# Machine Learning Talk IV

## Effective Dimension in High-Dimensional Problems

Axel G. R. Turnquist

NJIT Department of Mathematical Sciences

October 16, 2020

# High-Dimensional Bounds: A Case for Probability Theory

Often in high dimensions, bounds can be improved by looking at the expectation. From "Probability in High Dimensions" by Ramon van Handel pg. 129:

- Estimate via direct methods:

$$|X_f - X_g| \leq 2||f - g||_\infty, \text{ a.s.} \tag{1}$$

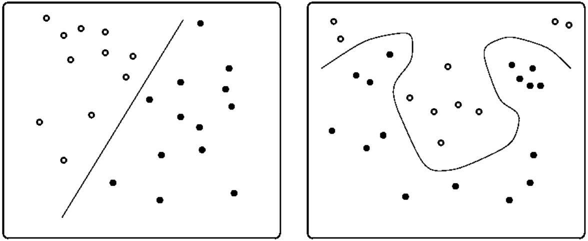
- Estimate of the expectation:

$$\mathbb{E}|X_f - X_g| \leq n^{-1/2}||f - g||_\infty \tag{2}$$

**Takeaway**: Bounds that depend on expectation can sometimes be asymptotically tighter in high dimensions! (Same thing is true in $L^p$ spaces)

# Dimension Reduction

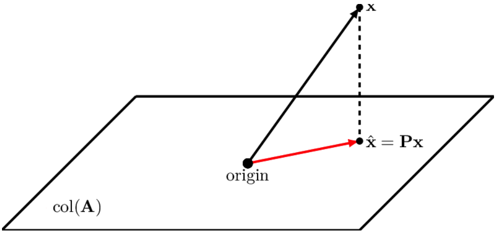
Think of the **classification problem**:



**Goal**: reduce dimension and keep data “fidelity”

# Goals

- Separating hyperplane theorem requires the notion of **orthogonality**
- Want the notion of **distance** to be the same, so we can quantify error of fitted model as in ambient space
- Would like to apply dimension reduction **randomly**
- Have the reduction only depend somehow on the ambient dimension $n$ and the number of sampled points $N$



# Isometry

Suppose you have two metric spaces, $\mathcal{X}, \mathcal{Y}$ with metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, respectively and you have a mapping $T : \mathcal{X} \rightarrow \mathcal{Y}$ between them. An **isometry** is the most ideal way of comparing the two spaces, if such a mapping is possible. An isometry guarantees:

- Unique points in $\mathcal{X}$ are mapped to unique points in $\mathcal{Y}$, i.e. this is an isomorphic mapping
- The "size" of **vectors** is the same:

$$d_{\mathcal{X}}(x_1, x_2) = d_{\mathcal{Y}}(y_1, y_2) \tag{3}$$

  for $x_1, x_2 \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$.
- Vectors that are orthogonal in $\mathcal{X}$ are orthogonal in $\mathcal{Y}$, i.e. angles have the same meaning in both metric spaces.

# Johnson-Lindenstrauss Lemma

Let $\mathcal{X}$ be a set of $N$ points in $\mathbb{R}^n$ and $\epsilon > 0$. Assume that

$$m \geq (C/\epsilon^2) \log N \tag{4}$$

Consider a random $m$-dimensional subspace $E$ in $\mathbb{R}^n$ uniformly distributed in $G_{n,m}$. Denote the orthogonal projection onto $E$ by $P$. then, with probability at least $1 - 2\exp(-c\epsilon^2 m)$, the scaled projection:

$$Q := \sqrt{\frac{n}{m}} P \tag{5}$$

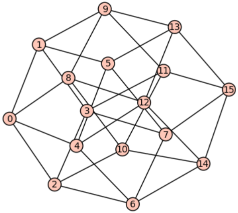
is an approximate isometry on $\mathcal{X}$:

$$(1-\epsilon)||x-y||_2 \leq ||Qx - Qy||_2 \leq (1+\epsilon)||x-y||_2 \tag{6}$$

for all $x, y \in \mathcal{X}$.

# The Continuous Case: What is going on geometrically?

- Concentration of area on spheres in high dimension $\{x : ||x||_2 = 1\}$. Most of the mass is located around every "equator".
- What about cubes in high dimensions $\{x : ||x||_\infty = 1\}$? Most of the volume is located near the vertices (many vertices).
- What about $\{x : ||x||_1 = 1\}$? This object appears much smaller than it actually is in high-dimensions (very little mass concentrates about the vertices). Very spiky.



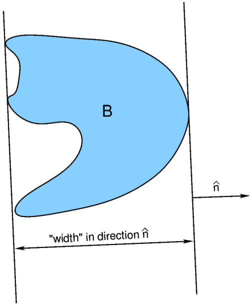
# Spherical Width (Mean Width)

The **spherical width** of a subset $T \subset \mathbb{R}^n$ is defined as:

$$w_s(T) := \mathbb{E} \sup_{x \in T} \langle \theta, x \rangle \tag{7}$$

where $\theta \sim \text{Unif}(\mathbb{S}^{n-1})$.

$$w_s(B_1^n) \sim \sqrt{\frac{\log n}{n}} \tag{8}$$



# Size of Random Projections

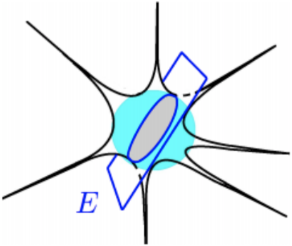
Consider a bounded set $T \subset \mathbb{R}^n$. Let $P$ be a projection in $\mathbb{R}^n$ onto a random $m$-dimensional subspace $E \sim \mathrm{Unif}(G_{n,m})$. Then, with probability at least $1 - 2e^{-m}$, we have:

$$\mathrm{diam}(PT) \leq C\left(w_s(T) + \sqrt{\frac{m}{n}}\mathrm{diam}(T)\right) \tag{9}$$

or, equivalently,

$$\mathrm{diam}(PT) \leq C \max\left(w_s(T), \sqrt{\frac{m}{n}}\mathrm{diam}(T)\right) \tag{10}$$

which represents a kind of "phase transition". We see that the mean width governs the diameter of random projections in high-dimensions and this happens at the "effective" dimension $d(T) \sim \frac{n w_s(T)^2}{\mathrm{diam}(T)^2} \sim \frac{w(T)^2}{\mathrm{diam}(T)^2}$.


E

# Using Gaussian Processes to Learn about Geometry

In geometry, one can "study" the topology of a manifold by:

- Find the eigenvalues of the Laplace-Beltrami operator
- Define certain smooth functions (Morse theory) on the manifold and find their critical points

Can we learn something about the geometry here by using a Markov process? Yes.

$$w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle, \text{ where } g \sim N(0, I_n) \tag{11}$$

Recall the "effective" dimension above is: $d(T) \sim \frac{w(T)^2}{\mathrm{diam}(T)^2}$.

# Discretization of Sets to Accuracy $\epsilon$

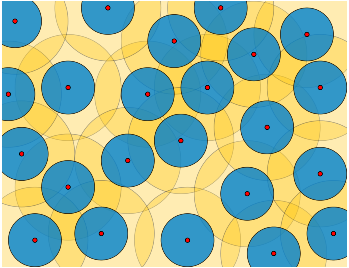
**Another Idea**: Maybe we can learn something about "effective" dimension by discretization parametrized by $\epsilon$, and noting how the complexity of the set changes as $\epsilon \to 0$.

Specify the points in a set $K$ in a metric space $(T, d)$ to accuracy $\epsilon$ in the metric $d$. Then, the number of bits by $\mathcal{C}$, can be bounded by a quantity called the **metric entropy** of the set $K$:

$$\log_2 \mathcal{N}(K, d, \epsilon) \leq \mathcal{C} \leq \log_2 \mathcal{N}(K, d, \epsilon/2) \qquad (12)$$

# $\epsilon$-nets

- $\epsilon$-**net**: Let $(T, d)$ be a metric space. Consider a subset $K \subset T$ and let $\epsilon > 0$. A subset $\mathcal{N} \subset K$ is called an $\epsilon$-net of $K$ if every point in $K$ is within a distance $\epsilon$ of some point of $\mathcal{N}$
- **Covering number**: The smallest possible cardinality of an $\epsilon$-net of $K$ is called the covering number of $K$ and is denoted by $\mathcal{N}(K, d, \epsilon)$



# Relation Between Metric Entropy and Stable Dimension

**Theorem (Fernique)**
Let $\{X_t\}_{t \in T}$ be a stationary separable Gaussian process. Then, $\exists c_1, c_2$ s.t.:

$$c_1 \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon \leq \mathbb{E}\left[\sup_{t \in T} X_t\right] \leq c_2 \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon \quad (13)$$

**Conclusion**: Another interpretation of "effective" dimension:

$$d(T) \sim \left( \frac{\int_0^\infty \sqrt{\log \mathcal{N}}}{\mathrm{diam}(T)} \right)^2 \quad (14)$$

# Questions?

# Some Useful Resources

- “High-Dimensional Probability” Vershynin, Roman.
- “Pattern Recognition and Machine Learning” Christopher M. Bishop
- “Probability in High Dimensions” Ramon van Handel. APC 550 Lecture Notes Princeton University.

# Future Talks

**Further potential topics**:

- Adversarial attacks
- Data augmentation
- ???

Oct 23: TBD