqrt{t}} f^T \tilde{L}(t,X) f = 1/N\sqrt{\pi/t} \sum_{x\in X_1}\sum_{y\in X_2} \frac{G^t(x,y)}{\{(\sum_{z\neq x} G^t(x,z))(\sum_{z\neq y} G^t(y,z))\}^{1/2}}.$$

If, instead, we had in the denominator of the right side

$$\sqrt{\int\limits_M \rho(z)G^t(x,z)dz \int\limits_M \rho(z)G^t(y,z)dz},$$

using the linearity of Expectation,

$$E\left[1/N\sqrt{\pi/t}\sum_{x\in X_1}\sum_{y\in X_2}\frac{G^t(x,y)}{\sqrt{\left(\int\limits_M \rho(z)G^t(x,z)dz\right)\left(\int\limits_M \rho(z)G^t(y,z)dz\right)}}\right] = \alpha.$$

Using Chernoff bounds, we can show that with high probability, for all $x \in X$,

$$\frac{\sum_{z\neq x} G^t(x,z)}{N-1} \approx \int\limits_M \rho(z)G^t(x,z)dz.$$

Putting the last two facts together and using the Generalization of McDiarmid's inequality from [49, 50], the result follows. Since the exact details require fairly technical calculations, we present them later.

Theorem 10.3.3:

The quantity

$$\alpha := \sqrt{\frac{\pi}{t}}\int\limits_{S_1}\int\limits_{S_2} G^t(x,y)\psi_t(x)\psi_t(y)dxdy$$

is similar to the heat that would flow from one part to another if the first were heated proportional to ρ in time t. Intuitively, the heat that would flow from one part to the other

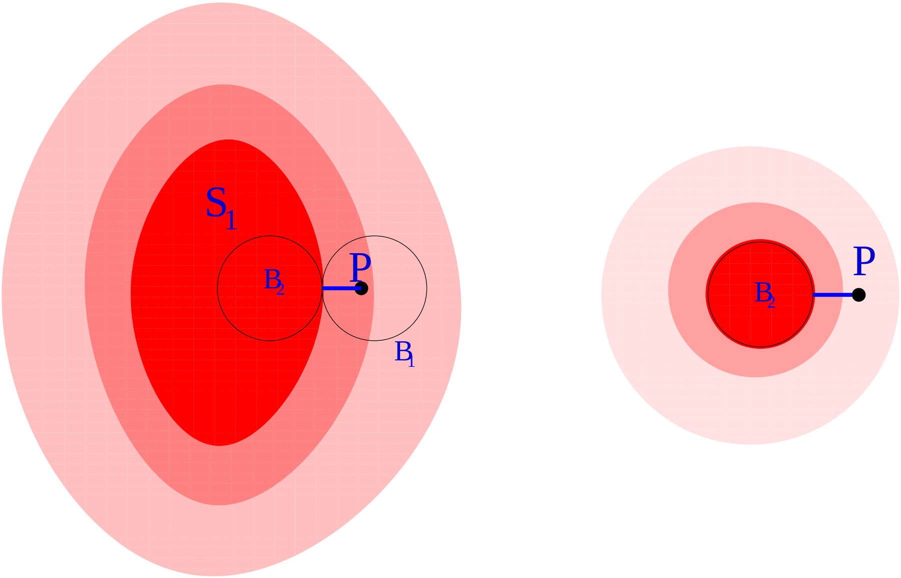

Figure 10.4: Comparing the total heat density at P with the portion due to diffusion from B_2 alone.

in a small interval ought to be related to the volume of the boundary between these two parts, which in our setting is $\int_S \rho(s)ds$. To prove this relationship, we bound α both above and below in terms of the weighted volume and condition number of the boundary. These bounds are obtained by making comparisons with the "worst case", given condition number $\frac{1}{\tau}$, which is when S is a sphere of radius τ. In order to obtain a lower bound on α, we observe that if B_2 is the nearest ball of radius τ contained in S_1 to a point P in S_2 that is within τ of S_1,

$$\int\limits_{S_1} G^t(x,P)\psi_t(x)\psi_t(P)dx \geq \int\limits_{B_2} G^t(x,P)\psi_t(x)\psi_t(P)dx,$$

as in Figure 4. Similarly, to obtain an upper bound on α, we observe that if B_1 is a ball or radius τ in S_2, tangent to B_2 at the point of S nearest to P,

$$\int\limits_{S_1} G^t(x,P)\psi_t(x)\psi_t(P)dx \leq \int\limits_{B_1^c} G^t(x,P)\psi_t(x)\psi_t(P)dx.$$

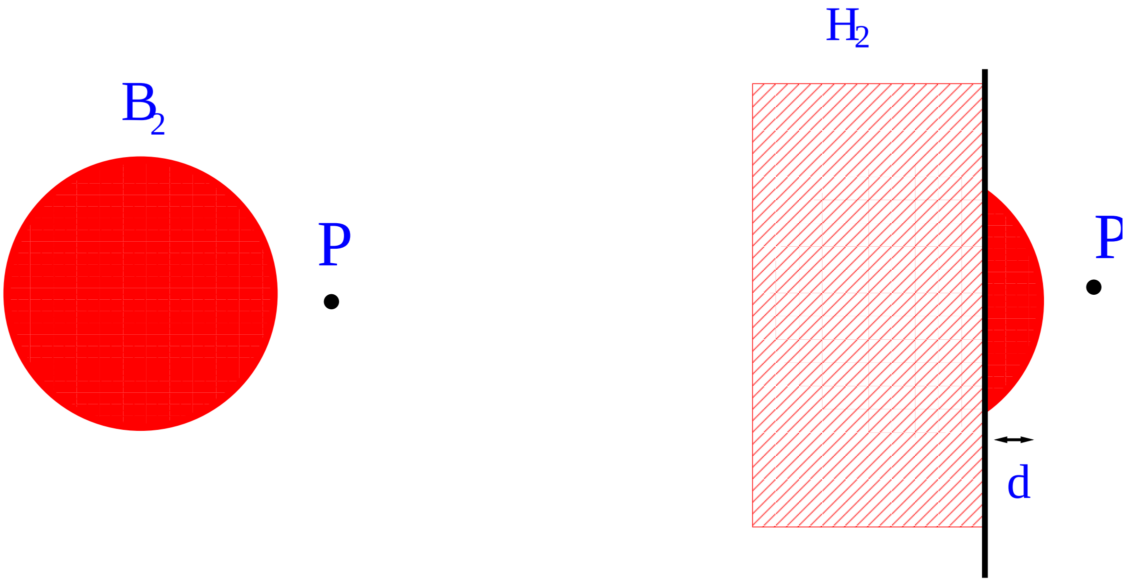

Figure 10.5: The integral over $B_2 \approx$ that over H_2

We now indicate how a lower bound is obtained for

$$\int\limits_{B_2} G^t(x,P)\psi_t(x)\psi_t(P)dx.$$

A key observation is that for $R = \sqrt{2dt\ln(1/t)}$, $\int\limits_{\|x-P\|>R} G^t(x,P)dx << 1$. For this reason, only the portions of B_2 near P contribute to the the integral

$$\int\limits_{B_2} G^t(x,P)\psi_t(x)\psi_t(P)dx.$$

It turns out that a good lower bound can be obtained by considering the integral over H_2 instead, where H_2 is as in figure 10.4.

An upper bound for

$$\int\limits_{B_1^c} G^t(x,P)\psi_t(x)\psi_t(P)dx$$

is obtained along similar lines.

10.5 Proof of Theorem 10.3.1

This follows from Theorem 10.3.4 (which is proved in a later section), by setting μ to be equal to $\frac{1-2\epsilon}{2d+2}$.

10.6 Proof of Theorem 10.3.2

In the proof we will use a generalization of McDiarmid's inequality from [49, 50]. We start with the with the following

Definition 10.6.1. *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_1^m \Omega_k$ and let Y be a random variable on Ω. We say that Y is strongly difference-bounded by (b, c, δ) if the following holds: there is a "bad" subset $B \subset \Omega$, where $\delta = Pr(\omega \in B)$. If $\omega, \omega' \in \Omega$ differ only in the kth coordinate, and $\omega \notin B$, then*

$$|Y(\omega) - Y(\omega')| \leq c.$$

Furthermore, for any ω and ω' differing only in the kth coordinate,

$$|Y(\omega) - Y(\omega')| \leq b.$$

Theorem 10.6.1. *([49, 50]) Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_1^m \Omega_k$ and let Y be a random variable on Ω which is strongly difference-bounded by (b, c, δ). Assume $b \geq c > 0$. Let $\mu = E(Y)$. Then for any $r > 0$,*

$$Pr(|Y - \mu| \geq r) \leq 2 \left(\exp\left(\frac{-r^2}{8mc^2}\right) + \frac{mb\delta}{c} \right).$$

By Hoeffding's inequality

$$
\begin{aligned}
P[|\frac{\sum_{z\neq x} G^t(x,z)}{N-1} - E(G^t(x,z))| > \epsilon_1 E(G^t(x,z))] \quad &< \quad \mathrm{e}^{-\frac{2(N-1)E(G^t(x,z))^2\epsilon_1^2}{M_t^2}} \\
&\leq \quad \mathrm{e}^{-\frac{2(N-1)\rho_{min}^2\epsilon_1^2}{M_t^2}}.
\end{aligned}
$$

We set ϵ_1 to be $M_t/N^{\frac{1-\delta}{2}}$. Let $\mathrm{e}^{-\frac{2(N-1)\rho_{min}^2\epsilon_1^2}{M_t^2}}$ be δ/N. By the union bound, the probability that the above event happens for some $x \in X$ is $\leq \delta$. The set of all $\omega \in \Omega$ for which this occurs shall be denoted by B. Also, for any X, the largest possible value that

$$
1/N\sqrt{\pi/t}\sum_{x\in X_1}\sum_{y\in X_2}\frac{G^t(x,y)}{\{(\sum_{z\neq x} G^t(x,z))(\sum_{z\neq y} G^t(y,z))\}^{1/2}}
$$

could take is $\sqrt{\pi/t}(N-1)$. Then,

$$
|E[\beta]-\alpha| < |1-(1-\epsilon_1)^{-1}|\alpha + \delta\sqrt{\pi/t}(N-1). \tag{10.5}
$$

Let $q = (\rho_{min}/M_t)^2$. β is *strongly difference-bounded* by (b,c,δ) where $c = O((qN\sqrt{t})^{-1})$, $b = O(N/\sqrt{t})$. We now apply the generalization of McDiarmid's inequality in Theorem 10.6.1. Using the notation of Theorem 10.6.1,

$$
\Pr[|\beta - E[\beta]| > r] \leq 2\left(\exp\left(\frac{-r^2}{8mc^2}\right) + \frac{Nb\delta}{c}\right) \tag{10.6}
$$

$$
\leq 2\left(\exp\left(-O(Nr^2q^2t)\right) + O\left(N^3q\exp\left(-O(Nq\epsilon_1^2)\right)\right)\right). \tag{10.7}
$$

Putting this together with the relation between $E[\beta]$ and α in (10.5), the theorem is proved. We note that in (10.5), the rate of convergence of $E[\beta]$ to α is controlled by ϵ_1, which is $M_t/N^{\frac{1-\mu}{2}}$, and in (10.6), the rate of convergence of β to $E[\beta]$ depends on r, which we set to be

$$M_t^2/\sqrt{tN^{1-\mu}}.$$

We note that in (10.6), the dependence on r of the *probability* is exponential. Since we have assumed that $u = M_t^2/\sqrt{(}tN^{1-\mu}) = o(1)$, $M_t/N^{\frac{1-\mu}{2}} = O(t^{\frac{d+1}{2}}u)$. Thus the result follows.

$\square$

10.7 Proof of Theorem 10.3.3

We shall prove theorem 10.3.3 through a sequence of lemmas.

Without a loss of generality we can assume that $\tau = 1$ by rescaling, if necessary.

Let $R = \sqrt{2dt\ln(1/t)}$ and $\epsilon = \int_{\|z\|>R} G^t(0,z)dx$. Using the inequality

$$\int_{\|z\|>R} G^t(0,z)dx \leq \left(\frac{2td}{R}\right)^{-d/2} \mathrm{e}^{-\frac{R^2}{4t}+\frac{d}{2}} = (\mathrm{e}t\ln(1/t))^{d/2} \tag{10.8}$$

we know that $\epsilon \leq (\mathrm{e}t\ln(1/t))^{d/2}$. For any positive real t,

$$\ln(1/t) \leq t^{\frac{-1}{\mathrm{e}}}.$$

Therefore the assumption that

$$\frac{t}{\tau} \in \left(0, \frac{1}{(2d)^{\frac{\mathrm{e}}{\mathrm{e}-1}}}\right)$$

implies that $R \leq \sqrt{2dt^{1-1/\mathrm{e}}} < 1$.

Let the point y (represented as A in figures 10.7 and 10.7) be at a distance $r < R$ from M. Let us choose a coordinate system where $y = (r, 0, \ldots, 0)$ and the point nearest to it on M is the origin. There is a unique such point since $r < R < 1$. Let this point be C. Let

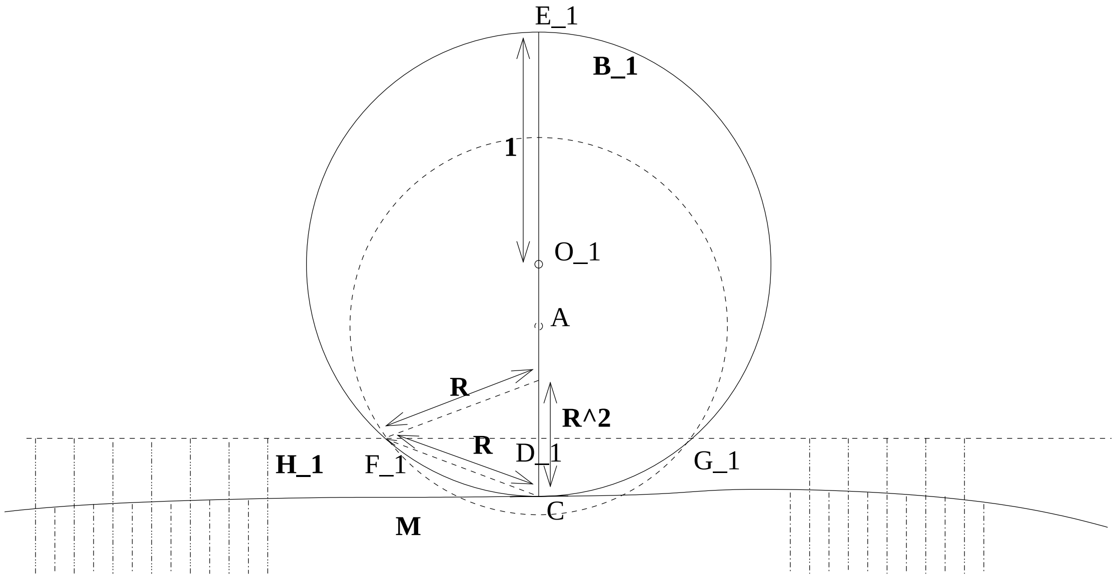

Figure 10.6: A sphere of radius 1 outside S_1 that is tangent to S

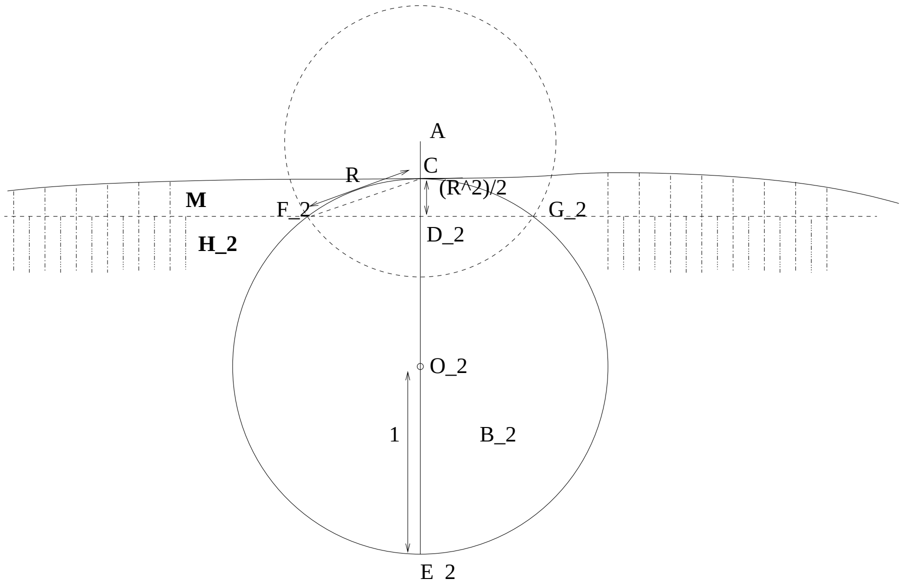

Figure 10.7: A sphere of radius 1 inside S_1 that is tangent to S

D_1 lie on the segment AC, at a distance $R^2/2$ from C. Let D_2 lie on the extended segment AC, at a distance $R^2/2$ from C. Thus C is the midpoint of D_1D_2.

Definition 10.7.1. 1. *Denote the ball of radius* 1 *tangent to* ∂M *at* C *that is outside* M *by* B_1.

2. *Denote the ball of radius* 1 *tangent to* ∂M *at* C *which is inside* M *by* B_2.

3. *Let* H_1 *be the halfspace containing* C *bounded by the hyperplane perpendicular to* AC *and passing through*

   D_1.

4. *Let* H_2 *be the halfspace not containing* C *bounded by the hyperplane perpendicular to* AC *and passing through* D_2.

5. *Let* H_3 *be the halfspace not containing* A, *bounded by the hyperplane tangent to* ∂M *at* C.

6. *Let* B_1' *be the ball with center* $y = A$, *whose boundary contains the intersection of* H_1 *and* B_1.

7. *Let* B_2' *be the ball with center* $y = A$, *whose boundary contains the intersection of* H_2 *and* B_2.

Definition 10.7.2. 1. $h(r) := \int_{H_3} G^t(x, y)dx$.

2. $f(r) := \int_{H_2 \cap B_2'} G^t(x, y)dx$.

3. $g(r) := \int_{H_1 \cap B_1'} G^t(x, y)dx$.

It follows that

$$\int_{H_1} G^t(x, y)dx = h(r - R^2/2)$$

and

$$\int_{H_2} G^t(x, y)dx = h(r + R^2/2).$$

Observation 10.7.1. *Although $h(r)$ is defined by an d-dimensional integral, this can be simplified to*

$$h(r) = \int_{x_1<0} \frac{\mathrm{e}^{-(r-x_1)^2/4t}}{\sqrt{4\pi t}} dx_1,$$

by integrating out the coordinates $x_2, \dots, x_d$.

Lemma 10.7.1. *If $r > R^2$, the radius of B_1' is $\geq R$.*

Proof: By the similarity of triangles CF_1D_1 and CE_1F_1 in figure 10.7, it follows that $\frac{CF_1}{CE_1} = \frac{CD_1}{CF_1}$. $|CE_1| = 2$ and $|CD_1| = R^2/2$. Therefore $CF_1 = R$. Since CD_1F_1 is right angled at D_1, and $|CD_1| = R^2/2$, this proves the claim. □

Lemma 10.7.2. *The radius of B_2' is $\geq R$.*

Proof: By the similarity of triangles CF_2E_2 and CD_2F_2 in figure 10.7 $|CF_2| = R$. However, the distance of point $y := A$ from F_2 is $\geq |CF_2|$. Therefore, the radius of B_2' is $\geq R$. □

Definition 10.7.3. *Let the set of points x such that $B(x, 10R) \subseteq M$ be denoted by M^0. Let $S_1 \cap M^0$ be S_1^0 and $S_2 \cap M^0$ be S_2^0. Let $M - M^0 = M^1$, $S_1 \cap M^1$ be S_1^1 and $S_2 \cap M^1$ be S_2^1. We shall denote $(1 + L/\rho_{min})R)$ by ℓ.*

Consider a point $x \in M^0$, Then,

$$\begin{aligned}\int_M G^t(x,y)\rho(y)dy &\geq \int_{\|y-x\|<R} G^t(x,y)\rho(y)dy \\ &\geq (1-\epsilon)(\rho(x) - LR) \\ &\geq (1-\epsilon)\rho(x)(1 - LR/\rho_{min}) \\ &= \rho(x)(1 - O(\ell))\end{aligned}$$

On the other hand,

$$\begin{aligned}\int_M G^t(x,y)\rho(y)dy &\leq \int_{\|y-x\|\leq 2R} G^t(x,y)\rho(y)dy + \int_{\|y-x\|>2R} G^t(x,y)\rho(y)dy \\ &\leq \rho(x)(1+2\ell) + G^t(0,2R) \\ &= \rho(x)(1 + O((1 + L/\rho_{min})R))\end{aligned}$$

Therefore, $\psi_t(x) = \sqrt{\rho(x)}(1 \pm O((1 + \frac{L}{\rho_{min}})R))$.

Lemma 10.7.3. *$B(x, 5R) \subseteq M$ implies that $\frac{d}{dx}\int G^t(x,z)\rho(z)dy = O(L)$.*

Proof: Consider the function ρ', which is equal to ρ on M, but which has a larger support and is $L-$Lipshitz as a function on $\mathbb{R}^d$. $\int G^t(x,z)\rho'(z)dy$ is $L-$Lipshitz and on points x where $B(x, 5R) \subseteq M$, the contribution of points z outside M, is $o(1)$. Therefore $\frac{d}{dx}\int G^t(x,z)\rho(z)dy = O(L)$. □

This implies that on the set of points x such that $B(x, 5R) \subseteq M$, $\psi_t(X)$ is $O(L)-$Lipshitz. We now estimate $\int_{S_1} G^t(y,z)\rho(z)dz$ for $y \in S_2^0$.

Definition 10.7.4. *For a point $y \in S_2^0$, such that $d(y, S_1) < R < \tau = 1$ let $\pi(y)$ be the nearest point to y in S.*

Note that by the assumption that the condition number of S is 1, since R is smaller than 1, there is a unique candidate for $\pi(y)$. Let y be as in Definition 10.7.4.

Lemma 10.7.4.

$$h(r + R^2/2) - \epsilon < f(r) \leq \int_{S_1} G^t(y, z)dz.$$

Proof:

$$\begin{aligned}
\int_{S_1} G^t(x, y)dx &\geq \int_{H_2 \cap B_2'} G^t(x, y)dx (\text{since } H_2 \cap B_2' \subseteq S_1) \\
&> \int_{H_2} G^t(x, y)dx - \int_{B_2'^c} G^t(x, y)dx \\
&> h(r + R^2/2) - \epsilon
\end{aligned}$$

The last inequality follows from Lemma 10.7.2. □

Lemma 10.7.5. $\int_{S_1} G^t(x, y)\psi_t(x)\psi_t(y)dx > \rho(\pi(y))(1 - O(\ell))(h(r + R^2/2) - \epsilon)$.

Proof:

$$\begin{aligned}
\int_{S_1} G^t(x, y)\psi_t(x)\psi_t(y)dx &\geq \int_{H_2 \cap B_2'} G^t(x, y)\psi_t(x)\psi_t(y)dx (\text{since } H_2 \cap B_2' \subseteq S_1) \\
&> \rho(\pi(y))(1 - O(\ell))(h(r + R^2/2) - \epsilon).
\end{aligned}$$

□

Lemma 10.7.6. *Let $r > R^2$. Then,*

$$\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dx < (1 + O(\ell))(h(r - R^2/2)\rho(\pi(y)) + \epsilon\rho_{max}).$$

Proof:

$$\begin{aligned}
\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dx &\leq \int_{\mathbb{R}^d - B_1} G^t(x,y)\psi_t(x)\psi_t(y)dx \\
&\leq \int_{H_1 \cup \mathbb{R}^d - B_1'} G^t(x,y)\psi_t(x)\psi_t(y)dx \\
&< \int_{H_1 \cap B_1'} G^t(x,y)\psi_t(x)\psi_t(y)dx + \int_{B_1'^c} G^t(x,y)\psi_t(x)\psi_t(y)dx \\
&< h(r - R^2/2)\rho(\pi(y))(1 + O(\ell)) + \epsilon\rho_{max}(1 + O(\ell)) \\
&< (1 + O(\ell))(h(r - R^2/2)\rho(\pi(y)) + \epsilon\rho_{max})
\end{aligned}$$

The last inequality follows from Lemma 10.7.1. □

Definition 10.7.5. *Let $[S_1]_r$ denote the set of points at a distance of $\leq r$ to $[S_1]$. Let π_r be map from $\partial[S_1]_r$ to $\partial[S_1]$ that takes a point P on $\partial[S_1]_r$ to the foot of the perpendicular from P to ∂S_1. (This map is well-defined since $r < \tau = 1$.)*

Lemma 10.7.7. *Let $y \in \partial[S_1]_r$. Let the Jacobian of a map f be denoted by Df.*

$$(1 - r)^{d-1} \leq |D\pi_r(y)| \leq (1 + r)^{d-1}.$$

Proof: Let $\widehat{PQ}$ be a geodesic arc of infinitesimal length ds on ∂S_1 joining P and Q. Let $\pi_r^{-1}(P) = P'$ and $\pi_r^{-1}(Q) = Q'$ (see Figure 10.7.) The radius of curvature of $\widehat{PQ}$ is ≥ 1. Therefore the distance between P' and Q' is in the interval $[ds(1 - r), ds(1 + r)]$. This

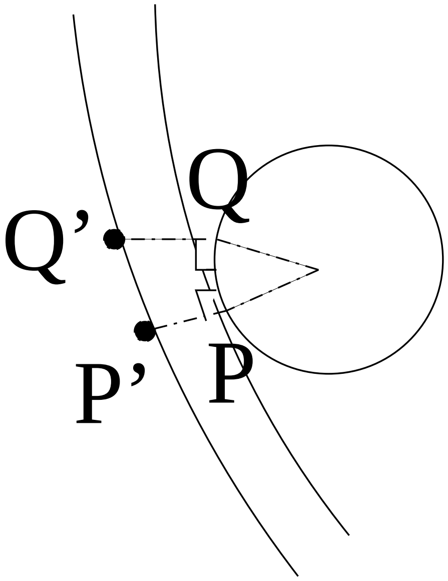

Figure 10.8: The correspondence between points on ∂S_1 and $\partial [S_1]_r$

implies that the Jacobian of the map π_r has a magnitude that is always in the interval $[(1+r)^{1-d}, (1-r)^{1-d}]$. □

Lemma 10.7.8.

$$\int_{\mathbb{R}^d \setminus [S_1]_R} \int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdy \leq \epsilon vol\ S_1 \rho_{max}(1+O(\ell)).$$

Proof:

$$\begin{aligned}
\int_{\mathbb{R}^d \setminus [S_1]_R} \int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdy &= \int_{S_1} \int_{\mathbb{R}^d \setminus [S_1]_R} G^t(x,y)\psi_t(x)\psi_t(y)dydx \\
&\leq \int_{S_1} \int_{\|z\|>R} G^t(0,z)\rho_{max}(1+O(\ell))dzdx \\
&< \text{vol } S_1 \rho_{max}(1+O(\ell)).
\end{aligned}$$

$\leq$ in line 2 holds because the distance between x and y in the double integral is always $\geq R$. □

Lemma 10.7.9.

$$(1 - \mathrm{e}^{-\alpha^2/4t})\sqrt{\pi/t} \leq \int_0^\alpha h(r)dr \leq \sqrt{\pi/t}.$$

Proof: Using observation 10.7.1,

$$\int_\alpha^\infty h(r)dr = \int_\alpha^\infty \int_{-\infty}^0 \frac{\mathrm{e}^{-(x_1-y_1)^2/4t}}{\sqrt{4\pi t}} dx_1 dy_1.$$

Setting $y_1 - x_1 := r$, this becomes

$$\int_\alpha^\infty \int_\alpha^r \frac{\mathrm{e}^{-r^2/4t}}{\sqrt{4\pi t}} dy_1 dr = \int_\alpha^\infty \frac{\mathrm{e}^{-r^2/4t}}{\sqrt{4\pi t}}(r-\alpha)dr.$$

Making the substitution $r - \alpha := z$, we have

$$\begin{aligned} \int_0^\infty \frac{\mathrm{e}^{-(z+\alpha)^2/4t}}{\sqrt{4\pi t}} z dz &\leq \int_0^\infty \frac{\mathrm{e}^{-\alpha^2/4t}\mathrm{e}^{-z^2/4t} z dz}{\sqrt{4\pi t}} \\ &= \sqrt{\frac{t}{\pi}}\mathrm{e}^{-\alpha^2/4t} \end{aligned}$$

Equality holds in the above calculation if and only if $\alpha = 0$. Hence the proof is complete. □

Definition 10.7.6. *Let $[S_2]^0 \cap \partial[S_1]_r$ be ∂M_r. Let $[S_2]^1 \cap \partial[S_1]_r$ be ∂M_r^1 and $[S_2]^1 \cap [S_1]_r$ be M_r^1.*

We assume that $\mathrm{vol}(M_R^1 - S_1) < C'R^2$ for some absolute constant C'. Since the thickness of $(M_R^1 - S_1)$ is $O(R)$ in two dimensions, this is a reasonable assumption to make. The assumption that ∂M has a $d-1-$dimensional volume implies that $\mathrm{vol} S_2^1 = O(R)$.

Proof of theorem 10.3.3:

$$\begin{aligned}
\int_{S_2^1}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdy &= \int_0^\infty \int_{\partial M_r^1}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdydr \\
&\leq O(\rho_{max}^2/\rho_{min})\int_0^\infty \int_{\partial M_r^1}\int_{S_1} G^t(x,y)dxdydr \\
&\leq O(\rho_{max}^2/\rho_{min})\left(\int_0^R \int_{\partial M_r^1}\int_{S_1} G^t(x,y)dxdydr + \text{vol}S_2^1\epsilon\right) \\
&\leq O(\rho_{max}^2/\rho_{min})\left(\text{vol }(M_R^1 - S_1) + \epsilon\text{vol}S_2^1\right). \\
&\leq O(\rho_{max}^2/\rho_{min})\left(C't^{1-\mu} + O(t^{1-\mu}\text{vol }\partial M)\right)
\end{aligned}$$

$$\begin{aligned}
\int_{S_2^0}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdy &= \int_0^\infty \int_{\partial M_r}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdydr \\
&= (\int_0^R \int_{\partial M_r}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdydr) \\
&+ (\int_R^\infty \int_{\partial M_r}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdydr) \\
&\leq (\int_0^R \int_{\partial M_r}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdydr) \\
&+ \underbrace{\epsilon\text{vol } S_1\rho_{max}(1+O(\ell))}_{E}.
\end{aligned}$$

(from lemma 10.7.8)

$$\begin{aligned}
&\leq \int_0^{R^2}\int_{\partial M_r}\rho_{max}(1+O(\ell))dydr \\
&+ \int_{R^2}^R\int_{\partial M_r}(1+O(\ell))(h(r-R^2/2)\rho(\pi(y)+\epsilon\rho_{max})+E
\end{aligned}$$

The last line follows from Lemma 10.7.6.

$$\begin{aligned}
&\leq \; E + R^2(1+R^2)^{d-1}\rho_{max}(1+O(\ell))\text{vol }(\partial M_0)dr(\text{from lemma 10.7.7})\\
&+ \; (1+R)^{d-1}(1+O(\ell))(\int_0^R h(r)dr\int_{\partial M_0}\rho(y)dy + \epsilon\rho_{max}R).\\
&\leq \; (1+O(\ell))(\sqrt{t/\pi}\int_{\partial M_0}\rho(y)dy + \rho_{max}((R^2+\epsilon R)\text{vol }(\partial M_0) + \epsilon\text{vol } S_1)).\\
&\leq \; (1+O(\ell))(\int_{\partial M_0}\rho(y)dy(\sqrt{t/\pi} + \frac{\rho_{max}}{\rho_{min}}o(t^{1-\mu})) + \rho_{max}\epsilon\text{vol } S_1)
\end{aligned}$$

Similarly, we see that

$$\int_{S_2^0}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdy$$

$$\begin{aligned}
&= \; \int_0^\infty\int_{\partial M_r}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdydr\\
&> \; (\int_0^R\int_{\partial M_r}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdydr)\\
&> \; \int_0^R\int_{\partial M_r}\rho(\pi(y))(1+O(\ell))f(r)dxdr\\
&> \; \int_0^R(1-R)^{d-1}(1-O(\ell))(\int_{\partial M_0}\rho(y)dy)(h(r+R^2/2)-\epsilon)dr\\
&> \; (1-R)^{d-1}(1-O((1-L/\rho_{min})R))(\int_{\partial M_0}\rho(y)dy)((\int_0^R(h(r)dr) - \epsilon R - R^2/2)\\
&\geq \; (1-O(\ell))(\int_{\partial M_0}\rho(y)dy)((1-\mathrm{e}^{-R^2/4t})\sqrt{t/\pi} - \epsilon R - R^2/2)\\
&\geq \; (1-O(\ell))(\int_{\partial M_0}\rho(y)dy)(\sqrt{t/\pi} - o(t^{1-\mu})).
\end{aligned}$$

Noting only the dependence of the rate on t, and introducing the condition number τ,

$$\sqrt{\frac{\pi}{t}}\int_{S_2}\int_{S_1} G^t(x,y)\psi_t(x)\psi_t(y)dxdy = \left(1+o((t/\tau^2))^{\frac{1-\mu}{2}}\right)\int_S \rho(s)ds.$$

□

Proof of Theorem 10.3.4: This follows directly from Theorem 10.3.2 and Theorem 10.3.3. The only change made was that the $t^{\frac{d+1}{2}}$ term was eliminated since it is dominated by t^{ϵ} when t is small.

CHAPTER 11

SAMPLE COMPLEXITY OF LEARNING SMOOTH CUTS ON A MANIFOLD

11.1 Introduction

Over the last several years, manifold based methods have been developed and applied to a variety of problems. Much of this work is empirically and algorithmically oriented and there is a need to better understand the learning-theoretic foundations of this class of machine learning problems. This chapter is a contribution in this direction with the hope that it will better delineate the possibilities and limitations.

In the manifold setting, one is canonically interested in learning a function $f : \mathcal{M} \to \mathbb{R}$ (regression) or $f : \mathcal{M} \to \{0, 1\}$ (classification/clustering). For regression therefore, the natural objects of study are classes of real valued functions on the manifold leading one to eventually consider functional analysis on the manifold. Thus, for example, the Laplace-Beltrami operator and its eigenfunctions have been studied with a view to function learning [6, 16].

Our interest in this chapter is the setting for classification or clustering where the function is 0/1 valued and therefore divides the manifold into two disjoint pieces $\mathcal{M}_1$ and $\mathcal{M}_2$. A natural class of such functions may be associated with smooth cuts on the manifold. We will consider *smooth cuts* where each cut corresponds to a submanifold (say $P \subset \mathcal{M}$) that divides $\mathcal{M}$ into two pieces. Since P is a submanifold of $\mathcal{M}$ and hence $\mathbb{R}^m$, one can associate to it a measure of complexity given by its condition number $1/\tau$. The condition number is defined as follows.

Definition 11.1.1 (Condition Number). *Let $\mathcal{M}$ be a smooth $d-$dimensional submanifold of $\mathbb{R}^m$. We define the condition number $c(\mathcal{M})$ to be $\frac{1}{\tau}$, where τ*

is the largest number to have the property that for any $r < \tau$ no two normals of length r that are incident on $\mathcal{M}$ at different points intersect.

Given two linear subspaces V, W, let $\sphericalangle(V, W)$ be the angle between V and W, defined as

$$\sphericalangle(V, W) = \arccos\left(\sup_{v\in V} \inf_{w\in W} \frac{v \cdot w}{\|v\|\|w\|}\right). \tag{11.1}$$

For any manifold $\mathcal{M}$,

$$c(\mathcal{M}) = \inf_{x,y\in L} \frac{2\sin(\frac{\sphericalangle(T_x,T_y)}{2})}{\|x-y\|}, \tag{11.2}$$

where the infimum is taken over distinct points $x, y \in \mathcal{M}$ and T_x and T_y are the tangent spaces at x and y.

We can define the following function class (concept class in PAC terminology.)

Definition 11.1.2. *Let*

$$\mathcal{S}_\tau := \left\{ S \middle| S = \overline{S} \subseteq \mathcal{M} \text{ and } c(S \cap \overline{\mathcal{M} \setminus S}) \leq \frac{1}{\tau} \right\},$$

where $\overline{S}$ is the closure of S. Let

$$\mathcal{C}_\tau := \left\{ f \middle| f : \mathcal{M} \to \{0, 1\} \text{ and } f^{-1}(1) \in \mathcal{S}_\tau \right\}.$$

Thus, the concept class $\mathcal{C}_\tau$ is the collection of indicators of all closed sets in $\mathcal{M}$ whose boundaries are $1/\tau$-conditioned $d-1$ dimensional submanifolds of $\mathbb{R}^m$.

Note that when $\tau = \infty$, $\mathcal{C}_\tau$ contains the indicators of all affine half-subspaces of dimension d that are contained in $\mathcal{M}$. By letting τ vary, we obtain a structured family of cuts. We now consider the following basic question.

[Question:] Let $\mathcal{M}$ be a d-dimensional submanifold of $\mathbb{R}^m$ and let $\mathcal{C}_\tau$ be a concept class of 0/1 valued functions corresponding to a family of smooth cuts with condition number $\frac{1}{\tau}$. Then what is the sample complexity of learning the elements of $\mathcal{C}_\tau$?

Our contributions in this chapter are as follows.

1. We show that distribution-free learning of $\mathcal{C}_\tau$ is impossible in general since for some $\mathcal{M}$, it is a space of infinite VC dimension. We prove that this is the case for a natural embedding in $\mathbb{R}^m$ of the $d-$dimensional sphere of radius $\kappa > \tau$.

2. On the the other hand, it is possible to provide *distribution-specific* sample complexity bounds that hold uniformly for a large class of probability measures on $\mathcal{M}$. These are the measures for which there exists a Radon Nikodym derivative with respect to the uniform measure on $\mathcal{M}$ such that there is an upper bound ρ_{max} on the associated density function. The sample complexity is seen to depend on the intrinsic dimension d, curvature bounds τ and κ, density bound ρ_{max}, but is independent of the ambient dimension m.

3. The proof technique used for obtaining these distribution specific bounds (Poissonization etc.) may be useful to prove distribution specific learning in other settings.

Our sample complexity bounds depend on an upper bound $\rho_{max} \geq 1$ on the maximum density of $\mathcal{P}$ with respect to the volume measure, (normalized to be a probability measure), the curvatures and the intrinsic dimension of $\mathcal{M}$ and the class boundary P, but are independent of the ambient dimension m. We also show that the dependence on the maximum density ρ_{max} of $\mathcal{P}$ is unavoidable by proving that for any fixed τ the VC-dimension of the function class associated with cuts that are submanifolds with a condition number $\frac{1}{\tau}$ is infinite (Lemma 11.3.2) for certain compact submanifolds.

11.2 Preliminaries

Suppose that $\mathcal{P}$ is a probability measure supported on a d-dimensional Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^m$ having condition number $\leq \frac{1}{\kappa}$. Suppose that data samples $\{x_i\}_{i\geq 1}$ are randomly drawn from $\mathcal{P}$ in an i.i.d fashion. Let each data point x be associated with a label $f(x) \in \{0, 1\}$.

Definition 11.2.1 (Annealed Entropy)**.** *Let $\mathcal{P}$ be a probability measure supported on a manifold $\mathcal{M}$. Given a class of indicator functions Λ and a set of points $Z = \{z_1, \dots, z_\ell\} \subset \mathcal{M}$, let $N(\Lambda, Z)$ be the the number of ways of partitioning $z_1, \dots, z_\ell$ into two sets using indicators belonging to Λ . We define $G(\Lambda, \mathcal{P}, \ell)$ to be the expected value of $N(\Lambda, Z)$. Thus*

$$G(\Lambda, \mathcal{P}, \ell) := \mathbb{E}_{Z \vdash \mathcal{P}^{\times \ell}} N(\Lambda, Z),$$

where expectation is with respect to Z and $\vdash$ signifies that Z is drawn from the Cartesian product of ℓ copies of $\mathcal{P}$. The annealed entropy of Λ with respect to ℓ samples from $\mathcal{P}$ is defined to be

$$H_{ann}(\Lambda, \mathcal{P}, \ell) := \ln G(\Lambda, \mathcal{P}, \ell).$$

Definition 11.2.2. *The risk $R(\alpha)$ of a classifier α is defined as the probability that α misclassifies a random data point x drawn from $\mathcal{P}$. Formally, $R(\alpha) := \mathbb{E}_{\mathcal{P}}[\alpha(x) \neq f(x)]$. Given a set of ℓ labeled data points $(x_1, f(x_1)), \dots, (x_\ell, f(x_\ell))$, the empirical risk is defined to be $R_{emp}(\alpha, \ell) := \frac{\sum_{i=1}^{\ell} \mathcal{I}[\alpha(x_i) \neq f(x_i)]}{\ell}$, where $\mathcal{I}[\cdot]$ denotes the indicator of the respective event and $f(x)$ is the label of point x.*

Theorem 11.2.1 (Vapnik [100], Thm 4.2)**.** *For any ℓ the inequality*

$$\mathbb{P}\left[\sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{emp}(\alpha, \ell)}{\sqrt{R(\alpha)}} > \epsilon\right] \quad < \quad 4e^{\left(\frac{H_{ann}(\Lambda, \mathcal{P}, 2\ell)}{\ell} - \frac{\epsilon^2}{4}\right)\ell}$$

holds true, where random samples are drawn from the distribution $\mathcal{P}$.

11.2.1 Remarks

Our setting is the natural generalization of halfspace learning applied to data on a $d-$dimensional sphere. In fact, when the sphere has radius τ, $\mathcal{C}_\tau$ corresponds to halfspaces, and the VC dimension is $d+2$. However, when $\tau < \kappa$, as we show in Lemma 11.3.2, on a $d-$dimensional sphere of radius κ, the VC dimension of $\mathcal{C}_\tau$ is infinite. Interestingly, for these spheres, if $\tau > \kappa$, $\mathcal{C}_\tau$ contains only the function that always takes value 1 and the function that always takes value 0, since there are normals of length κ from center of the sphere to any point of a submanifold embedded in the sphere. In this case, the VC dimension is 1.

If the decision surface is not thin, but there is a margin within which misclassification is not penalized, our results can be adapted to show that the VC dimension is finite.

Our results pertain to the sample complexity of classification of smooth cuts, and does not address algorithmic issues. We are not aware of a way to generate arbitrary $\frac{1}{\tau}-$conditioned cuts. One direction towards addressing algorithmic issues would be to prove bounds on the annealed entropy of the family of linear classifiers in Gaussian Hilbert space. Since the Hilbert space of Gaussians with a fixed width has infinite VC dimension, distribution independent bounds cannot be found and annealed entropy could be a useful tool. Since SVMs based on Gaussian kernels are frequently used for classification, such a result would have algorithmic implications as well.

11.3 Learning Smooth Class Boundaries

Following Definition 11.1.2, let $\mathcal{C}_\tau$ be the collection of indicators of all open sets in $\mathcal{M}$ whose boundaries are $1/\tau$-conditioned submanifolds of $\mathbb{R}^m$ of dimension $d-1$.

Our main theorem is the following.

Definition 11.3.1 (Packing number). *Let $N_p(\epsilon_r)$ be the largest number N such that $\mathcal{M}$ contains N disjoint balls $B_{\mathcal{M}}(x_i, \epsilon_r)$, where $B_{\mathcal{M}}(x, \epsilon_r)$ is a geodesic ball in $\mathcal{M}$ around x of radius ϵ_r.*

Notation 11.3.1. *Without loss of generality, let ρ_{max} be greater or equal to 1. Let $\epsilon_r = \min(\frac{\tau}{4}, \frac{\kappa}{4}, 1)\epsilon/(2\rho_{max})$. For some sufficiently large universal constant C, let*

$$\ell := C \left(\frac{\ln \frac{1}{\delta} + N_p(\epsilon_r/2) d \ln(d\rho_{max}/\epsilon)}{\epsilon^2} \right).$$

Theorem 11.3.1. *Let $\mathcal{M}$ be a $d-$dimensional submanifold of $\mathbb{R}^m$ whose condition number is $\leq \frac{1}{\kappa}$. Let $\mathcal{P}$ be a probability measure on $\mathcal{M}$, whose density relative to the uniform probability measure on $\mathcal{M}$ is bounded above by ρ_{max}. Then the number of random samples needed before the empirical risk and the true risk are close uniformly over $\mathcal{C}_\tau$ can be bounded above as follows. Let ℓ be defined as in Notation 11.3.1. then*

$$\mathbb{P}\left[\sup_{\alpha \in \mathcal{C}_\tau} \frac{R(\alpha) - R_{emp}(\alpha, \ell)}{\sqrt{R(\alpha)}} > \sqrt{\epsilon} \right] \quad < \quad \delta$$

Proof. The proof follows from Theorem 11.3.2 and Theorem 11.2.1. The former provides a bound on the annealed entropy of $\mathcal{C}_\tau$ with respect to samples from $\mathcal{P}$. The latter relates the sample complexity of learning an element of a class of indicators such as $\mathcal{C}_\tau$ using random samples drawn from a distribution $\mathcal{P}$, to the annealed entropy of that class. □

Lemma 11.3.1 provides a lower bound on the sample complexity that shows that some dependence on the packing number cannot be avoided in Theorem 11.3.1. Further, Lemma 11.3.2 shows that it is impossible to learn an element of $\mathcal{C}_\tau$ in a distribution-free setting in general.

Lemma 11.3.1. *Let $\mathcal{M}$ be a $d-$dimensional sphere in $\mathbb{R}^m$. Let the $\mathcal{P}$ have a uniform density*

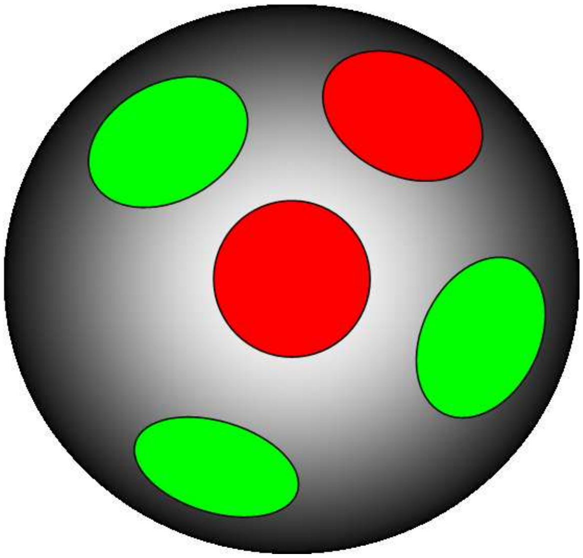

Figure 11.1: This illustrates the distribution from Lemma 11.3.1. The intersections of $f^{-1}(1)$ and $f^{-1}(0)$ with the support of $\mathcal{P}$ are respectively green and red.

over the disjoint union of $N_p(2\tau)$ identical spherical caps

$$S = \{B_{\mathcal{M}}(x_i, \tau)\}_{1 \le i \le N_p(2\tau)}$$

of radius τ, whose mutual distances are all $\ge 2\tau$. Then, if $s < (1-\epsilon)N_p(2\tau)$,

$$\mathbb{P}\left[\sup_{\alpha \in \mathcal{C}_\tau} \frac{R(\alpha) - R_{emp}(\alpha, s)}{\sqrt{R(\alpha)}} > \sqrt{\epsilon}\right] \;=\; 1.$$

Proof. Suppose that the labels are given by $f : \mathcal{M} \to \{0,1\}$, such that $f^{-1}(1)$ is the union of some of the caps in S as depicted in Figure 1. Suppose s random samples $z_1, \dots, z_s$ are chosen from $\mathcal{P}$. Then at least $\epsilon N_p(2\tau)$ of the caps in S do not contain any of the z_i. Let X be the union of these caps. Let $\alpha : \mathcal{M} \to \{0,1\}$ satisfy $\alpha(x) = 1 - f(x)$ if $x \in X$ and $\alpha(x) = f(x)$ if $x \in \mathcal{M} \setminus X$. Note that $\alpha \in \mathcal{C}_\tau$. However, $R_{emp}(\alpha, s) = 0$ and $R(\alpha) \ge \epsilon$. Therefore $\frac{R(\alpha) - R_{emp}(\alpha,s)}{\sqrt{R(\alpha)}} > \sqrt{\epsilon}$, which completes the proof. □

Lemma 11.3.2. *For any $m > d \ge 2$, and $\tau > 0$, there exist compact $d-$dimensional*

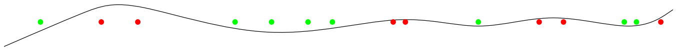

Figure 11.2: The 1−dimensional submanifold P of $\mathbb{R}^2$ traced out by all points $(x, f_S(x))$.

manifolds on which the VC dimension of $\mathcal{C}_\tau$ is infinite. In particular, this is true for the standard $d-$dimensional Euclidean sphere of radius κ embedded in $\mathbb{R}^m$, where $m > d \geq 2$ and $\kappa > \tau$.

Proof. First consider the two dimensional plane $\mathbb{R}^2$. Suppose that for $i = 0$ to $n-1$, $x_i = (i\tau/n, 0)$. If there is no bound on the condition number, we make the following claim.

Claim 11.3.1. *For every subset $S \subseteq [n]$ there exists a boundary given by a graph $(x, f_S(x))$, $f_S : \mathbb{R} \to \mathbb{R}$ such that the following hold.*

1. *$f_S(x_i) > 0$ if $i \in S$ (see Figure 2) and $f_S(x_i) < 0$ if $i \in [n] \setminus S$.*
2. *f is thrice continuously differentiable.*
3. *For all $x \in \mathbb{R}$, $|f''_S(x)| < \frac{1}{\gamma} := \frac{1}{2M\tau}$ for some large constant $M >> 1$ and for all x such that $|x| \geq \tau$, $f_S(x) = 0$.*

It is clear that for any S a function g_S exists that satisfy the first two conditions. We will use g_S to obtain f_S.

To see this, note that the radius of curvature at any point $(x, g_S(x))$ is given by $\frac{(1+g'_S(x)^2)^{\frac{3}{2}}}{|g''_S(x)|}$. Now, let

$$\alpha = \sup_{S\subseteq[n], x\in[-\tau,\tau]} \max(|g'_S(x)|, |g''_S(x)|).$$

Let $f_S(x) = \frac{g_S(x)}{\gamma\alpha}$. The 1−dimensional submanifold P of $\mathbb{R}^2$ traced out by all points $(x, f_S(x))$ has curvature $\leq \frac{1}{\gamma}$ because for all $x \in [-\tau, \tau]$, for all S,

$$\frac{\alpha\gamma\left(1 + \left(\frac{g'_S(x)}{\alpha\gamma}\right)^2\right)^{3/2}}{|g''_S(x)|} \geq \Omega(\gamma).$$

Let $S^2_\kappa = \{(x,y,z)\big|x^2+y^2+(z-\kappa)^2 = \kappa^2\}$ be the 2−sphere of radius $\kappa > \tau$ tangent to the (x,y) plane at the origin. Consider the stereographic projection υ_κ of $S^2_\kappa \setminus \{0,0,2\kappa\}$ onto $\mathbb{R}^2$ (embedded in $\mathbb{R}^3$), defined by

$$\upsilon_\kappa(x,y,z) := \left(\frac{2\kappa x}{2\kappa - z}, \frac{2\kappa y}{2\kappa - z}, 0\right).$$

Let B be the ball of radius 1 centered at the origin in the image of υ_κ. As $M \to \infty$, $\upsilon_\kappa^{-1}(B \cap P)$ tends uniformly to a great circle, and its tangent spaces (see (11.1) tend uniformly to the corresponding tangent spaces of the great circle in terms of the angle. Therefore, (by (11.2)) for sufficiently large M, the condition number of $\upsilon_\kappa^{-1}(P)$ is less than $\frac{1}{\tau}$, completing the proof. This argument carries over to when $S^2_\kappa \in \mathbb{R}^m$ for $m > 3$. Now, we may extend the copy of $\mathbb{R}^2$ that we considered to $\mathbb{R}^d$ by taking the canonical embedding $\mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^{d-2}$. The 1−dimensional manifold P can be similarly extended to obtain a $m-1$−dimensional submanifold $P \times \mathbb{R}^{d-2}$. We can then consider as we did in the case of $\mathbb{R}^2$, the stereographic projection that maps the d−sphere

$$S^d_\kappa = \{(x,y,z_1,z_2,\ldots,z_{d-1})\big|$$

$$x^2+y^2+z_1^2+\ldots+(z_{d-1}-\kappa)^2 = \kappa^2\}$$

onto $\mathbb{R}^d$ by the map

$$\upsilon_\kappa(x,y,z_1,\ldots,z_{d-1}) :=$$

$$\left(\frac{2\kappa x}{2\kappa - z_{d-1}}, \frac{2\kappa y}{2\kappa - z_{d-1}}, \frac{2\kappa z_1}{2\kappa - z_{d-1}}, \ldots, \frac{2\kappa z_{d-2}}{2\kappa - z_{d-1}}, 0\right),$$

and the same argument carries through. □

We shall nonetheless uniformly bound from above, the annealed entropy of $\mathcal{C}_\tau$ with respect to any distribution $\mathcal{P}$ on $\mathcal{M}$, whose density (with respect to the uniform probability measure) on $\mathcal{M}$ is bounded above by ρ_{max}. The number of samples that need to be taken before the empirical risk is within ϵ of the true risk, uniformly over $\mathcal{C}_\tau$ with probability $1-\delta$ is determined by the annealed entropy of $\mathcal{C}_\tau$ w.r.t $\mathcal{P}$. We have the following theorem that bounds the annealed entropy from above.

Theorem 11.3.2. *Let $\mathcal{M}$ be a $d-$dimensional submanifold of $\mathbb{R}^m$ whose condition number is $\leq \frac{1}{\kappa}$. Let $\mathcal{P}$ be a probability measure on $\mathcal{M}$, whose density relative to the uniform probability measure on $\mathcal{M}$ is bounded above by ρ_{max}. When the number n of random samples from $\mathcal{P}$ is large, the annealed entropy of $\mathcal{C}_\tau$ can be bounded from above as follows. Let $\epsilon_r = \min(\frac{\tau}{4}, frac\kappa 4, 1)\epsilon/(2\rho_{max})$. Suppose*

$$n \geq N_p(\epsilon_r/2)\frac{d\ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon^2},$$

then,

$$H_{ann}(\mathcal{C}_\tau, \mathcal{P}, \lfloor n - \sqrt{n\ln(2\pi n)}\rfloor) \quad \leq \quad 4\epsilon n + 1.$$

11.3.1 Overview of the Proof of Theorem 11.3.2

Our strategy is as follows.

1. Cut the manifold into small pieces $\mathcal{M}_i$ that are almost Euclidean, such that the restrictions of any cut hypersurface is almost linear.

2. Let the probability measure $\frac{\mathcal{P}|_{\mathcal{M}_i}}{\mathcal{P}(\mathcal{M}_i)}$ be denoted $\mathcal{P}_i$ for each i. Lemma 11.3.7 allows us to show, roughly, that

$$\frac{H_{ann}(\mathcal{C}_\tau, \mathcal{P}, n)}{n} \lesssim \sup_i \frac{H_{ann}(\mathcal{C}_\tau, \mathcal{P}_i, \lfloor n\mathcal{P}(\mathcal{M}_i)\rfloor)}{\lfloor n\mathcal{P}(\mathcal{M}_i)\rfloor},$$

thereby allowing us to focus on a single piece $\mathcal{M}_i$.

3. We use a projection π_i, to map $\mathcal{M}_i$ orthogonally onto the tangent space to $\mathcal{M}_i$ at a point $x_i \in \mathcal{M}_i$ and then reduce the question to a sphere inscribed in a cube $\Box$ of Euclidean space.

4. We cover $\mathcal{C}_\tau\big|_\Box$ by the union of classes of functions that are constant outside a thin slab (see Definition 11.3.5 and Figure 3).

5. Finally, we bound the annealed entropy of each of these classes using Lemma 11.3.8.

The rest of this chapter is devoted to a detailed treatment of the proof of Theorem 11.3.2.

11.3.2 Volumes of balls in a manifold

Let $\mathcal{M} \subseteq \mathbb{R}^m$ be a d-dimensional Riemannian manifold and let P be a $d-1-$dimensional submanifold of $\mathcal{M}$. Let $V_x^{\mathcal{M}}(r)$ be defined to be the volume of a ball of radius r (in the intrinsic metric) around a point $x \in \mathcal{M}$. The sectional curvature of a manifold at a point x depends on a two-dimensional plane in the tangent space at x. A formal definition of sectional curvature can be found in most textbooks of differential geometry (for example, [46]). The volumes of balls can be estimated using sectional curvatures. The Bishop-Günther inequalities tell us that if the sectional curvature $K^{\mathcal{M}}$ is upper bounded by λ, then the volume of the ball of radius r around x, $V_x^{\mathcal{M}}$ is bounded from below as follows (section 3.5, [28]).

$$V_x^{\mathcal{M}}(r) \geq \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \int_0^r \left(\frac{\sin(t\sqrt{\lambda})}{\sqrt{\lambda}}\right)^{d-1} dt,$$

where $\Gamma(x)$ is Euler's Γ function.

This allows us to get an explicit upper bound on the packing number $N_p(\epsilon_r/2)$, namely

$$N_p(\epsilon_r/2) \leq \frac{\mathrm{vol}\mathcal{M}}{\frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \int_0^{\epsilon_r/2} \left(\frac{\sin(t\sqrt{\lambda})}{\sqrt{\lambda}}\right)^{d-1} dt}.$$

11.3.3 Partitioning the Manifold

The next step is to partition the manifold $\mathcal{M}$ into disjoint pieces $\{\mathcal{M}_i\}$ such that each piece $\mathcal{M}_i$ is contained in the geodesic ball $B_{\mathcal{M}}(x_i, \epsilon_r)$. Such a partition can be constructed by the following natural greedy procedure.

- Choose $N_p(\epsilon_r/2)$ disjoint balls $B_{\mathcal{M}}(x_i, \epsilon_r/2)$, $1 \leq i \leq N_p(\epsilon_r/2)$ where $N_p(\epsilon_r/2)$ is the packing number as in Definition 11.3.1.
- Let $\mathcal{M}_1 := B_{\mathcal{M}}(x_1, \epsilon_r)$.
- Iteratively, for each $i \geq 2$, let $\mathcal{M}_i := B_{\mathcal{M}}(x_i, \epsilon_r) \setminus \{\cup_{k=1}^{i-1} \mathcal{M}_k\}$.

11.3.4 Constructing charts by projecting onto Euclidean Balls

In this section, we show how the question can be reduced to Euclidean space using a family of charts. The strategy is the following. Let ϵ_r be as defined in Notation 11.3.1. Choose a set of points $X = \{x_1, \dots, x_N\}$ belonging to $\mathcal{M}$ such that the union of geodesic balls in $\mathcal{M}$ (measured in the intrinsic Riemannian metric) of radius ϵ_r centered at these points in $\mathcal{M}$ covers all of $\mathcal{M}$.

$$\bigcup_{i \in [N]} B_{\mathcal{M}}(x_i, \epsilon_r) = \mathcal{M}.$$

Definition 11.3.2. *For each $i \in [N_p(\epsilon_r/2)]$, let the $d-$dimensional affine subspace of $\mathbb{R}^m$ tangent to $\mathcal{M}$ at x_i be denoted $\mathcal{A}_i$, and let the d-dimensional ball of radius ϵ_r contained in $\mathcal{A}_i$, centered at x_i be $B_{\mathcal{A}_i}(x_i, \epsilon_r)$. Let the orthogonal projection from $\mathbb{R}^m$ onto $\mathcal{A}_i$ be denoted π_i.*

Lemma 11.3.3. *The image of $B_{\mathcal{M}}(x_i, \epsilon_r)$ under the projection π_i is contained in the corresponding ball $B_{\mathcal{M}}(x_i, \epsilon_r)$ in $\mathcal{A}_i$.*

$$\pi_i(B_{\mathcal{M}}(x_i, \epsilon_r)) \subseteq B_{\mathcal{A}_i}(x_i, \epsilon_r).$$

Proof. This follows from the fact that the length of a geodesic segment on $B_{\mathcal{M}}(x_i, \epsilon_r)$ is greater or equal to the length of its image under a projection. □

Let P be a smooth $1/\tau$-conditioned boundary (i. e. $c(P) \leq \frac{1}{\tau}$) separating $\mathcal{M}$ into two parts. and $c(\mathcal{M}) \leq \frac{1}{\kappa}$.

Lemma 11.3.4. *Let $\epsilon_r \leq \min(1, \tau/4, \kappa/4)$. Let $\pi_i(B_{\mathcal{M}}(x_i, \epsilon_r) \cap P)$ be the image of P restricted to $B_{\mathcal{M}}(x_i, \epsilon_r)$ under the projection π_i. Then, the condition number of $\pi_i(B_{\mathcal{M}}(x_i, \epsilon_r) \cap P)$ is bounded above by $\frac{2}{\tau}$.*

Proof. Let $T_{\pi_i(x)}$ and $T_{\pi_i(y)}$ be the spaces tangent to L at $\pi_i(x)$ and $\pi_i(y)$ respectively. Then, for any $x, y \in B_{\mathcal{M}}(x_i, \epsilon_r) \cap P$, because the kernel of π_i is nearly orthogonal to $T_{\pi_i(x)}$ and $T_{\pi_i(y)}$,

$$\sphericalangle(T_{\pi_i(x)}, T_{\pi_i(y)}) \leq \sqrt{2} \sphericalangle(T_x, T_y). \tag{11.3}$$

$B_{\mathcal{M}}(x_i, \epsilon_r) \cap P$ is contained in a neighborhood of the affine space tangent to $B_{\mathcal{M}}(x_i, \epsilon_r) \cap P$ at x_i, which is orthogonal to the kernel of π_i. After some calculation, this can be used to show that for all $x, y \in B_{\mathcal{M}}(x_i, \epsilon_r) \cap P$,

$$\frac{1}{\sqrt{2}} \leq \frac{\|\pi_i(x) - \pi_i(y)\|}{\|x - y\|} \leq 1. \tag{11.4}$$

The lemma follows from (11.2). □

11.3.5 Proof of Theorem 11.3.2

We shall organize this proof into several Lemmas, which will be proved immediately after their respective statements. The following Lemma allows us to work with a random rather than deterministic number of samples. The purpose of allowing the number of samples to be a Poisson random variable is that we are able make the set of numbers of samples $\{\nu_i\}$ from different $\mathcal{M}_i$, a collection of independent random variables.

Lemma 11.3.5 (Poissonization)**.** *Let ν be a Poisson random variable with mean λ, where $\lambda > 0$. Then, for any $\epsilon > 0$ the expected value of the annealed entropy of a class of indicators with respect to ν random samples from a distribution $\mathcal{P}$ is asymptotically greater or equal to the annealed entropy of $\lfloor (1-\epsilon)\lambda \rfloor$ random samples from the distribution $\mathcal{P}$. More precisely, for any $\epsilon > 0$,* $\ln \mathbb{E}_\nu G(\Lambda, \mathcal{P}, \nu) \geq \ln G(\Lambda, \mathcal{P}, \lfloor \lambda(1-\epsilon) \rfloor) - \exp\left(-\epsilon^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right)$.

Proof.

$$\begin{aligned}
\ln \mathbb{E}_\nu G(\Lambda, \mathcal{P}, \nu) &= \ln \sum_{n \in \mathbb{N}} \mathbb{P}[\nu = n] H_{ann}(\Lambda, \mathcal{P}, n) \\
&\geq \ln \sum_{n \geq \lfloor \lambda(1-\epsilon) \rfloor} \mathbb{P}[\nu = n] G(\Lambda, \mathcal{P}, n).
\end{aligned}$$

$G(\Lambda, \mathcal{P}, n)$ is monotonically increasing as a function of n. Therefore the above expression can be lower bounded by $\ln \mathbb{P}[\nu \geq \lfloor \lambda(1-\epsilon) \rfloor] G(\Lambda, \mathcal{P}, \nu) \geq H_{ann}(\Lambda, \mathcal{P}, \lfloor \lambda(1-\epsilon) \rfloor)$
$- \exp\left(-\epsilon^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right)$. □

Definition 11.3.3. *For each $i \in [N_p(\epsilon_r/2)]$, let $\mathcal{P}_i$ be the restriction of $\mathcal{P}$ to $\mathcal{M}_i$. Let $|\mathcal{P}_i|$ denote the total measure of $\mathcal{P}_i$. Let λ_i denote $\lambda|\mathcal{P}_i|$. Let $\{\nu_i\}$ be a collection of independent Poisson random variables such that for each $i \in [N_p(\epsilon_r/2)]$, the mean of ν_i is λ_i.*

The following Lemma allows us to focus our attention to small pieces $\mathcal{M}_i$ which are almost Euclidean.

Lemma 11.3.6 (Factorization)**.** *The quantity $\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu)$ is less or equal to the sum over i of the corresponding quantities $\mathcal{C}_\tau$ with respect to ν_i random samples from $\mathcal{P}_i$. i. e.*

$$\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu) \leq \sum_{i \in N_p(\epsilon_r/2)} \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i).$$

Proof.

$$G(\mathcal{C}_\tau, \mathcal{P}, \ell) := \ln \mathbb{E}_{X \vdash \mathcal{P}^{\times \ell}} N(\mathcal{C}_\tau, X),$$

where expectation is with respect to X and $\vdash$ signifies that X is drawn from the Cartesian product of ℓ copies of $\mathcal{P}$. The number of ways of splitting $X = \{x_1, \dots, x_k, \dots, x_\ell\}$ using elements of $\mathcal{C}_\tau$, $N(\mathcal{C}_\tau, X)$ satisfies a sub-multiplicative property, namely

$$N(\mathcal{C}_\tau, \{x_1, \dots, x_\ell\}) \leq$$

$$N(\mathcal{C}_\tau, \{x_1, \dots, x_k\}) N(\mathcal{C}_\tau, \{x_{k+1}, \dots, x_\ell\}).$$

This can be iterated to generate inequalities where the right side involves a partition with any integer number of parts. Note that $\mathcal{P}$ is a mixture of the $\mathcal{P}_i$, and can be expressed as

$$\mathcal{P} = \sum_i \frac{\lambda_i}{\lambda} \mathcal{P}_i.$$

A draw from $\mathcal{P}$ of a Poisson number of samples can be decomposed as the union of independently chosen sets of samples. The i^{th} set is a draw of size ν_i from $\mathcal{P}_i$, ν_i being a Poisson random variable having mean λ_i. These facts imply that
$\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu) \leq \sum_{i \in N_p(\epsilon_r/2)} \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)$.

□

Lemma 11.3.6 can be used together with an upper bound on annealed entropy based on the number of samples to obtain

Lemma 11.3.7 (Localization)**.** *For any* $\epsilon' > 0$

$$\frac{\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu)}{\lambda} \leq \sup_{i \ s.t \ |\mathcal{P}_i| \geq \frac{\epsilon'}{N_p(\epsilon_r/2)}} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} + \epsilon'.$$

Proof. Lemma 11.3.7 allows us to reduce the question to a single $\mathcal{M}_i$ in the following way.

$$\frac{\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu)}{\lambda} \leq \sum_{i \in N_p(\epsilon_r/2)} \frac{\lambda_i}{\lambda} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i}$$

Allowing all summations to be over i s.t $|\mathcal{P}_i| \geq \frac{\epsilon'}{N_p(\epsilon_r/2)}$, the right side can be split into

$$\sum_i \frac{\lambda_i}{\lambda} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} + \sum_i \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i).$$

$G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)$ must be less or equal to the expression obtained in the case of complete shattering, which is 2^{ν_i}. Therefore the second term in the above expression can be bounded above as follows,

$$\begin{aligned}\sum_i \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i) &\leq \sum_i \ln \mathbb{E}_{\nu_i} 2^{\nu_i} \\ &= \sum_i \lambda_i \\ &\leq \epsilon'.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu)}{\lambda} &\leq \sum_i \frac{\lambda_i}{\lambda} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} + \epsilon' \\ &\leq \sup_i \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} + \epsilon'.\end{aligned}$$

□

As mentioned earlier, Lemma 11.3.7 allows us to reduce the proof to a question concerning a single piece $\mathcal{M}_i$. This is more convenient because $\mathcal{M}_i$ can be projected onto a single Euclidean ball in the way described in Section 11.3.4 without incurring significant distortion.

By Lemmas 11.3.3 and 11.3.4, the question can be transferred to one about the annealed entropy of the induced function class $\mathcal{C}_\tau \circ \pi_i^{-1}$ on chart $B_{\mathcal{A}_i}(x_i, \epsilon_r)$ with respect to ν_i random samples from the projected probability distribution $\pi_i(\nu_i)$. $\mathcal{C}_\tau \circ \pi_i^{-1}$ is contained in $\mathcal{C}_{\tau/2}(\mathcal{A}_i)$ which is the analogue of $\mathcal{C}_{\tau/2}$ on $\mathcal{A}_i$. For simplicity, henceworth we shall abbreviate $\mathcal{C}_{\tau/2}(\mathcal{A}_i)$ as $\mathcal{C}_{\tau/2}$. Then,

$$\begin{aligned}\frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} &= \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau \circ \pi_i^{-1}, \pi_i(\mathcal{P}_i), \nu_i)}{\lambda_i} \\ &\leq \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_{\tau/2}, \pi_i(\mathcal{P}_i), \nu_i)}{\lambda_i}.\end{aligned}$$

We inscribe $B_{\mathcal{A}_i}(x_i, \epsilon_r)$ in a cube of side $2\epsilon_r$ for convenience, and proceed to find the desired upper bound on

$G(\mathcal{C}_{\tau/2}, \pi_i(\mathcal{P}_i), \nu_i)$. We shall indicate how to achieve this using covers. For convenience, let this cube be dilated until we have the cube of side 2. The measure $\pi_i(\mathcal{P}_i)$ assigns to it must be scaled to a probability measure that we call $\mathcal{P}_\circ$, which is actually supported on the inscribed ball. We shall normalize all quantities appropriately when the calculations are over. The $\tau_\Box$ that we shall work with below is a rescaled version of the original, $\tau_\Box = \tau/\epsilon_r$. Let B_∞^d be the cube of side 2 centered at the origin and ι_∞^d be its indicator. Let B_2^d be the unit ball inscribed in B_∞^d.

Definition 11.3.4. *Let $\tilde{C_{\tau_\Box}}$ be defined to be the set of all indicators of the form $\iota_\infty^d \cdot \iota$, where ι is the indicator of some set in $\mathcal{C}_{\tau_\Box}$.*

In other words, $\tilde{\mathcal{C}}_{\tau_\Box}$ is the collection of all functions that are indicators of sets that can be expressed as the intersection of the unit cube and an element of $\mathcal{C}_{\tau_\Box}$.

$$\tilde{C}_{\tau_\Box} = \{f \mid \exists c \in \mathcal{C}_{\tau_\Box}, \text{ for which } f = \mathcal{I}_c \cdot \iota_\infty^d\}, \tag{11.5}$$

where $\mathcal{I}_c$ is the indicator of c.

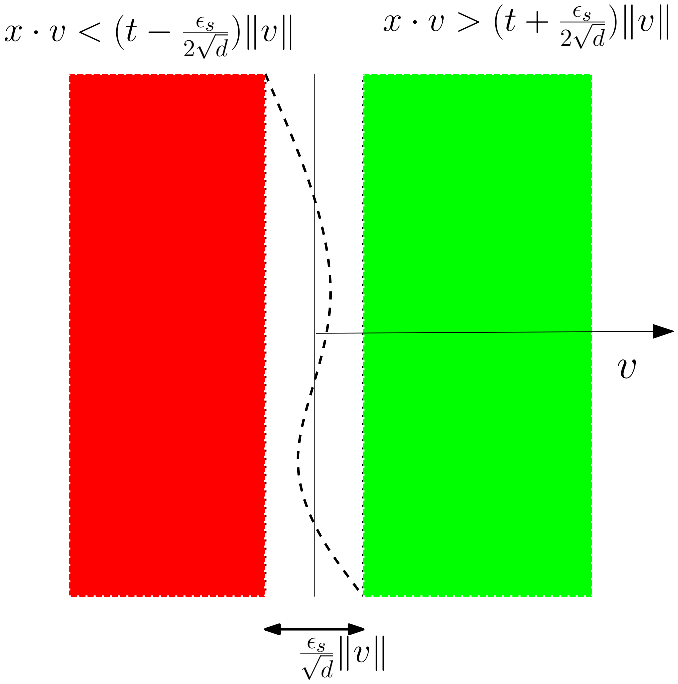
$x \cdot v < (t - \frac{\epsilon_s}{2\sqrt{d}})\|v\|$
$x \cdot v > (t + \frac{\epsilon_s}{2\sqrt{d}})\|v\|$
v
$\frac{\epsilon_s}{\sqrt{d}}\|v\|$

Figure 11.3: Each class of the form $\tilde{\mathcal{C}}_{\epsilon_s}^{(v,t)}$ contains a subset of the set of indicators of the form $\mathcal{I}_c \cdot \iota_\infty^d$

Definition 11.3.5. *For every $v \in \mathbb{R}^d$ where $\|v\| = 1$, $t \in \mathbb{R}$ and $\epsilon > 0$ and $\epsilon_s = \epsilon^2/\rho_{max}$. Let $\tilde{\mathcal{C}}_{\epsilon_s}^{(v,t)}$ be a class of indicator functions consisting of all those measurable indicators ι that satisfy the following.*

1. *$x \cdot v < (t - \frac{\epsilon_s}{2\sqrt{d}})\|v\|$ or $x \notin B_\infty^d \Rightarrow \iota(x) = 0$ and*
2. *$x \cdot v > (t + \frac{\epsilon_s}{2\sqrt{d}})\|v\|$ and $x \in B_\infty^d \Rightarrow \iota(x) = 1$.*

The VC dimension of the above class is clearly infinite since any samples lying within the slab of thickness $\epsilon_s/\sqrt{d}$ get shattered. However if a distribution is sufficiently uniform, most samples would lie outside the slab and so the annealed entropy can be bounded from above. We shall construct a finite set W of tuples (v, t) such that the union of the corresponding classes $\tilde{\mathcal{C}}_{\epsilon_s}^{(v,t)}$ contains $\tilde{\mathcal{C}}_{\tau_\square}$. Let tv take values in an $\frac{\tau_\square}{2}$-grid contained in B_∞^d, i. e. $tv \in \frac{\epsilon_s}{2\sqrt{d}}\mathbb{Z}^d \cap B_\infty^d$. It is then the case (see Figure 3) that any indicator in $\tilde{\mathcal{C}}_{\tau_\square}$ agrees over B_2^d

with a member in some class $\tilde{C}^{(v,t)}_{\epsilon_s}$, if $\epsilon_s \geq \frac{2}{\tau_\Box}$, i. e.

$$\tilde{\mathcal{C}}_{\tau_\Box} \subseteq \bigcup_{tv \in \frac{\epsilon_s}{2\sqrt{d}}\mathbb{Z}^d \cap B^d_\infty} \tilde{C}^{(v,t)}_{\epsilon_s}.$$

A bound on the volume of the band where $(t - \frac{\epsilon_s}{2\sqrt{d}})\|v\| < x \cdot v < (t + \frac{\epsilon_s}{2\sqrt{d}})\|v\|$ in B^d_2 follows from the fact that the maximum volume hyperplane section is a bisecting hyperplane, whose volume is $< 2\sqrt{d}\,\mathrm{vol}(B^d_2)$.

This allows us to bound the annealed entropy of a single class $\tilde{C}^{(v,t)}_{\epsilon_s}$ in the following lemma, where ρ_{max} is the same maximum density with respect to the uniform density on B^d_2. (Re-scaling was unnecessary because that was with respect to the Lebesgue measure normalized to be a probability measure).

Lemma 11.3.8. *The logarithm of the expected growth function of a class $\tilde{C}^{(v,t)}_{\epsilon_s}$ with respect to $\nu_\circ$ random samples from $\mathcal{P}_\circ$, is $< 2\epsilon_s\rho_{max}\lambda_\circ$, where $\nu_\circ$ is a Poisson random variable of mean $\lambda_\circ$; i. e.*

$$\ln \mathbb{E}_{\nu_\circ} G(\mathcal{C}_{\tau_\Box}, \mathcal{P}_\circ, \nu_\circ) < 2\epsilon_s\rho_{max}\lambda_\circ.$$

Proof. A bound on the volume of the band where $(t - \frac{\epsilon_s}{2\sqrt{d}})\|v\| < x \cdot v < (t + \frac{\epsilon_s}{2\sqrt{d}})\|v\|$ in B^d_2 follows from the fact that the maximum volume hyperplane section is a bisecting hyperplane, whose $d-1$-dimensional volume is $< 2\sqrt{d}\,\mathrm{vol}(B^d_2)$. Therefore, the number of samples that fall in this band is a Poisson random variable whose mean is less than $2\epsilon_s\rho_{max}\lambda_\circ$. This implies the Lemma. □

Therefore the expected annealed entropy of

$$\bigcup_{tv \in \frac{\epsilon_s}{2\sqrt{d}}\mathbb{Z}^d \cap B^d_\infty} \tilde{C}^{(v,t)}_{\epsilon_s}$$

with respect to $\nu_\circ$ random samples from $\mathcal{P}_\circ$ is bounded above by $2\epsilon_s\rho_{max}\lambda_\circ + \ln |\frac{\epsilon}{2\sqrt{d}}\mathbb{Z}^d \cap$

$B_\infty^d|$. Putting these observations together,

$$\begin{aligned} \ln \mathbb{E}_\nu G\left(\mathcal{C}_\tau, \mathcal{P}, \nu\right)/\lambda &\leq \frac{\ln \mathbb{E}_{\nu_\circ} G(\mathcal{C}_{\tau\Box}, \mathcal{P}_\circ, \nu_\circ)}{\lambda_\circ} + \epsilon \\ &\leq 2\epsilon_s \rho_{max} + \frac{d \ln(2\sqrt{d}/\epsilon_s)}{\lambda_\circ} + \epsilon \end{aligned}$$

We know that $\lambda_\circ N_p(\epsilon_r/2) \geq \epsilon\lambda$. Then,

$$2\epsilon_s \rho_{max} + \frac{d \ln(2\sqrt{d}/\epsilon_s)}{\lambda_\circ} + \epsilon \leq$$

$$2\epsilon + N_p(\epsilon_r/2)\frac{d \ln(2\sqrt{d}\rho_{max}/\epsilon_s)}{\epsilon\lambda} + \epsilon,$$

which is

$$\leq 2\epsilon + N_p(\epsilon_r/2)\frac{d \ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon\lambda} + \epsilon.$$

Therefore, if $\lambda \geq N_p(\epsilon_r/2)\frac{d\ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon^2}$, then,

$$\ln \mathbb{E}_\nu G\left(\mathcal{C}_\tau, \mathcal{P}, \nu\right)/\lambda \leq 4\epsilon.$$

Together with Lemma 11.3.5, this shows that for any $\epsilon_1 > 0$,

if

$$\lambda \geq N_p(\epsilon_r/2)\frac{d \ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon^2},$$

then

$$\begin{aligned} H_{ann}(\Lambda, \mathcal{P}, \lfloor\lambda(1-\epsilon_1)\rfloor) &\leq \ln \mathbb{E}_\nu G(\Lambda, \mathcal{P}, \nu) \\ &+ \exp\left(-\epsilon_1^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right) \\ &\leq 4\epsilon\lambda + \exp\left(-\epsilon_1^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right). \end{aligned}$$

Setting ϵ_1 to $\sqrt{\frac{\ln(2\pi\lambda)}{\lambda}}$, $\exp\left(-\epsilon_1^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right)$ is less than 1. Therefore,

$$H_{ann}(\Lambda, \mathcal{P}, \lfloor \lambda - \sqrt{\lambda \ln(2\pi\lambda)} \rfloor) \quad \leq \quad 4\epsilon\lambda + 1.$$

This completes the proof of Theorem 11.3.2.

CHAPTER 12

CONCLUDING REMARKS

In this thesis, the dual interpretations of diffusion as an evolution of densities satisfying a version of the heat equation on the one hand, and as the aggregate effect of a multitude of random walks on the other, were systematically used in a number of settings.

In Chapter 2, a model of language evolution was considered in which a network of interacting agents each of which, at each time step, produces a word based on its belief and updates its belief on the basis of the words produced by its neighbors. In our analysis, we interpreted the belief of an agent as a weighted linear combination of all the words it had heard over time. Then, we studied the evolution of beliefs by tracing backward in time, the trajectory of the words that influenced the beliefs.

In Chapter 4, we used the electrical flows in a network with unit resistors to construct a multicommodity flow. When a unit current is injected in into a node v in an electrical network with unit resistors and extracted out of a node u, the currents can be as follows [23]. Suppose a negatively charged electron does a random walk from u until it hits v. For any edge $e = (a, b)$, let i_{ab} be the expected number of times the electron traverses e from b to a minus the expected number of times it traverses e from a to b. Then i_{ab} is the current in the edge e flowing from a to b. This fact was used to relate the competitive ratio (Definition 4.3.2) of this routing scheme to the mixing time of the graph.

In Chapter 5, we interpreted the amount of heat leaving a uniformly heated convex body K in time t in terms of the probability that a Brownian particle, whose initial position is chosen uniformly at random inside K is outside K after time t. This probability was then estimated by making repeated trials.

In Chapter 6, we traced Brownian particles whose initial positions were chosen independently and uniformly at random, approximated their trajectories by straight line segments, and in the event that they exited the body, output the intersection of the line segment with the boundary as an approximately random sample from the surface.

Chapter 7 investigated halfplane capacity of a hull in the upper half plane,and related it to the area of a neighborhood of the hull. The halfplane capacity can be interpreted in terms of harmonic functions or Brownian motion. The proof used properties of Brownian motion in the upper half plane.

Markov Chain Monte Carlo methods are algorithms for sampling from probability distributions by designing a Markov chain whose stationary distribution is the desired distribution. In Chapter 8 and Chapter 9, we constructed such Markov Chains for sampling from Riemannian manifolds defined using charts and polytopes respectively. In Chapter 9, we also used a random walk on polytopes to design an interior point algorithm for linear programming. This random walk had a drift that guided it towards the optimal point. While a deterministic analogue of this algorithm due to Dikin [22] does not have polynomial time guarantees, our randomized version did.

In Chapter 3, we designed a distributed algorithm for averaging the values held by nodes in a graph motivated by a diffusion taking place on two scales. Chapter 10 and Chapter 11 are discussed in Section 12.4 below.

Finally, we discuss directions for future work and mention some open questions based on the results of this thesis.

12.1 Sampling manifolds

While there is a large body of work devoted to sampling convex sets from log-concave densities, much less is known about sampling manifolds. It would be interesting to attempt to extend the existing framework for sampling convex sets to more general settings. We made some progress towards this goal in Chapter 8 where we presented a class of Markov chains that could be interpreted as approximations to Brownian motion with drift. Algorithms for sampling manifolds have a number of interesting applications, including Goodness-of-Fit tests in statistics. Another interesting application of sampling manifolds could be to sample compact Lie groups from the Haar measure. Random matrices have a number of applications in statistics, and more recently in wireless technology, and their distributions in several important cases (such as random matrices from the orthogonal group) are, in fact, from the Haar measure of some compact Lie group. Thus, this question is of more than purely theoretical interest.

12.2 Randomized interior point methods for convex optimization

Self concordant barriers are real valued functions defined on a convex set, that "blow up" as one approaches the boundary of the set. These functions have been used to develop interior point algorithms for convex optimization. Their role is to ensure that the path taken by the algorithm does not approach too close to the boundary. The algorithm in Chapter 9 may be interpreted as a discretization of Brownian motion with drift, on a manifold whose metric tensor is derived from the Hessian of a barrier function. It would be interesting to study global aspects of its geometry such as its isoperimetric constant. A better understanding of global aspects of the geometry of these manifolds could lead to better average case analysis of interior point methods (which are frequently used in practice).

12.3 Sampling polytopes and computing volumes

One of the most basic quantities that can be associated with a polytope is its volume. It can be shown ([27], [2]) that any deterministic algorithm to approximate the volume of a convex body given by a membership oracle, within a constant factor in $\mathbb{R}^n$, needs time that is exponential in the dimension n. Remarkably, randomized algorithms turn out to be more powerful. In their path breaking paper [25] Dyer, Frieze and Kannan gave the first randomized polynomial time algorithm to approximate the volume of a convex body to arbitrary accuracy. Since then a considerable body of work has been devoted to improving the complexity of volume computation culminating with the recent best of $O^*(n^4)$ due to Lovász and Vempala [61]. In the case of $n-$dimensional polytopes defined by m constraints, in Chapter 9, we developed a Markov Chain whose mixing time was $O(mn)$ from a starting density that was upper bounded by a constant with respect to the uniform density, i. e. from a warm start. When $m = O(n^{1.62})$, the number of arithmetic operations taken to produce an almost uniform sample by the algorithm that uses this chain is less than that for other algorithms. There seem to be some avenues for improvement. The mixing time we obtained was $O(mn^2)$ from a fixed point and $O(mn)$ from a warm start. It may be possible to bridge this gap if instead of using $\mathcal{L}_2$ bounds, we could use Log-Sobolev inequalities. Secondly, in some important applications, such as sampling contingency tables, $m = n + o(n)$. It may be possible to improve the mixing time from a warm start to $O((m-n)n)$ by proving a better isoperimetric inequality.

It would be interesting to obtain an algorithm to compute the volume of a polytope that performs better under similar conditions. A natural way to proceed would be to devise an efficient annealing strategy starting from a simple polytope whose volume can easily be estimated, ending with the polytope we are interested in, along the lines of [25, 61].

12.4 Learning on manifolds

Manifold learning has emerged as a new important paradigm for modeling high dimensional data. The underlying intuition is that while modern data often lie in very high dimensions, the number of degrees of true freedom is usually much less. An example of this is the case of human speech, where the waveforms lie in an infinite dimensional space, but the undulations of our vocal chords have, in essence, far fewer degrees of freedom. Many basic questions on learning over manifolds remain open. It would be interesting to develop a clustering algorithm that finds a cut that minimizes the total amount of heat diffusing out of it interpreted appropriately. Diffusion maps and random projections onto low dimensional vector spaces seem to be natural tools in this context. Concepts from functional analysis, in particular reproducing kernel Hilbert spaces are likely to be useful in tackling this question. Chapter 10 was an attempt to address the sample complexity of learning the Cheeger cut of a probability density on $\mathbb{R}^n$. Towards this end, we proved that for any fixed hypersurface satisfying appropriate smoothness conditions, its weighted surface area (Equation 10.1) is the limit of the weights of the induced cuts, on the data dependent graphs of Theorem 10.3.1, as the number of samples tends to ∞. In order to take the results of Chapter 10 to their natural conclusion, we would need to prove a uniform bound over all sufficiently smooth hypersurfaces, relating the weighted surface area of the hypersurfaces to the normalized weights of the corresponding cuts, i. e. a uniform analogue of Theorem 10.3.1. Then, we would be able to claim that the Cheeger cut can be obtained by structural risk minimization [100] over the class of all smooth cuts, where the class of cuts is made progressively richer by gradually increasing the permissible condition number (see Definition 6.3.1).

In Chapter 11, we proved a bound on the number of samples needed to learn a smooth partition on a manifold separating data into two classes. Our results were based on a quantity known as annealed entropy, that measures the complexity of a class of indicator functions with respect to a probability measure. In our setting, the more commonly used notion known as VC dimension could not be bounded, since it is potentially infinite. The bounds that we obtain are not optimal, and to prove better bounds is a subject for future research.

REFERENCES

[1] J. Abernethy, E. Hazan and A. Rakhlin, " Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization," *21st Annual Conference on Learning Theory, (COLT)*, 2008, pp. 263-274 118

[2] I. Bárány and Z. Furedi, "Computing the Volume is Difficult," *Discrete and Computational Geometry 2*, 1987, pp. 319-326. 60, 200

[3] Y. Bartal and S. Leonardi, "On-line routing in all optimal networks," *Theoretical Computer Science*, 1997, 516-526. 41

[4] W. Baur and V. Strassen, "The Complexity of Partial Derivatives," *Theoretical Computer Science*, **22** (1983) pp. 317-330 120

[5] M. Belkin, H. Narayanan and P. Niyogi, "Heat Flow and a Faster Algorithm to Compute the Surface Area of a Convex Body," *Proc. of the 44th IEEE Foundations of Computer Science (FOCS) '06*, pp. 47-56 6, 82, 83

[6] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, June 2003, pp. 1373-1396. 177

[7] A. Belloni and R. Freund, "Projective re-normalization for improving the behavior of a homogeneous conic linear system," *Mathematical Programming*, **118(2)**, pp. 279-299 115

[8] D. Bertsekas and J. Tsitsiklis, "Parallel and Distributed Computation: Numerical Methods," *Prentice-Hall, 1989; republished in 1997 by Athena Scientific.* 4, 17, 19, 30, 32

[9] D. Bertsimas and S. Vempala, "Solving convex programs by random walks," *Journal of the ACM (JACM)*, 2004, **51(4)**. 61, 78

[10] B. de Boer, "Evolution of Speech and its Acquisition," *Adaptive Behavior*, **13(4)**, pp. 281-292 15

[11] C. Borell, "The Brunn-Minkowski inequality in Gauss space," *Inventiones Math.*, 1975 **30**, pp. 205-216 88

[12] V. S. Borkar, "Stochastic approximation with two time-scales," *Systems and Control letters*, **29(5)**, February 1997, pp. 291 - 294. 30

[13] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, " Gossip algorithms : Design, analysis and applications," *Proceedings of the 24th Conference of the IEEE Communications Society (INFOCOM 2005)*, 2005, pp. 1653-1664. 4, 30, 32

[14] J. Bybee, "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change," *Language Variation and Change* **14**, pp. 261-290. 19

[15] O. Chapelle and A. Zein,"Semi-supervised Classification by Low Density Separation," *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS 2005)*, pp. 57-64 154

[16] R.R. Coifman, S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets*, **21**, July 2006, pp 5-30. 177

[17] F. Cucker, S. Smale, "Emergent behavior in flocks," *IEEE transactions on Automatic Control*, May 2007, **52(5)**, pp. 852-862. 19

[18] F. Cucker, S. Smale, D. X. Zhou, "Modelling Language evolution," *Foundations of Computational Mathematics*, July 2004, **4(3)**, pp. 315-343. 20

[19] P. Diaconis, "Generating random points on a Manifold," *Berkeley Probability Seminar,* 2008 (Talk based on joint work with S. Holmes and M. Shahshahani) 10, 103

[20] P. Diaconis and B. Efron, "Testing for independence in a two-way table: new interpretations of the chi-square statistic," *Annals of Statistics*, 1995, **13**, pp. 845-913. 11, 115

[21] P. Diaconis and D. Stroock, "Geometric bounds for eigenvalues of Markov Chain," *Annals of Applied Probability*, 1991, **1**, pp. 36-61. 57

[22] I. I. Dikin, "Iterative solutions of problems of linear and quadratic programming," *Soviet Math. Dokl*, 1967, **8**, pp. 674-675. 112, 115, 118, 199

[23] P. G. Doyle and J. L. Snell, "Random Walks and Electric Networks," *Mathematical Association of America*, 1984. 55, 198

[24] B. Duplantier, "Brownian Motion", "Diverse and Undulating", in *Einstein, 1905-2005. Poincaré Seminar 2005,* Birkhaeuser Verlag, Basel, 2006, pp. 201-293, `http://arxiv.org/abs/0705.1951`. 1

[25] M. Dyer, A. Frieze and R. Kannan, "A random polynomial time algorithm for approximating the volume of convex sets," 1991, *Journal of the Association for Computing Machinary,* **38**, pp. 1-17, 60, 77, 112, 113, 200, 201

[26] M. Dyer, P. Gritzmann and A. Hufnagel, "On the complexity of computing Mixed Volumes," *SIAM J. COMPU.*, April 1998, **27(2)**, pp. 356-400. 60, 61

[27] G. Elekes, "A Geometric Inequality and the Complexity of Computing Volume," *Discrete and Computational Geometry 1*, 1986, pp. 289-292. 60, 200

[28] A. Gray, "Tubes," *Birkhauser Verlag*, 2004. 187

[29] M. Grötschel, L. Lovász, and A. Schrijver, "Geometric algorithms and combinatorial optimization," *Springer-Verlag*, Berlin, 1988. 60

[30] M. Hajiaghayi, R. Kleinberg, T. Leighton, and H. Räcke "New lower bounds for oblivious routing in undirected graphs," *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006, pp. 918-927. 41, 42

[31] C. Harrelson, K. Hildrum, and S. B. Rao, "A polynomial-time tree decomposition to minimize congestion," *in Proceedings of the 15th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*, 2003, pp. 34-43. 40

[32] P. Harsha, T. Hayes, H. Narayanan, H. Räcke, and J. Radhakrishnan, "Minimizing average latency in Oblivious routing", *(SODA)*, 2008, pp. 200-207. 41

[33] M.D. Hauser (1996), "The evolution of communication", **MIT Press**, Cambridge, MA. 15

[34] R. R. Hudson, "Gene Genealogies and the Coalescent Process," *Oxford Surveys in Evolutionary Biology*, 1991, **7**, pp. 1-44. 19, 22

[35] Itô, K., and McKean, Jr., H. P., "Diffusion Processes and Their Sample Paths." *Grundlehren Math.Wiss. 125, Springer-Verlag, Berlin 1965*; *Classics in Math., Springer-Verlag, Berlin 1996.* 2

[36] M. Jackson, "The Economics of Social Networks", *Chapter 1 in Volume I of Advances in Economics and Econometrics, Theory and Applications*, 2006. 17

[37] M. R. Jerrum, L. G. Valiant and V. V. Vazirani "Random generation of Combinatorial structures from a uniform distribution," *Theoretical Computer Science*, 1986, **43**, pp. 169-188. 68

[38] A. Kalai and S. Vempala. "Simulated Annealing for Convex Optimization," *Math of OR*, February 2006, **31(2)**, pp. 253-266. 61

[39] R. Kannan, L. Lovász and M. Simonovits, "Random walks and an $O^*(n^5)$ volume algorithm for convex bodies," *Random Structures and Algorithms*, August 1997, **11(1)**, pp. 1-50. 112, 113

[40] R. Kannan and H. Narayanan, "Random walks on polytopes and an affine interior point method for linear programming," *Proceedings of the ACM Symposium on Theory of Computing,* 2009, pp. 561-570. 11

[41] R. Kannan, S. Vempala, "Sampling Lattice points," *Proceedings of the ACM Symposium on Theory of Computing,* 1997, pp. 696-700. 11, 114

[42] N. K. Karmarkar, "A new polynomial-time algorithm for linear programming," *Combinatorica*, 1984, **4**, pp. 373-395. 115, 117

[43] R. M. Karp and M. Luby, (1983). "Monte-Carlo algorithms for enumeration and reliablility problems," *Proc. of the 24th IEEE Foundations of Computer Science (FOCS '83)*, 1983, pp. 56-64. 68

[44] Kingman, J.F.C. “On the Genealogy of Large Populations,” *Journal of Applied Probability* 1982, **19A**, pp. 27-43. 19, 22

[45] S. Kirby, “Function Selection and Innateness: the Emergence of Language Universals,” *Oxford University Press. Innateness and culture in the evolution of language.* 15, 20

[46] S. Kobayashi and K. Nomizu, “Foundations of differential geometry,” *Interscience Tracts in Pure and Applied Math.*, 1963, **15**. 187

[47] M. Kojima and Y. Ye, “Recovering optimal dual solutions in Karmarkar’s polynomial algorithm for linear programming,” *Mathematical Programming,* 1987, **39**, pp. 305-317. 145

[48] V. Konda and J. Tsitsiklis, “Convergence rate of a linear two time scale stochastic approximation,” *Annals of Applied Probability* 2004, pp. 1671-1702. 30

[49] Samuel Kutin, Partha Niyogi, “Almost-everywhere Algorithmic Stability and Generalization Error,” *Uncertainty in Artificial Intelligence*, 2002, pp. 275-282. 159, 160, 163

[50] S. Kutin, “Extensions to McDiarmid’s inequality when differences are bounded with high probability,” *Technical report TR-2002-04 at the Department of Computer Science, University of Chicago.* 159, 160, 163

[51] G. Lawler, “Conformally Invariant Processes in the Plane,” *American mathematical Society*, 2005 5, 9, 95, 97, 99, 101

[52] G. Lawler and V. Limic, “Symmetric Random Walk,”
`http://www.math.uchicago.edu/∼lawler/srwbook.pdf` 38

[53] G. Lawler and H. Narayanan, “Mixing times and ℓ_p bounds for Oblivious Routing”, *Workshop on Analytic Algorithmics and Combinatorics*, (ANALCO ’09) 4

[54] G. Lawler, W. Werner and O. Schramm, “The dimension of the planar brownian frontier is 4/3,” *Mathematical Research Letters*, 2001, **8**, pp. 401-411. 8, 95

[55] M. Ledoux, “The concentration of measure phenomenon,” *Mathematical Surveys and Monographs*,2001, **89**, American Mathematical Society. 128

[56] Mark Liberman, “The Lexical Contract: Modeling the emergence of word pronunciations,” http://www.ldc.upenn.edu/myl/abm/index.html 3, 15

[57] L. Lovász, “Hit-and-run mixes fast,” *Math. Programming, series A,* 1999, **86**, pp. 443-461. 112, 120, 141

[58] L. Lovász and M. Simonovits, “Random walks in a convex body and an improved volume algorithm,” *Random structures and algorithms,* 1993, **4**, pp. 359-412. 103, 143, 147

[59] L. Lovász and S. Vempala, "Fast Algorithms for Logconcave Functions: Sampling, Rounding, Integration and Optimization," *Proc. of the 47th IEEE Symposium on Foundations of Computer Science* (FOCS '06),2006, pp. 57-68. 115

[60] L. Lovász and S. Vempala, "Hit-and-run from a corner," *SIAM J. Comput.*, 2006, **4**, pp. 985-1005. 63, 84, 103, 112

[61] L. Lovász and S. Vempala, "Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm" *Proc. of the 44th IEEE Foundations of Computer Science* (FOCS '03), 2003, pp. 392-417. 60, 61, 63, 64, 66, 73, 77, 201

[62] L. Lovász and S. Vempala, "The geometry of logconcave functions and sampling algorithms," *Random Structures and Algorithms,*, May 2007, **3**,pp. 307 - 358. 113, 150

[63] M. Maier, U. von Luxburg, M. Hein, "Influence of graph construction on graph-based clustering measures," *Advances in Neural Information Processing Systems*, 2009. 12

[64] P. Matthews, "Mixing Rates for Brownian Motion in a Convex Polyhedron," *Journal of Applied Probability,* June 1990, **27(2)**, pp. 259-268. 77

[65] P. Matthews, "Covering Problems for Brownian Motion on Spheres," *Annals of Probability,* Januart 1988, **16(1)**, pp. 189-199. 77

[66] B.S. Mityagin, "An interpolation theorem for modular spaces," *Mat. Sb. (N.S.)* , 1965, **66**, pp. 473482. 46

[67] B. Morris, "Improved bounds for sampling contingency tables," *Random Structures and Algorithms*, September 2002, **21(2)**. 11, 57, 115

[68] S. Muthukrishnan, B. Ghosh and M. H. Schultz, "First and Second-Order Diffusive Methods for Rapid, Coarse, Distributed Load Balancing," *Theory of Computing Systems,* December 1998, **31(4)**, pp. 331–354. 30

[69] H. Narayanan, "Geographic Gossip on Geometric Random Graphs via Affine Combinations," *Principles of Distributed Computing (PODC)*, 2007, pp. 388 - 389. 30

[70] H. Narayanan, " Distributed Averaging in the presence of a Sparse Cut", *Principles of Distributed Computing (PODC),* 2008. 3

[71] H. Narayanan, M. Belkin and P. Niyogi, "On the relation between low density separation, spectral clustering and graph cuts", *20th Annual Conference on Neural Information Processing Systems (NIPS)*, December 2006 12

[72] H. Narayanan and P. Niyogi, "Sampling Hypersurfaces through Diffusion", *Neural Information Processing Systems (NIPS)*, December 2006. 7

[73] H. Narayanan and P. Niyogi, "On the sample complexity of learning smooth cuts on a manifold", *22nd Annual Conference on Learning Theory (COLT)*, June 2009. 13

[74] Y. E. Nesterov and A. S. Nemirovskii, "Interior point polynomial algorithms in convex programming," *SIAM Publications. SIAM, Philadelphia*, USA, 1994. 117

[75] Y. E. Nesterov and M. J. Todd, "Self-Scaled Barriers and Interior-Point Methods for Convex Programming," *Mathematics of Operations Research,* February 1997, **22(1)**, pp. 1-42. 147

[76] P. Niyogi, "The Computational Nature of Language Learning and Evolution," *MIT Press,* April 2006. 15

[77] P.Niyogi, S. Weinberger, S. Smale, "Finding the Homology of Submanifolds with High Confidence from Random Samples," *Discrete and Computational Geometry*, 2004. 85

[78] M. A. Nowak, and N.L. Komarova (2001), "Towards an evolutionary theory of language," *Trends in Cognitive Sciences* **5(7)**, pp. 288-295. 20

[79] M. A. Nowak, and K. Sigmund, "Evolutionary Dynamics of Biological Games, *Science 6*, February 2004, **303(5659)**, pp. 793–799. 20

[80] V. Y. Pan, Z. Chen and A. Zheng, "The Complexity of the Algebraic Eigenproblem," *MSRI Preprint 1998-71,* Mathematical Sciences Research Institute, Berkeley, California, 1998, pp.507 - 516. 66, 81

[81] H. Räcke, "Minimizing congestion in general networks," *in Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science (FOCS)*, 2002, pp. 43–52. 40

[82] H. Räcke, "Optimal Hierarchical Decompositions for Congestion Minimization in Networks," *In Proc. of the 40th STOC,* 2008, pp. 255–264. 40, 41

[83] J. Renegar, "A polynomial-time algorithm, based on Newton's method, for linear programming," *Mathematical Programming,* 1988, **40**, pp. 59-93. 113, 117

[84] M. Riesz, "Sur les maxima des formes bilinéaires et sur les fonctionelles linéaires," *Acta Math.*, 1926, **49**, pp. 465 - 497. 46

[85] M. Rudelson, "Random vectors in the isotropic position," *J. of Functional Analysis,* 164 1999,**1**, pp. 60-72. 81, 83

[86] R. Schneider, "Convex bodies: The Brunn-Minkowski Theory," *Encyclopedia of Mathematics and its Applications,* Cambridge University Press 1993. 69, 71, 72

[87] A. Sinclair " Improved Bounds for Mixing Rates of Markov Chains and Multicommodity Flow," *Combinatorics, Probability and Computing 1,* 1992, pp. 351-370. 57

[88] S. Smirnov, "Conformal invariance in random cluster models. I. Holomorphic fermions in the Ising model," *to appear in Ann. Math.* 8, 95

[89] J. M. Smith, “Evolution and the Theory of Games,” *Cambridge University Press*, 1982. 20

[90] L. Steels. “The puzzle of language evolution,” *Kognitionswissenschaft,* **8(4)**, 1999. 15

[91] D. W. Stroock, “An Introduction to tha Analysis of Paths on a Riemannian Manifold,” *American Mathematical Soceity*, 1991. 8

[92] O. Schramm, “Scaling limits of loop-erased random walks and uniform spanning trees,” *Israel J. Math.,* 2000, **118**, pp. 221-288. 8, 95

[93] P. Tetali, “Random walks and the effective resistance of networks,” *Journal of Theoretical Probability,* January 1991, **4(1)**. 48

[94] G. O. Thorin, “Convexity theorems generalizing those of M. Riesz and Hadamard with some applications,” *Comm. Sem. Math. Univ. Lund = Medd. Lunds Univ. Sem. 9*, 1948, pp. 1 - 58. 46

[95] J. N. Tsitsiklis, D. P. Bertsekas and M. Athans, “Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms,” *IEEE Transactions on Automatic Control,* 1986, **31(9)**, pp. 803-812. 18

[96] P. Vaidya, “A new algorithm for minimizing convex functions over convex sets,” *Mathematical Programming,* 1996, **73**, pp. 291-341. 61, 113, 117

[97] P. Vaidya, “An algorithm for linear programming which requires $O(((m+n)n^2+(m+n)^{1.5}n)L)$ arithmetic operations,” *Mathematical Programming,* 1990, **47**, pp. 175-201. 140

[98] L. G. Valiant and G. J. Brebner, “Universal schemes for parallel communication,” *in Proceedings of the 13th ACM Symposium on Theory of Computing (STOC),* 1981, pp. 263–277. 40

[99] R. J. Vanderbei and J. C. Lagarias, “I. I. Dikin’s convergence result for the affine-scaling algorithm,” *Contemporary mathematics,* 1990, **114**, pp. 285-300. 115

[100] V. Vapnik, “Statistical Learning Theory,” *Wiley*, 1998. 180, 202

[101] S. Vempala, “Personal Communication”. 84

[102] S. Vempala (2005), “Geometric Random Walks: A Survey,” *Combinatorial and Computational Geometry, MSRI Publications,* 2005, **52**. 64, 66