Random hyperplane tessellations and dimension reduction

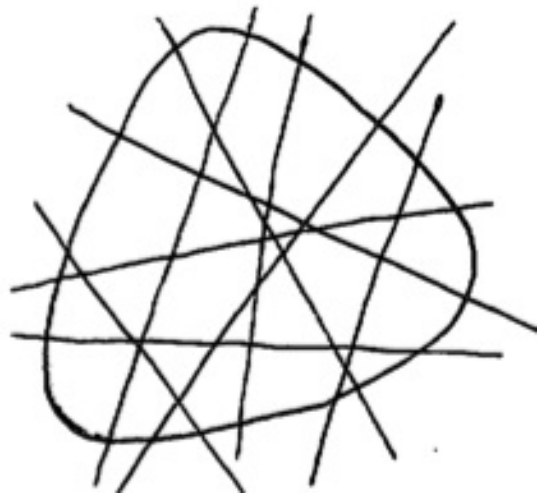
Roman Vershynin

University of Michigan, Department of Mathematics

Phenomena in high dimensions in geometric analysis, random matrices and computational geometry
Roscoff, France
June 25, 2012

Joint work with Yaniv Plan, University of Michigan, Mathematics

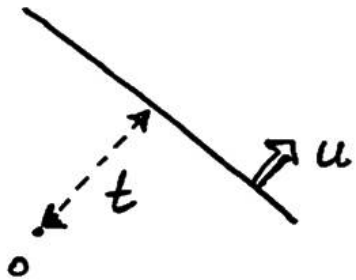
Cutting Problem. How many hyperplanes are needed to cut a given set $K \subset \mathbb{R}^n$ into small pieces, say of diameter $\leq \delta$?

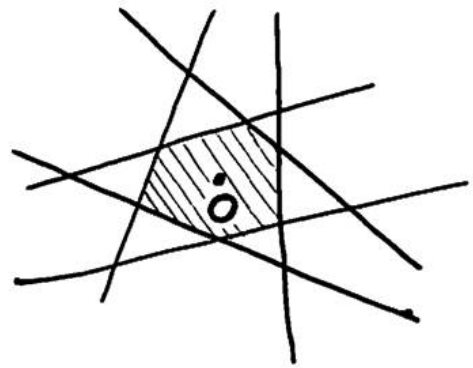


The problem is non-trivial even for $K = S^{n-1}$ and for fixed $\delta = 0.1$.

Random hyperplanes $\longrightarrow$ **stochastic geometry**:

Random hyperplane tessellation $=$ Poisson process of hyperplanes in $\mathbb{R}^n$ with random directions $u \in S^{n-1}$, random shifts $t \in \mathbb{R}_+$.



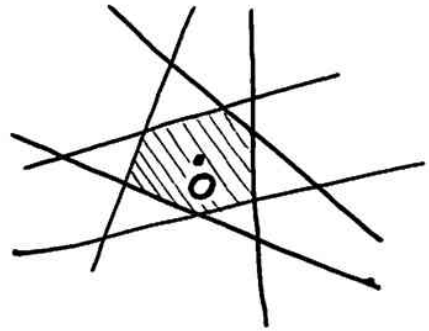


Main interests:

(a) the **shape** of the typical (or a given) cell, e.g. the zero cell;

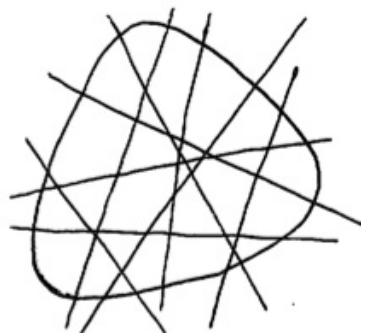
(b) the **statistics** of the cells, e.g. how many have large volume.

Example: Kendall's conjecture (1940's). Consider a stationary, isotropic Poisson process of hyperplanes in $\mathbb{R}^n$. If the volume of the zero cell $\to 0$, then the zero cell $\to$ round ball.



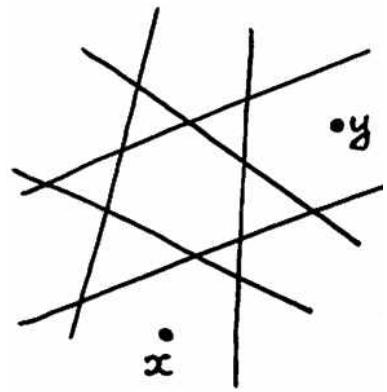
Proved by Kovalenko'97 ($n = 2$), Hug, Reitzner, Schneider '04 ($n \geq 2$).

Does **not** answer the cutting problem – need to control **all cells uniformly**.



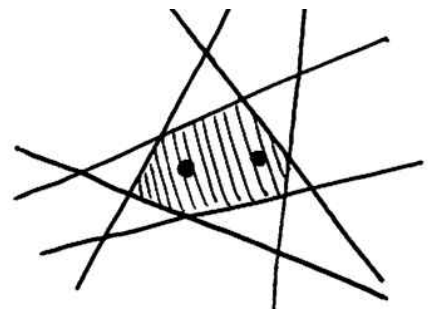
We may ask for an even stronger property than small cells – that the hyperplanes **cut** K **as evenly as possible**.

Definition. A hyperplane tessellation of K is a uniform tessellation if the fraction of the hyperplanes separating $\forall x, y \in K$ is proportional ($\pm\delta$) to the Euclidean distance between x, y.



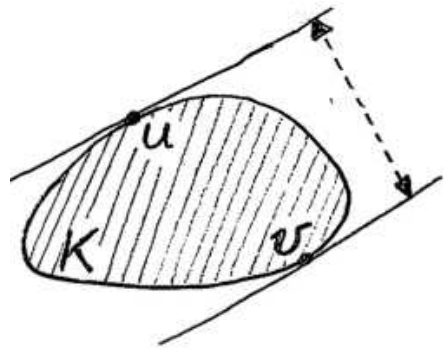
Problem. **How many** hyperplanes are needed to make a δ-uniform tessellation of K?

This is stronger than the cutting problem: all **cells** of a δ-uniform tessellation have diameter $\leq \delta$.



The answer will depend on the Gaussian mean width of K:

$$w(K) = \mathbb{E}\left[\sup_{u\in K}\langle g, u\rangle - \inf_{v\in K}\langle g, v\rangle\right] = \mathbb{E}\sup_{x\in K-K}\langle g, x\rangle$$



Examples of mean widths of sets $K \subseteq S^{n-1}$:

(a) $K = S^{n-1} \Rightarrow w(K) \sim \sqrt{n}$.

(b) $\dim(K) = k \Rightarrow w(K) \lesssim \sqrt{k}$.

(c) K finite $\Rightarrow w(K) \lesssim \sqrt{\log|K|}$.

Hence: $w(K)^2 \approx$ effective dimension. It is $\lesssim n$; often $\ll n$.

Theorem. Let $K \subseteq S^{n-1}$ and $\delta > 0$. Let

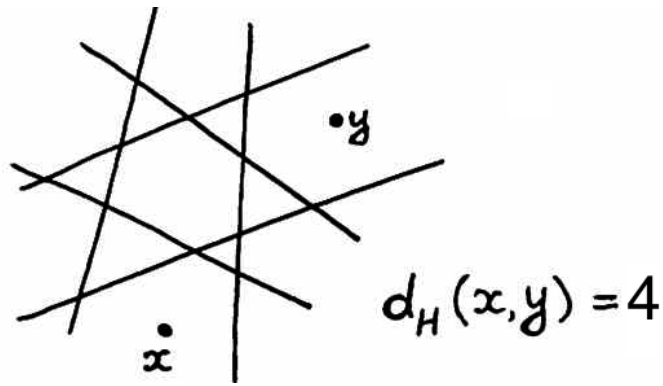
$$m \geq C\delta^{-6} w(K)^2.$$

Then m independent random hyperplanes (uniform in the Grassmanian) form a δ-uniform tessellation of K with high probability $(1 - e^{-c\delta^2 m})$.

Precisely, a "δ-uniform tessellation" means:

$$\left| \frac{1}{m} d_H(x, y) - \frac{1}{\pi} d(x, y) \right| \leq \delta \qquad \forall x, y \in K$$

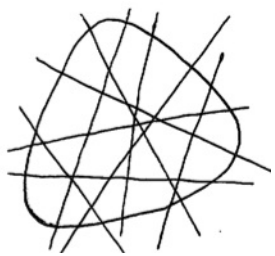
where $d_H(x, y) = \#$ separating hyperplanes, $d(x, y) =$ geodesic distance.



Theorem. Let $K \subseteq S^{n-1}$. Then $m \sim \delta^{-6} w(K)^2$ random hyperplanes form a δ-uniform tessellation of K with high probability.

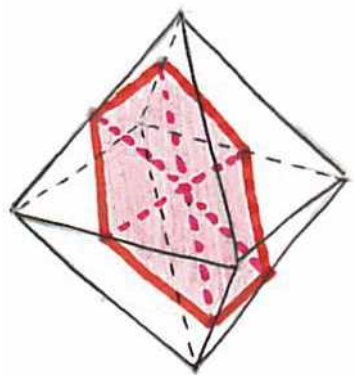
Corollary (Cutting). These hyperplanes cut K into pieces bounded by δ in diameter.

Corollary. For $K = S^{n-1}$, $m \sim n$ hyperplanes suffice.



Corollary. $m \sim \delta^{-4} n$ random hyperplanes cut S^{n-1} into pieces of diameter $\leq \delta$.

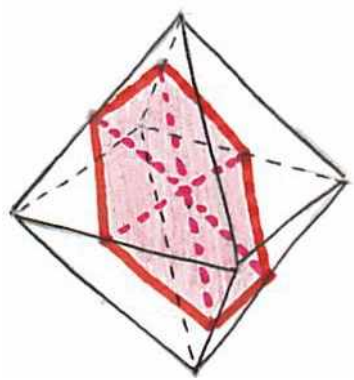
Direct proof of Cutting Corollary
from **Dvoretzky-Milman Theorem** for ℓ_1^n:



D-M Theorem. A random n-dimensional section of the unit ball of ℓ_1^m is ε-close to the round ball, for $m \sim \varepsilon^{-2} n$.

Proof of Corollary.
The D-M section is an n-dimensional polytope defined by m hyperplanes.
The polytope is ε-close to the ball $\Rightarrow$ all faces have diameter $\leq \varepsilon^{1/2} =: \delta$.
This gives $m \sim \delta^{-4} n$. $\square$

Corollary. $m \sim \delta^{-4} n$ random hyperplanes cut S^{n-1} into pieces of diameter $\leq \delta$.



The argument seems tight. The unusual dependence δ^{-4} can be optimal.

However: the optimal dependence in Dvoretzky-Milman theorem is unknown.

This argument can be generalized for cutting arbitrary $K \subseteq S^{n-1}$.
But it **can not** prove the full theorem (for uniform tessellations):

Theorem. Let $K \subseteq S^{n-1}$. Then $m \sim \delta^{-6} w(K)^2$ random hyperplanes form a δ-uniform tessellation of K with high probability.

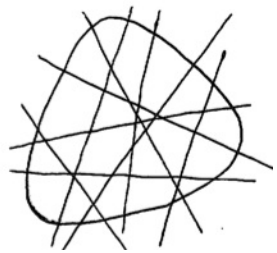
Difficulty here: **discontinuity** of the Hamming distance.
Naive ε-net arguments fail.
A way around: **soft** Hamming distance.

Remarks on the Theorem:

(a) m is always linear in the effective dimension of K.

(b) Dependence on $w(K)$ is optimal.

(c) Dependence on δ is **not** optimal; conjectured δ^{-4}.



Theorem. Let $K \subseteq S^{n-1}$. Then $m \sim \delta^{-6} w(K)^2$ random hyperplanes form a δ-uniform tessellation of K with high probability.

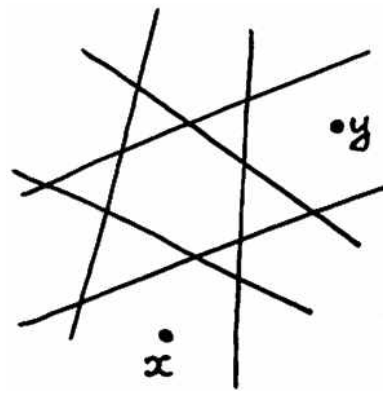
Application to dimension reduction

For $x \in K$, record the orientations of x with respect to the m hyperplanes:

$$\Phi(x) = \text{the vector of orientations} \in \{-1, 1\}^m.$$

Key: # of separating hyperplanes = Hamming distance in $\{-1, 1\}^m$:

$$\text{dist}(\Phi(x), \Phi(y)) = d_H(x, y).$$

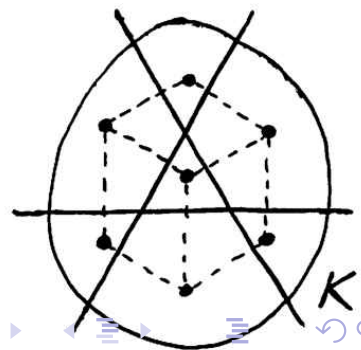


The conclusion of the Theorem is:

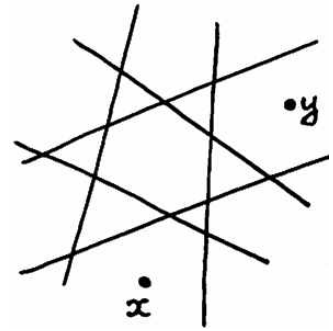
$$\left| \frac{1}{m} d_H(x, y) - \frac{1}{\pi} d(x, y) \right| \le \delta \qquad \forall x, y \in K$$

This means that the map $\Phi : K \hookrightarrow \{-1, 1\}^m$
is an almost isometric embedding (δ-almost isometry).

Visualize this embedding $\longrightarrow$



Application to dimension reduction



Oriented hyperplane $\leftrightarrow$ normal $a \in \mathbb{R}^n$.

Random hyperplane $\leftrightarrow$ random normal $a \sim N(0, I_n)$.

m random hyperplanes $\leftrightarrow$ $m \times n$ Gaussian random matrix

$$A = \begin{bmatrix} \cdots & a_1 & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & a_m & \cdots \end{bmatrix}$$

Vector of orientations of $x \in \mathbb{R}^n$:

$$\Phi(x) = (\text{sign}\langle a_i, x\rangle)_{i=1}^m = \text{sign}(Ax).$$

Summarizing: Φ is a coarsely **quantized linear map**.

Restate:

Theorem. Let $K \subseteq S^{n-1}$ and $m \sim \delta^{-6} w(K)^2$. Consider an $m \times n$ Gaussian random matrix A. Then the quantized map

$$\Phi(x) = \text{sign}(Ax)$$

provides a δ-isometric embedding of K into the Hamming cube $\{-1, 1\}^m$.

Compare with **Johnson-Lindenstrauss Lemma**:
A itself is an nearly isometric embedding $K \hookrightarrow \ell_2^m$ and ℓ_1^m.

(For finite K, $m \sim \log |K|$, so this is the original J-L Lemma.
For general K, proved by Klartag-Mendelson'05, Schechtman'06.)

Theorem = **J-L Lemma with extreme quantization**.

Dimension of K gets reduced to its effective dimension $m \sim w(K)^2$.
JL Lemma achieves this by **projecting** K;
Theorem achieves this by **cutting** K.

$\Phi : K \hookrightarrow \{-1, 1\}^m$, $m \sim w(K)^2$ is an almost isometric embedding.

No universality.

$\Phi(x) = \text{sign}(Ax)$ works for a random **Gaussian** matrix A.
But this fails for a random **Bernoulli** matrix A:

$$x = (1, \frac{1}{2}, 0, \ldots, 0), \quad y = (1, -\frac{1}{2}, 0, \ldots, 0) \qquad \Rightarrow \qquad \Phi(x) = \Phi(y)$$

for **any** target dimension m, even though x and y are not close.

Data recovery

Problem. Estimate $x \in K$ from $y = \Phi(x) = \text{sign}(Ax) \in \{-1, 1\}^m$.

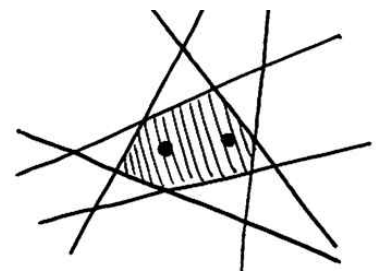
Suppose K is convex.
(If not, pass to $\text{conv}(K)$; the mean width $w(K)$ won't change.)

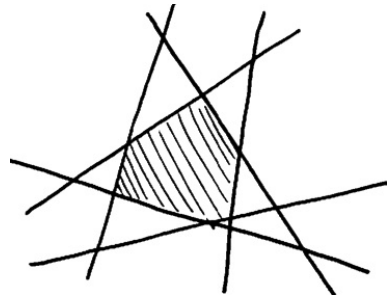
Proposition (Recovery). One can accurately estimate $x \in K$ from $y = \Phi(x)$ by solving the *convex feasibility program*

$$\text{Find } x' \in K \text{ subject to } y = \Phi(x').$$

Indeed, the solution satisfies $\|x' - x\|_2 \le \delta$.

Proof. The feasible set of the program is some **cell** of K.
Both x and x' belong to this cell.
But as we know, all cells have **diameter** $\le \delta$. Q.E.D. □





Drawback: the recovery is **not robust**. Flip a few bits of y, get an infeasible program.

Still, robust recovery is possible:

Robust Recovery Problem. Estimate $x \in K$ from $y = \text{sign}(Ax) \in \{-1, 1\}^m$ after some proportion of bits of y are corrupted (flipped).

Answer: maximize **correlation** with the data:

$$\max \langle Ax', y \rangle \quad \text{subject to} \quad x' \in K.$$

K convex $\Rightarrow$ *convex program*.

Theorem (Robust recovery). Let $x \in K$ and $y = \text{sign}(Ax)$. We corrupt τm bits of y, getting $\widetilde{y}$. One can still accurately estimate x from $\widetilde{y}$ by solving the convex program above (with $\widetilde{y}$). Indeed, the solution satisfies

$$\|x' - x\|_2 \lesssim \delta + \tau \log(1/\tau).$$

Proof is based on the full power of the Theorem on uniform tessellations. □

Applications in sparse recovery

Suppose the unknown signal x is sparse: few non-zeros, $\|x\|_0 \leq s$.
Signal set:

$$S_{n,s} = \{x \in \mathbb{R}^n : \ \|x\|_2 \leq 1, \ \|x\|_0 \leq s\}.$$

But $S_{n,s}$ is not convex. Convexify:

$$\text{conv}(S_{n,s}) \approx K_{n,s} = \{x \in \mathbb{R}^n : \ \|x\|_2 \leq 1, \ \|x\|_1 \leq \sqrt{s}\}.$$

$K_{n,s} = \{\text{approximately sparse vectors}\}$.

Simple computation yields:

$$w(K_{n,s}) \sim \sqrt{s \log(n/s)}.$$

Thus we have a **dimension reduction** $K \hookrightarrow \{-1, 1\}^m$,

$$m \sim w(K)^2 \sim s \log(n/s).$$

Note: m is linear in the sparsity s.

Approx. sparse: $K_{n,s} = \{x \in \mathbb{R}^n : \|x\|_2 \le 1, \ \|x\|_1 \le \sqrt{s}\}$. $\quad m \sim w(K)^2 \sim s \log(n/s)$.

Specialize the Robust Recovery Theorem to the sparse case:

Corollary (Sparse recovery). Let x be approximately sparse: $x \in K_{n,s}$, and let $y = \text{sign}(Ax) \in \{-1, 1\}^m$ where

$$m \sim_\delta s \log(n/s).$$

We can estimate x from y by solving the *convex program*

$$\max \langle Ax', y \rangle \quad \text{subject to} \quad x' \in K.$$

Indeed, the solution satisfies $\|x' - x\|_2 \le \delta$.

The recovery is robust as before (i.e. one can flip bits of y).

Single-bit compressed sensing

- Traditional compressed sensing: recover an s-sparse signal $x \in \mathbb{R}^n$ from m linear measurements given by $y = Ax \in \mathbb{R}^m$.
  Available results: recovery by convex programming, $m \sim s \log(n/s)$.
- Single-bit compressed sensing: recover an s-sparse signal $x \in \mathbb{R}^n$ from m single-bit measurements given by $y = \text{sign}(Ax) \in \{-1, 1\}^m$.
  An extreme way of measurement quantization, A/D conversion.
- [Boufounos-Baraniuk'08] formulated single-bit CS, connections with embeddings into the Hamming cube, algorithms.
  +[Gupta-Nowak-Recht'10, Jacques-Laska-Boufounos-Baraniuk'11]
- No **tractable algorithms** have been known (unless x has constant dynamic range, or for adaptive measurements).
- Present work: robust sparse recovery via **convex programming**.

Applications in Statistics: sparse binomial regression

Our model of m one-bit measurements was

$$y_i = \text{sign}\langle a_i, x\rangle, \ i = 1, \dots, m. \qquad (a_i \in N(0, I_n))$$

More general stochastic model: $y_i = \pm 1$ r.v's independent given $\{a_i\}$,

$$\mathbb{E}\, y_i = \theta(\langle a_i, x\rangle), \ i = 1, \dots, m.$$

Here $\theta(u)$ is some function satisfying the **correlation assumption:**

$$\mathbb{E}\, \theta(g)g =: \lambda > 0 \quad g \sim N(0, 1).$$

Reason: ensures *positive correlation with data*. Since $\langle a_i, x\rangle \sim N(0, 1)$,

$$\mathbb{E}\, y_i\langle a_i, x\rangle = \mathbb{E}\, \theta(g)g =: \lambda.$$

This is the generalized linear model (GLM) with link function θ^{-1}.

Example. $\theta(z) = \tanh(z/2)$: logistic regression

$$\mathbb{P}\{y_i = 1\} = f(\langle a_i, x\rangle), \qquad f(z) = \frac{e^z}{e^z + 1}.$$

Statistical notation: x = coefficient vector (β) (unknown), y_i = binary response variables (known), a_i = independent variables (x_i) (known);

$$\mathbb{P}\{y_i = 1\} = f\Big(\sum_{j=1}^{n} \beta_j x_{ij}\Big), \quad i = 1, \ldots, m$$

Recent work on sparse logistic regression:
[Negahban-Ravikumar-Wainwright-Yu'11, Bunea'08, Van De Geer'08, Bach'10, Ravikumar-Wainwright-Lafferty'10, Meier-Van De Geer-Bühlmann'08, Kakade-Shamir-Sridharan-Tewari'11]

Theorem (Sparse binomial regression). Suppose we have a GLM where the coefficient vector $x \in \mathbb{R}^n$, $\|x\|_2 = 1$ is approximately s-sparse, $x \in K_{n,s}$. If the sample size is

$$m \sim_\delta s \log(n/s)$$

then we can estimate x by solving the **convex program**

$$\max \sum_{i=1}^{m} y_i \langle a_i, x' \rangle \quad \text{subject to} \quad x' \in K_{n,s}.$$

Indeed, the solution satisfies $\|x' - x\|_2 \le \sqrt{\delta/\lambda}$ w.h.p.

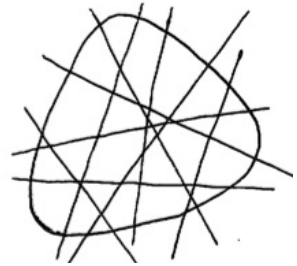
In statistics notation, the sample size is $n \sim s \log(p/s)$, thus $n \ll p$.

New, unusual feature? The knowledge of the link function θ is not needed (unlike in max-likelihood approaches). Here the form of GLM may be unknown. The solution is non-parametric.

Summary:

JL lemma: dimension reduction $K \hookrightarrow \mathbb{R}^m$ by projecting K onto m-dimensional subspace.

Alternative way: dimension reduction $K \hookrightarrow \{-1, 1\}^m$ is by cutting K into small pieces by m hyperplanes.



Dimension reduction map:

$$y = \text{sign}(Ax) \quad \text{where } A \text{ is an } m \times n \text{ random Gaussian matrix.}$$

Target dimension $m \sim w(K)^2 \sim$ the effective dimension of K.
If $K = \{\text{approximately } s\text{-sparse vectors}\}$, then $m \sim s \log(n/s)$.

One can accurately and robustly estimate x from y by a convex program.

More generally, one can accurately estimate a sparse solution to GLM $\mathbb{E}\, y = \theta(Ax)$, and without even knowing the link function θ.