Summer School on Hashing'14

Dimension Reduction

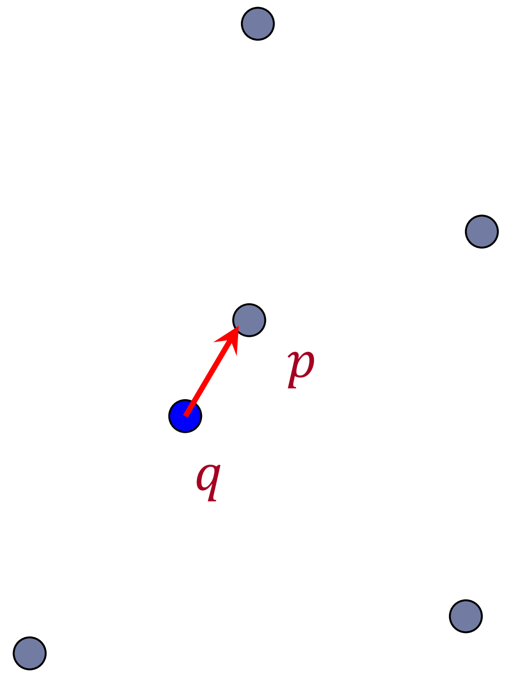
Alex Andoni

(Microsoft Research)

Nearest Neighbor Search (NNS)

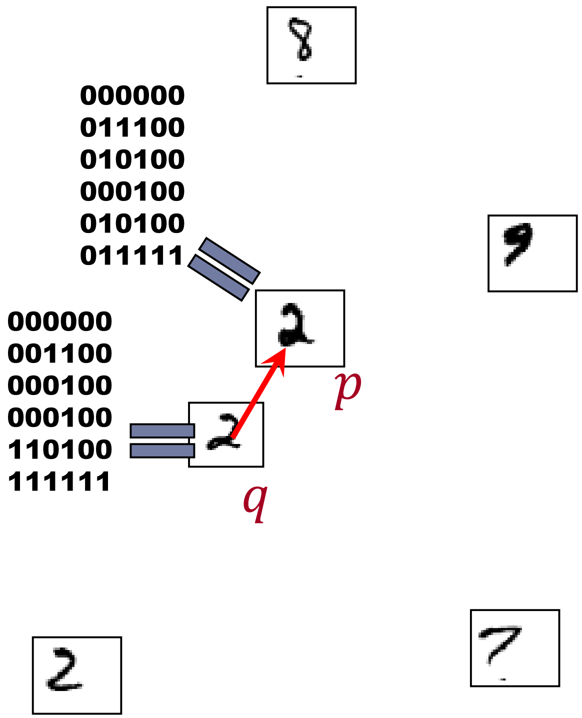
- **Preprocess:** a set D of points

- **Query:** given a query point q, report a point $p \in D$ with the smallest distance to q



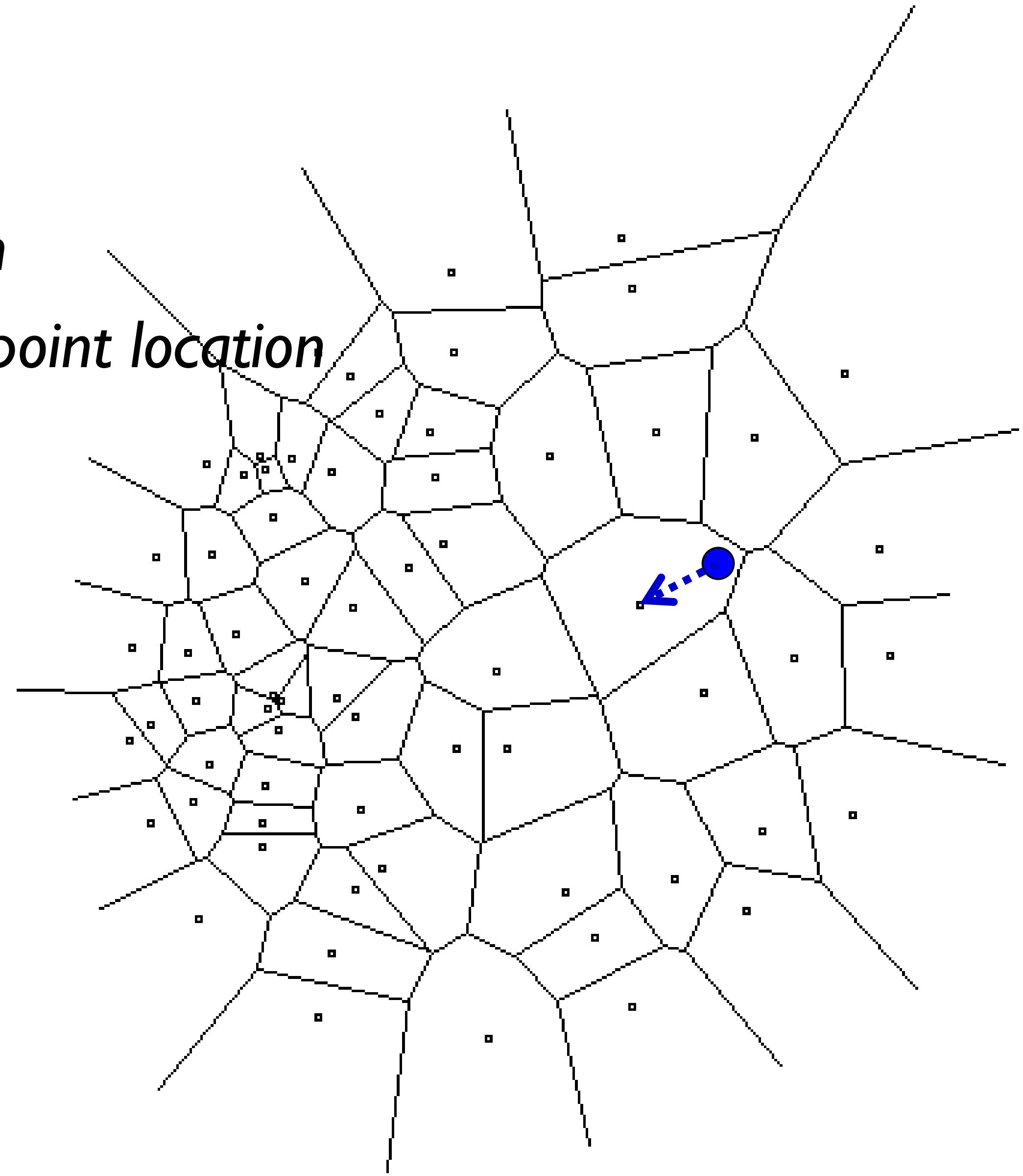
Motivation

- Generic setup:
  - Points model *objects (e.g. images)*
  - Distance models *(dis)similarity measure*
- Application areas:
  - machine learning: k-NN rule
  - speech/image/video/music recognition, vector quantization, bioinformatics, etc…
- Distance can be:
  - Hamming, Euclidean,
    edit distance, Earth-mover distance, etc…
- Primitive for other problems:
  - find the similar pairs in a set D, clustering…



2D case

- Compute *Voronoi diagram*
- Given query q, perform *point location*
- Performance:
  - Space: $O(n)$
  - Query time: $O(\log n)$



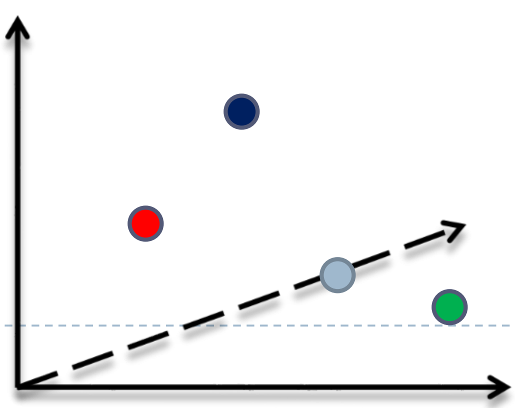
High-dimensional case

- All exact algorithms degrade rapidly with the dimension d

Algorithm	*Query time*	*Space*
Full indexing	$O(\log n \cdot d)$	$n^{O(d)}$ (Voronoi diagram size)
No indexing – linear scan	$O(n \cdot d)$	$O(n \cdot d)$

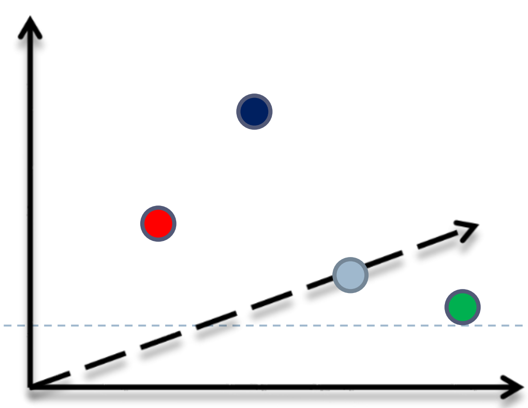
Dimension Reduction

- If high dimension is an issue, reduce it?!
  - "flatten" dimension d into dimension $k \ll d$
- Not possible in general: packing bound
- But can if: for a fixed subset of $\Re^d$
  - Johnson Lindenstrauss Lemma [JL'84]
- Application: NNS in $\Re^d$
  - Trivial scan: $O(n \cdot d)$ query time
  - Reduce to $O(n \cdot k) + T_{dim-red}$ time if preprocess, where $T_{dim-red}$ time to reduce dimension of the query point



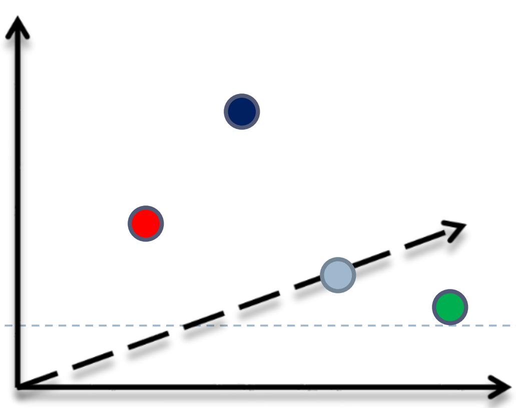
Johnson Lindenstrauss Lemma

- There is a randomized linear map $F: \ell_2^d \to \ell_2^k, k \ll d$, that preserves distance between two vectors x, y
  - up to $1 + \epsilon$ factor:
    $$||x - y|| \leq ||F(x) - F(y)|| \leq (1 + \epsilon) \cdot ||x - y||$$
  - with $1 - e^{-C\epsilon^2 k}$ probability (C some constant)
- Preserves distances among n points for $k = O\left(\frac{\log n}{\epsilon^2}\right)$
- Time to compute map: $T_{dim-red} = O(kd)$

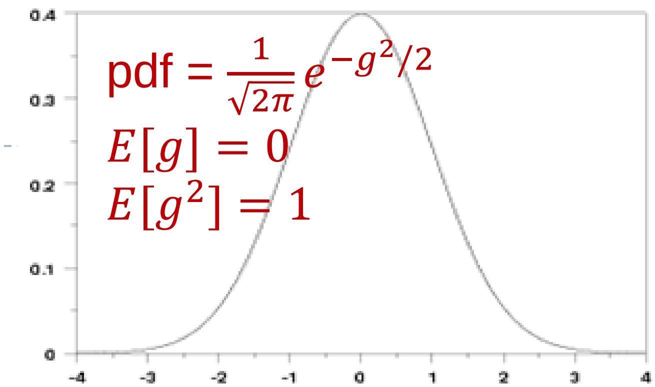


Idea:

- Project onto a *random* subspace of dimension k!

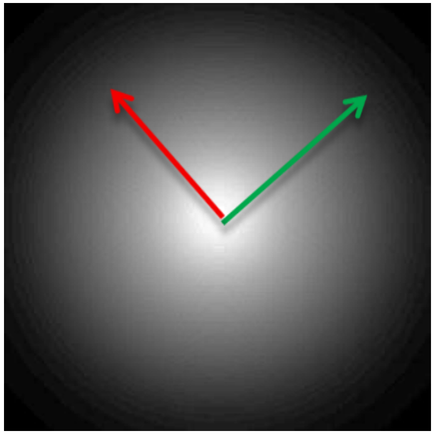


1D embedding



- How about one dimension ($k = 1$) ?
- Map $f: \ell_2^d \to \mathfrak{R}$
  - $f(x) = \sum_i g_i \cdot x_i$,
    - where g_i are iid normal (Gaussian) random variable
- Why Gaussian?
  - Stability property: $\sum_i g_i \cdot x_i$ is distributed as $||x|| \cdot g$, where g is also Gaussian
  - Equivalently: $\langle g_1, \dots, g_d \rangle$ is centrally distributed, i.e., has random direction, and projection on random direction depends only on length of x

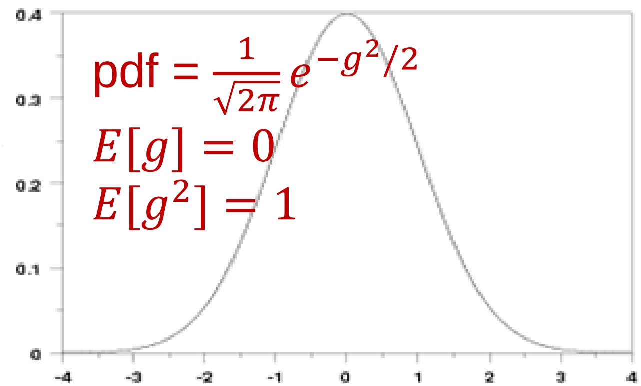
$$P(a) \cdot P(b) =$$
$$= \frac{1}{\sqrt{2\pi}} e^{-a^2/2} \frac{1}{\sqrt{2\pi}} e^{-b^2/2}$$
$$= \frac{1}{2\pi} e^{-(a^2+b^2)/2}$$



1D embedding

- Map $f(x) = \sum_i g_i \cdot x_i$,
  - for any $x, f(x) \sim ||x|| \cdot g$
  - Linear: $f(x) - f(y) = f(x-y)$
- Want: $|f(x) - f(y)| \approx \|x-y\|$
- Ok to consider $z = x - y$ since f linear
  - $|f(z)|^2 \approx \|z\|^2$
- Claim: for any $x, y \in \Re^d$, we have
  - Expectation: $\mathrm{E}[|f(z)|^2] = \|z\|^2$
  - Standard deviation:
    - $\sigma[|(f(z)|^2] = O(\|z\|^2)$
- Proof:
  - Expectation $= \mathrm{E}\left[\left(f(z)\right)^2\right] = \mathrm{E}[\|z\|^2 \cdot g^2]$

$$= \|z\|^2$$



pdf = $\frac{1}{\sqrt{2\pi}} e^{-g^2/2}$

$E[g] = 0$

$E[g^2] = 1$

Full Dimension Reduction

- Just repeat the 1D embedding for k times!
  - $F(x) = (g_1 \cdot x, g_2 \cdot x, \dots g_k \cdot x)/\sqrt{k} = \frac{1}{\sqrt{k}} Gx$
  - where G is $k \times d$ matrix of Gaussian random variables

- Again, want to prove:
  - $\|F(z)\| = (1 \pm \epsilon) * \|z\|$
  - for fixed $z = x - y$
  - with probability $1 - e^{-\Omega(\epsilon^2 k)}$

Concentration

- $F(z)$ is distributed as
  - $\frac{1}{\sqrt{k}}(||z|| \cdot a_1, ||z|| \cdot a_2, \ldots ||z|| \cdot a_k)$
  - where each a_i is distributed as Gaussian
- Norm $||F(z)||^2 = ||z||^2 \cdot \frac{1}{k} \sum_i a_i^2$
  - $\sum_i a_i^2$ is called chi-squared distribution with k degrees
- Fact: chi-squared very well concentrated:
  - Equal to $1 + \epsilon$ with probability $1 - e^{-\Omega(\epsilon^2 k)}$
  - Akin to central limit theorem

Johnson Lindenstrauss: wrap-up

- $F(x) = (g_1 \cdot x, g_2 \cdot x, \ldots g_k \cdot x) / \sqrt{k} = \frac{1}{\sqrt{k}} Gx$
- $||F(x)|| = (1 \pm \epsilon)||x||$ with high probability

- Beyond Gaussians:
  - Can use ± 1 instead of Gaussians [AMS'96, Ach'01, TZ'04…]

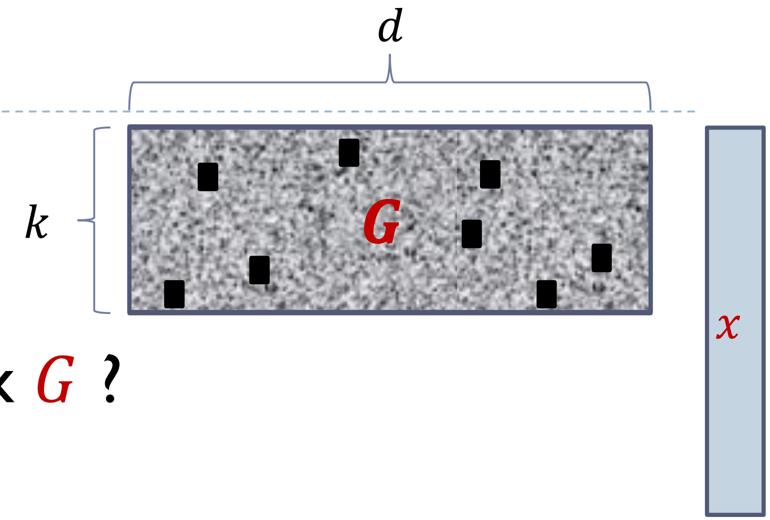
Faster JL ?

- Time: $O(kd)$
  - To compute Gx
  - $O(d+k)$ time ?

- Yes!
  - [AC'06, AL'08'11, DKS'10, KN'12…]

- Will show: $O(d\log d + k^3)$ time [Ailon-Chazelle'06]

Fast JL Transform

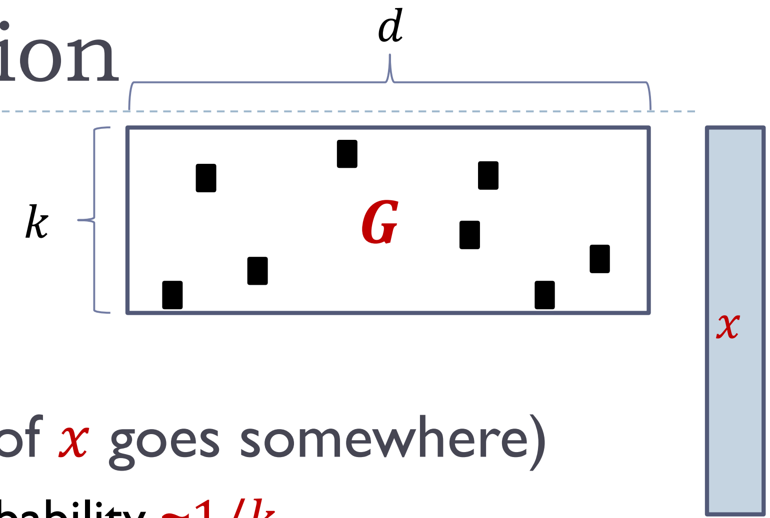


- $z = Gx$
- Costly because G is dense
- Meta-approach: use *sparse* matrix G ?
- Suppose sample s entries/row
- Analysis of one row:
  - $h: [d] \to \{0,1\}$ s.t. $h(i) = 1$ with probability s/d
  - $z_1 = \eta \cdot \sum_{i=1}^{d} h(i) \cdot g_i x_i$
  - Expectation of z_1^2:
  - $E[z_1^2] = \eta^2 \, E\left[\sum_i h(i) g_i^2 \, x_i^2\right] = \eta^2 \cdot \frac{s}{d} \cdot \|x\|^2$
  - What about variance??

Set $\eta = \sqrt{d/s}$

normalization constant

Fast JLT: sparse projection



- Variance of z_1 can be large ☹
  - Bad case: x is sparse
    - think: $x = e_1 - e_2$
  - Even for $s \approx d/k$ (each coordinate of x goes somewhere)
    - two coordinates collide (bad) with probability $\sim 1/k$
    - want exponential in k failure probability
    - really would need $s \approx d$
- But, take away: may work if x is "spread around"
- New plan:
  - "spread around" x
  - use sparse G

FJLT: full

$$z = PHD \cdot x$$

Projection:
sparse matrix

“spreading around”

Hadamard
(Fast Fourier Transform)

Diagonal

- D = matrix with ± 1 r.v. on diagonal
- H= Hadamard matrix:
  - Hx can be computed in time $O(d \cdot \log d)$
  - H composed of $\pm\frac{1}{\sqrt{d}}$ only
- P = sparse matrix as before, size $k' \times d$, with $k' \approx k^2$

Spreading around: intuition

$$z = PHDx$$

“spreading around”

Projection: sparse matrix

Hadamard (Fast Fourier Transform)

Diagonal

- $y = HDx$
- Idea for Hadamard/Fourier Transform:
  - “Uncertainty principle”: if the original x is sparse, then the transform is dense!
  - Though can “break” x’s that are already dense

Spreading around: proof

- $y = HDx$
- Suppose $||x|| = 1$
- Ideal spreading around:
  - $y_i = \pm 1/\sqrt{d}$
- **Lemma:** $y_i^2 \leq O\left(\log \frac{1}{\delta}\right) \cdot 1/d$ with probability at least $1 - \delta$, for each coordinate i
- Proof:
  - $y_i = H_i D x = g x$
    - for some $\pm \frac{1}{\sqrt{d}}$ vector $g = H_i D$
  - Hence y_i is approx. $\frac{1}{\sqrt{d}} \times$Gaussian (in fact, a bit better)
  - Hence $y_i^2 \leq O\left(\log \frac{1}{\delta}\right) \cdot 1/d$ with probability at least $1 - \delta$

Why projection P?

$$z = PHDx$$

- Why aren't we done?
  - choose first few coordinates of $y = HDx$
  - each has same distribution: $||x|| \times$ gaussian
  - Issue: $y_1, y_2, \ldots$ are not independent
- Nevertheless:
  - $||y|| = ||x||$ since H is a change of basis (rotation in $\mathfrak{R}^d$)

Projection P

$$z = PHDx$$

- Have: $y = HDx$
  - $m = \max y_i^2 \le O\left(\log\frac{1}{\delta}\right) \cdot 1/d$ with probability $1 - d\delta$
- P = projection onto just k' random coordinates!
  - $s = 1$
- Proof: standard concentration
  - $y_1^2 + y_2^2 + \cdots + y_d^2 = ||x||^2 = 1$
  - Chernoff: enough to sample $O\left(dm \cdot \frac{1}{\epsilon^2} \cdot \log\frac{1}{\delta}\right)$ terms for $1 + \epsilon$ approximation
  - Hence $k' = O\left(\frac{1}{\epsilon^2} \cdot \log^2\frac{1}{\delta}\right)$ suffices

FJLT: wrap-up

$$z = PHDx$$

- Obtain:
  - $||z||^2 = (1 \pm \epsilon)||x||^2$ with probability at least $1 - 2d\delta$
  - dimension of z is $k' = O\left(\frac{1}{\epsilon^2} \cdot \log^2 \frac{1}{\delta}\right)$
  - time: $O(d \log d + k')$
- Dimension not optimal: apply regular (dense) JL on z to reduce further to $k = O\left(\frac{1}{\epsilon^2} \cdot \log \frac{1}{\delta}\right)$
- Final time: $O(d \log d + k^3)$
- [AC'06, AL'08'11, WDLSA'09, DKS'10, KN'12, BN, NPW'14]

Dimension Reduction: beyond Euclidean

- Johnson-Lindenstrauss: for Euclidean space
  - $O_\epsilon(\log n)$ dimension, oblivious

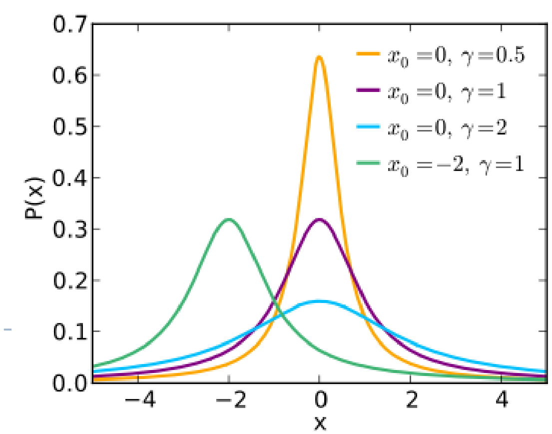
- Other norms, such as ℓ_1?
  - Essentially no: [CS'02, BC'03, LN'04, JN'10…]
  - For n points, D approximation: between $n^{\Omega(1/D^2)}$ and $O(n/D)$ [BC03, NR10, ANN10…]
    - even if map depends on the dataset!
  - But can do *weak* dimension reduction

Towards dimension reduction for ℓ_1

- Can we do the "analog" of Euclidean projections?
- For ℓ_2, we used: Gaussian distribution
  - has stability property:
  - $g_1 x_1 + g_2 x_2 + \cdots g_d x_d$ is distributed as $g \cdot ||x||$
- Is there something similar for 1-norm?
  - Yes: Cauchy distribution!
  - 1-stable:
  - $c_1 x_1 + c_2 x_2 + \cdots c_d x_d$ is distributed as $c \cdot ||x||_1$

$$pdf(s) = \frac{1}{\pi(s^2+1)}$$

- What's wrong then?
  - Cauchy are **heavy-tailed…**
  - doesn't even have finite expectation



Weak embedding [Indyk'00]

- Still, can consider map as before
  - $f(x) = (c_1 x, c_2 x, \ldots, c_k x)$
  - Each coordinate distributed as $||x||_1$ ×Cauchy
  - $||f(x)||_1$ does not concentrate at all, but…
- Can estimate $||x||_1$ by:
  - Median of absolute values of coordinates!
  - Concentrates because abs($||x||_1$ ×Cauchy) is in the correct range for 90% of the time!
- Gives a *sketch*
  - OK for nearest neighbor search