Introduction

Ilya Razenshteyn (MIT CSAIL)

Geometric structure of data

Datasets often have (implicit or explicit) geometric structure

CSAIL at MIT @MIT_CSAIL · 21 Dec 2016
Happy holidays from the robots (and humans) of MIT CSAIL!

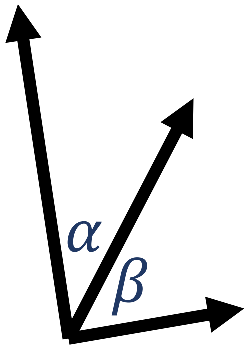
bag of words

…

0	cats
1	humans
0	birds
1	robots
0	hamsters

…

feature vectors



Data analysis task

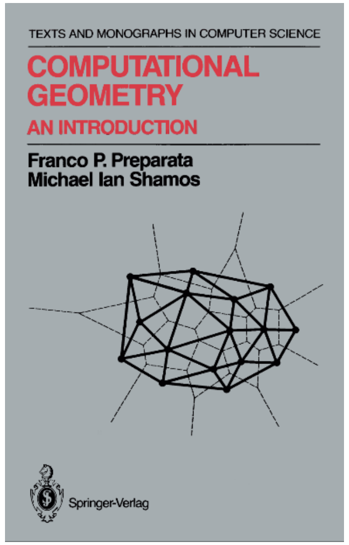
Geometric problem

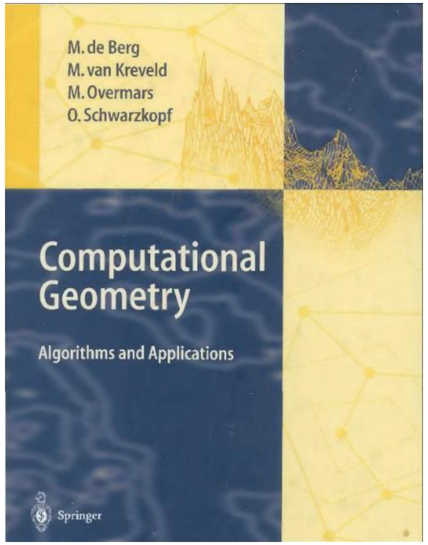
Example: find a cluster of similar documents

Scale

High-dimensional data requires a new theory

- Twitter example: **200B** tweets per year, **100K** dimensions
- **Massive high-dimensional** geometric data (100 is high too)
- Computational geometry: typically, **exponential** dependence on the dimension
- **Curse of dimensionality**

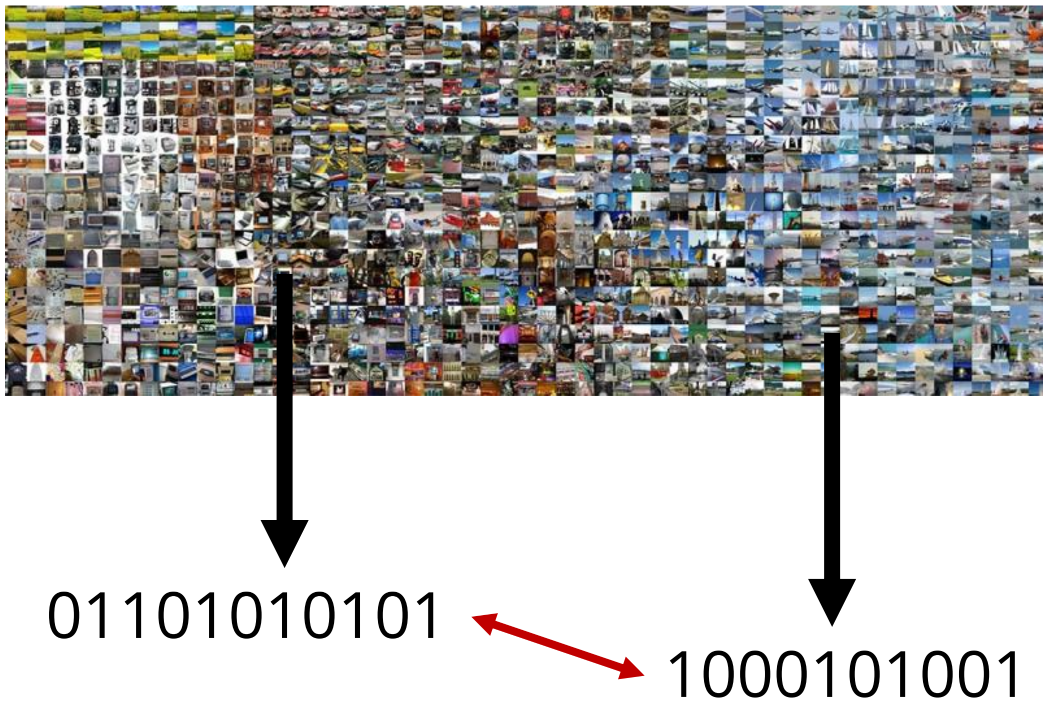




…

The toolbox

- **Efficient representations** of data
  - Randomized hashing
  - Sketching (summarization)
  - Dimension reduction
  - Metric embeddings
  - …



Plan

- Dimension reduction: theory and practice
- Nearest Neighbor Search: theory
- Nearest Neighbor Search: practice
- Fast linear algebra (big maybe)
- **Michael will cover streaming and sketching**

Meta

- Heavily biased towards my PhD research
- Complicated proofs but practical algorithms
  - See, e.g., **https://falconn-lib.org/**
- Not everything I will say has a proof
- Randomness is the key to everything
- Most important: flavor of modern algorithms research
- Interact!
- Lots of cool open problems

But let's first take a detour...

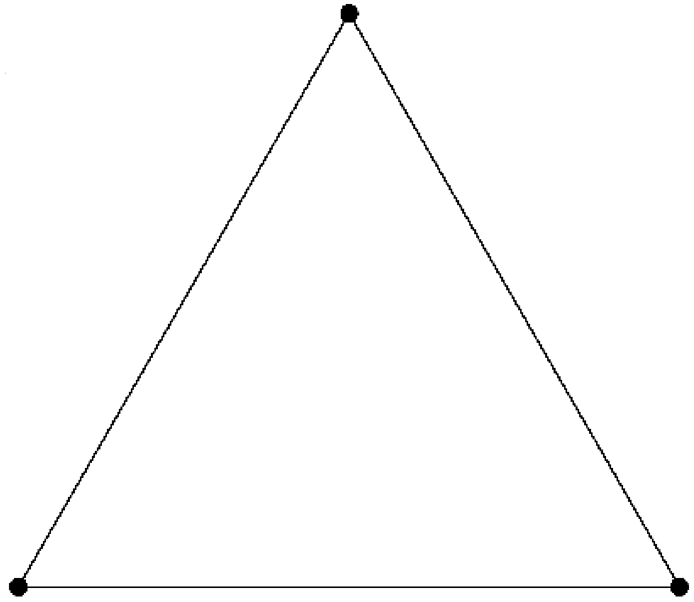
Chapter 0: Measure Concentration

How to think about high dimensions?

- How to think about geometric objects in R^d when d is large?
- Counterintuitively:
  - **Geometry** barely helps
  - **Probability** and **analysis** are extremely useful
- **Concentration of measure** is ubiquitous

Case study

- Case study: as many as possible points in such that all pairwise distances are
  - Exactly 1
  - Between $1 - \varepsilon$ and $1 + \varepsilon$
- For $d = 2$, the answer is 3 for both
- What happens if $d \to \infty$?



The exact case

- Maximum number of equidistant points in R^d is $d + 1$
- Let points be $v_0, v_1, \ldots, v_t$
- **Exercise:** show that $v_1 - v_0, v_2 - v_0, \ldots, v_t - v_0$ are linearly independent
- Hence, $t \leq d + 1$
- **Tight:** a regular simplex

The approximate case

- Unlike the exact case, can have as many as $2^{\Omega(\varepsilon^2 d)}$ points with pairwise distances between $1-\varepsilon$ and $1+\varepsilon$
- Counterintuitive!
- A special case of **dimension reduction** (will see later)

The probabilistic method

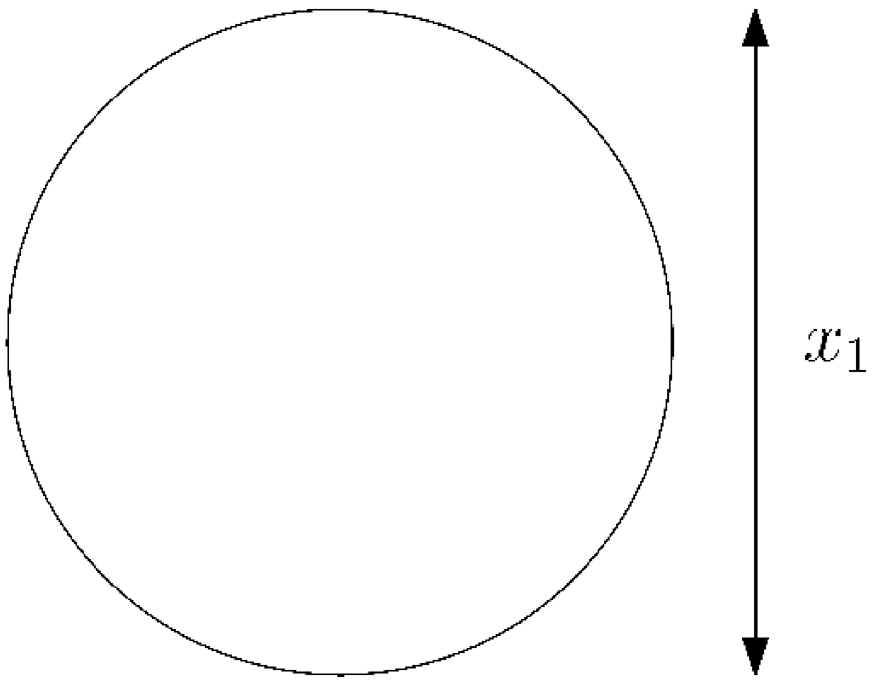
- **To prove that some object exists, show that a random object has desired properties with positive probability**
- Pioneered by Paul Erdős
- Allows to import probabilistic techniques into combinatorics and geometry
- Alon, Spencer, “Probabilistic Method”, **319pp.**

Are random points ~ equidistant?

- Want: n points in R^d with distances between $1 - \varepsilon$ and $1 + \varepsilon$
  - With $n = 2^{\Omega(\varepsilon^2 d)}$
- Proof idea: choose n points uniformly and independently from $S^{d-1} \subset R^d$, with high probability pairwise distances are close, rescale
- Simple but powerful trick: use **union bound**
- $\Pr[\text{some pair is bad}] \leq n^2 \cdot \Pr[\text{a fixed pair is bad}] < (?)1$
- Enough to understand the distribution of distances between **two** random points!

Concentration of measure on the sphere

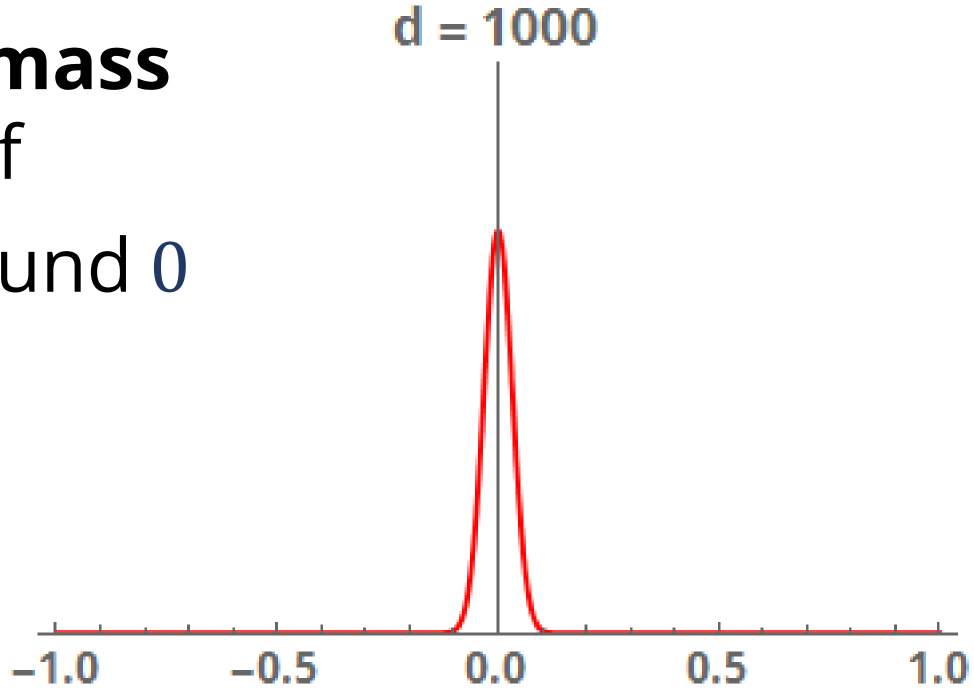
- Let $x, y \in S^{d-1} \subset R^d$ be two uniform random points
- Understand the random variable $\|x - y\|$
- The same if $y = (1, 0, 0, \ldots, 0)$, and x is random
- Thus, need to understand x_1 for a random $x \in S^{d-1}$



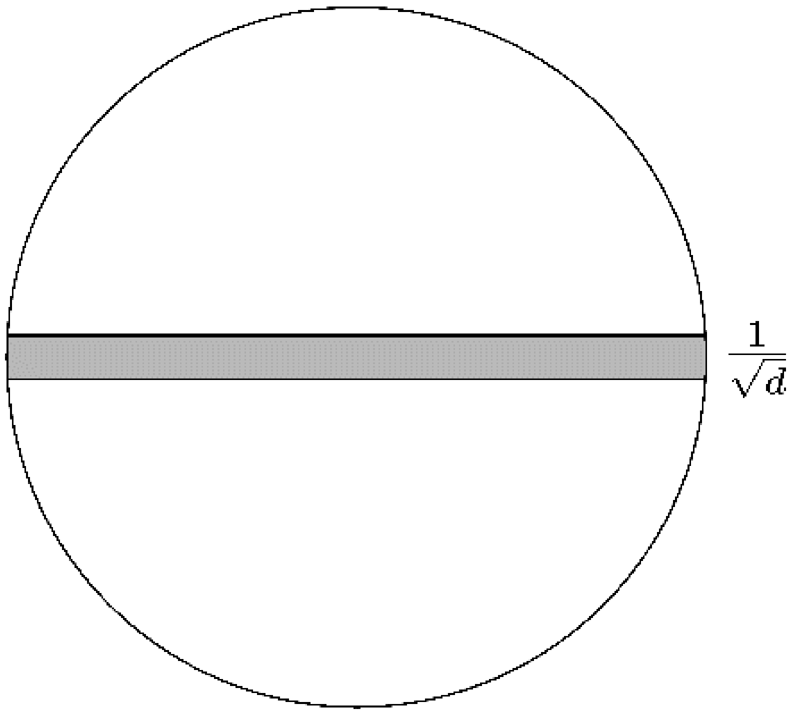
The distribution of x_1

- **Exercise:** compute it for $d = 3$ and get surprised

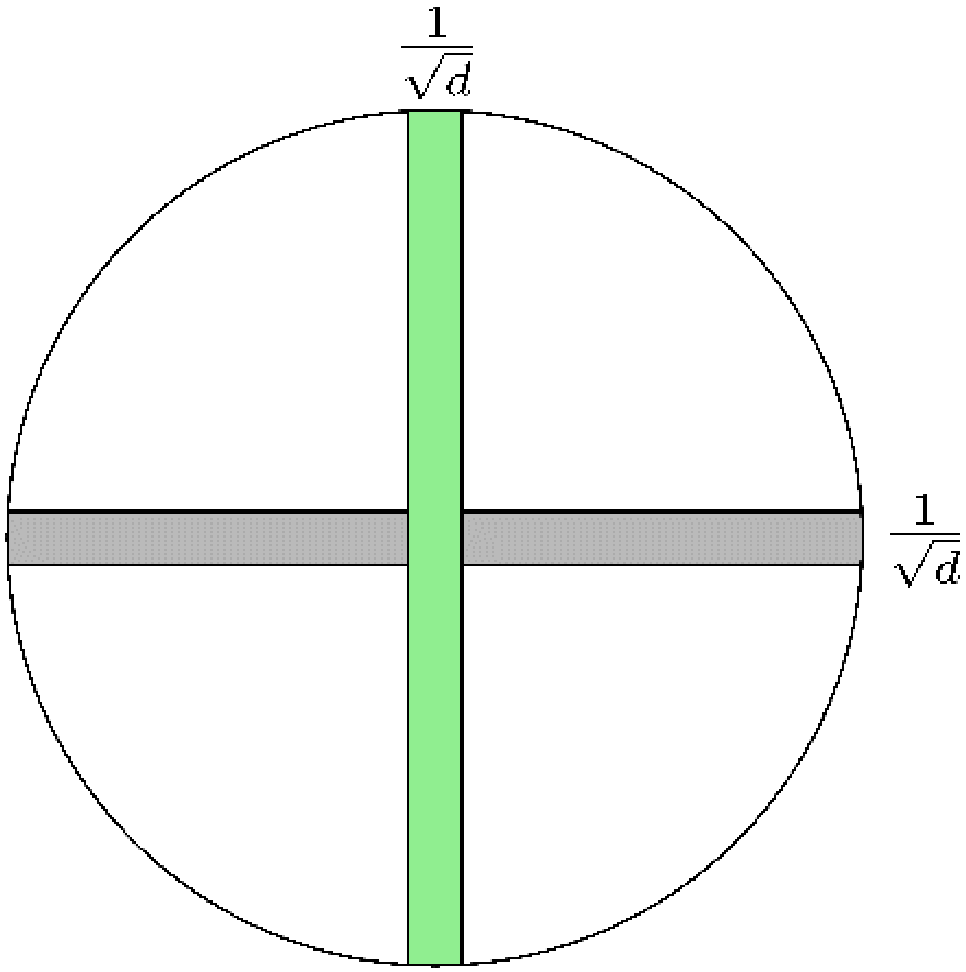
Almost all the mass is in the stripe of width $\Theta\left(\frac{1}{\sqrt{d}}\right)$ around 0



Almost everything is near an equator...



Any equator!



Near-orthogonal vectors

- With probability $1 - 1/10n^2$, one has $|x_1| \leq O\left(\sqrt{\frac{\log n}{d}}\right)$
- At most ε if $n \leq 2^{O(\varepsilon^2 d)}$

Fast forward: dimension reduction

- Given n points in R^d, map them into $R^{d'}$ with $d' \ll d$ such that pairwise distances are preserved up to $1 \pm \varepsilon$
- The above argument shows that for a regular simplex we can obtain $d' = O\left(\frac{\log n}{\varepsilon^2}\right)$
- **Johnson–Lindenstrauss lemma**: the same bound for an **arbitrary** set of points (will see later today)

What to think of it?

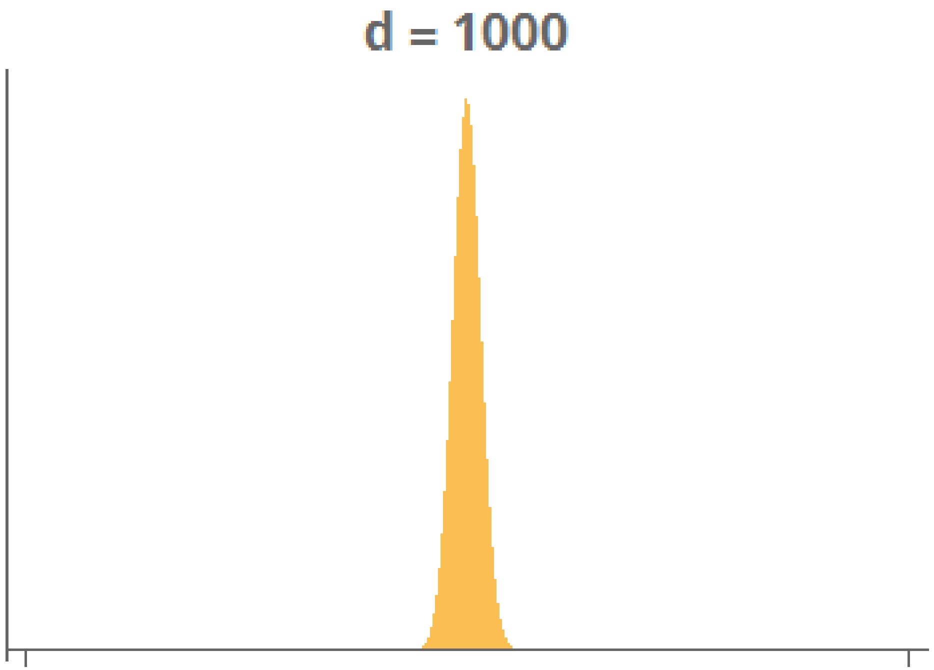
- Quite counter-intuitive, and might not make sense when you first think about it…

The number of heads

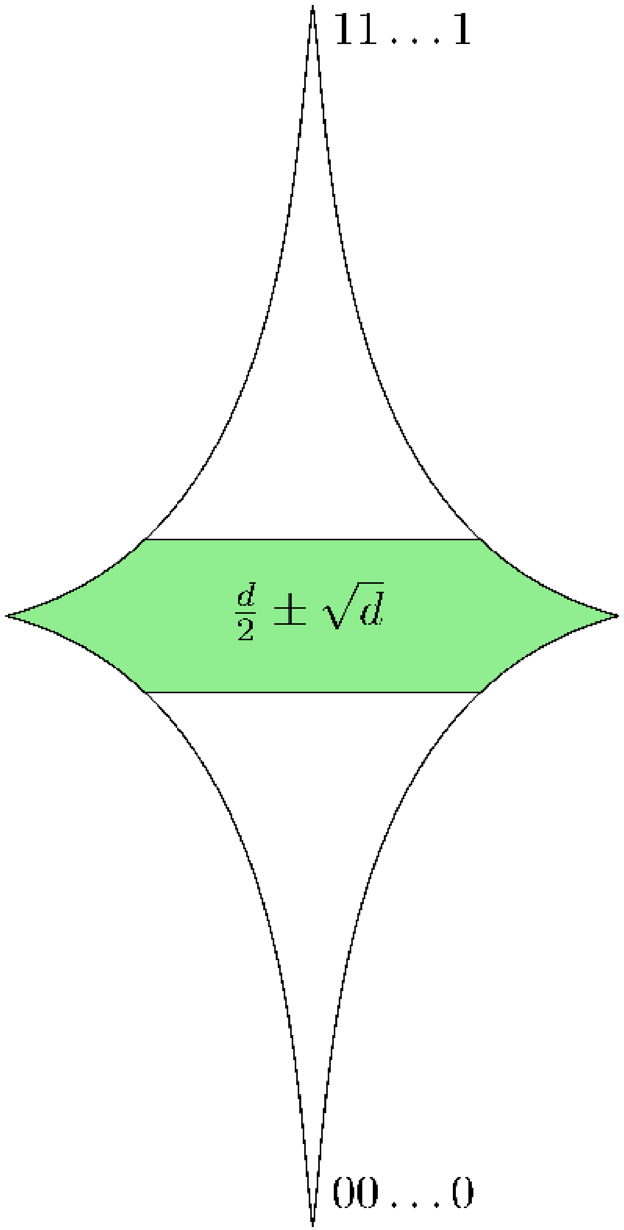
- Toss 1000 fair coins
- What is the number of heads?
- With high probability,

$$\frac{d}{2} \pm O(\sqrt{d})$$

- Enables error-correcting codes etc.



The hypercube



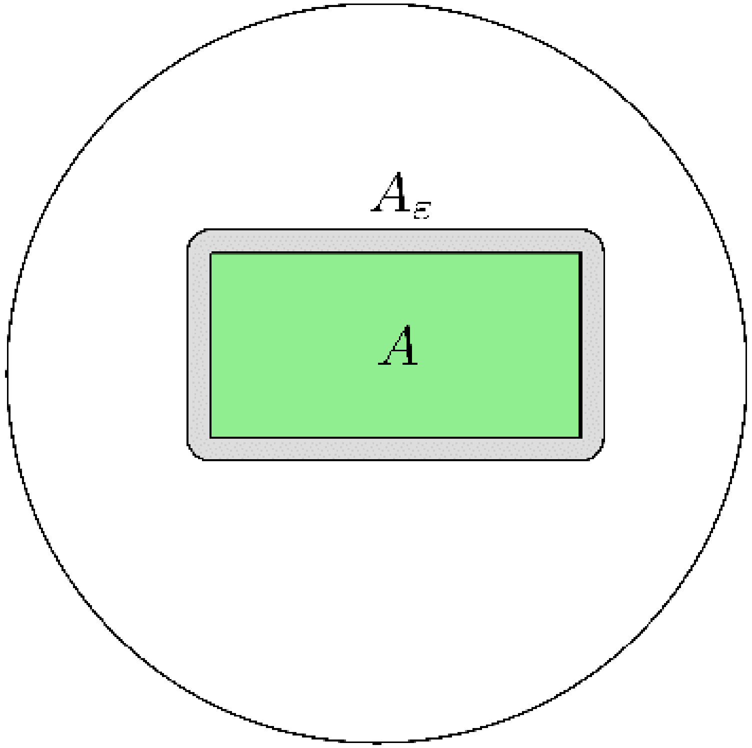
Central limit theorem

- Let $X_1, X_2, \ldots, X_d$ — independent random variables with **zero mean and variance one**
- Then, $\frac{X_1+X_2+\cdots+X_d}{\sqrt{d}} \rightarrow N(0,1)$
- Weak convergence
- Lots of work put into showing similar results, when X_i's are (mildly) dependent
- And understanding the rate of convergence (“finitary” statements)

A bit more advanced material...

Isoperimetric inequality

- Shape in R^d of unit volume with the smallest surface
  - A ball of an appropriate radius
- What if we live on a unit sphere?
- $\mu(A_\varepsilon)$ is minimized for a fixed $\mu(A)$ iff A is a spherical cap of appropriate size **[Levy]**
- **Corollary**: if $\mu(A) = 1/2$, then $\mu(A_\varepsilon) \geq 1 - e^{-\Omega(\varepsilon^2 d)}$



Concentration of Lipschitz functions

- Let $f: S^{d-1} \to R$ such that $|f(x) - f(y)| \leq \|x - y\|$
- Then, f is sharply concentrated around the **median**
- $\mu(f(x) \leq \text{med } f) = 1/2$
- $f(x) > \text{med } f + \varepsilon$ implies x is at distance at least ε
- Use the previous slide to get:
  - $\mu(f(x) > \text{med } f + \varepsilon) \leq e^{-\Omega(\varepsilon^2 d)}$

What to read next?

- Matousek, “Lectures on Discrete Geometry”
- Ball, “An Elementary Introduction to Modern Convex Geometry”
- Milman, Talagrand, …