A concentration theorem for projections

Sanjoy Dasgupta, Daniel Hsu, and Nakul Verma
University of California, San Diego

Abstract

Suppose the random vector $X \in \mathbb{R}^D$ has mean zero and finite second moments. We show that there is a precise sense in which almost all linear projections of X into $\mathbb{R}^d$ (for $d < D$) look like a scale-mixture of spherical Gaussians—specifically, a mixture of distributions $N(0, \sigma^2 I_d)$ where the σ values follow the same distribution as $\|X\|/\sqrt{D}$. The extent of this effect depends upon the ratio of d to D, and upon a particular coefficient of eccentricity of X's distribution.

1 Introduction

Let $X \in \mathbb{R}^D$ be a random vector with mean zero and finite second moments.[1] In this paper, we examine the behavior of "typical" linear projections of X into $\mathbb{R}^d, d < D$.

The first step is to specify a distribution γ over linear projections from $\mathbb{R}^D$ to $\mathbb{R}^d$. Suppose a $d \times D$ random matrix Θ has entries which are i.i.d. standard normals. It is well-known that with high probability, the rows of this matrix are approximately orthogonal and have length approximately $\sqrt{D}$; for more details and proof techniques see, for instance, Dasgupta and Gupta (2003). The projection we will use is thus:

$$X \;\mapsto\; \frac{1}{\sqrt{D}}\,\Theta X.$$

An alternative distribution over projection matrices would be to take the first d basis vectors of a random orthonormal basis of $\mathbb{R}^D$. The distribution we use is quite close to this, and is more convenient to work with analytically and algorithmically.

For any specific projection θ, let f_θ denote the distribution of the projection of X, a probability measure on $\mathbb{R}^d$. As we shall see, for any (measurable) $S \subset \mathbb{R}^d$, the expected probability mass of that set under a random choice of Θ is

$$\mathbb{E}_\Theta f_\Theta(S) \;=\; \int f_\theta(S)\,\gamma(d\theta) \;=\; \int \nu_\sigma(S)\,\mu(d\sigma) \;\stackrel{\triangle}{=}\; \overline{f}(S)$$

where ν_σ is a shorthand for density of the spherical Gaussian $N(0, \sigma^2 I_d)$, and μ is the distribution of $\|X\|/\sqrt{D}$ (a probability measure on $\mathbb{R}$).[2] In other words, the average projected distribution is a scale-mixture of spherical Gaussians, $\overline{f} = \int \nu_\sigma\,\mu(d\sigma)$.

Given the lack of assumptions on X, the individual projected distributions f_θ could all, for instance, have discrete support. We will show, however, that with high probability over the choice of θ, the distribution f_θ is close to $\overline{f}$ in the following sense: it assigns roughly the same probability mass to every ball $B \subset \mathbb{R}^d$ as does $\overline{f}$. The precise statement is in Theorem 13 but reads approximately like this: for almost all θ,

$$\sup_{\text{balls } B \text{ in } \mathbb{R}^d} |f_\theta(B) - \overline{f}(B)| \;\leq\; \widetilde{O}\left(\frac{\mathrm{ecc}(X) d^2}{D}\right)^{1/4},$$

[1] The spaces we consider are $\mathbb{R}^k$ for various k, under the corresponding Borel σ-algebras. The symbol S always denotes a Borel set, and by "ball" we always mean open ball. We use the notation $\mathbb{P}_Y$ to generically denote the distribution of random variable Y and likewise $\mathbb{E}_Y$ to denote expectation over Y.

[2] Take ν_0 to be a point mass at the origin.

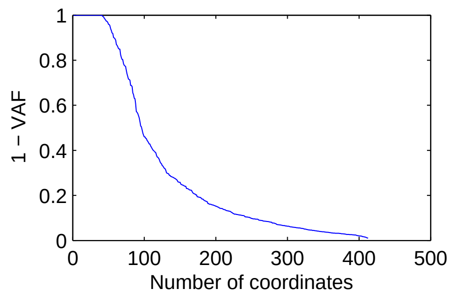

Figure 1: For each coordinate in the MNIST dataset of handwritten "1" digits, this plot shows the fraction of its variance unaccounted for by the best affine combination of the preceding coordinates. The ordering of the coordinates is chosen greedily, by least variance accounted-for.

where $\mathrm{ecc}(X)$ is a specific measure of how eccentric the distribution of X is ($\lambda_{max}/\sigma_\epsilon^2$ in the theorem statement) and the $\widetilde{O}$ notation hides some lower-order terms. We'll see examples of the eccentricity value in the next section.

Implications

Apart from its general insights into data distributions and the enterprise of projection pursuit, we have sought this result for two rather specific reasons.

The first is curiosity about a widely-observed empirical fact, that a Gaussian distribution is often an accurate density model for one-, two-, or three-dimensional data, but very rarely for high-dimensional data. From the birth weight of babies (Clemens and Pagano, 1999) to the calendar dates of hail and thunder occurrences (Hey and Waylen, 1986), many natural phenomena follow a normal distribution. And yet high-dimensional data is unlikely to be Gaussian, in part because of the high degree of independence this demands (after all, a Gaussian is merely a rotation of a distribution with completely independent coordinates). In a typical application, it might be possible to find a few features that are roughly independent, but as more features are added, the dependencies between them will inevitably grow. See Figure 1 for an illustration of this effect.

The result we prove gives a plausible explanation for how high-dimensional distributions that are very far from Gaussian can have low-dimensional projections which are almost Gaussian; and moreover, we quantify the rate at which this effect drops off with increasing d.

Our second motivation has to do with the analysis of statistical procedures, and it also explains the particular notion of closeness in distribution. Many learning algorithms do not look too closely at the data but, rather, look only at low-order statistics of the data distribution restricted to simple geometric regions in space. For instance, consider the k-means clustering algorithm, whose updates depend only on the zero- and first-order statistics of Voronoi regions determined by the current centers. Its behavior on general data sets is hard to characterize, but its performance on data with Gaussian clusters is much better understood (e.g., Dasgupta and Schulman, 2000). Likewise, there has been a recent spate of clustering algorithms which are specifically geared towards data whose clusters look approximately Gaussian in terms of their zero-order statistics on balls in space; and which can be rigorously analyzed in this case (e.g., Dasgupta, 1999; Arora and Kannan, 2001).

One of the motivations of the present paper is to give a *randomized reduction* from data distributions with fairly general clusters to distributions with better-behaved clusters, and thereby generalize results about the performance of learning algorithms which previously applied only to approximately-Gaussian data. This can be thought of in two ways. Either: the initial process of feature selection can be modeled as being itself a sort of random projection, and thus yielding data whose clusters resemble scale-mixtures of Gaussians in their low-order statistics. Or: random projection can be used as an explicit preprocessing step to specifically produce well-behaved data.

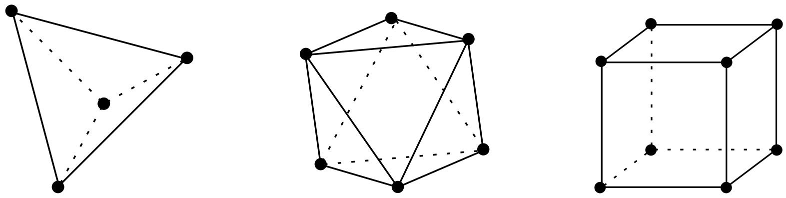

Figure 2: The three-dimensional discrete simplex, cross-polytope, and cube.

Previous work

Our work follows a string of previous results, and draws heavily upon them. The seminal work of Diaconis and Freedman (1984) established this same effect in the case where X has independent coordinates and the projection is to $d = 1$ dimension; in such cases, the projected distribution is close to a single Gaussian (as opposed to a scale-mixture). They also gave asymptotic results (for $d = 1$ and $D \uparrow \infty$) for a variety of other cases, including general distributions in which most pairs of data points are approximately orthogonal and most data points have approximately unit length.

Sudakov (1978), von Weisäcker (1997), Bobkov (2003), and Naor and Romik (2003) have studied the problem for more general distributions of X. These works focus upon $d = 1$ (except for Naor and Romik, who consider general d but define a notion of closeness in distribution which makes the problem essentially one-dimensional), and are based upon various different assumptions on X. We closely follow Bobkov's method, and also use ideas from von Weisäcker and Sudakov. Our result is more general than the union of these earlier works in two ways, both of which are crucial for the algorithmic applications mentioned above: (1) we have no constraints, other than finiteness, on the second moments of X (this particular generalization is straightforward), and (2) we accommodate the case $d > 1$ (this takes some doing).

2 Examples

Our main result says that most linear projections of $X \in \mathbb{R}^D$ are close to $\overline{f}$, a scale-mixture of Gaussians which is determined only by the distribution of $\|X\|/\sqrt{D}$. We will call this latter distribution the *profile* of X.

2.1 Three discrete distributions

We start with three particular examples: uniform distributions over the vertices of a simplex, a cross-polytope, and a cube in $\mathbb{R}^D$ (Figure 2). In each case, almost all linear projections are near-Gaussian.

The simplex

This is perhaps the most surprising of the three examples: a discrete distribution in $\mathbb{R}^D$ whose support is of size just $D + 1$, the smallest possible full-dimensional support.

For concreteness, let the vertices be $\{x_0, x_1, \ldots, x_D\}$, where

$$x_0 = \frac{1 - \sqrt{D+1}}{\sqrt{D}} \cdot 1_D \quad \text{and} \quad x_i = \sqrt{D} e_i \quad \text{for } i = 1, \ldots, D.$$

Here, 1_D is the all-ones vector in $\mathbb{R}^D$ and e_i is the ith coordinate basis vector.

The crucial fact is that each vertex has the same squared distance $D^2/(D + 1)$ to the mean of the distribution and thus the profile μ puts all of its mass at a single point (Figure 3). This means most linear projections will look Gaussian, rather than a more general scale-mixture.

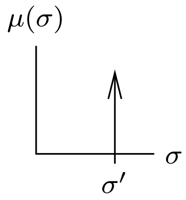

Figure 3: The profile μ for the uniform distributions over the discrete simplex ($\sigma' = D/(D+1)$), cross-polytope and cube ($\sigma' = 1$).

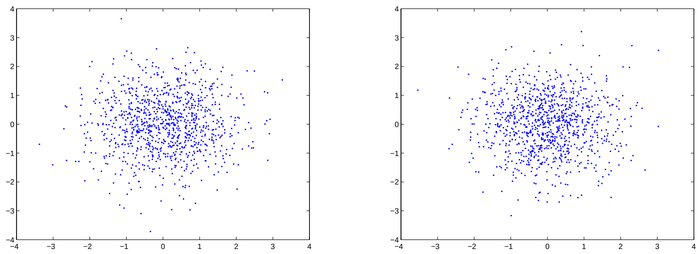

Figure 4: One is the plot of a 2-d projection of the vertices of a 1000-dimensional simplex; the other is the plot of 1001 points sampled from $N(0, I_2)$. Which is which?[1]

Specifically, the covariance matrix of the high-dimensional distribution is $(D/(D+1))\, I_D$, and the coefficient of eccentricity is 1. A direct application of Theorem 13 reveals that most projections are close to a single Gaussian, in the sense that the discrepancy on any ball is $\widetilde{O}((d^2/D)^{1/4})$. Figure 4 illustrates this effect.

Notice that the projected distribution has a discrete support of size at most $D+1$. Yet it is almost Gaussian, in the sense that a random sample from this distribution looks just like a random sample from a Gaussian, if you count the number of points in any ball.

In this specific case, we can tighten the bound on the discrepancy. A random projection of the vertices $x_1, x_2, \ldots, x_D$ (ignore x_0 for now) is distributed as D independent draws from $N(0, I_d)$: the projection of x_i is

$$\frac{1}{\sqrt{D}} \Theta \left(\sqrt{D} e_i \right) = \Theta_i,$$

the ith column of Θ, which has a $N(0, I_d)$ distribution. A standard VC-dimension argument then implies that the fraction of these D projected vertices which fall in any ball is within $O(\sqrt{d(\log D)/D})$ of the probability mass assigned to that ball by $N(0, I_d)$; we use the fact that the class of balls in $\mathbb{R}^d$ has VC dimension $d+1$ (Dudley, 1979). The projection of the remaining vertex x_0 can only increase the error by $O(1/D)$.

The cross-polytope and cube

The uniform distributions over the discrete cross-polytope $\{\pm\sqrt{D}e_i : i = 1, \ldots, D\}$ and discrete cube $\{\pm 1\}^D$ are similar: each has covariance I_D and vertices at squared distance D from the center. Again, the profile has mass only at a single point, 1 in these cases (Figure 3). And again the coefficient of eccentricity is 1, so Theorem 13 shows that most projections are close to a single Gaussian, with the discrepancy on any ball being $O((d^2/D)^{1/4})$.

As with the simplex, a tighter bound for discrepancy can be given in the case of the cross-polytope. We can think of a random projection of the vertices $\sqrt{D}e_i$ as D independent draws from $N(0, I_d)$, call them $\{\theta_1, \ldots, \theta_D\}$; the projections of the remaining D vertices are the negations $\{-\theta_1, \ldots, -\theta_D\}$. With high probability, each half taken

[1]The second plot shows the Gaussian samples.

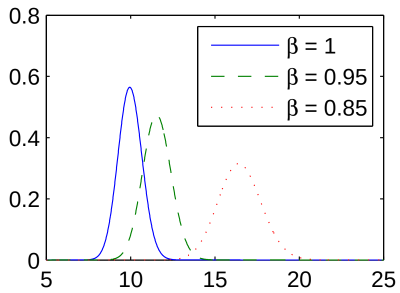

Figure 5: The profile μ for the power-exponential distributions in $\mathbb{R}^{100}$, parameterized by β. The Gaussian has $\beta = 1$, while heavier-tailed distributions have $\beta < 1$.

separately is close to Gaussian in the sense of being within $\epsilon = O(\sqrt{d(\log D)/D})$ on any ball. So the two halves together are within 2ϵ on any ball.

The uniform distribution over the vertices of the cube $\{-1, +1\}^D$ is different from the previous two examples in that it is a product distribution: its coordinates are independent. Such cases permit special arguments (Diaconis and Freedman, 1984) which show that for 1-d projections, the discrepancy from Gaussian is $O(1/\sqrt{D})$ on any interval of the real line.

2.2 Spherically symmetric distributions

Next, we consider the general class of spherically symmetric distributions. This class includes distributions such as the Gaussian, the power-exponential distribution, and Hotelling's T-square distribution. Practitioners in the sciences and engineering often prefer this class over the specific case of the Gaussian because it allows for tails that are "heavier" than that of the Gaussian (e.g. Gales and Olson, 1999; Lindsey and Jones, 2000; and see Figure 5).

If X has a spherically symmetric distribution centered at the origin, then it can also be written in the form $X = UT$, where U is a random vector uniformly drawn from the surface of the D-dimensional sphere S^{D-1}, and T is a scalar random variable whose distribution is the profile of X (scaled appropriately). We've seen that a random projection will preserve the profile (and therefore the heavy tail) of X. This raises an interesting question: can *any* linear projection can alter the tail T of X? No, because a linear projection (with orthonormal rows) $\Phi : \mathbb{R}^D \to \mathbb{R}^d$ merely sends

$$X \ \mapsto \ \Phi X \stackrel{d}{=} (\Phi U)T$$

where ΦU is uniformly distributed over S^{d-1}, so the tail T is preserved exactly.

2.3 Other examples

OCR, text, and speech data

Next, we look at low-dimensional projections of three data sets well-known in the machine learning literature: the MNIST database of handwritten digits, the Reuters database of news articles and Mel-frequency cepstral coefficients of the TIMIT data set. Restricting attention to just one cluster from each dataset, we note that projecting the data onto its top principal components suggests the existence of non-Gaussian projections, even though most random projections still look like scale-mixtures of Gaussians (see Figure 6).

Clustered data

A distribution with well-separated clusters is unlikely to look like a single scale-mixture upon projection; indeed, its high coefficient of eccentricity renders the bound on the discrepancy for a particular ball effectively meaningless

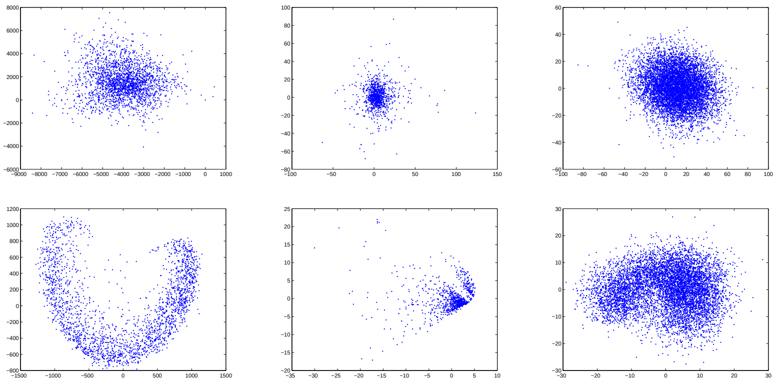

Figure 6: Above: "Typical" two-dimensional projections of handwritten 1's images, word counts of Reuters news articles about Canada, and Mel-frequency cepstral coefficients of the spoken phoneme 's'. Below: The corresponding two-dimensional PCA projections.

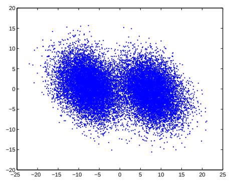

Figure 7: A typical linear projection of a two-cluster (highly eccentric) distribution.

(Figure 7). In many such cases, the Johnson-Lindenstrauss theorem (1984) dictates that a typical projection will keep the clusters apart, and the result of this paper can more usefully be applied to the individual clusters.

3 Proof

3.1 Preliminaries

We assume the random vector $X \in \mathbb{R}^D$ has mean zero and finite second moments. Let μ denote the distribution of $\|X\|/\sqrt{D}$. Writing ν_σ for the density of the d-dimensional spherical Gaussian $N(0, \sigma^2 I_d)$, let $\overline{f}$ be the scale-mixture

$$\overline{f} = \int \nu_\sigma \; \mu(d\sigma).$$

For any fixed $d \times D$ matrix θ, let f_θ denote the distribution of the projection $\frac{1}{\sqrt{D}}\theta X$. And let $f_\theta(B)$ (which we'll sometimes write $f(\theta, B)$) be the probability mass that f_θ assigns to an open ball $B \subset \mathbb{R}^d$.

We will consistently use $\|\cdot\|$ to denote Euclidean norm:

$$\|A\|^2 = \begin{cases} \sum_i A_i^2 & \text{if } A \text{ is a vector} \\ \sum_{i,j} A_{ij}^2 & \text{if } A \text{ is a matrix} \end{cases}$$

One last piece of jargon: a function $g : \mathbb{R}^N \to \mathbb{R}$ is C-Lipschitz if for all $x, y \in \mathbb{R}^N$,

$$|g(x) - g(y)| \;\le\; C\|x - y\|.$$

3.2 Overview

The first part of our proof, following Bobkov (2003), rests crucially upon recent results on concentration of measure, so we start with a brief overview of these.

The familiar Chernoff and Hoeffding bounds say that the *average* of n i.i.d. random variables $X_1, X_2, \ldots, X_n$ is tightly concentrated around its mean, provided the X_i are bounded and n is sufficiently large. But what is so special about the average; what about other functions $g(X_1, \ldots, X_n)$? It turns out that the relevant feature of the average yielding tight concentration is that it is *Lipschitz*.

The following concentration bound applies to *any* Lipschitz function of i.i.d. normal random variables. One good reference for this is Ledoux (2001, page 41, 2.35).

Theorem 1 (Concentration bound) *Let γ_N denote the distribution $N(0, I_N)$. Suppose the function $g : \mathbb{R}^N \to \mathbb{R}$ is C-Lipschitz. Then*

$$\gamma_N\{z : g(z) \geq \mathbb{E}[g] + r\} \;\leq\; e^{-r^2/2C^2}.$$

(To bound the probability that $g(z) \leq \mathbb{E}[g] - r$, use $-g$, which is also C-Lipschitz.)

In our case, the random variable with a $N(0, I_N)$ distribution is the matrix Θ (so $N = dD$). Here is an outline of our argument.

1. Fix a ball $B \subset \mathbb{R}^d$. The first observation is that $\mathbb{E}_\Theta f_\Theta(B) = \overline{f}(B)$: *in expectation,* f_Θ assigns the desired probability mass to B.
2. We would like to conclude that $f_\theta(B)$ is very close to $\overline{f}(B)$ for typical θ, but this doesn't immediately follow from the concentration bound since $f(\theta, B)$ may not be Lipschitz in θ.
3. So instead, as was done for one-dimensional projections in Bobkov (2003), we introduce a smoothed version of f_θ. We call it $\widetilde{f}_\theta$, and we show that *it* is concentrated around its expected value.
4. Then we need to relate $\widetilde{f}_\theta$ to f_θ; this is the main technical portion of the proof.
5. So for a fixed ball B, for almost all θ, $f_\theta(B) \approx \overline{f}(B)$. But we want to show that $f_\theta(B) \approx \overline{f}(B)$ for *all* balls $B \subset \mathbb{R}^d$ simultaneously. To do so, we explicitly construct a finite set of balls $B_1, \ldots, B_M$ with the property that if f_θ is close to $\overline{f}$ on these balls, then it is close to $\overline{f}$ on all balls. We finish by taking a union bound over the B_i.

3.3 The expectation of f_Θ

Let Θ be a $d \times D$ matrix with i.i.d. $N(0, 1)$ entries; we will denote this distribution over matrices by γ. Recall f_θ is the distribution of the projected random variable

$$X \;\mapsto\; \frac{1}{\sqrt{D}}\,\theta X.$$

Lemma 2 *Fix any $x \in \mathbb{R}^D$. The distribution of $\frac{1}{\sqrt{D}}\Theta x$ (for Θ chosen at random according to γ) is $N(0, \frac{\|x\|^2}{D} I_d)$.*

Proof. Any linear transformation of a Gaussian is Gaussian, so $\frac{1}{\sqrt{D}}\Theta x$ has a Gaussian distribution. Its mean and second moments are easily checked. ∎

For any ball $B \subset \mathbb{R}^d$, define $f_\theta(B)$ to be the probability mass of B under projection θ, that is,

$$f_\theta(B) \;\triangleq\; f(\theta, B) \;\triangleq\; \mathbb{P}_X\left[\frac{\theta X}{\sqrt{D}} \in B\right] \;=\; \mathbb{E}_X\left[\mathbb{1}\left(\frac{\theta X}{\sqrt{D}} \in B\right)\right],$$

where $\mathbb{1}(\cdot)$ is the indicator function.

Lemma 3 *Fix any ball $B \subset \mathbb{R}^d$. Then $\mathbb{E}_\Theta f_\Theta(B) = \overline{f}(B)$.*

Proof. Applying Fubini's theorem, and using P for the distribution of X,

$$\mathbb{E}_\Theta f(\Theta, B) \;=\; \int\int \mathbb{1}\left(\frac{\theta x}{\sqrt{D}} \in B\right) P(dx)\, \gamma(d\theta) \;=\; \int\int \mathbb{1}\left(\frac{\theta x}{\sqrt{D}} \in B\right) \gamma(d\theta)\, P(dx).$$

By the previous lemma, the inner expectation is $\nu_{\|x\|/\sqrt{D}}(B)$, a function only of $\|x\|$. Thus

$$\mathbb{E}_\Theta f(\Theta, B) \;=\; \int \nu_{\|x\|/\sqrt{D}}(B)\, P(dx) \;=\; \int \nu_\sigma(B)\, \mu(d\sigma)$$

under the change of variable $\sigma = \|x\|/\sqrt{D}$. ∎

Fix some ball B. We can't directly apply the concentration bound to show that $f(\cdot, B)$ is tightly concentrated around its expectation because this function may not be Lipschitz in θ. To see this, suppose that X is uniformly distributed over k support points, and that under projection θ, exactly one of these support points falls in B. Then $f(\theta, B) = 1/k$. However, if this projected point is right at the boundary of B, even a tiny perturbation $\theta \to \theta'$ could cause it to fall outside B, whereupon $f(\theta', B) = 0$. So $|f(\theta, B) - f(\theta', B)|$ cannot be upper-bounded in terms of $\|\theta - \theta'\|$.

3.4 A smoothed version of f_θ

Fix a ball $B \subset \mathbb{R}^d$ and a projection θ. Consider an experiment in which a point X is randomly drawn and is assigned a score of 1 if its projection happens to fall in B; and a score of 0 otherwise. Then $f(\theta, B) = \mathbb{E}_X\left[\mathbb{1}(\frac{\theta X}{\sqrt{D}} \in B)\right]$ is the expected score achieved. To get a smoother version of this function, we will assign a fractional score if the projected point doesn't fall exactly in B but is nonetheless close by.

For some value $\Delta > 0$ to be determined, define the function $h_B : \mathbb{R}^d \to [0, 1]$ as follows:

$$h_B(z) = \begin{cases} 1 & \text{if } d(z, B) = 0 \\ 1 - (d(z, B)/\Delta) & \text{if } 0 < d(z, B) \le \Delta \\ 0 & \text{if } d(z, B) > \Delta \end{cases}$$

where $d(z, B) = \inf_{y \in B} \|y - z\|$ is the distance from point z to ball B. Clearly h_B is $(1/\Delta)$-Lipschitz.

Now, define the smoothed function $\widetilde{f}(\theta, B)$ as

$$\widetilde{f}(\theta, B) \;=\; \mathbb{E}_X\left[h_B\left(\frac{\theta X}{\sqrt{D}}\right)\right].$$

A one-dimensional version of the following lemma was used by Sudakov (1978).

Claim 4 *Fix a ball $B \subset \mathbb{R}^d$. Then $\widetilde{f}(\cdot, B)$ is $\sqrt{\lambda_{max}/D\Delta^2}$-Lipschitz, where λ_{max} is the largest eigenvalue of the covariance matrix $\mathbb{E}_X\left[XX^T\right]$.*

Proof. For any θ, θ',

$$\begin{aligned}
|\widetilde{f}(\theta, B) - \widetilde{f}(\theta', B)| &= \left| \mathbb{E}_X \left[h_B\left(\frac{\theta X}{\sqrt{D}}\right) - h_B\left(\frac{\theta' X}{\sqrt{D}}\right) \right] \right| \\
&\leq \mathbb{E}_X \left| h_B\left(\frac{\theta X}{\sqrt{D}}\right) - h_B\left(\frac{\theta' X}{\sqrt{D}}\right) \right| \\
&\leq \frac{1}{\Delta} \cdot \mathbb{E}_X \left\| \frac{\theta X}{\sqrt{D}} - \frac{\theta' X}{\sqrt{D}} \right\| \quad (h_B \text{ is } (1/\Delta)\text{-Lipschitz}) \\
&= \frac{1}{\Delta\sqrt{D}} \cdot \mathbb{E}_X \|(\theta - \theta')X\| \\
&\leq \frac{1}{\Delta\sqrt{D}} \cdot \sqrt{\mathbb{E}_X \|(\theta - \theta')X\|^2} \\
&\leq \frac{1}{\Delta\sqrt{D}} \cdot \sqrt{\lambda_{max}\|\theta - \theta'\|^2} \;=\; \frac{\sqrt{\lambda_{max}}}{\Delta\sqrt{D}} \cdot \|\theta - \theta'\|,
\end{aligned}$$

as claimed. ▮

The concentration bound (Theorem 1) gives

Claim 5 *Fix any ball $B \subset \mathbb{R}^d$, and any $0 < \epsilon < 1$. When Θ is picked at random according to γ,*

$$\mathbb{P}_\Theta \left[|\widetilde{f}(\Theta, B) - \mathbb{E}_\Theta \widetilde{f}(\Theta, B)| \geq \epsilon \right] \;\leq\; 2e^{-\epsilon^2\Delta^2 D/2\lambda_{max}}.$$

The problem is that we are interested in the original functions f_θ rather than their smoothed counterparts. To relate the two, we use:

$$f_\theta(B) \;\leq\; \widetilde{f}_\theta(B) \;\leq\; f_\theta(B_\Delta)$$

where B_Δ is a shorthand for the Minkowski sum $B + B(0, \Delta)$ (to put it simply, grow the radius of B by Δ). By abuse of notation, let $B_{-\Delta}$ be the ball with the same center as B but whose radius is smaller by Δ (this might be the empty set). Then:

Corollary 6 *Fix any ball $B \subset \mathbb{R}^d$, and any $0 < \epsilon < 1$. When Θ is picked at random according to γ,*

$$\mathbb{P}_\Theta \left[\overline{f}(B_{-\Delta}) - \epsilon \leq f(\Theta, B) \leq \overline{f}(B_\Delta) + \epsilon \right] \geq \;1 - 2e^{-\epsilon^2\Delta^2 D/2\lambda_{max}}.$$

It is necessary, therefore, to relate $\overline{f}(B)$ to $\overline{f}(B_\Delta)$.

3.5 Relating the probability mass of B to that of B_Δ

Recall $\overline{f}$ is the scale-mixture

$$\overline{f} \;=\; \int \nu_\sigma \, \mu(d\sigma)$$

where ν_σ is the spherical Gaussian $N(0, \sigma^2 I_d)$. As a first step towards relating $\overline{f}(B)$ and $\overline{f}(B_\Delta)$, we relate $\nu_\sigma(B)$ and $\nu_\sigma(B_\Delta)$.

If Δ is small enough, then $\nu_\sigma(B_\Delta)$ will not be too much larger than $\nu_\sigma(B)$. But how small exactly does Δ need to be? There are two effects that come into play.

1. The Gaussian ν_σ has significant mass at radius $\sigma\sqrt{d}$. So it is important to deal properly with balls of radius approximately $\sigma\sqrt{d}$.

2. If B has radius r then $\text{vol}(B_\Delta) \;=\; \text{vol}(B) \cdot \left(\frac{r+\Delta}{r}\right)^d$. Therefore, if we want the probability mass of B_Δ to be at most $(1 + \epsilon)$ times that of B, we need $\Delta = O(r\epsilon/d)$.

These two considerations tell us that we need $\Delta \leq \epsilon\sigma/\sqrt{d}$. The second also implies that any value of Δ we choose will only work for balls of radius $> \Delta d$. To get around this, we make sure that Δ is sufficiently small that any ball of radius less than Δd has insignificant (less than ϵ) probability mass.

The following key technical lemma is proved in the appendix. Notice that the bound on Δ is roughly $\epsilon\sigma/\sqrt{d}$.

Lemma 7 *Pick any $0 < \epsilon < 1$ and any $\sigma > 0$. If*

$$\Delta \leq \frac{\sigma}{\sqrt{d}} \cdot \ln\left(1 + \frac{\epsilon}{4}\right) \cdot \frac{1}{2 + \sqrt{\frac{2}{d}\ln\frac{4}{\epsilon}}},$$

then $\nu_\sigma(B_\Delta) \leq \nu_\sigma(B) + \epsilon$ for any ball B.

Finally, we consider the scale-mixture rather than just individual components ν_σ.

Corollary 8 *Pick any $0 < \epsilon < 1$ and a threshold $\sigma_\epsilon > 0$ such that $\mu\{\sigma : \sigma < \sigma_\epsilon\} \leq \epsilon$. If*

$$\Delta \leq \frac{\sigma_\epsilon}{\sqrt{d}} \cdot \ln\left(1 + \frac{\epsilon}{4}\right) \cdot \frac{1}{2 + \sqrt{\frac{2}{d}\ln\frac{4}{\epsilon}}},$$

then $\overline{f}(B_\Delta) \leq \overline{f}(B) + 2\epsilon$ for any ball B.

Proof. We can rewrite $\overline{f}$ as $\mathbb{E}_\sigma[\nu_\sigma]$, where the expectation is taken over σ drawn according to μ.

$$\begin{aligned}
\overline{f}(B_\Delta) - \overline{f}(B) &= \mathbb{E}_\sigma[\nu_\sigma(B_\Delta)] - \mathbb{E}_\sigma[\nu_\sigma(B)] \\
&\leq \mathbb{E}_\sigma[\nu_\sigma(B_\Delta) - \nu_\sigma(B) \mid \sigma \geq \sigma_\epsilon] + \mathbb{P}_\sigma(\sigma < \sigma_\epsilon) \\
&\leq 2\epsilon,
\end{aligned}$$

as claimed. ∎

At this stage, we have shown that for any given B, almost all projections θ have $f_\theta(B) \approx \overline{f}(B)$. Putting together Corollaries 6 and 8, we get:

Theorem 9 *Pick any $0 < \epsilon < 1/2$ and $\sigma_\epsilon > 0$ such that $\mu\{\sigma : \sigma < \sigma_\epsilon\} \leq \epsilon$. Pick any ball $B \subset \mathbb{R}^d$. Then*

$$\mathbb{P}_\Theta\left[|f_\Theta(B) - \overline{f}(B)| > \epsilon\right] \leq \exp\left\{-\Omega\left(\frac{\epsilon^4 D}{d} \cdot \frac{\sigma_\epsilon^2}{\lambda_{max}} \cdot \frac{1}{\ln 1/\epsilon}\right)\right\}$$

It remains to prove this for all balls *simultaneously*.

3.6 Uniform convergence for all balls

We follow a standard method for proving uniform convergence: we carefully choose a small finite set of balls $B_0, \ldots, B_M \subset \mathbb{R}^d$ such that if the concentration property (Theorem 9) holds on these B_i's, then it holds for *all* balls in $\mathbb{R}^d$. Specifically, our B_i's have the following property:

> For any ball $B \subset \mathbb{R}^d$ there exist B_i, B_j such that $B_i \subset B \subset B_j$ and $\overline{f}(B_j) - \overline{f}(B_i) \leq 2\epsilon$.

It follows that if $f_\theta(B_i) \approx \overline{f}(B_i)$ for the finite set of balls B_i, then $f_\theta(B) \approx \overline{f}(B)$ for all balls $B \subset \mathbb{R}^d$.

Actually, things are just slightly more complicated than this. There is no finite collection of B_i's that can possibly satisfy this criterion given that the balls B can be arbitrarily far from the origin. The saving grace is that almost all the probability mass of $\overline{f}$ lies within $B(0, c\sqrt{d})$ for some suitable constant c, and so we need only make sure that for all balls $B \subset \mathbb{R}^d$, there exist $B_i \subset B_j$ such that:

- $B_i \subset B$ and $B \cap B(0, c\sqrt{d}) \subset B_j$.

- $\overline{f}(B_j) - \overline{f}(B_i) \leq 2\epsilon$.

The balls B_i will be centered at grid points in $[-C\sqrt{d}, C\sqrt{d}]^d$, for some $C > c$.

Here's the construction of $B_1, \ldots, B_M$, for some parameters C, ϵ_o to be determined:

1. Place a grid with resolution (spacing) $2\epsilon_o$ on $[-C\sqrt{d}, C\sqrt{d}]^d$.
2. At each point on the grid, create a set of balls centered at that point, with radii $\epsilon_o\sqrt{d}, 2\epsilon_o\sqrt{d}, \ldots, (2C + 2\epsilon_o)\sqrt{d}$.

The total number of balls is then $M = \left(\frac{C\sqrt{d}}{\epsilon_o} + 1\right)^d \cdot \frac{2C+2\epsilon_o}{\epsilon_o}$. For good measure, add in two final balls: $\emptyset$ and $\mathbb{R}^d$.

The first step is to confirm that most of $\overline{f}$ indeed lies close to the origin.

Lemma 10 *Suppose $c \geq \sqrt{\lambda_{avg}/\epsilon}$, where λ_{avg} is the average eigenvalue of $\mathbb{E}[XX^T]$. Then $\overline{f}(B(0, c\sqrt{d})) \geq 1 - \epsilon$.*

Proof. Let Z be a random draw from $\overline{f} = \int \nu_\sigma \mu(d\sigma)$.

$$\mathbb{E}\|Z\|^2 \;=\; \int \sigma^2 d\, \mu(d\sigma) \;=\; \frac{d}{D}\mathbb{E}\|X\|^2 \;=\; d\lambda_{avg}.$$

Since $c^2 \geq \lambda_{avg}/\epsilon$, by Markov's inequality

$$\mathbb{P}\left[\|Z\| \geq c\sqrt{d}\right] \;\leq\; \frac{\mathbb{E}\|Z\|^2}{c^2 d} \;\leq\; \epsilon$$

and so $\overline{f}(B(0, c\sqrt{d})) = 1 - \mathbb{P}[\|Z\| \geq c\sqrt{d}] \geq 1 - \epsilon$. ∎

Now we show that all balls $B \subset \mathbb{R}^d$ centered in $B(0, C\sqrt{d})$ are well approximated by the B_i.

Lemma 11 *Suppose $c \geq \sqrt{\lambda_{avg}/\epsilon}$ and $\epsilon_o \leq \Delta/(4\sqrt{d})$. Pick any ball $B \subset \mathbb{R}^d$ centered in $B(0, C\sqrt{d})$. Then there exist B_i, B_j such that*

$$B_i \subset B \subset B_j$$

and $\overline{f}(B_j) - \overline{f}(B_i) \leq 2\epsilon$.

Proof. Say $B = B(x, r)$, with $x \in B(0, C\sqrt{d})$. There are two cases to consider.

Case 1: $r \leq 2C\sqrt{d}$.

By construction, there is a grid point $\tilde{x}$ which differs from x by at most ϵ_o on each coordinate, so $\|x - \tilde{x}\| \leq \epsilon_o\sqrt{d}$. Let $B^{in} = B(\tilde{x}, r_1)$ be the largest of the B_i's centered at $\tilde{x}$ and *contained inside* B (if necessary take $r_1 = 0$ so that $B^{in} = \emptyset$). Likewise let $B^{out} = B(\tilde{x}, r_2)$ be the smallest of the B_i's centered at $\tilde{x}$ and *containing* B. Again by construction,

$$\begin{aligned} r_1 &\;\geq\; r - \|x - \tilde{x}\| - \epsilon_o\sqrt{d} \;\geq\; r - 2\epsilon_o\sqrt{d} \\ r_2 &\;\leq\; r + \|x - \tilde{x}\| + \epsilon_o\sqrt{d} \;\leq\; r + 2\epsilon_o\sqrt{d} \end{aligned}$$

Since $\epsilon_o \leq \Delta/(4\sqrt{d})$, we have $B^{out} \subset B^{in}_{\Delta}$ and thus, by corollary 8, $\overline{f}(B^{out}) - \overline{f}(B^{in}) \leq 2\epsilon$.

Case 2: $r > 2C\sqrt{d}$.

In this case, B is contained in $\mathbb{R}^d$ and contains the ball B_i which is centered at the origin and has radius $C\sqrt{d}$. The previous lemma shows that $\overline{f}(B_i) \geq 1 - \epsilon$. ∎

Finally, we handle balls centered outside $B(0, C\sqrt{d})$.

Lemma 12 *Suppose that* $c \geq \sqrt{\lambda_{avg}/\epsilon}$, $\epsilon_o \leq \Delta/(5\sqrt{d})$, *and* $C \geq c + (c^2/2\epsilon_o)$. *For any ball* B *centered outside* $B(0, C\sqrt{d})$, *there exist* $B_i \subset B_j$ *such that*

$$B_i \subset B \quad \textit{and} \quad B \cap B(0, c\sqrt{d}) \subset B_j$$

and $\overline{f}(B_j) - \overline{f}(B_i) \leq 2\epsilon$.

Proof. Suppose $B = B(x, r)$ for $\|x\| > C\sqrt{d}$. If B either contains $B(0, c\sqrt{d})$ or doesn't intersect it at all, then we're done: in the first case, take $B_i = B(0, c\sqrt{d}), B_j = \mathbb{R}^d$, and in the second case, take $B_i = B_j = \emptyset$.

So assume B intersects $B(0, c\sqrt{d})$ but doesn't contain it. This means $-c\sqrt{d} < \|x\| - r < c\sqrt{d}$. We'll approximate B by a ball $B(x', r')$ centered on the surface of $B(0, C\sqrt{d})$. To this end, let $x' = \frac{C\sqrt{d}}{\|x\|}x$ and $r' = r - \|x - x'\|$. In particular, $\|x'\| = C\sqrt{d}$ and $r' = r - \|x\| + \|x'\| \in ((C - c)\sqrt{d}, (C + c)\sqrt{d})$.

To see that $B(x', r')$ is a close approximation to $B(x, r)$, notice first that $B(x', r') \subset B(x, r)$ because for any $z \in B(x', r')$,

$$\|z - x\| \;\leq\; \|z - x'\| + \|x - x'\| \;\leq\; r' + \|x - x'\| \;=\; r.$$

In the other direction, $B(x, r) \cap B(0, c\sqrt{d}) \subset B(x', r' + \epsilon_o\sqrt{d})$. To see this, pick any $z \in B(x, r) \cap B(0, c\sqrt{d})$, and examine the components of $z - x'$ in the direction of x and perpendicular to x. Letting $\hat{x}$ be the unit vector in the direction of x,

$$(x' - z) \cdot \hat{x} = (x' - x) \cdot \hat{x} + (x - z) \cdot \hat{x} \leq \|x'\| - \|x\| + \|x - z\| \leq r' - r + r = r'.$$

And the component of $z - x'$ perpendicular to $\hat{x}$ is the same as the component of z perpendicular to $\hat{x}$, which is in turn at most $\|z\| \leq c\sqrt{d}$. Combining both components, we get

$$\|z - x'\| \;\leq\; \sqrt{(r')^2 + c^2 d} \;=\; r'\sqrt{1 + \frac{c^2 d}{(r')^2}} \;\leq\; r'\left(1 + \frac{c^2 d}{2(r')^2}\right) \;\leq\; r' + \frac{c^2 d}{2(C - c)\sqrt{d}} \;\leq\; r' + \epsilon_o\sqrt{d}.$$

Since $\|x'\| \leq C\sqrt{d}$ and $r' \leq (C + c)\sqrt{d} < 2C\sqrt{d}$, we can proceed as in case 1 of the previous lemma. This gives us a grid point $\tilde{x}$ and radii r_1, r_2 which are multiples of $\epsilon_o\sqrt{d}$ such that

$$B(\tilde{x}, r_1) \subset B(x', r') \subset B(x', r' + \epsilon_o\sqrt{d}) \subset B(\tilde{x}, r_2)$$

and $r_2 - r_1 \leq 5\epsilon_o\sqrt{d}$. Since $\epsilon_o \leq \Delta/(5\sqrt{d})$, we have from Corollary 8 that $\overline{f}(B(\tilde{x}, r_2)) - \overline{f}(B(\tilde{x}, r_1)) \leq 2\epsilon$. ∎

All the pieces are now in place. We apply Theorem 9 to each of the balls B_i, taking an union bound over the error probabilities. And then, using the last three lemmas, we conclude uniform convergence for all balls $B \subset \mathbb{R}^d$.

Theorem 13 *Suppose* X *has mean zero and finite second moments. Define* $f_\theta, \mu, \overline{f}$ *as above. Pick any* $0 < \epsilon < 1/2$ *and* $\sigma_\epsilon > 0$ *such that* $\mu\{\sigma : \sigma < \sigma_\epsilon\} \leq \epsilon$. *Then*

$$\mathbb{P}_\Theta\left[\sup_{\textit{balls } B \subset \mathbb{R}^d} |f_\Theta(B) - \overline{f}(B)| > \epsilon\right] \;\leq\; \left(O\left(\frac{d^2}{\epsilon^3}\frac{\lambda_{avg}}{\sigma_\epsilon^2}\ln\frac{1}{\epsilon}\right)\right)^d \exp\left\{-\Omega\left(\frac{\epsilon^4 D}{d} \cdot \frac{\sigma_\epsilon^2}{\lambda_{max}} \cdot \frac{1}{\ln 1/\epsilon}\right)\right\}$$

where λ_{avg} *and* λ_{max} *are the average and maximum eigenvalues of the covariance* $\mathbb{E}[XX^T]$.

Acknowledgements

We are grateful to Yoav Freund for suggesting this problem, to the anonymous reviewers for their extremely helpful feedback, and to the NSF for grant IIS-0347646.

References

S. Arora and R. Kannan (2001). Learning mixtures of arbitrary Gaussians. *ACM Symposium on Theory of Computing*, 247–257.

S. Bobkov (2003). On concentration of distributions of random weighted sums. *Annals of Probability*, 31(1):195–215.

T. Clemons and M. Pagano (1999). Are babies normal? *The American Statistician*, 53(4):298–302.

S. Dasgupta (1999). Learning mixtures of Gaussians. *IEEE Symposium on Foundations of Computer Science*, 634–644.

S. Dasgupta and A. Gupta (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65.

S. Dasgupta and L. Schulman (2000). A two-round variant of EM for Gaussian mixtures. *Sixteenth Conference on Uncertainty in Artificial Intelligence*, 152–159.

P. Diaconis and D. Freedman (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12:793–815.

R.M. Dudley (1979). Balls in R^k do not cut all subsets of $k+2$ points. *Advances in Mathematics*, 31:306-308.

M. Hey and P.Waylen (1986). Probabilities of daily hail and thunder occurrences. *Journal of Applied Meteorology*, 26(8):1014–1020.

M.J.F. Gales and P.A. Olsen (1999). Tail distribution modelling using the richter and power exponential distributions. *Eurospeech*, 1507–1510.

W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

M. Ledoux (2001). *The concentration of measure phenomenon*. American Mathematical Society.

J.K. Lindsey and B. Jones (2000). Modeling pharmacokinetic data using heavy-tailed multivariate distributions. *Journal of Biopharmaceutical Statistics*, 10(3):369–381.

A. Naor and D. Romik (2003). Projecting the surface measure of the sphere of l_p^n. *Annales de l'Institut Henri Poincaré (B), Probability and Statistics*, 39:241–261.

V.N. Sudakov (1978). Typical distributions of linear functionals in finite-dimensional spaces of high dimension. *Soviet Math. Doklady*, 19(6):1578–1582.

H. von Weizsäcker (1997). Sudakov's typical marginals, random linear functionals and a conditional limit theorem. *Probability Theory and Related Fields*, 107:313–324.

4 Appendix: proof of Lemma 7

Let $B = B(x,r)$, so that $B_\Delta = B(x, r+\Delta)$, and fix any $\sigma > 0$. We will compare $\nu_\sigma(B)$ with $\nu_\sigma(B_\Delta)$. First, note that

$$\begin{aligned}
\nu_\sigma(B_\Delta) &= \int_{B(0,r+\Delta)} \nu_\sigma(x+z)\,dz \\
&\leq \int_{\substack{z\in B(0,r+\Delta)\\ \|x+z\|\leq C}} \nu_\sigma(x+z)\,dz \;+\; \nu_\sigma(\mathbb{R}^d \setminus B(0,C)) \\
&= \int_{\substack{y\in B(0,r)\\ \|x+y+(\Delta/r)y\|\leq C}} \nu_\sigma\left(x+y+\frac{\Delta y}{r}\right)\left(\frac{r+\Delta}{r}\right)^d dy \;+\; \nu_\sigma(\mathbb{R}^d \setminus B(0,C))
\end{aligned}$$

under the change of variable $y = z \cdot \frac{r}{r+\Delta}$. This integral can be upper-bounded in terms of $\nu_\sigma(B)$.

Lemma 14 *If $\|y\| \leq r$ and $\|x+y+\frac{\Delta}{r}y\| \leq C$ then $\nu_\sigma(x+y+\frac{\Delta}{r}y) \;\leq\; \nu_\sigma(x+y)\cdot e^{(C+\Delta)\Delta/\sigma^2}$.*

Proof. The length bounds tell us that $\|\frac{\Delta}{r}y\| \leq \Delta$ and that $\|x+y\| \leq \|x+y+\frac{\Delta}{r}y\| + \|\frac{\Delta}{r}y\| \leq C+\Delta$. Therefore,

$$\begin{aligned}
\|x+y+\tfrac{\Delta}{r}y\|^2 &\geq \|x+y\|^2 - 2\|x+y\|\|\tfrac{\Delta}{r}y\| \\
&\geq \|x+y\|^2 - 2(C+\Delta)\Delta
\end{aligned}$$

whereupon

$$\begin{aligned}
\nu_\sigma(x+y+\tfrac{\Delta}{r}y) &= \frac{1}{\sigma^d(2\pi)^{d/2}} e^{-\|x+y+(\Delta/r)y\|^2/2\sigma^2} \\
&\leq \frac{1}{\sigma^d(2\pi)^{d/2}} e^{-\|x+y\|^2/2\sigma^2} e^{(C+\Delta)\Delta/\sigma^2},
\end{aligned}$$

as claimed. ∎

We also need to bound the term $\nu_\sigma(\mathbb{R}^d \setminus B(0,C))$.

Lemma 15 *If the random variable Z is distributed as $N(0, I_d)$, then for any $c > 0$,*

$$\mathbb{P}[\|Z\| \geq c\sqrt{d}] \;\leq\; e^{-(c-1)^2 d/2}.$$

Proof. This is just a chi-squared tail bound, but can also be quickly derived from Theorem 1, since $\|\cdot\|$ is 1-Lipschitz and $\mathbb{E}\|Z\| \leq \sqrt{\mathbb{E}\|Z\|^2} = \sqrt{d}$. ∎

Lemma 16 *For any $0 < \epsilon < 1$, choose*

$$\Delta \;\leq\; \frac{\sigma}{\sqrt{d}} \cdot \ln(1+\epsilon) \cdot \frac{1}{1 + \frac{1}{d} + \sqrt{\frac{2}{d}\ln\frac{1}{\epsilon}}}.$$

Then for all $x \in \mathbb{R}^d$ and all $r > 0$, we have $\nu_\sigma(B(x,r+\Delta)) \leq (1+\epsilon)\left(\frac{r+\Delta}{r}\right)^d \nu_\sigma(B(x,r)) + \epsilon$.

Proof. Continuing from our earlier expression for $\nu_\sigma(B(x, r+\Delta))$,

$$\begin{aligned}
\nu_\sigma(B(x,r+\Delta)) &\leq \int_{\substack{y\in B(0,r)\\ \|x+y+(\Delta/r)y\|\leq C}} \nu_\sigma\left(x+y+\frac{\Delta y}{r}\right)\left(\frac{r+\Delta}{r}\right)^d dy \;+\; \nu_\sigma(\mathbb{R}^d \setminus B(0,C)) \\
&\leq e^{(C+\Delta)\Delta/\sigma^2}\left(\frac{r+\Delta}{r}\right)^d \int_{\substack{y\in B(0,r)\\ \|x+y+(\Delta/r)y\|\leq C}} \nu_\sigma(x+y)\,dy \;+\; \nu_\sigma(\mathbb{R}^d \setminus B(0,C)) \\
&\leq e^{(C+\Delta)\Delta/\sigma^2}\left(\frac{r+\Delta}{r}\right)^d \nu_\sigma(B(x,r)) \;+\; e^{-((C/\sigma\sqrt{d})-1)^2 d/2},
\end{aligned}$$

where the last two inequalities follow from Lemmas 14 and 15, respectively. Now set $C = (1+\sqrt{\frac{2}{d}\ln\frac{1}{\epsilon}})\sigma\sqrt{d}$. ▮

This bound is reasonable if r is large. On the other hand, if r is tiny, then $\nu_\sigma(B(x, r+\Delta))$ is less than ϵ and the whole point is moot anyway. We now establish this formally.

Lemma 17 *For any $r > 0$ and $\sigma > 0$,*

$$\nu_\sigma(B(0,r)) \;\leq\; \left(\frac{2r}{\sigma\sqrt{d}}\right)^d.$$

Proof. This is surely known, but as we haven't been able to find a reference, we are including a quick proof. Without loss of generality $\sigma = 1$, so ν_σ is $N(0, I_d)$. The squared length of a random point from this distribution has chi-squared density with d degrees of freedom,

$$p(z) = \frac{z^{(d/2)-1}e^{-z/2}}{2^{d/2}\Gamma(d/2)}.$$

Therefore,

$$\begin{aligned}
\nu_1(B(0,r)) &= \int_0^{r^2} \frac{z^{(d/2)-1}e^{-z/2}}{2^{d/2}\Gamma(d/2)}dz \\
&\leq \frac{1}{2^{d/2}\Gamma(d/2)}\int_0^{r^2} z^{(d/2)-1}dz \\
&= \frac{(2/d)r^d}{2^{d/2}\Gamma(d/2)} \\
&= \frac{r^d}{2^{d/2}\Gamma((d/2)+1)} \\
&\leq \frac{r^d}{2^{d/2}(d/8)^{d/2}} \;=\; \left(\frac{2r}{\sqrt{d}}\right)^d,
\end{aligned}$$

where the last inequality is a consequence of the following fact.

Fact. $\Gamma((d/2)+1) \geq (d/8)^{d/2}$ for any integer $d \geq 1$.

One way to see this is by induction. The base cases $d = 1, 2$ are trivially checked, and for $d > 2$,

$$\begin{aligned}
\Gamma\left(\frac{d}{2}+1\right) &= \frac{d}{2}\,\Gamma\left(\frac{d}{2}\right) \;\geq\; \frac{d}{2}\left(\frac{d-2}{8}\right)^{(d-2)/2} \\
&= 4\cdot\left(\frac{d}{8}\right)^{d/2}\cdot\left(\frac{d-2}{d}\right)^{(d/2)-1} \;\geq\; \frac{4}{e}\left(\frac{d}{8}\right)^{d/2} \;\geq\; \left(\frac{d}{8}\right)^{d/2},
\end{aligned}$$

where the second-last inequality relies on one last fact.

Fact. $\left(1 - \frac{2}{d}\right)^{(d/2)-1} \geq 1/e$ for any $d > 2$.

This comes from rewriting the familiar $e^x \geq 1 + x$ using $x = 2/(d-2)$:

$$e^{2/(d-2)} \;\geq\; 1 + \frac{2}{d-2} \;=\; \frac{d}{d-2}.$$

Flipping this, $e^{-2/(d-2)} \leq 1 - \frac{2}{d}$, and so $\left(1 - \frac{2}{d}\right)^{(d/2)-1} \geq \left(e^{-2/(d-2)}\right)^{(d/2)-1} = e^{-1}$. ∎

As mentioned earlier, the cases when r is very large or very small are easy to handle. The more involved case is when r is of intermediate size. Together, they complete the proof of lemma 7.

Lemma 7 *Pick any* $0 < \epsilon < 1$ *and any* $\sigma > 0$. *If*

$$\Delta \leq \frac{\sigma}{\sqrt{d}} \cdot \ln\left(1 + \frac{\epsilon}{4}\right) \cdot \frac{1}{2 + \sqrt{\frac{2}{d}\ln\frac{4}{\epsilon}}},$$

then $\nu_\sigma(B_\Delta) \leq \nu_\sigma(B) + \epsilon$ *for any ball* B.

Proof. Pick any $B = B(x, r)$. By Lemma 16, the choice of Δ implies that

$$\nu_\sigma(B(x, r+\Delta)) \leq (1 + \epsilon/4)\left(\frac{r+\Delta}{r}\right)^d \nu_\sigma(B(x,r)) + \epsilon/4.$$

We'll look at three cases for the value of r.

Case 1 (large values): $r \geq \frac{1}{2}\sigma\sqrt{d}$.
The lower bound on r implies $\Delta d \leq r\ln(1 + \epsilon/4)$, and thus that

$$\left(\frac{r+\Delta}{r}\right)^d \;\leq\; \left(1 + \frac{\ln(1+\epsilon/4)}{d}\right)^d \;\leq\; 1 + \epsilon/4,$$

whereupon $\nu_\sigma(B_\Delta) \leq (1+\epsilon/4)^2\nu_\sigma(B) + \epsilon/4 \leq \nu_\sigma(B) + \epsilon$.

Case 2 (small values): $r + \Delta \leq \frac{1}{2}\epsilon\sigma\sqrt{d}$.
From Lemma 17:

$$\nu_\sigma(B_\Delta) \;\leq\; \nu_\sigma(B(0, r+\Delta)) \;\leq\; \left(\frac{2(r+\Delta)}{\sigma\sqrt{d}}\right)^d \;\leq\; \epsilon \;\leq\; \nu_\sigma(B) + \epsilon.$$

Case 3 (intermediate values): $\frac{1}{2}\sigma\sqrt{d} \geq r \geq \frac{1}{2}\epsilon\sigma\sqrt{d} - \Delta$.
First observe that $\Delta \leq \epsilon\sigma/(8\sqrt{d})$. This means

$$r \;\geq\; \frac{1}{2}\epsilon\sigma\sqrt{d} - \frac{\epsilon\sigma}{8\sqrt{d}} \;\geq\; \frac{3}{8}\epsilon\sigma\sqrt{d},$$

and so

$$\left(\frac{r+\Delta}{r}\right)^d \;\leq\; \left(1 + \frac{\epsilon\sigma/(8\sqrt{d})}{(3/8)\epsilon\sigma\sqrt{d}}\right)^d \;=\; \left(1 + \frac{1}{3d}\right)^d \;\leq\; e^{1/3} \;\leq\; \frac{7}{5}.$$

Using the inequality $(1+x)^d \leq 1 + dx(1+x)^d$ for integers $d \geq 1$ and $0 < x < 1/d$ (see Lemma 18 below), we can also write

$$\left(\frac{r+\Delta}{r}\right)^d \ \leq \ 1 + \frac{\Delta d}{r}\left(\frac{r+\Delta}{r}\right)^d \ \leq \ 1 + \frac{7\Delta d}{5r}.$$

Thus

$$\begin{aligned}
\nu_\sigma(B_\Delta) - \nu_\sigma(B) &\leq \left(1+\frac{\epsilon}{4}\right)\left(\frac{r+\Delta}{r}\right)^d \nu_\sigma(B) + \frac{\epsilon}{4} - \nu_\sigma(B) \\
&\leq \left(\left(\frac{r+\Delta}{r}\right)^d - 1\right)\nu_\sigma(B) + \left(\frac{r+\Delta}{r}\right)^d \frac{\epsilon}{4} + \frac{\epsilon}{4} \\
&\leq \frac{7\Delta d}{5r}\left(\frac{2r}{\sigma\sqrt{d}}\right)^d + \frac{7}{5}\cdot\frac{\epsilon}{4} + \frac{\epsilon}{4} \\
&= \frac{14\Delta\sqrt{d}}{5\sigma}\left(\frac{2r}{\sigma\sqrt{d}}\right)^{d-1} + \frac{3\epsilon}{5} \\
&\leq \frac{14}{5}\cdot\frac{\epsilon}{8}\cdot 1 + \frac{3\epsilon}{5} \quad < \quad \epsilon.
\end{aligned}$$

∎

There is one last technical detail to tie up.

Lemma 18 *For any integer $d \geq 1$ and any $0 < x < 1/d$,*

$$(1+x)^d \ \leq \ \frac{1}{1-dx}.$$

Proof. This is immediate from the series expansions of both sides:

$$\begin{aligned}
(1+x)^d &= 1 + dx + \binom{d}{2}x^2 + \binom{d}{3}x^3 + \cdots + \binom{d}{d}x^d \\
&\leq 1 + dx + (dx)^2 + (dx)^3 + \cdots \\
&= \frac{1}{1-dx}.
\end{aligned}$$

∎