On the Utility of Privacy-Preserving Histograms

Shuchi Chawla   Cynthia Dwork   Frank McSherry   Kunal Talwar

Abstract

In a census, individual respondents give private information to a trusted party (the census bureau), who publishes a sanitized version of the data. There are two fundamentally conflicting requirements: privacy for the respondents and utility of the sanitized data. Note that this framework is inherently noninteractive.

Recently, Chawla et al. (TCC'2005) initiated a theoretical study of the census problem and presented an intuitively appealing definition of privacy breach, called *isolation*, together with a formal specification of what is required from a data sanitization algorithm: access to the sanitized data should not increase an adversary's ability to isolate any individual. They also showed that if the data are drawn uniformly from a highdimensional hypercube then recursive histogram sanitization can preserve privacy with a high probability.

We extend these results in several ways. First, we develop a method for computing a privacy-preserving histogram sanitization of "round" distributions, such as the uniform distribution over a high-dimensional ball or sphere. This problem is quite challenging because, unlike for the hypercube, the natural histogram over such a distribution may have long and thin cells that hurt the proof of privacy. We then develop techniques for randomizing the histogram constructions both for the hypercube and the hypersphere. These permit us to apply known results for approximating various quantities of interest (e.g., cost of the minimum spanning tree, or the cost of an optimal solution to the facility location problem over the data points) from histogram counts – in a privacy-preserving fashion.

1 Introduction

In a census, individual respondents give private information to a trusted party (the census bureau), who publishes a sanitized version of the data. Note that this framework is inherently noninteractive. There are two fundamentally conflicting requirements: privacy for the respondents and utility of the sanitized data. Very roughly, the sanitization should permit the data analyst to identify strong stereotypes, while preserving the privacy of individuals.

Recently, Chawla et al. [11] initiated a theoretical study of the census problem and presented a definition of privacy that captures the compelling concept of protection from being brought to the attention of others. As the philosopher Ruth Gavison points out, not only is such protection inherently valuable, but, its compromise invites further invasion of privacy, as every action of the target of attention can now be scrutinized. Thus, in [11] the goal of the adversary is to single out, or *isolate*, an individual. Chawla et al. also gave a formal specification of what is required from a data sanitization algorithm: access to the sanitized data should not increase an adversary's ability to isolate any individual.

Histograms are natural candidates for privacy-preserving data sanitization, and as such are frequently encountered in official statistics. Indeed, there is a vast literature on methods of perturbing contingency tables to ensure privacy. They are considered useful, essentially by definition. A central result in [11] states that if the data are drawn uniformly from a high-dimensional hypercube, then a simple technique, *recursive histogram sanitization*, preserves privacy. More precisely, with overwhelming probability over the choice of the database, in the absence of auxiliary information, the probability that the adversary can isolate even one database point is exponentially small in the dimension. The probability space is over the random choices made by the sanitization algorithm and the adversary. The proof is quite robust and can tolerate many, but not all, kinds of auxiliary information[1].

The proof extends to mixtures of well-separated hypercubes, but not to distributions such as high-dimensional spheres, Gaussians, and balls. Extending the argument to such "round" distributions is quite challenging. The heart of the argument involves showing that the volume of the intersection of a ball with a histogram cell grows exponentially with the radius of the ball. The proof in [11] is for the case that cells are hypercubes. Unlike for the hypercube, the natural histogram over a round distribution may have long, thin cells, for which the proof would break down (think of a ball and a strand of spaghetti: increasing the radius of the ball gives only a linear growth in the volume of the intersection with the spaghetti strand). Privacy-preserving histograms for round distributions are the first contribution of this work.

Let us now return to the case in which the data are drawn from the unit d-dimensional cube. A complete recursive histogram is computed by subdividing the cube into 2^d sub-cubes, of equal size, in the natural fashion, recursing until each cell contains at most a single data point[2]. On the one hand, this resembles embedding into a tree [25, 6, 5, 10, 23]: subdivision on the cube corresponds to branching in the tree, on which distances are defined. This is a popular technique because computation in a tree metric is typically simpler than in more general metrics and because randomization techniques can be used to obtain expected low distortion embeddings.

[1] A companion paper discusses impossibility results for privacy-preserving data sanitizers [12].

[2] The privacy-preserving recursive histogram construction stops a bit short of this, and cells with sufficiently small population are not subdivided

However, our focus differs in that we are charged with maintaining privacy and not (necessarily) with ensuring computational simplicity. The main difficulty is to ensure that the randomization does not produce cells that are long and thin (i.e., have a bad aspect ratio), as these interfere with the privacy proof. On the other hand, we need not think only in terms of the tree, but are free to employ geometric information. This leads to reduced expected distortion (at the cost of computational simplicity). Despite these differences, our techniques draw heavily from the literature on tree embeddings.

Things are even more difficult in the case of histograms for round distributions, where the shapes of the cells are irregular[3]. We give two different constructions of histograms for this case — the first is deterministic and preserves privacy, but does not provide the required properties for a good embedding; the second achieves a good embedding, while obtaining slightly worse bounds for privacy.

Below we describe our results in greater detail.

1.1 Summary of Results

Privacy. Chawla et al. [11] investigated the privacy of the following recursive histogram sanitization procedure, in which the data lie in the d-dimensional hypercube of side-length 2, centered at the origin.

The histogram sanitization procedure with privacy parameter $t \geq 2$ is easily described: First, cut the top-level hypercube into 2^d sub-cubes, of equal size, by splitting along the midpoint of each side. Recurse on every sub-cube containing at least $2t$ points. This process results in a set of d-dimensional hypercubes of varying sizes. The sanitization is a description of the cuts made and the exact population of every resulting cell. The adversary suceeds with respect to privacy parameters c and t, if it produces a point $q \in \mathcal{R}^d$ (q need not be a real database point) such that within a ball of radius c times the distance to the nearest database point x to q, there are fewer than t real database points (see Section 2 for the intuition). This event is called *c-isolation*.

Theorem 1.1. [11] *Suppose that* RDB *consists of n points drawn i.i.d. and uniformly from the cube $[-1, 1]^d$. There exists a constant c such that the probability that an adversary, given a recursive histogram sanitization as described above, can c-isolate an* RDB *point is at most $2^{-\Omega(d)}$, independent of n.*

The proof in [11] of this theorem relies on a lemma stating that for appropriate values of r, the volume of the intersection of a ball $B(q, r)$ with a cell in the histogram grows exponentially with r (the upper bound on r depends on the diameter of the cell in question). It is for this reason that we need histogram cells to be “well-rounded”.

In the proof, volume is used as a surrogate for probability density: the adversary is given information about the number of database points in a cell, but from the adversary’s perspective the location of these points within a cell is uniform. The same is true for all the arguments in the current paper except one of our histogram sanitizations for round distributions (Section 3.4), in which the ball or sphere is embedded into a cube. Because the embedding suffers a distortion, the density of points in the cube is not uniform.

[3] Methods appropriate for our needs that yield regularly shaped cells may exist; we have not found any.

In this work, we describe three methods for creating histograms for round distributions, specifically, for high-dimensional balls and spheres, obtaining similar bounds for privacy as in [11], albiet with a larger value of c. In Section C (in the Appendix) we discuss extensions to high-dimensional Gaussians. Our constructions replace the subcubes-based subdivision used for hypercubes in [11] by an appropriate algorithm to subdivide "nice" cells into smaller "nice" cells. Chawla et al. prove that if for all points q, all radii r, and all cells C in the histogram, one of the following two conditions hold, then the probability that the adversary (c,t)-isolates any point in RDB is at most ϵ. Here $P(C)$ denotes the parent cell of C.

$$B(q, cr) \supseteq P(C) \quad \text{or} \quad \frac{\text{Vol}(B(q,r) \cap C)}{\text{Vol}(B(q,cr) \cap C)} < \epsilon$$

In Section 3, we prove that for the subdivision algorithms defined below, the above conditions hold with a constant value of c and $\epsilon = 2^{-O(d)}$. We now describe the subdivision techniques.

In the first method, we use a deterministic decomposition technique based on *nets*. Roughly speaking, we choose "centers" in succession from the cell so that they are *well-spread* (the distance between any two centers is not "too small") and the cell is tightly covered (the dis tance from any point in the cell to its nearest center is not "too large"). The subcells are then given by the Voronoi partition of the cell created by the centers. We argue that the Voronoi regions are relatively well-rounded, and so this technique gives nice privacy constants; however, the constants deteriorate with the depth of recursion, intuitively, because the cells become increasingly less well-rounded.

In the second, we again subdivide cells by picking a set of centers and constructing a Voronoi diagram over these. However in this case, a carefully chosen number of centers is picked uniformly at random from the cell. The randomization allows us to obtain a good embedding of the dataset (as described in the following section). However, the bound on the privacy parameter c in this case, is worse than that for the previous construction.

In the third method, we embed the d-dimensional ball into the d-dimensional hypercube, and then apply the histogram sanitization from [11] to the cube. The resulting embedding has a distortion of $O(d)$. In this case, recursion is not a problem, but the privacy parameter c becomes $\Omega(d^2)$.

Utility. As well as preserving privacy, we would like our histograms to provide substantial utility as well. The deterministic nature of previous histogram work allows for bad examples in which we are unable to compute fairly simple quantities accurately. For example, a simple arrangement of 2^d points near the center of the cube can ensure that in one step each sample is placed in its own cell and recursion thus terminates, reporting one element in each cell. While we might have liked to learn the diameter of the data set, or the cost of a minimum spanning tree, the adversarial arrangement prevents approximation to within any factor.

As has been noted elsewhere, many quantities of interest about a data set can be accurately approximated using *randomized* hierarchical subdivisions. The randomization prevents the adversarial arrangements and can give strong approximation guarantees about approximate solutions.

We give two examples of randomized histogram constructions in Section 4, one using nested cubes for data that lies uniform on the unit cube, and one based on nested spherical regions intended for data that is drawn from spherical distributions. By defining the histogram distance $d_H(x, y)$ between two points x and y to be the maximum possible distance between the two cells, we ensure that (letting Δ_x and Δ_y be the diameters of these cells),

$$\|x - y\| \quad \leq \; d_H(x, y) \; \leq \quad \|x - y\| + \Delta_x + \Delta_y \; .$$

Understanding and bounding the diameters of the final cells then gives us fairly accurate preservation of pairwise distances. We will see in Section 4 that much as randomization enables the preservation of distances in previous work on randomized embeddings, randomization will ensure a strong connection between Δ_x and the t-radius of x.

1.2 Additional Related Work

There is a vast literature on statistical disclosure control and data sanitization. [11] contains a brief discussion of statistical techniques such as suppression, aggregation and perturbation of contingency tables [16, 22], input perturbation (see, for example, [32, 33, 28] in the statistics literature and [3, 2, 17] from the data mining literature), imputation [29], k-anonymity [30], and cryptographic approaches to privacy (see, e.g., [20] and [21]). There is also mention of interactive solutions, such as query auditing [26] (see also [14] for a pessimistic view) and output perturbation ([13, 8], and, the more careful modern treatment [14, 15]). An excellent survey of older work on statistical disclosure control is [1].

To our knowledge, the only work on data sanitization, other than [11], that explicitly takes into account auxiliary information[4] is the lovely paper of Efvimievski et al. [17].

Moving to utility, there is a strong algorithmic literature on the power of probabilistic embeddings(see e.g. [25, 4, 6, 7, 9, 10, 18, 5, 23, 31]). Applications include approximation algorithms, online algorithms and geometric data stream algorithms. We refer the reader to [24, 6, 18, 27] for surveys of some these applications.

2 Definitions

The Isolating Adversary. We assume a distribution $\mathcal{D}$ on databases. The real database will be denoted RDB containing n points in d-dimensional space $\mathcal{R}^d$. A sanitization algorithm takes as input the real database and produces as output a sanitized database, denoted SDB. The sanitization algorithm may be randomized. The SDB may be of essentially any form, such as some number n' of points, in some space (not necessarily the same space as the data), or a summary of the dataset RDB or the distribution $\mathcal{D}$.

The goal of the adverary is, intuitively, to single out, or *isolate*, someone from the crowd, and so the adversary is denoted $\mathcal{I}$. The isolator has two inputs: the sanitized database SDB and *auxiliary* information z. This is any information that the adversary may have access to *other than* the SDB itself. For example, it may contain complete or partial information for many of the points.

On input (SDB, z) the isolator produces a point $q \in \mathcal{R}^d$. We can think of q as a description of a candidate database point (q need not actually be in the RDB; think of it as the adversary's guess as to what someone in the RDB might look like). The isolator may flip coins.

[4] Auxiliary information is information other than what is in the sanitized database that may be available to the adversary.

We now define success for the isolator. There are two parameters.

Definition 2.1. **((c,t)-isolation)** *Let y be any RDB point, and let $\delta_y = \|q-y\|$. We say that q* (c,t)-isolates y *if $B(q, c\delta_y)$ contains fewer than t points in the* RDB*, i.e.,* $|B(q, c\delta_y) \cap \mathrm{RDB}| < t$.

We frequently omit explicit mention of t, and speak of c-isolation. It is an easy consequence of the definitions that if $q = \mathcal{I}(\mathrm{SDB}, z)$ fails to c-isolate the nearest RDB point to q, then it fails to c-isolate even one RDB point.

Perfect Sanitization. In order to evaluate a data sanitization algorithm, we compare how well the isolator can do with how well an isolator simulator—not having access to the sanitized database—can do. The isolator simulator is denoted $\mathcal{I}'$. Unlike the isolator, $\mathcal{I}'$ only has access to the auxiliary information.

More formally, a *database sanitizer*, or simply *sanitizer* for short, is a randomized algorithm that takes as input a real database of some number n of points in $\mathcal{R}^d$, and outputs a sanitized database SDB. A sanitizer is *perfect* if for all isolating adversaries $\mathcal{I}$, and for every distribution $\mathcal{D}$ over $\mathcal{R}^{n \times d}$ from which the real database points are drawn, there exists an adversary simulator $\mathcal{I}'$ such that with high probability over choice of RDB, for all auxiliary information strings z, the probability that $\mathcal{I}(\mathrm{SDB}, z)$ succeeds minus the probability that $\mathcal{I}'(z)$ succeeds is small. The probabilities are over the coin tosses of the sanitization and isolation algorithms. We allow the sanitizer to depend on the parameters c, t, and also allow $\mathcal{I}$ and $\mathcal{I}'$ to have access to $\mathcal{D}$.

More precisely, let ϵ be a parameter (for example, $\epsilon = 2^{-d/2}$). For parameters c and t we require that for all $\mathcal{I}$ there exists an $\mathcal{I}'$ such that, if we first pick a real database $\mathrm{RDB} \in_R \mathcal{D}$, then with overwhelming probability over RDB, for all z,

$$\forall S \subseteq \mathrm{RDB}\, |Pr[\exists x \in S : I(SDB, z)(c,t)\mathrm{-isolates}\ x] - Pr[\exists x \in S : I'(z)(c,t)\mathrm{-isolates}\ x]| < \epsilon$$

where the probabilities are over the choices made by $\mathcal{I}$, $\mathcal{I}'$, and the sanitization algorithm.

In a companion paper we show that a nontrivial perfect sanitizer does not exist [12]. (The "killer" is the auxiliary information.) Nonetheless, the definition lays the foundation for comparing among sanitization techniques.

Notation. Throughout the paper, we use $\|x-y\|$ to denote the distance between two points x and y and $|x|$ to denote the length of a vector $\vec{x}$. $B(x,r)$ denotes a ball of radius r around point x. We use t_x to denote the t-radius of x, i.e. $t_x = \min\{r \mid |B(x,r) \cap \mathrm{RDB}| \geq t\}$.

3 Histogram Sanitizations for Round Distributions

We first argue that histogram sanitizations with well-rounded cells preserve privacy (Section 3.1). We then define two parameterized properties of collections of points and show that if a collection of centers (points chosen by the sanitization algorithm, not database points) satisfies these properties, then the Voronoi regions induced by the collection are well-rounded (Section 3.2). The degree of well-roundedness depends on the values of the parameters. We then discuss two techniques for generating centers and prove that both yield collections of centers satisfying the two properties with reasonably good values of the parameters.

3.1 Privacy for well-rounded cells

In this section, we show that histogram sanitizations with well-rounded cells preserve privacy. We start with a definition of well-rounded cells.

Definition 3.1. *A cell C is said to be k-*well-rounded *with radius R and center p iff C is convex and $B(p, \frac{R}{k}) \subseteq C \subseteq B(p, R)$.*

We now present the key lemma of this section.

Lemma 3.1. *Let C be a k-well-rounded cell, $k \geq 1$, of size R with center p, and let $c' = \max\{2k, 2\sqrt{2}\}$. Then, for any point $q \in C$ and radius $r < \frac{R}{kc'}$, the following holds:*

$$\frac{Vol(B(q, r) \cap C)}{Vol(B(q, c'r) \cap C)} < 2^{-(\frac{d}{2}-1)}d$$

Proof. First consider the case when q lies in the ball $B(p, \frac{R}{k})$. In this case, we will show that the volume of $B(q, c'r) \cap B(p, \frac{R}{k})$ is large. In particular this quantity is larger than the volume of a cap of $B(q, c'r)$ that subtends an angle of $\delta = \pi/3$ at the center q.[5]

This volume can be computed as an integral over the $d-1$ dimensional volume of balls of radius ranging from 0 to $c'r \sin\delta$. Some calculation shows that this volume is at least $\frac{\sin^d \delta}{\sqrt{\pi d}} \mathrm{Vol}(B(q, c'r)) > \frac{2^{-d}}{2d} \mathrm{Vol}(B(q, c'r))$. This implies that $\mathrm{Vol}(B(q, c'r) \cap C) \geq \mathrm{Vol}(B(q, c'r) \cap B(p, \frac{R}{k})) > \frac{2^{-d}}{2d} \mathrm{Vol}(B(q, c'r))$.

Likewise, $\mathrm{Vol}(B(q, r) \cap C) \leq \mathrm{Vol}(B(q, r)) = (2\sqrt{2})^{-d} \mathrm{Vol}(B(q, c'r))$. Combining the two expressions, we get that for $q \in B(p, \frac{R}{k})$,

$$\frac{\mathrm{Vol}(B(q, r) \cap C)}{\mathrm{Vol}(B(q, c'r) \cap C)} < 2^{-(\frac{d}{2}-1)}d$$

Next consider the case when q lies outside the ball $\mathcal{B} = B(p, \frac{R}{k})$. Consider the convex hull H_q of q and the ball $\mathcal{B}$. This lies entirely inside C, as C is convex. Furthermore, apart from the ball $\mathcal{B}$, the convex hull contains a cone Λ_q formed by tangents from q to $\mathcal{B}$. Note that this cone subtends a large solid angle at q. In particular, the angle between any tangent and the line joining q and p is at least $\theta = \sin^{-1} \frac{1}{k}$ (by the well-roundedness of C).

Now we can compute the volume $\mathrm{Vol}(B(q, c'r) \cap H_q)$ as the union of two terms—the intersection with Λ_q, and the intersection with $\mathcal{B}$. Note that when q is far from the center p (in particular, farther than $\frac{\sqrt{2}R}{k}$), then the intersection of $B(q, c'r)$ and H_q is contained entirely inside the open-ended cone defined by Λ_q. This intersection has volume larger than the volume of a cap of $B(q, c'r)$ that subtends an angle θ at q, which is at least $\frac{1}{\sqrt{\pi d} k^d} \mathrm{Vol}(B(q, c'r))$. On the other hand, when q is closer than $\frac{\sqrt{2}R}{k}$ to p, the intersection of $B(q, c'r)$ and $\mathcal{B}$ has volume at least $\frac{1}{\sqrt{\pi d} 2^{\frac{d}{2}}} \mathrm{Vol}(B(q, c'r)) > \frac{1}{\sqrt{\pi d} (\sqrt{2}k)^d} \mathrm{Vol}(B(q, c'r))$. Therefore, we get that $\mathrm{Vol}(B(q, c'r) \cap H_q) > \frac{1}{\sqrt{\pi d} (\sqrt{2}k)^d} \mathrm{Vol}(B(q, c'r))$.

[5] A cap of a d-dimensional ball subtending an angle δ at the center of the ball is defined by an axis through the center, and contains all points on the surface of the ball that subtend an angle at most δ with the axis. The volume of a cap is the volume of its d-dimensional convex hull.

Putting everything together, we get

$$\frac{\mathrm{Vol}(B(q,r)\cap C)}{\mathrm{Vol}(B(q,c'r)\cap C)} < \frac{(2k)^{-d}\mathrm{Vol}(B(q,c'r))}{(\sqrt{\pi}d(\sqrt{2}k)^d)^{-1}\mathrm{Vol}(B(q,kr))} < 2^{-(\frac{d}{2}-1)}d$$

□

The following corollary follows from the lemma by observing that for a point q outside C, the ratio of volumes is bounded above by the corresponding ratio for an appropriate point q' in C, albeit with $(c-1)$ in the place of c.

Corollary 3.2. *Let C be a k-well-rounded cell of size R with center p. Then for any point q and radius r, and for $c = 4k^2$, either $cr \geq R$, or,*

$$\frac{Vol(B(q,r)\cap C)}{Vol(B(q,cr)\cap C)} < 2^{-(\frac{d}{2}-1)}d$$

3.2 Voronoi-based histograms

Let S be the region of interest, say, the d-dimensional unit ball or sphere. We start with the set S as the level 0 cell in the histogram. At step $i = 1, 2, \ldots$, we consider all level $i-1$ cells C that contain more than t points. For each of these cells, we obtain level i cells by subdividing the cell as follows: we pick a set of centers in the cell, and construct a Voronoi diagram over these. The next level cells are then given by the partition of the cell induced by the Voronoi partition. We assume that cells created at level i are k_i-rounded. Different techniques for choosing the centers yield different values for k_i.

We continue the recursion for a constant number of steps, or until all the cells have fewer than t points, and release the exact counts of points in each cell. Assume the procedure runs for s steps. Then each cell in this histogram is k_s-well-rounded for a constant k_s. Therefore, using the argument in the previous sections, privacy is achieved for a constant $4k_s^2$.

Definition 3.2. *A set of points $\{p_1, \cdots, p_m\}$ is said to r_1-cover a set C, if $C \in \cup_i B(p_i, r_1)$. It is said to be r_2-well-spread if for every pair of points p_i and p_j, $i \neq j$, $\|p_i - p_j\| \geq r_2$.*

The following lemma gives conditions under which Voronoi cells are well-rounded. A proof appears in the appendix.

Lemma 3.3. *Let C be a k-well-rounded cell of size R, and $\{p_1, \cdots, p_m\}$ be a set of r_2-well-spread points that r_1-cover C with $r_2 \leq r_1 < R$. Let V_i be the region containing p_i in the voronoi partition of the points $\{p_1, \cdots, p_m\}$. Then the subcells $V_i \cap C$ are $O(\frac{r_1 k}{r_2})$-well-rounded with radius $2r_1$.*

3.3 Two Methods for Choosing Centers

We now describe two methods for picking centers in cells so as to achieve the well-spread and covering properties. This along with the results in the previous section implies that the Voronoi-based histograms constructed from these centers preserve privacy.

Method (1): Picking Well-Spread Centers Directly. We describe the procedure as applied to a cell C. We pick points $p_1, p_2, \cdots$ in C in succession as follows: the point p_i is picked arbitrarily such that it is at distance at least $R/4$ from each of the points p_j for $j < i$. We stop when every point in C is within distance $R/4$ of at least one of the points p_i. By construction, this gives $R/4$-well-spread centers that $R/4$-cover the cell C.

As mentioned earlier, we do not know how to randomize this construction so as to be able to prove that the probability of a cell boundary lying between two real database points is proportional to the distance between the points. This is addressed in Method (2), although the parameter k for well-roundedness deteriorates more quickly with recursion than the corresponding one for Method (1).

Method (2): Picking Centers Uniformly at Random As described earlier, in the second method, we pick centers from the cell at random. The following lemma shows that we obtain well-rounded Voronoi regions with a high probability. The proof appears in the appendix.

Lemma 3.4. *Let C be a k-well-rounded cell of size R, and let $p_1, \cdots, p_m$ be $m = 4d8^d$ points picked uniformly at random from C. Then with probability at least $1 - exp(-d)$, the points are $\Omega(R/k)$-well-spread, and $R/4$-cover C.*

We get the following corollary:

Corollary 3.5. *Let C be a k-well-rounded cell of size R, and let $p_1, \cdots, p_m$ be $m = 4d8^d$ points picked uniformly at random from C. Then with probability at least $1 - exp(-d)$ the subcells $V_i \cap C$ are $O(k^2)$-well-rounded with radius $R/2$.*

3.4 An Alternative Sanitization: Embedding in a Hypercube

We first describe a simple embedding of the unit d-dimensional ball into the d-dimensional hypercube centered at the origin and having side length 2. We argue that the distortion of the mapping is at most d. Finally, we show that if we first carry out any embedding into the cube with some distortion α and then apply recursive histogram sanitization to the cube, then the probability that an adversary can $(\alpha^2 c, t)$-isolate is exponentially small in d. Here, c is the constant obtained by Chawla et al. described in Theorem 1.1.

The main technical claim for this section is the following bound on distortion. A proof is given in the appendix.

Lemma 3.6. *There exists a bijection $f : B(0,1) \to [-1,1]^d$ such that for all pairs x and y, $\frac{1}{\alpha}\|x-y\| \leq \|f(x) - f(y)\| \leq \beta\|x-y\|$ with $\alpha\beta = O(d)$.*

Finally, this implies that given a uniform distribution on the sphere, we can use the map f defined above to map it to a d^d-uniform distribution over the cube.

Lemma 3.7. *For any point $x \in B(0,1)$, let $g(x)$ denote the probability density at the point x under the uniform distribution over the ball. Let $g'(x)$ denote the probability density at point $f(x)$ in the cube. Then, $\alpha^{-d} g(x) \leq g'(x) \leq \beta^d g(x)$, where α and β are as in the previous lemma.*

The proof of this lemma is straightforward from the previous lemma on distortion, and we omit it for brevity. It follows that a (c, t) privacy-presererving scheme with probability $(1 - 2^{-\Omega(d)})$ on the cube leads to a (cd^2, t) privacy preserving scheme with probability $(1 - 2^{-\Omega(d)})$.

4 Randomized Histograms

We now discuss several randomizations of histogram construction. Deterministic histograms have the slight defect that their rigid structure allows for adversarial inputs on which the histogram may misrepresent aspects of the data. While cell counts will always be accurate (given a correct histogram), we might hope for more: distances between points that have been histogram sanitized could reflect their actual distances in some capacity.

In the randomized histogram constructions we now consider, we will provide bounds on the expected distance between any two points in the histogram. Bounds on these expectations are useful in a large class of optimization problems for which the cost function is a simple function of inter-sample distances. Problems such as minimum spanning tree, facility location, and minimum weight matching fall into this framework.

Stated generally, the optimal solution for any such problem is a valid solution when distances are computed from a histogram, but the cost will likely have increased. The expected cost of this solution (it is a random quantity) can be bounded using the bounds that we establish on the increase in distances. Now, if we compute an optimal solution using histogram distances we are ensured that it is no more than the cost of the original optimal solution with its histogram distances. Of course, this histogram-optimal solution is a valid solution using the original distances, and as the histogram can only increase distances, returning from it will never increase the cost of a solution. Therefore, in expectation a histogram-optimal solution costs at most the optimal solution plus its histogram expansion. As we shall see, the histogram expansion for an edge (x, y) has expectation $\alpha(t_x + t_y)$ for some constant α independent of n. Depending on the problem at hand, we can use this to get a better estimate.

For example, it is shown in [19] that there is always a spanning tree with maximum degree 3 and cost no more than $\frac{5}{3}$ times the optimal. Thus the expected cost of our tree is bounded above by $\frac{5}{3}MST + \sum_x 3\alpha t_x$. Since MST has expectation $O(n^{(1-\frac{1}{d})})$ and t_x is expected to be $O(d(\frac{t}{n}^{\frac{1}{d}})$ the error in our estimate is expected to be within a constant factor of the optimum. Similar results can be shown for other problems.

4.1 Randomized Histograms on the Unit Hypercube

We now look at a modification of the histogram approach of Chawla et al. [11], who construct histograms using recursive subdivision of hypercubes. Their approach subdivides any hypercube that contained at least $2t$ samples into its 2^d constituent hypercubes of half the edge length. This process continues until each cube contains no more than $2t$ samples, at which point the counts of samples in each cube are released. [11] prove that this subdivision preserves privacy when the prior distribution on the points is uniform on the unit hypercube.

To randomize this construction and yield bounds on the expected increase in distances, we start from a construction that has been previously studied in the metric embeddings literature. The main idea is to conduct a standard hypercube subdivision, but to translate the center of the main hypercube by a vector chosen uniformly at random from the unit hypercube. The hypercube should also be inflated by a factor of two in each dimension to ensure that it still covers the entire hypercube.

As we will see in Theorem 4.1, this random translation will provide us with utility, but it may lack privacy. Specifically, for any cell in the histogram the points are distributed uniformly over the intersection of the cell and the unit hypercube. This is not a problem for interior cells where the intersection is the cell itself, a shapely hypercube, but those cells that intersect the surface of the hypercube may have highly non-uniform dimensions: distributions that have low effective dimension and are easy to attack. To correct this, for every level of the histogram, all cells intersecting the hypercube surface are disbanded, and the interior cells exposed are expanded to cover the area of any adjacent discarded cells. The resulting subdivision into hyperrectangular regions has guaranteed aspect ratios, accomodating the privacy proofs in [11].

4.1.1 The Expected Radius of a Containing Cell

The distance between any two points x and y as captured by the histogram, taken to be the maximum distance between the two smallest cells containing x and y, can be fairly easily described by the diameters of these cells. Letting Δ_x be the diameter of the smallest histogram cell containing x, and likewise for Δ_y and y, the triangle inequality tells us that

$$\|x-y\| \leq \quad d_H(x,y) \quad \leq \|x-y\| + \Delta_x + \Delta_y \tag{1}$$

As such, we are interested in bounding Δ_x and Δ_y. As noted previously, bounding their expectation suffices, which we do now. Recall that t_x denotes the t-radius of a point x.

Theorem 4.1. *For any sample x, letting Δ_x be the diameter of the smallest cell containing x.*

$$E[\Delta_x] \quad \leq \quad 2\min\{d^{\frac{3}{2}}, td\}t_x \log(1/t_x)$$

Proof. The diameter of the smallest cell is, up to a factor of 2, dominated by the diameter of the smallest cell in the traditional randomized histogram construction, where we do not collapse perimeter cells. As such, we shift our attention to that construction. We will consider the contribution to the expectation from each level of the histogram. The total expectation will then be their sum. At level i, we can upper bound the probability that the recursion will terminate from lack of neighbors by the probability that each of the samples within t_x of x are separated into a different cell. This is at most the probability that the neighborhood $B(x,t_x)$ of x is cut by the decomposition. To bound this probability, notice that it is at most the sum of the probabilities that it happens in each of the dimensions. In each dimension, the probability that $B(x,t_x)$ is cut is simply the length of the projection of $B(x,t_x)$ along that dimension, divided by 2^{-i}, as the separating lines are dropped uniformly at random. Thus we get an upper bound of $\frac{dt_x}{2^{-i}}$ on the probability of recursion terminating.

Alternately, taking a simple union bound, this is at most t times the probability that it happens to any one of the t nearest neighbors. To bound this probability, notice that it is at most the sum of the probabilities that it happens in each of the dimensions. In each dimension, the probability that x is separated from y is simply the absolute value of the length of $x-y$ in that dimension, divided by 2^{-i}, as the separating lines are dropped uniformly at random. Summing these absolute values gives us $\|x-y\|_1/2^{-i}$, which for each of the t-neighbors is at most $d^{1/2}t_x/2^{-i}$. Thus the probability that the t nearest neighbors of x do not land in the same level i cell is also bounded by $td^{1/2}t_x/2^{-i}$.

If this event occurs, the contribution to Δ_x would be $d^{1/2}2^{-i}$, and so the expected contribution is $\min\{d^{\frac{3}{2}}, td\}t_x$. There are at most $\log(1/t_x)$ levels we must worry about, yielding the stated bound and completing the proof. □

4.2 Randomized Histograms for Round Distributions

In this section, we revisit the randomized histogram construction from Section 3.3, and show that these imply a low-distortion embedding of the dataset into trees.

Recall that at any level in the recursion, we pick $m = 4d \cdot 8^d$ points $p_1, p_2, \ldots, p_m$ uniformly at random from the cell C. The clusters are the Voronoi cells defined by these centers, i.e. cluster C_i consists of all points $x \in C$ such that $\|x - p_i\| \leq \|x - p_j\|$ for all $j \neq i$ (breaking ties arbitrarily). In Section 3.3 we showed that these cells are well-rounded.

We now show that for any point x and any r, $B(x, r)$ is cut with probability proportional to the r. The main idea behind the proof is that if the distance r is much smaller than the distance from x to its closest center, then any point $y \in B(x, r)$ gets assigned to a different center only if this center falls within a thin shell of thickness r around $B(x, r)$, and the probability of this event is proportional to r.

Given a clustering $C_1, \ldots, C_k$ and a set $S \subseteq \Re^d$, we say S is *cut* by the clustering if there are distinct indices i and j such $S \cap C_i$ and $S \cap C_j$ are both non-empty. The proof of the following lemma is given in Appendix B.

Lemma 4.2. *For any convex set C of radius ρ and $x \in C$ and for any r,*

$$Pr[B(x,r) \textit{ is cut by a subdivision of } C] = O(dr/\rho)$$

The lemma states that at any level i, the probability that a ball of radius t_x around x is cut is proportional to $\frac{dt_x}{2^{-i}}$. When this happens, x lies in a cell of diameter about 2^{-i}. Note that if a pair of points is cut at level i, it incurs a distance of $O(2^{-i})$ in the resulting tree, which is roughly the size of the cells at this level. Then by the argument in the proof of theorem 4.1, the contribution of this level to the expected cell diameter is $O(dt_x)$. Since the number of levels in the recursion is a constant, the error term in the analysis above has expectation $O(dt_x)$.

References

[1] N. R. Adam and J. C. Wortmann, Security-Control Methods for Statistical Databases: A Comparative Study, *ACM Computing Surveys* 21(4): 515-556 (1989).

[2] D. Agrawal and C. Aggarwal, On the Design and Quantification of Privacy Preserving Data Mining Algorithms, *Proceedings of the 20th Symposium on Principles of Database Systems*, 2001.

[3] R. Agrawal and R. Srikant, Privacy-preserving data mining, *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 439–450, 2000.

[4] N. Alon, R. M. Karp, D. Peleg, and D. West. A graph-theoretic game and its application to the k-server problem. *SIAM Journal on Computing*, 24(1):78–100, Feb. 1995.

[5] S. Arora. Polynomial time approximation schemes for Euclidean TSP and other geometric problems. In *37th Annual Symposium on Foundations of Computer Science*, pages 2–11, Burlington, Vermont, 14–16 Oct. 1996. IEEE.

[6] Y. Bartal. Probabilistic approximations of metric spaces and its algorithmic applications. In *IEEE Symposium on Foundations of Computer Science*, pages 184–193, 1996.

[7] Y. Bartal. On approximating arbitrary metrics by tree metrics. In *STOC*, 1998.

[8] Beck, L., A security machanism for statistical database, *ACM Transactions on Database Systems (TODS)*, 5(3), p.316-3338, 1980.

[9] M. Charikar, C. Chekuri, A. Goel, and S. Guha, and S. Plotkin, Approximating a finite metric by a small number of tree metrics, FOCS pp. 379 – 388, 1998.

[10] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. A. Plotkin. Approximating a finite metric by a small number of tree metrics. In *IEEE Symposium on Foundations of Computer Science*, pages 379–388, 1998.

[11] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, Toward Privacy in Public Databases, *To Appear, Proc. Theory of Cryptography Conference*, 2005 available at `http://research.microsoft.com/research/sv/DatabasePrivacy/`

[12] S. Chawla, C. Dwork, M. Naor, et al., On the Limits of Privacy-Preserving Data Sanitization, *in preparation*, 2004.

[13] Denning, D., Secure statistical databases with random sample queries, *ACM Transactions on Database Systems (TODS)*, 5(3), p.291-315, 1980.

[14] I. Dinur and K. Nissim, Revealing information while preserving privacy, *Proceedings of the Symposium on Principles of Database Systems*, pp. 202-210, 2003.

[15] C. Dwork and K. Nissim, Privacy-Preserving Datamining on Vertically Partitioned Databases, *Proc. CRYPTO 2004*.

[16] A. Dobra and S.E. Fienberg, and M. Trottini, Assessing the risk of disclosure of confidential categorical data, *Bayesian Statistics 7*, pp. 125–14, Oxford University Press, 2000.

[17] A. V. Evfimievski, J. Gehrke and R. Srikant, Limiting privacy breaches in privacy preserving data mining, *Proceedings of the Symposium on Principles of Database Systems*, pp. 211-222, 2003.

[18] J. Fakcharoenphol, S. Rao, and K. Talwar, A tight bound on approximating arbitrary metrics by tree metrics, *Proc. 35th ACM Symposium on Theory of Computing,* pp. 228 – 455, 2003.

[19] S. Khuller, B. Raghavachari, and N. Young, Low degree spanning trees of small weight. *SIAM Journal on Computing*, 25(2):355–368, 1996.

[20] W. Gasarch, A Survey on Private Information Retrieval. *BEATCS Computational Complexity Column*, 82, pp. 72-107, Feb 2004.

[21] O. Goldreich, *The Foundations of Cryptography - Volume 2.* Cambridge University Press, 2004.

[22] D. Gusfield, A Graph Theoretic Approach to Statistical Data Security, *SIAM Journal on Computing 17*(3), pp. 552–571, 1988

[23] P. Indyk, Algorithms for Dynamic Geometric Problems over Data Streams *Proc. STOC 2004*

[24] P. Indyk, J. Matousek. Low Distortion Embeddings of Finite Metric Spaces. In *CRC Handbook of Discrete and Computational Geometry.* To appear.

[25] R. Karp, A $2k$-competitive Algorithm for the Circle, *Manuscript*, August, 1989.

[26] J. M. Kleinberg, C. H. Papadimitriou, and P. Raghavan, Auditing Boolean Attributes, *J. Comput. Syst. Sci. 66*(1), pp. 244–253, 2003

[27] S. Muthukrishnan. Data streams: Algorithms and applications (invited talk at soda'03). Available at http://athos.rutgers.edu/m̃uthu/stream-1-1.ps, 2003.

[28] G. Roque. Application and Analysis of the Mixture-of-Normals Approach to Masking Census Public-use Microdata. Manuscript, 2003.

[29] D. B. Rubin, Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics 9*(2), 1993, pp. 461–468.

[30] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10* (5), 2002; 571-588.

[31] K. Talwar. Bypassing the Embedding: Algorithms for low dimensional metrics. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 281–290, 2004.

[32] W. E. Winkler. Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems. In *Proc. Privacy in Statistical Databases* 2004, Springer LNCS 3050.

[33] W. E. Winkler. Re-identification Methods for Masked Microdata. In *Proc. Privacy in Statistical Databases* 2004, Springer LNCS 3050.

A Proofs for Sections 3

Proof of Lemma 3.3: It is immediate from the covering property of the points that each subcell $V_i \cap C$ is contained within the ball $B(p_i, r_1)$. Therefore, for any point $q \in V_i \cap C$, we have $V_i \cap C \subseteq B(q, 2r_1)$.

Note that by the well-spreading property of the points, we have $B(p_i, \frac{r_2}{2}) \subseteq V_i$. We now show that there is a point $p'_i \in V_i \cap C$ such that $B(p'_i, \frac{r_2}{2(k+1)}) \subseteq V_i \cap C$. This along with the above property $B(p'_i, 2r_1) \supseteq V_i \cap C$ implies the lemma.

Let x be the center of C, that is, $B(x, \frac{R}{k}) \subseteq C$ holds. We consider two cases. First suppose that $p_i \in B(x, \frac{R}{k})$. Then, it is easy to see that $B(x, \frac{R}{k}) \cap B(p_i, \frac{r_2}{2})$ contains a ball of radius $\min\{\frac{R}{2k}, \frac{r_2}{4}\}$, which is at least $\frac{r_2}{2(k+1)}$ for $k \geq 1$, with center $p_i' = (x + p_i)/2$. This intersection is contained inside $V_i \cap C$ and therefore in this case, the above claim holds.

Next suppose that $p_i \notin B(x, \frac{R}{k})$. Then, consider the cone Λ_{p_i} defined by tangents from p_i to $B(x, \frac{R}{k})$, and note that this cone subtends an angle at least $\sin^{-1}(1/k)$ at p_i. The intersection of Λ_{p_i} with $B(p_i, \frac{r_2}{2})$ is contained entirely inside $V_i \cap C$. Furthermore, this intersection contains a ball of radius $\frac{r_2}{2(k+1)}$ (details omitted for brevity). Therefore, again the claim holds, and the lemma follows. □

Proof of Lemma 3.4: Consider a collection of balls of radius $R/8$ that cover C. Note that the smallest such collection contains at most 16^d balls[6]. If there is at least one point in each ball in the collection, then the entire set C is $R/4$-covered. To show that this happens with high probability, note that if we pick m points u.a.r. from C, the probability that no points fall into a particular ball of radius $R/8$ is at most $(1 - 8^{-d})^m$. The probability that any of the 16^d balls are left empty is at most $16^d(1 - 8^{-d})^m$. Taking $m = 4d8^d$, this probability is exponentially small.

Secondly, for any two points picked u.a.r. from C, the probability that these points fall within distance $R/100k$ of each other is at most 100^{-d}. Therefore, the probability that any two of the picked points fall within distance $R/16$ of each other is at most $m^2 100^{-d}$ which is again exponentially small in d. □

Proof of Lemma 3.6: Consider the map f defined as follows. For a point x on the surface of the unit ball $B(0, 1)$, we map it to the point x' on the surface of the cube $[-1, 1]^d$ such that the $\langle x, x' \rangle = |x||x'|$, i.e. x and x' are parallel vectors. For a point $y \in B(0, 1)$, such that $y = \alpha x$, where α is a scalar and x is on the surface of the sphere, we map y to $y' = \alpha x'$. Finally the origin is mapped to itself.

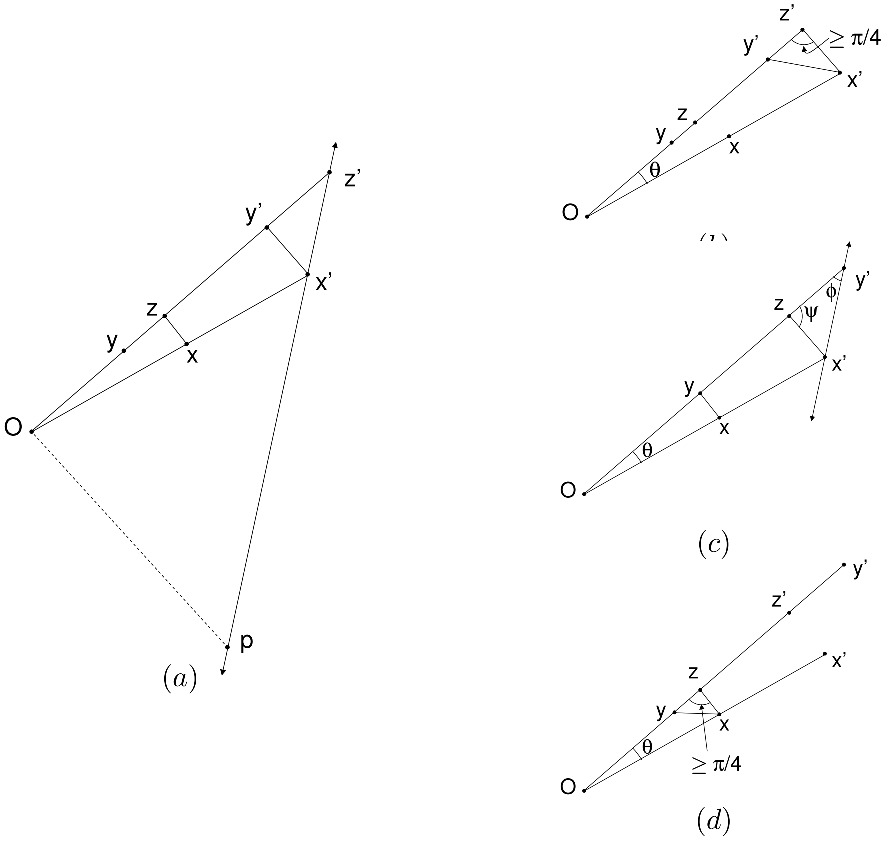
It is easy to see that this map is bijective. We now argue that f has low distortion.

Claim A.1. *There exists a universal constant c such that for any $x, y \in B(0, 1)$, $\|f(x) - f(y)\| \geq c\|x - y\|$.*

Proof. First consider the case when x and y are parallel. Since $B(0, 1)$ is contained in $[-1, 1]^d$, clearly $|x'| \geq |x|$ and hence for x, y parallel, the claim holds.

Next consider the case when $|x'| = |y'|$ and assume without loss of generality that x' lies on the surface of the cube(see Figure 1(a)). Let $z = \frac{|x|}{|y|} y$ be the vector parallel to y with length $|x|$. By triangle inequality, $\|x - y\| \leq \|x - z\| + \|z - y\|$. Since $\|x - z\| = \frac{|x|}{|x'|}\|x' - y'\|$, the first term is at most $\|x' - y'\|$. Let z' be the point on the surface of the cube in the direction of y (note that z' is also $f(z)$) . Clearly $\|z - y\| = |z| - |y|$. But $|z| = |x| = 1$ and $|y| = \frac{|y'|}{|z'|}|z| = \frac{|y'|}{|z'|}$. Thus $\|z - y\| = \frac{1}{|z'|}(|z'| - |y'|) = \frac{\|y' - z'\|}{|z'|}$. Now consider the point p on the face of the cube such that the triangles $z'y'x'$ and $z'Op$ are similar. Thus $\frac{\|y' - z'\|}{|z'|} = \frac{\|x' - y'\|}{|p|}$. Since p lies on a face, $|p|$ is at least 1, and hence $\|x - y\| \leq 2\|x' - y'\|$ in this case.

[6] To see this, note that if we cover the cell with as many non-overlapping balls of radius $R/16$ as possible, then the cell can be covered by balls of radius $R/8$ around the same centers. There are at most 16^d such balls.



(a)

(c)

(d)

Figure 1: Proofs of claims A.1 and A.2

Now we turn our attention to the general case when $|x'| \neq |y'|$. Assume without loss of generality that $|x| \geq |y|$ (see Figure 1(b)). First note that whenever the angle θ is at least $\frac{\pi}{2}$, $\|x - y\| \leq |x| + |y| \leq |x'| + |y'| \leq \sqrt{2}\|x' - y'\|$ and hence the claim holds. Let $z' = \frac{|x'|}{|y'|}y'$ be the vector parallel to y with norm $|x'|$ and let z be its preimage. Then $\|x - y\| \leq \|x - z\| + \|z - y\|$. By the previous two cases considered, this is at most $2\|x' - z'\| + \|y' - z'\|$. On the other hand, in the triangle $x'z'y'$, the angle at z is at least $\frac{\pi}{4}$ and hence a simple argument using the cosine rule shows that $\|x' - y'\| \geq \sqrt{\frac{7}{8}}\max\{\|x' - z'\|, \|y' - z'\|\}$. Hence it follows that $\|x - y\| \leq 3\sqrt{\frac{8}{7}}\|x' - y'\|$. Hence the claim.

Note that we assumed that x' and y' are defined be the same face of the hypercube. When this is not the case, we can split the shortest x'-y' path into segments each of which is defined by a single face, and then use the triangle inequality. □

Claim A.2. *There exists a universal constant c' such that for any $x, y \in B(0,1)$, $\|f(x) - f(y)\| \leq c'd\|x - y\|$.*

Proof. We once again prove the claim in steps. First note that any point p on the surface of the cube satisfies $|p| \leq \sqrt{d}$. Thus for any x, $\|0 - x'\| \leq \sqrt{d}\|0 - x\|$. This also implies that whenever x and y are parallel, $\|x' - y'\| \leq \sqrt{d}\|x - y\|$.

Now consider the case when $|x| = |y|$. Without loss of generality, x and y are unit vectors and x' and y' thus lie on the surface of the hypercube(see Figure 1(c)). When the angle θ is at least $\frac{\pi}{2}$ a simple argument analogous to the previous case applies. Let $z = \frac{|x'|}{|y'|}y'$. Then clearly $\frac{\|x'-z\|}{\|x-y\|} = \frac{|x'|}{|x|}$ and this ratio is at most $\sqrt{d}$. The ratio of $\|x' - y'\|$ to $\|x' - z\|$ depends on the ratio of the sines of angles ϕ and ψ. Since ψ is between $\frac{\pi}{2}$ and $\frac{3\pi}{4}$ and $sin(\phi)$ is at least $\frac{1}{\sqrt{d}}$, $\|x' - y'\|$ is at most $\sqrt{2d}\|x' - z\|$. Thus $\|x' - y'\| \leq d\sqrt{2}\|x - y\|$.

Next we consider the case when $|x| \neq |y|$ and assume without loss of generality that $|x| \geq |y|$ (see Figure 1(d)). Let $z = \frac{|x|}{|y|}y$ and let $z' = f(z)$. Once again the case $\theta \geq \frac{\pi}{2}$ is easy. Now $\|x' - y'\| \leq \|x' - z'\| + \|y' - z'\|$. By the previous two case, this is at most $d\sqrt{2}\|x - z\| + \sqrt{d}\|z - y\|$. Since the angle at z in the triangle xzy is at least $\frac{\pi}{4}$, cosine rule implies that $\|x - y\| \geq \sqrt{\frac{7}{8}} \max\{\|x - z\|, \|z - y\|\}$. Hence $\|x' - y'\| < 3d\|x - y\|$.

Note that once again we have implicitly assumed that x' and y' are defined by a the same face of the cube. The bound on the expansion in the general case follows by splitting the shortest x-y path into segments x-p_1,p_1-p_2,...,p_k-y, such that the image of each segment is defined by a single face, and using the triangle inequality. □

The two claims together show that the distortion of the map is $O(d)$ and this completes the proof of the lemma. □

B Proofs for Section 4

The proofs in this section rely on the following simple observation.

Observation B.1. *Let S be any convex region containing x. Then for any $r < \frac{R}{d}$*

$$\frac{Vol(B(x, R + r) \cap S)}{Vol(B(x, R) \cap S)} \leq 1 + O(\frac{dr}{R})$$

Proof. Without loss of generality, let x be the origin. First consider the case that S contains $B(x, R + r)$ so that the relevant ratio is $\frac{\text{Vol}(B(x,R+r))}{\text{Vol}(B(x,R))}$. This ratio is simply $(1 + \frac{r}{R})^d$ which is bounded by $(1 + O(\frac{dr}{R}))$ as required.

For any body S, let $f_S(r)$ denote the ratio $\frac{\text{Vol}(B(O,r)\cap S)}{\text{Vol}(B(O,r))}$, that is, the fraction of $B(O, r)$ that lies in S. Note that this is just an average over all length r vectors $\vec{v}$, of the fraction of $\vec{v}$ that lies in S. Whenever S is a convex body containing the origin, this fraction decreases (or stays the same) as the length r of $\vec{v}$ increases. Therefore, it is immediate that $f_S(r)$ is a non-increasing function of r. Now the ratio $\frac{\text{Vol}(B(x,R+r)\cap S)}{\text{Vol}(B(x,R)\cap S)}$ is simply $\frac{f_S(R+r)}{f_S(R)}(1 + \frac{r}{R})^d \leq (1 + \frac{r}{R})^d$, and hence the claim follows. □

Proof of Lemma 4.2: We carry out the proof assuming $C = B(0, 1)$. The proof is based on bounding ratios of volumes of intersections of balls with the set C. Therefore, Observation B.1 implies that the result holds for an arbitrary convex C as well.

Let the closest center to x be c and let R denote the random variable $\|x - c\|$. For any point y in $B(x, r)$, $\|y - c\| \leq R + r$. Thus for $B(x, r)$ to be cut, there must be a center c' such that for some $y \in B(x,r)$, $\|y - c'\|$ is less than $R + r$. That is, $\|x - c'\| \leq R + 2r$. Hence a center c' must fall in $\big(B(x, R+2r) \setminus B(x,r)\big)$. We shall upper bound the probability of this event.

Let R^* be such that in a ball of radius R^* around x, the expected number of centers is 1. Let V^* be the volume of this ball. Thus the probability of any particular center falling in the ball $B(x, R^*)$ is exactly $\frac{1}{t} = \frac{V^*}{\text{Vol}(B(0,1))}$.

Let X denote the bad event "$B(x, R+2r)$ contains a center $c' \neq c$". Let R be as above, let V denote the volume of $B(x,R)$ inside $B(0,1)$ and let V' denote the volume of $\big(B(x, R+2r) \setminus B(x,R)\big)$ inside $B(0,1)$. Note that $\frac{V'}{V}$ is at most $O(dr/R)$.

Thus X is the event that a region of volume V' contains a center. This is more likely when V (and hence V') is large, but V being large itself is unlikely.

So let Y_p denote the event "$V \leq pV^*$". Then $(Y_p)^c$ happens when all of the t centers lie outside a ball of volume pV^*. Each of the t centers falls in this ball with probability $\frac{p}{t}$. Thus the probability of $(Y_p)^c$ is at most $(1 - \frac{p}{t})^t \approx e^{-p}$.

Let Z_p denote the event $Y_p \cap Y_{p-1}^c$. Clearly, $Pr[Z_p] \leq Pr[(Y_{p-1})^c] \leq e^{-(p-1)}$. Conditioned on Z_p, X happens when a region of volume V' contains a center. Since the expected number of centers in a region of volume V' is at most $(\frac{V'}{V^*}) = (\frac{V}{V^*})(\frac{V'}{V}) = O(pdr/R)$, the probability of X given Z_p is at most $O(pdr/R)$. Note that for $p = 1$, we can directly bound $\frac{V'}{V^*}$ by $O(dr/R^*)$ and for $p > 1$, $R \geq R^*$ and hence the above bound is $O(pdr/R^*)$. Thus

$$\begin{aligned} Pr[X] &= \sum_{p=1}^{\infty} Pr[Z_p] Pr[X \mid Z_p] \\ &\leq \sum_{p=1}^{\infty} e^{-(p-1)} p(dr/R^*) \\ &= O(dr/R^*) \sum_{p=1}^{\infty} \frac{p}{e^{p-1}} \end{aligned}$$

The last summation converges to a constant and the proof follows by noting that R^* is a constant times the radius of $B(0,1)$ (or C). □

C Open Questions and Future Work

An immediate next step is to extend the results on histograms for round disributions to high-dimensional Gaussians.

We might be able to construct histograms for Gaussians by dividing the Gaussian into shells with appropriate thickness such that the density in each shell varies only by a constant factor. Now apply Method (1) or Method (2) to each shell separately. The resulting cells will be "flat," but well-rounded in $d - 1$ dimensions. This turns out to be sufficient for the proof in Chawla et al [11].

Chawla et al. examined a very natural type of sanitization in which random spherical Gaussian noise is added to each database point, and these noisy versions of the points form the sanitized database. The variance of the Gaussian is a function of the t-radius of the point, that is, the distance to its t'th nearest neighbor, so that densely clustered points are perturbed less than remote points are. They obtained two useful learning results: a worst-case result for minimum-diameter k-clustering and an average case result for learning mixtures of Gaussians. However, as they explained, proving privacy is problematic for this sanitization technique. As a result, they proposed (and proved private) the following cross-training approach: the data are randomly partitioned into two sets, call them A and B. A histogram sanitization is computed for B and the results are released. In addition, for every real database point $v \in \mathcal{R}$, a perturbation radius for v is chosen according to the diameter of the cell in which v lies in the histogram sanitization of B. Each such v is then perturbed by adding Gaussian noise with variance equal to the square of the perturbation radius, and the noisy version of v is released. Privacy for the points in B follows immediately from the privacy of histogram sanitization and the fact that the only information about B used in computing the sanitization of A is the histogram informaion for B. Privacy for the points in A required further argument.

Once we have good histogram techniques for round distributions we would like to apply cross training. This would allow release of the data in a way that would permit learning the Gaussian mixture.

Finally, we hope to find a way to carry out full recursion and ranodmization of histograms for round distributions, while protecting from c-isolation for small values of c.