Projections of Probability Distributions: A Measure-theoretic Dvoretzky Theorem

Elizabeth Meckes

Case Western Reserve University

Workshop on Concentration of Measure Phenomena
Simons Institute
October 2020

Marginals are normally Gaussian

Marginals are normally Gaussian

General phenomenon: if $X \in \mathbb{R}^d$ is a random vector and d is large, then (under some conditions on $\mathcal{L}(X)$), for a large measure of $\theta \in \mathbb{S}^{d-1}$, $\langle X, \theta \rangle$ is approximately Gaussian.

Marginals are normally Gaussian

General phenomenon: if $X \in \mathbb{R}^d$ is a random vector and d is large, then (under some conditions on $\mathcal{L}(X)$), for a large measure of $\theta \in \mathbb{S}^{d-1}$, $\langle X, \theta \rangle$ is approximately Gaussian.

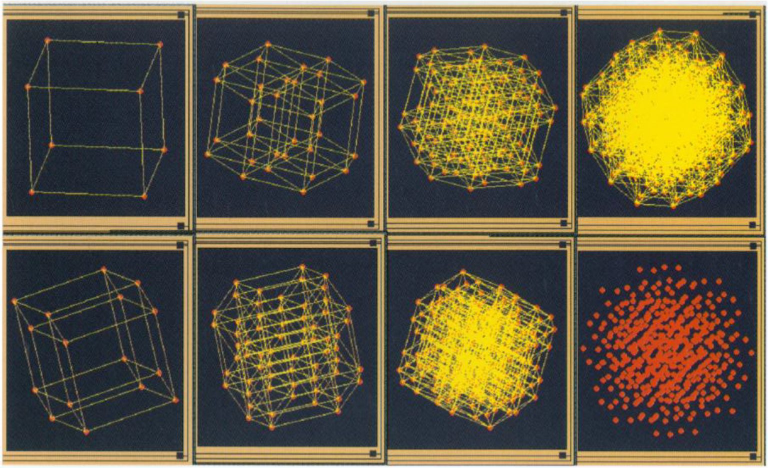

Figure from Buja, Cook, and Swayne “Interactive High-dimensional Data Visualization”, 1996.

The previous page is a series of pictures of the "Diaconis-Freedman effect", well-known to statisticians.

Diaconis and Freedman (1984) proved that, under some conditions, if

$$\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$$

is a data set (i.e., deterministic vectors with no assumptions on the process which generated them), θ is a uniform random point in the sphere $\mathbb{S}^{d-1}$, and

$$\mu_x^\theta := \frac{1}{n} \sum_{i=1}^{n} \delta_{\langle x_i, \theta \rangle}$$

is the empirical measure of the projection of the x_i in the θ-direction, then as $n, d \to \infty$, the measures μ_x^θ tend to $\mathcal{N}(0, \sigma^2)$ weakly in probability.

Many other authors (Sudakov, von Weiszäcker, Klartag, Bobkov, Dümbgen, Zerial...) have observed and contributed to the understanding of this phenomenon.

Many other authors (Sudakov, von Weiszäcker, Klartag, Bobkov, Dümbgen, Zerial...) have observed and contributed to the understanding of this phenomenon. In particular:

Theorem (Bobkov)

Suppose that X satisfies $\mathbb{E} X_i X_j = \delta_{ij}$ *and*

$$\mathbb{P}\left[\left|\frac{|X|}{\sqrt{d}} - 1\right| > \epsilon_d\right] \leq \epsilon_d.$$

Then

$$\sigma_{d-1}\left\{\theta \middle| d_\infty\left(\langle \theta, X\rangle, Z\right) \geq 4\epsilon_d + \delta\right\} \leq 4d^{3/8} e^{-cd\delta^4}.$$

Higher-dimensional marginals

Higher-dimensional marginals

A natural question: if $X \in \mathbb{R}^d$ is a random vector as before, are k-dimensional marginals close to Gaussian for fixed k?

Higher-dimensional marginals

A natural question: if $X \in \mathbb{R}^d$ is a random vector as before, are k-dimensional marginals close to Gaussian for fixed k?

Presumably.

Higher-dimensional marginals

A natural question: if $X \in \mathbb{R}^d$ is a random vector as before, are k-dimensional marginals close to Gaussian for fixed k?

Presumably.

If so, how can k grow with d? Logarithmically? Polynomially?

Higher-dimensional marginals

A natural question: if $X \in \mathbb{R}^d$ is a random vector as before, are k-dimensional marginals close to Gaussian for fixed k?

Presumably.

If so, how can k grow with d? Logarithmically? Polynomially?

Answer: $k < \frac{2\log(d)}{\log(\log(d))}$.

Random subspaces

Random subspaces

The Stiefel manifold is the set

$$\mathfrak{W}_{d,k} = \left\{ (\theta_1, \ldots, \theta_k) : \theta_j \in \mathbb{R}^d, \langle \theta_i, \theta_j \rangle = \delta_{ij} \right\}.$$

Random subspaces

The Stiefel manifold is the set

$$\mathfrak{W}_{d,k} = \Big\{ (\theta_1, \ldots, \theta_k) : \theta_j \in \mathbb{R}^d, \langle \theta_i, \theta_j \rangle = \delta_{ij} \Big\}.$$

$\mathfrak{W}_{d,k}$ has a rotation-invariant (Haar) probability measure:

random point $\theta \in \mathfrak{W}_{d,k}$ $\Longleftrightarrow$ first k columns of a Haar-distributed random orthogonal matrix in $\mathbb{O}(d)$.

Main result

Main result

Theorem (E.M.)

Let X be a random vector in $\mathbb{R}^d$ satisfying

Main result

Theorem (E.M.)

Let X be a random vector in $\mathbb{R}^d$ satisfying

- $\mathbb{E}X = 0, \mathbb{E}|X|^2 = \sigma^2 d,$ *and* $\sup_{\xi \in \mathbb{S}^{d-1}} \mathbb{E}\langle \xi, X \rangle^2 \leq L'$

Main result

Theorem (E.M.)

Let X be a random vector in $\mathbb{R}^d$ satisfying

- $\mathbb{E}X = 0$, $\mathbb{E}|X|^2 = \sigma^2 d$, *and* $\sup_{\xi \in \mathbb{S}^{d-1}} \mathbb{E}\langle \xi, X \rangle^2 \leq L'$
- $\mathbb{E}\left| |X|^2 \sigma^{-2} - d \right| \leq L \frac{d}{\sqrt{\log(d)}}$.

Main result

Theorem (E.M.)

Let X be a random vector in $\mathbb{R}^d$ satisfying

- $\mathbb{E}X = 0, \mathbb{E}|X|^2 = \sigma^2 d$, *and* $\sup_{\xi \in \mathbb{S}^{d-1}} \mathbb{E}\langle \xi, X\rangle^2 \leq L'$
- $\mathbb{E}\left||X|^2\sigma^{-2} - d\right| \leq L\frac{d}{\sqrt{\log(d)}}$.

For θ in the Stiefel manifold $\mathfrak{W}_{d,k}$, let X_θ denote the projection of X onto the span of θ.

Main result

Theorem (E.M.)

Let X be a random vector in $\mathbb{R}^d$ satisfying

- $\mathbb{E}X = 0$, $\mathbb{E}|X|^2 = \sigma^2 d$, *and* $\sup_{\xi \in \mathbb{S}^{d-1}} \mathbb{E}\langle \xi, X\rangle^2 \le L'$
- $\mathbb{E}\left| |X|^2 \sigma^{-2} - d \right| \le L \frac{d}{\sqrt{\log(d)}}$.

For θ in the Stiefel manifold $\mathfrak{W}_{d,k}$, let X_θ denote the projection of X onto the span of θ. Fix $\delta \in (0,2)$, and let $k = \delta \frac{\log(d)}{\log(\log(d))}$.

Main result

Theorem (E.M.)

Let X be a random vector in $\mathbb{R}^d$ satisfying

- $\mathbb{E}X = 0$, $\mathbb{E}|X|^2 = \sigma^2 d$, *and* $\sup_{\xi \in \mathbb{S}^{d-1}} \mathbb{E}\langle \xi, X \rangle^2 \leq L'$
- $\mathbb{E}\left| |X|^2 \sigma^{-2} - d \right| \leq L \frac{d}{\sqrt{\log(d)}}$.

For θ in the Stiefel manifold $\mathfrak{W}_{d,k}$, let X_θ denote the projection of X onto the span of θ. Fix $\delta \in (0,2)$, and let $k = \delta \frac{\log(d)}{\log(\log(d))}$. Then there is a $c > 0$ depending only on δ, L and L' such that for $\epsilon = \frac{2}{[\log(d)]^c}$, there is a subset $\mathfrak{T} \subseteq \mathfrak{W}_{d,k}$ with $\mathbb{P}_{d,k}[\mathfrak{T}^c] \leq Ce^{-c' d \epsilon^2}$, such that for all $\theta \in \mathfrak{T}$,

$$d_{BL}(X_\theta, \sigma Z) \leq C' \epsilon.$$

Sharpness

Sharpness

Let X be uniform among $S := \{\pm\sqrt{d}e_1, \ldots, \pm\sqrt{d}e_d\} \subseteq \mathbb{R}^d$.

Sharpness

Let X be uniform among $S := \{\pm\sqrt{d}e_1, \ldots, \pm\sqrt{d}e_d\} \subseteq \mathbb{R}^d$.
Let $c > 2$ and let E be a subspace of $\mathbb{R}^d$ with
$dim(E) = c\frac{\log(d)}{\log(\log(d))}$.

Sharpness

Let X be uniform among $S := \{\pm\sqrt{d}e_1, \ldots, \pm\sqrt{d}e_d\} \subseteq \mathbb{R}^d$.
Let $c > 2$ and let E be a subspace of $\mathbb{R}^d$ with
$dim(E) = c\frac{\log(d)}{\log(\log(d))}$.
Define $f : E \to \mathbb{R}$ by $f(x) := (1 - d(x, \pi_E(S)))_+$.

Sharpness

Let X be uniform among $S := \{\pm\sqrt{d}e_1, \ldots, \pm\sqrt{d}e_d\} \subseteq \mathbb{R}^d$.
Let $c > 2$ and let E be a subspace of $\mathbb{R}^d$ with
$dim(E) = c\frac{\log(d)}{\log(\log(d))}$.
Define $f : E \to \mathbb{R}$ by $f(x) := (1 - d(x, \pi_E(S)))_+$. Then
$\|f\|_{BL} \leq 1$ and

$$\int f d\mu_{\pi_E(S)} = 1$$

but

$$\int f d\gamma_E \xrightarrow{d\to\infty} 0.$$

That is, for this choice of k, $d_{BL}(X_\theta, \sigma Z) \approx 1$ for all choices of
$\theta \in \mathfrak{W}_{d,k}$.

The example shows that $k_c = \frac{2\log(d)}{\log(\log(d))}$ is a sharp cut-off such that if X is a random vector in $\mathbb{R}^d$ satisfying some natural conditions on $\mathcal{L}(X)$, then most k-dimensional margins of X are approximately Gaussian for $k < k_c$ and this need not be true for $k > k_c$.

Dvoretzky’s Theorem

Dvoretzky's Theorem

Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ such that the maximum volume ellipsoid in its unit ball is a dilate of the sphere. Let $\epsilon > 0$ be fixed.

Dvoretzky's Theorem

Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ such that the maximum volume ellipsoid in its unit ball is a dilate of the sphere. Let $\epsilon > 0$ be fixed. Then there is some rescaling of $\|\cdot\|$ and a constant $C(\epsilon)$ (depending only on ϵ!) such that if

$$k \leq C(\epsilon)\log(d)$$

and if E is a random subspace of $\mathbb{R}^d$ of dimension k, then with probability tending to 1,

$$|v| \leq \|v\| \leq (1+\epsilon)|v|$$

for all $v \in E$.

Dvoretzky's Theorem

Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ such that the maximum volume ellipsoid in its unit ball is a dilate of the sphere. Let $\epsilon > 0$ be fixed. Then there is some rescaling of $\|\cdot\|$ and a constant $C(\epsilon)$ (depending only on ϵ!) such that if

$$k \leq C(\epsilon)\log(d)$$

and if E is a random subspace of $\mathbb{R}^d$ of dimension k, then with probability tending to 1,

$$|v| \leq \|v\| \leq (1+\epsilon)|v|$$

for all $v \in E$.

That is, if $k \leq C(\epsilon)\log(d)$, then most k-dimensional subspaces of the normed space $(\mathbb{R}^d, \|\cdot\|)$ look very similar to k-dimensional Euclidean space $(\mathbb{R}^k, |\cdot|)$.

The analogy

The analogy

- In both theorems, an additional structure is imposed on $\mathbb{R}^n$ (a norm in the case of Dvoretzky's theorem; a probability measure in our context);

The analogy

- In both theorems, an additional structure is imposed on $\mathbb{R}^n$ (a norm in the case of Dvoretzky's theorem; a probability measure in our context);
- in either case, there is a particularly nice way to do this (the Euclidean norm and the Gaussian distribution, respectively).

The analogy

- In both theorems, an additional structure is imposed on $\mathbb{R}^n$ (a norm in the case of Dvoretzky's theorem; a probability measure in our context);
- in either case, there is a particularly nice way to do this (the Euclidean norm and the Gaussian distribution, respectively).
- If you reduce the dimension sufficiently, what typically happens is that all of the original structure is lost and all you see is this canonical nice (or boring) space.

Dvoretzky dimension

Dvoretzky dimension

Under extra assumptions on the norm $\| \cdot \|$, it may be that k can be larger as a function of d. In particular:

Dvoretzky dimension

Under extra assumptions on the norm $\|\cdot\|$, it may be that k can be larger as a function of d. In particular:

- Figiel, Lindenstrauss and V. Milman showed that if a d-dimensional Banach space X has cotype $q \in [2, \infty)$, then X has subspaces of dimension of the order $d^{\frac{2}{q}}$ which are approximately Euclidean.

Dvoretzky dimension

Under extra assumptions on the norm $\|\cdot\|$, it may be that k can be larger as a function of d. In particular:

- Figiel, Lindenstrauss and V. Milman showed that if a d-dimensional Banach space X has cotype $q \in [2, \infty)$, then X has subspaces of dimension of the order $d^{\frac{2}{q}}$ which are approximately Euclidean.
- Szarek showed that if X has bounded volume ratio, then X has nearly Euclidean subspaces of dimension $\frac{d}{2}$.

Dvoretzky dimension

Under extra assumptions on the norm $\|\cdot\|$, it may be that k can be larger as a function of d. In particular:

- Figiel, Lindenstrauss and V. Milman showed that if a d-dimensional Banach space X has cotype $q \in [2, \infty)$, then X has subspaces of dimension of the order $d^{\frac{2}{q}}$ which are approximately Euclidean.
- Szarek showed that if X has bounded volume ratio, then X has nearly Euclidean subspaces of dimension $\frac{d}{2}$.

This is analogous to the difference between the main theorem and a result of Klartag, showing that if the random vector X has a log-concave distribution, then most projections are close to Gaussian for $k = d^{\epsilon}$ for a specific value of ϵ.

Outline of the proof of the main theorem

Outline of the proof of the main theorem

- The mean projection $X_\Theta = \langle X, \Theta \rangle$, when both X and Θ are random and independent, is approximately Gaussian.
  This is shown using Stein's method.

Outline of the proof of the main theorem

- The mean projection $X_\Theta = \langle X, \Theta \rangle$, when both X and Θ are random and independent, is approximately Gaussian.
  This is shown using Stein's method.
- The mean bounded-Lipschitz distance $\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta)$ is small.
  The bounded-Lipschitz distance is interpreted as the supremum of a stochastic process indexed by test functions. Concentration of measure on the Stiefel manifold implies that this process has subgaussian increments, allowing the expected supremum to be estimated via entropy methods.

Outline of the proof of the main theorem

- The mean projection $X_\Theta = \langle X, \Theta \rangle$, when both X and Θ are random and independent, is approximately Gaussian.
  This is shown using Stein's method.
- The mean bounded-Lipschitz distance $\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta)$ is small.
  The bounded-Lipschitz distance is interpreted as the supremum of a stochastic process indexed by test functions. Concentration of measure on the Stiefel manifold implies that this process has subgaussian increments, allowing the expected supremum to be estimated via entropy methods.
- The bounded-Lipschitz distance $d_{BL}(X_\theta, X_\Theta)$ is tightly concentrated near its mean.
  This also follows from concentration of measure on the Stiefel manifold.

More about step 1

Exchangeable pairs with infinitesimal symmetries:

More about step 1

Exchangeable pairs with infinitesimal symmetries: If $W \in \mathbb{R}^k$ is a random vector, and a family $(W, W_\epsilon)_{\epsilon>0}$ of exchangeable pairs can be constructed so that, for some deterministic $\lambda(\epsilon)$,

More about step 1

Exchangeable pairs with infinitesimal symmetries: If $W \in \mathbb{R}^k$ is a random vector, and a family $(W, W_\epsilon)_{\epsilon>0}$ of exchangeable pairs can be constructed so that, for some deterministic $\lambda(\epsilon)$,

- $\mathbb{E}[W_\epsilon - W|W] \approx -\lambda(\epsilon)W$
- $\mathbb{E}[(W_\epsilon - W)(W_\epsilon - W)^T|W] \approx 2\lambda(\epsilon)\sigma^2 \mathrm{I}_{k\times k}$
- $\mathbb{E}|W_\epsilon - W|^3 \ll \lambda(\epsilon)$

More about step 1

Exchangeable pairs with infinitesimal symmetries: If $W \in \mathbb{R}^k$ is a random vector, and a family $(W, W_\epsilon)_{\epsilon>0}$ of exchangeable pairs can be constructed so that, for some deterministic $\lambda(\epsilon)$,

- $\mathbb{E}[W_\epsilon - W|W] \approx -\lambda(\epsilon)W$
- $\mathbb{E}[(W_\epsilon - W)(W_\epsilon - W)^T|W] \approx 2\lambda(\epsilon)\sigma^2 \mathrm{I}_{k\times k}$
- $\mathbb{E}|W_\epsilon - W|^3 \ll \lambda(\epsilon)$

Then $W \approx \sigma Z$, where Z is a standard Gaussian random vector.

More about step 1

Exchangeable pairs with infinitesimal symmetries: If $W \in \mathbb{R}^k$ is a random vector, and a family $(W, W_\epsilon)_{\epsilon>0}$ of exchangeable pairs can be constructed so that, for some deterministic $\lambda(\epsilon)$,

- $\mathbb{E}[W_\epsilon - W|W] \approx -\lambda(\epsilon)W$
- $\mathbb{E}[(W_\epsilon - W)(W_\epsilon - W)^T|W] \approx 2\lambda(\epsilon)\sigma^2 \mathrm{I}_{k\times k}$
- $\mathbb{E}|W_\epsilon - W|^3 \ll \lambda(\epsilon)$

Then $W \approx \sigma Z$, where Z is a standard Gaussian random vector.

Here, we take $W = \langle X, \Theta \rangle$, where $\Theta \in \mathfrak{W}_{d,k}$ is uniform and independent of X.

$W = \langle X, \Theta \rangle$ is approximately Gaussian:

$W = \langle X, \Theta \rangle$ is approximately Gaussian:

To construct W_ϵ, rotate Θ by ϵ in a random direction:

$W = \langle X, \Theta \rangle$ is approximately Gaussian:

To construct W_ϵ, rotate Θ by ϵ in a random direction: if

$$\Theta = (\Theta_1, \ldots, \Theta_k),$$

then

$$\Theta_\epsilon = \big([UR_{1,2}(\epsilon)U^T]\Theta_1, \ldots, [UR_{1,2}(\epsilon)U^T]\Theta_k\big),$$

where U is an independently chosen random orthogonal matrix and $R_{1,2}(\epsilon)$ rotates by ϵ in the span of the first two basis elements.

$W = \langle X, \Theta \rangle$ is approximately Gaussian:

To construct W_ϵ, rotate Θ by ϵ in a random direction: if

$$\Theta = (\Theta_1, \ldots, \Theta_k),$$

then

$$\Theta_\epsilon = \big([UR_{1,2}(\epsilon)U^T]\Theta_1, \ldots, [UR_{1,2}(\epsilon)U^T]\Theta_k\big),$$

where U is an independently chosen random orthogonal matrix and $R_{1,2}(\epsilon)$ rotates by ϵ in the span of the first two basis elements.

The theorem on the last slide can be applied, and the result is that

$$d_{BL}(X_\Theta, \sigma Z) \leq \frac{C\sigma\sqrt{k}\mathbb{E}\big||X|^2\sigma^{-2} - d\big| + \sigma k}{d}.$$

Concentration of measure

Concentration of measure

Define the metric ρ on $\mathfrak{W}_{d,k}$ by

$$\rho(\theta, \theta') = \sqrt{\sum_{i=1}^{k} |\theta_i - \theta_i'|^2}.$$

Concentration of measure

Define the metric ρ on $\mathfrak{W}_{d,k}$ by

$$\rho(\theta, \theta') = \sqrt{\sum_{i=1}^{k} |\theta_i - \theta_i'|^2}.$$

There are constants C, c (independent of d, k) such that if $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ is Lipschitz with Lipschitz constant L,

$$\mathbb{P}\Big[\big|F(\Theta) - \mathbb{E}F(\Theta)\big| > L\epsilon\Big] \leq Ce^{-cd\epsilon^2}.$$

Concentration of measure

Define the metric ρ on $\mathfrak{W}_{d,k}$ by

$$\rho(\theta, \theta') = \sqrt{\sum_{i=1}^{k} |\theta_i - \theta_i'|^2}.$$

There are constants C, c (independent of d, k) such that if $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ is Lipschitz with Lipschitz constant L,

$$\mathbb{P}\Big[\big|F(\Theta) - \mathbb{E}F(\Theta)\big| > L\epsilon\Big] \leq Ce^{-cd\epsilon^2}.$$

It's straightforward to show that $F(\theta) := d_{BL}(X_\theta, \sigma Z)$ is Lipschitz with constant $\sqrt{L'}$; this is the whole content of step 3.

Step 2 – Average distance to average

Step 2 – Average distance to average

We need to estimate

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) = \mathbb{E}\left(\sup_{\|f\|_{BL}\le 1}\left|\mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta)\right|\right).$$

Step 2 – Average distance to average

We need to estimate

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) = \mathbb{E}\left(\sup_{\|f\|_{BL}\leq 1}\left|\mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta)\right|\right).$$

If the stochastic process $\{X_f\}_{\|f\|_{BL}\leq 1}$ is defined by

$$X_f := \mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta),$$

then what we want is $\mathbb{E}\sup_{\|f\|_{BL}\leq 1} X_f$.

Step 2 – Average distance to average

We need to estimate

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) = \mathbb{E}\left(\sup_{\|f\|_{BL}\le 1}\left|\mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta)\right|\right).$$

If the stochastic process $\{X_f\}_{\|f\|_{BL}\le 1}$ is defined by

$$X_f := \mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta),$$

then what we want is $\mathbb{E}\sup_{\|f\|_{BL}\le 1} X_f$.

Applying measure concentration to $F(\theta) := \mathbb{E}\left[(f-g)(X_\theta)\middle|\theta\right]$ shows that the process has the property:

$$\mathbb{P}\left[\left|X_f - X_g\right| > \epsilon\right] \le Ce^{-\frac{cd\epsilon^2}{\|f-g\|_{BL}^2}}.$$

Theorem (Dudley)

If a stochastic process $\{X_t\}_{t\in T}$ satisfies the a sub-Gaussian increment condition

$$\mathbb{P}\left[\left|X_t - X_s\right| > \epsilon\right] \leq Ce^{-\frac{\epsilon^2}{2\delta^2(s,t)}} \qquad \forall \epsilon > 0,$$

then

$$\mathbb{E} \sup_{t\in T} X_t \leq C \int_0^\infty \sqrt{\log N(T, \delta, \epsilon)} d\epsilon,$$

where $N(T, \delta, \epsilon)$ is the ϵ-covering number of T with respect to the distance δ.

Theorem (Dudley)

If a stochastic process $\{X_t\}_{t\in T}$ satisfies the a sub-Gaussian increment condition

$$\mathbb{P}\left[|X_t - X_s| > \epsilon\right] \leq Ce^{-\frac{\epsilon^2}{2\delta^2(s,t)}} \qquad \forall \epsilon > 0,$$

then

$$\mathbb{E}\sup_{t\in T} X_t \leq C\int_0^\infty \sqrt{\log N(T,\delta,\epsilon)}d\epsilon,$$

where $N(T,\delta,\epsilon)$ is the ϵ-covering number of T with respect to the distance δ.

Recall that our process satisfies

$$\mathbb{P}\left[|X_f - X_g| > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{\|f-g\|_{BL}^2}}.$$

The question, then, is: if $BL_1^k := \left\{ f : \mathbb{R}^k \to \mathbb{R} \middle| \|f\|_{BL} \leq 1 \right\}$, what is $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right)$?

The question, then, is: if $BL_1^k := \left\{ f : \mathbb{R}^k \to \mathbb{R} \middle| \|f\|_{BL} \leq 1 \right\}$, what is $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right)$?

Bad news: $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right) = \infty$.

The question, then, is: if $BL_1^k := \left\{ f : \mathbb{R}^k \to \mathbb{R} \middle| \|f\|_{BL} \leq 1 \right\}$, what is $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right)$?

Bad news: $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right) = \infty$.

But not to worry: approximating Lipschitz functions by piecewise affine functions and using volumetric estimates in the resulting finite-dimensional normed space of approximating functions does the job, and ultimately we get (with the simplification $L' = 1$)

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) \leq C \frac{k + \log(d)}{k^{\frac{2}{3}} d^{\frac{2}{3k+4}}}.$$

So:

So:

- $d_{BL}(X_\Theta, \sigma Z) \le \frac{C\sigma\sqrt{k}\mathbb{E}||X|^2\sigma^{-2} - d| + \sigma k}{d}$

So:

- $d_{BL}(X_\Theta, \sigma Z) \leq \frac{C\sigma\sqrt{k}\mathbb{E}||X|^2\sigma^{-2}-d|+\sigma k}{d}$
- $\mathbb{P}\left[\theta : \left|d_{BL}(X_\theta, X_\Theta) - \mathbb{E}d_{BL}(X_\theta, X_\Theta)\right| > \epsilon\right] \leq Ce^{-cd\epsilon^2}$.

So:

- $d_{BL}(X_\Theta, \sigma Z) \leq \frac{C\sigma\sqrt{k}\mathbb{E}||X|^2\sigma^{-2}-d|+\sigma k}{d}$
- $\mathbb{P}\left[\theta : \left|d_{BL}(X_\theta, X_\Theta) - \mathbb{E}d_{BL}(X_\theta, X_\Theta)\right| > \epsilon\right] \leq Ce^{-cd\epsilon^2}.$
- $\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) \leq C\frac{k+\log(d)}{k^{\frac{2}{3}}d^{\frac{2}{3k+4}}}.$

So:

- $d_{BL}(X_\Theta, \sigma Z) \le \frac{C\sigma\sqrt{k}\mathbb{E}||X|^2\sigma^{-2} - d| + \sigma k}{d}$
- $\mathbb{P}\left[\theta : \left| d_{BL}(X_\theta, X_\Theta) - \mathbb{E} d_{BL}(X_\theta, X_\Theta) \right| > \epsilon \right] \le Ce^{-cd\epsilon^2}$.
- $\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) \le C \frac{k + \log(d)}{k^{\frac{2}{3}} d^{\frac{2}{3k+4}}}$.

Choosing $k = \frac{\delta \log(d)}{\log(\log(d))}$ and $\epsilon = \frac{2}{\log(d)^c}$ (for a particular c which depends on δ) finishes the proof.

Thank you.