Sampling Hypersurfaces through Diffusion

Hariharan Narayanan and Partha Niyogi

Department of Computer Science, University of Chicago, USA,
{hari,niyogi}@cs.uchicago.edu

Abstract. We are interested in efficient algorithms for generating random samples from geometric objects such as Riemannian manifolds. As a step in this direction, we consider the problem of generating random samples from smooth hypersurfaces that may be represented as the boundary ∂A of a domain $A \subset \mathbb{R}^d$ of Euclidean space. A is specified through a membership oracle and we assume access to a blackbox that can generate uniform random samples from A. By simulating a diffusion process with a suitably chosen time constant t, we are able to construct algorithms that can generate points (approximately) on ∂A according to a (approximately) uniform distribution.
We have two classes of related but distinct results. First, we consider A to be a convex body whose boundary is the union of finitely many smooth pieces, and provide an algorithm (Csample) that generates (almost) uniformly random points from the surface of this body, and prove that its complexity is $O^*(\frac{d^4}{\epsilon})$ per sample, where ϵ is the variation distance. Next, we consider A to be a potentially non-convex body whose boundary is a smooth (co-dimension one) manifold with a bound on its absolute curvature and diameter. We provide an algorithm (Msample) that generates almost uniformly random points from ∂A, and prove that its complexity is $O(\frac{R}{\sqrt{\epsilon}\tau})$ where $\frac{1}{\tau}$ is a bound on the curvature of ∂A, and R is the radius of a circumscribed ball.

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 [8], [10]) 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 [11],[12]), as far as we can see, these do not directly lead to algorithms with complexity guarantees.

Our interest in sampling a manifold is motivated by several considerations from diverse areas in which such a result would be applicable. In machine learning, the problem of clustering may be posed as finding (on the basis of empirically drawn data points) a partition of the domain (typically $\mathbb{R}^d$) into a finite number of pieces. In the simplest form of this (partition into two pieces) the partition boundary (if smooth) may be regarded as a submanifold of co-dimension one and the best partition is the one with smallest volume (in a certain sense corresponding to a natural generalization of Cheeger's cut of a manifold). More generally, the area of *manifold learning* has drawn considerable attention in recent years within the machine learning community (see [5, 18] among others) and many of the questions may be posed as learning geometric and topological properties of a submanifold from randomly drawn samples on it. In scientific computing, one may be interested in numerical methods for integrating functions on a manifold by the Monte Carlo method. Alternatively, in many physical applications, one may be interested in solving partial differential equations where the domain of interest may have the natural structure of a manifold. In contrast to a finite element scheme on a deterministic triangulation (difficult to obtain in high dimensions), one may explore randomized algorithms by constructing a random mesh and solving such PDEs on such a mesh. Finally, in many applications to dynamical systems, one is interested in the topology of the space of attractors which have the natural structure of a manifold (see [13]). In statistics, one is interested in goodness of fit tests for a variety of multivariate random variables. For example, testing for a gamma distribution leads one to consider (positive real valued) random variables $X_1, \ldots, X_n$ such that $\sum_i X_i = a$ and $\prod_j X_j = b$. The set of all $(X_1, \ldots, X_n)$ under these constraints is the boundary of a convex body in the hyperplane defined by $\sum_i X_i = a$. Sampling this is a question that arises naturally in this setting (see [6], [7]).

Thus, we see that building an efficient sampler for a manifold is a problem of fundamental algorithmic significance. Yet, not much is known about this and as a step in this general direction, in the current paper, we address the problem of sampling manifolds that are boundaries of open sets in $\mathbb{R}^d$ from the measure induced by the Lebesgue measure. The particular setting we consider in this paper has direct applications to clustering and goodness of fit tests where co-dimension 1 manifolds naturally arise. In addition, we also provide an algorithm and obtain complexity bounds for sampling the surface of a convex body – a problem to which we have not seen a solution at the present moment.

1.1 Summary of Main Results

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 [4]).

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 $f : \mathbb{R}^d \to \mathbb{R}$. 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.

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 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 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 $n = O(m \, polylog(m))$. In the complexity analysis, we shall only consider the number of oracle calls made, as is customary in this literature.*

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 centred 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 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.)*

2.1 Algorithm Csample

Algorithm 1 *Csample*

1. *Estimate (see [15]) with confidence $> 1 - \epsilon$, the smallest eigenvalue κ 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 [16].)*

2. *Set*

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

3. *(a) Set* $p = \mathsf{Ctry}\ (t)$.
   (b) If $p = \emptyset$*, goto (3a). Else output* p*.*

Algorithm 2 *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* ℓ *the segment whose endpoints are* x *and* y*.*
4. *If* $y \notin K$ *output* $\ell \cap \partial K$*, else output* $\emptyset$*.*

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 $\mathsf{Csample}$. Let $|\mu|$ denote the total mass for any measure μ. We shall define a measure $\mu_{\partial K}$ on ∂K related to the "local diffusion" out of small patches. Formally, if Δ a subset of ∂K, the measure assigned to it by $\mu_{\partial K}$ is

$$\mu_{\partial K}(\Delta) := \int_{x\in S}\int_{y\in\mathbb{R}^d\setminus S} G^t(x,y)\mathcal{I}\left[\overline{xy}\cap\Delta\neq\emptyset\right] d\lambda(x)d\lambda(y) \tag{1}$$

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 3 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 [3]. This proves that

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

. □

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 ([16]). 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 1. *Lemma 5.5 in [3]]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 2 (Lemma 5.4 in [3]). *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 2, we see that

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

The probability that Ctry succeeds in one trial is

$$\begin{aligned}
\mathbb{P}[\textsf{Ctry } (t) \neq \emptyset] &= \sqrt{\frac{t}{\pi}}\frac{F_t}{V} && (2)\\
&> \sqrt{\frac{t}{\pi}}\frac{S}{V}(1 - O(\epsilon)) && (3)\\
&> \sqrt{\frac{t}{\pi}}\frac{1 - O(\epsilon)}{4r_{in}} \text{ (By Lemma 1)} && (4)\\
&> \Omega(\frac{\epsilon}{d}). && (5)
\end{aligned}$$

Therefore the expected number of calls to $\mathcal{B}$ and the membership oracle is $O^*(\frac{d}{\epsilon})$. By results of Lovász and Vempala ([9]) this number of random samples can be obtained using $O^*(\frac{d^4}{\epsilon})$ calls to the membership oracle. □

2.4 Extensions

S. Vempala [17] 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 $\partial K \cap H$.

3 Sampling Well Conditioned Hypersurfaces

3.1 Preliminaries and Notation

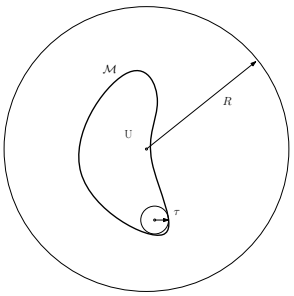

Fig. 1.

Let $\mathcal{M}$ be a (codimension one) hypersurface.

Definition 3 *Let $\mathcal{M}$ be a codimension 1 hypersurface. The condition number of $\mathcal{M}$ is defined as $\frac{1}{\tau}$ where τ is is the largest number with the following property: No two normals to $\mathcal{M}$ of length less than τ intersect.*

In fact $\frac{1}{\tau}$ is an upper bound on the curvature of $\mathcal{M}$ ([14]). In this paper, 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 $\mathcal{M}$.

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 distribution is close to the uniform in ℓ_∞, 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 2. *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 .*
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 $O(\frac{R(1+\frac{2}{d}\ln\frac{1}{\epsilon})}{\tau\sqrt{\epsilon}})$.*

Algorithm 3 *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.*

Algorithm 4 *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.)

3.3 Correctness

Proof of part (1) of Theorem 2: 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 to it by $\mu_{\mathcal{M}}$ is

$$\mu_{\mathcal{M}}(\Delta) := \int_{x\in U}\int_{y\in\mathbb{R}^d\setminus U} G^t(x,y)\mathcal{I}\left[\overline{xy}\cap\Delta\neq\emptyset\right]d\lambda(x)d\lambda(y) \tag{6}$$

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 3. *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 4. *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.

3.4 Complexity

Proof of part (2) of Theorem 2: 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 4 implies that

$$\mathbb{P}[\mathsf{Mtry}\ (t) \neq \emptyset] > \frac{S\sqrt{\frac{t}{\pi}}(1 - O(\epsilon))}{V} \tag{7}$$

$$\geq \frac{d}{R}\frac{\tau\sqrt{\epsilon}(1 - O(\epsilon))}{8(d + 2\ln\frac{1}{\epsilon})} \tag{8}$$

$$= \Omega(\frac{\tau\sqrt{\epsilon}}{R(1 + \frac{2}{d}\ln\frac{1}{\epsilon})}). \tag{9}$$

This completes the proof. □

In our proofs of Lemma 3 and Lemma 4, we shall use the following Theorem of C. Borell.

Theorem 3 (Borell, [2]). *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 3: 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 $\frac{d\mu_{\partial H}}{d\lambda_{\partial H}}$.

$$\frac{d\mu_{\partial H}}{d\lambda_{\partial H}} = \int_{\mathbb{R}^-}\int_{\mathbb{R}^+} G^t(x,y)d\lambda(x)d\lambda(y), \tag{10}$$

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. □

Proof of Lemma 4: 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.

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

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

$$\mathbb{P}_\tau := \lim_{|\Delta|\to 0} \frac{\int_{x\in B_1}\int_{y\in B_2} G^t(x,y)\,\mathcal{I}\left[\overline{xy}\cap\Delta\neq\emptyset\right]\,d\lambda(x)\,d\lambda(y)}{\int_{x\in H_z}\int_{y\in\mathbb{R}^d\setminus H_z} G^t(x,y)\,\mathcal{I}\left[\overline{xy}\cap\Delta\neq\emptyset\right]\,d\lambda(x)\,d\lambda(y)}. \tag{11}$$

Then

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

The proof now follows from

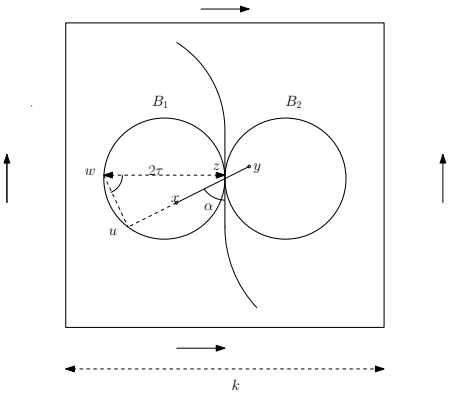

Fig. 2.

Lemma 5. $\mathbb{P}_\tau > 1 - O(\epsilon)$. □

Proof of Lemma 5: 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

$$\phi_k : \mathbb{R}^d \to \mathbb{T}_k \tag{12}$$

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

$$\mathbb{P}_\tau^{(k)} := \mathbb{P}[y_2^2 + \cdots + y_d^2 < |y_1| \tau < \tau^2 \,\big|\, 1 = |\phi_k(\overline{xy}) \cap \Xi_k|] \tag{13}$$

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 2 is slid along itself to the right until x occupies the position where z is now, y does not leave B_2. Figure 3 illustrates ball B_2 being slid, which is equivalent. In particular, this event would imply that $x \in B_1$ and $y \in B_2$.

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

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. This completes the proof of this proposition. □

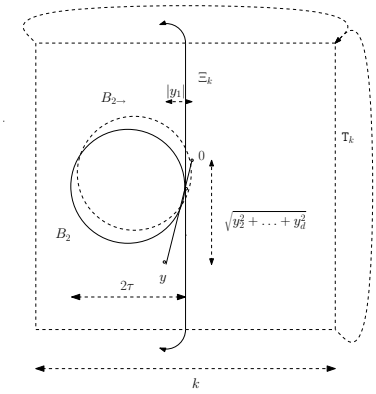

Fig. 3.

Lemma 6. *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,

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

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

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

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

$$\mathcal{I}_{\shortparallel} := \mathcal{I}\left[|\, y_1| \notin [\sqrt{\epsilon t}, \tau]\right] \mathcal{I}\left[E_k\right],$$

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 $\mathbb{E}[\mathcal{I}_{\shortparallel} | E_k] = O(\epsilon)$. Similarly $\mathbb{E}[\mathcal{I}_\perp | E_k] = O(\epsilon)$ follows from Theorem 3 and the fact mentioned below it. This Lemma is implied by the following claim. □

Claim.

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

Proof:

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

Therefore

$$\begin{aligned}
\mathcal{I}[E_k] - \mathcal{I}_{\shortparallel} - \mathcal{I}_\perp &\leq \mathcal{I}[E_k]\,[y_2^2 + \cdots + y_d^2 < \tau\sqrt{\epsilon t} < \tau|y_1|\,]\,\mathcal{I}[|y_1| < \tau] \quad (14) \\
&\leq \mathcal{I}_\tau \quad (15)
\end{aligned}$$

□

4 Acknowledgements

We are grateful to David Jerison and Emanuel Milman for numerous very helpful discussions and to Santosh Vempala for pointing out the extension in Section 2.4 and permitting us to include it here. The first author is grateful to AIM for its generous hospitality during the workshop on Algorithmic Convex Geometry and that on Fourier Analytic methods in Convex Geometric Analysis. We would like to thank the anonymous referee for carefully reading an earlier version and pointing out an error in the proof of Lemma 4.

References

1. K. Ball, An Elementary Introduction to Modern Convex Geometry, *Mathematical Sciences Research Institute Publications 31*, Cambridge Univ. Press, 1997, pp. 1-58
2. C. Borell, "The Brunn-Minkowski inequality in Gauss space." *Inventiones Math.* 30 (1975), 205-216
3. 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*
4. D. Bertsimas and S. Vempala, "Solving convex programs by random walks" *Journal of the ACM* (JACM) 51(4), 540–556, 2004. Proc. of the 34th ACM Symposium on the Theory of Computing (STOC '02), Montreal, 2002.
5. R.R. Coifman, S. Lafon, "Diffusion maps", *Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets,* Vol 21, July 2006, pp 5-30.
6. P. Diaconis, Generating random points on a Manifold, *Berkeley Probability Seminar,* (Talk based on joint work with S. Holmes and M. Shahshahani)
7. P. Diaconis, *Personal Communication.*
8. M. Dyer, A. Frieze and R. Kannan, "A random polynomial time algorithm for approximating the volume of convex sets" (1991) in *Journal of the Association for Computing Machinary,* 38:1-17,
9. L. Lovász and S. Vempala , "Hit-and-run from a corner" *Proc. of the 36th ACM Symposium on the Theory of Computing,* Chicago, 2004
10. 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), Boston, 2003.
11. P. Matthews, Mixing Rates for Brownian Motion in a Convex Polyhedron, *Journal of Applied Probability,* Vol. 27, No. 2 (Jun., 1990), pp. 259-268
12. P. Matthews, Covering Problems for Brownian Motion on Spheres, *Annals of Probability,* Vol. 16, No. 1 (Jan., 1988), pp. 189-199
13. Kaczynski T., Mischaikov K., and Mrozek M., Computational Homology, Springer, New York (2004), (Applied Math. Sci.; 157)
14. P.Niyogi, S. Weinberger, S. Smale (2004), "Finding the Homology of Submanifolds with High Confidence from Random Samples." *Discrete and Computational Geometry.*
15. 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).
16. M. Rudelson, Random vectors in the isotropic position, *J. of Functional Analysis,* 164 (1999) no. 1, 60-72. *Encyclopedia of Mathematics and its Applications,* Cambridge University Press 1993.
17. S. Vempala, *Personal Communication.*
18. A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete and Computational Geometry* (2004), 33 (2), pp. 247