The Johnson-Lindenstrauss Lemma for Linear and Integer Feasibility

Leo Liberti, Vu Khac Ky, Pierre-Louis Poirion
CNRS LIX Ecole Polytechnique, France

Rutgers University, 151118

The gist

- I want to solve huge LPs in standard form

$$\min\{c^\top x \mid Ax = b \wedge x \geq 0\}$$

- I am prepared to accept approximate solutions

- I want to randomly project the rows of $Ax = b \wedge x \geq 0$ to a lower dimensional space, while keeping the accuracy with high probability

- Then I can use the projected LP as a feasibility oracle

- *And I'd also like to do the same for ILPs*

Linear Membership Problems

Restricted Linear Membership

Restricted Linear Membership (RLM)
Given vectors $A_1, \dots, A_n, b \in \mathbb{R}^m$ and $X \subseteq \mathbb{R}^n$, is there $x \in X$ s.t.

$$b = \sum_{i \le n} x_i A_i \ ?$$

RLM is a fundamental problem, which subsumes:

- **Linear Feasibility Problem** (LFP)
  Given an $m \times n$ matrix A and $b \in \mathbb{R}^m$, is there an $x \in \mathbb{R}^n_+$ s.t. $Ax = b$?

- **Integer Feasibility Problem** (IFP)
  Given an $m \times n$ matrix A and $b \in \mathbb{R}^m$, is there an $x \in \mathbb{Z}^n_+$ s.t. $Ax = b$?

- Efficient solution methods for LFP/IFP directly yield fast bisection algorithms for (bounded) **Linear Programming** (LP) and **Integer Linear Programming** (ILP)

"Big data" RLM instances

- Most efficient deterministic methods:
  LFP: simplex method, ellipsoid algorithm
  IFP: SAT solution algorithms, Constraint Programming

- slow if m, n too large
  guarantees are useless if data not accurate/correct

- trade *guarantee* off for *efficiency*: randomized algorithms

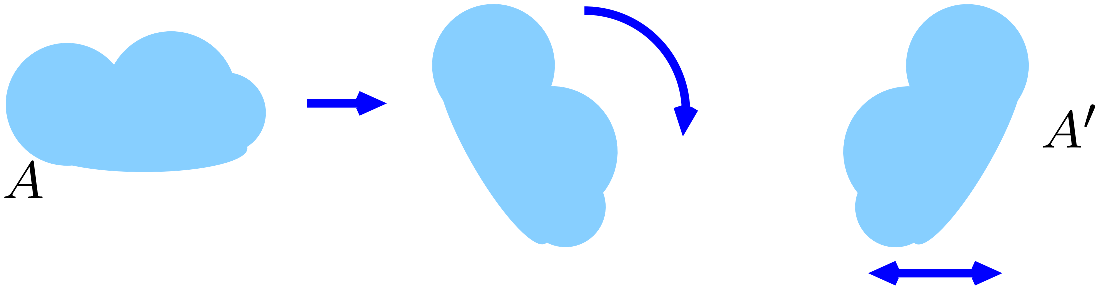
- The approach: decrease m (i.e., lose some dimensions)
  make sure geometry of projected prob. is similar to original

The shape of a cloud of points

- RLM: Is b a *restricted linear combination* of pts $A_1, \dots, A_n \in \mathbb{R}^m$?

- **Lose dimensions but not too much accuracy**
  can we find $k \ll m$ and pts $A'_1, \dots, A'_n \in \mathbb{R}^k$ s.t.
  $A = (A_i \mid i \leq n)$ and $A' = (A'_i \mid i \leq n)$ **"have the same shape"**?

- **What is the shape of a set of points?**



- Approximate congruence: A, A' have almost the same shape if

$$\forall i < j \leq n \quad (1-\varepsilon)\|A_i - A_j\| \leq \|A'_i - A'_j\| \leq (1+\varepsilon)\|A_i - A_j\|$$

  for some small $\varepsilon > 0$

Assume norms are all Euclidean

Losing dimensions in the RLM

Given $X \subseteq \mathbb{R}^n$ and $b, A_1, \ldots, A_n \in \mathbb{R}^m$, find $k \ll m$, $b', A'_1, \ldots, A'_n \in \mathbb{R}^k$ such that:

$$\underbrace{\exists x \in X \;\; b = \sum_{i \le n} x_i A_i}_{\textit{high dimensional}} \quad \text{iff} \quad \underbrace{\exists x \in X \;\; b' = \sum_{i \le n} x_i A'_i}_{\textit{low dimensional}}$$

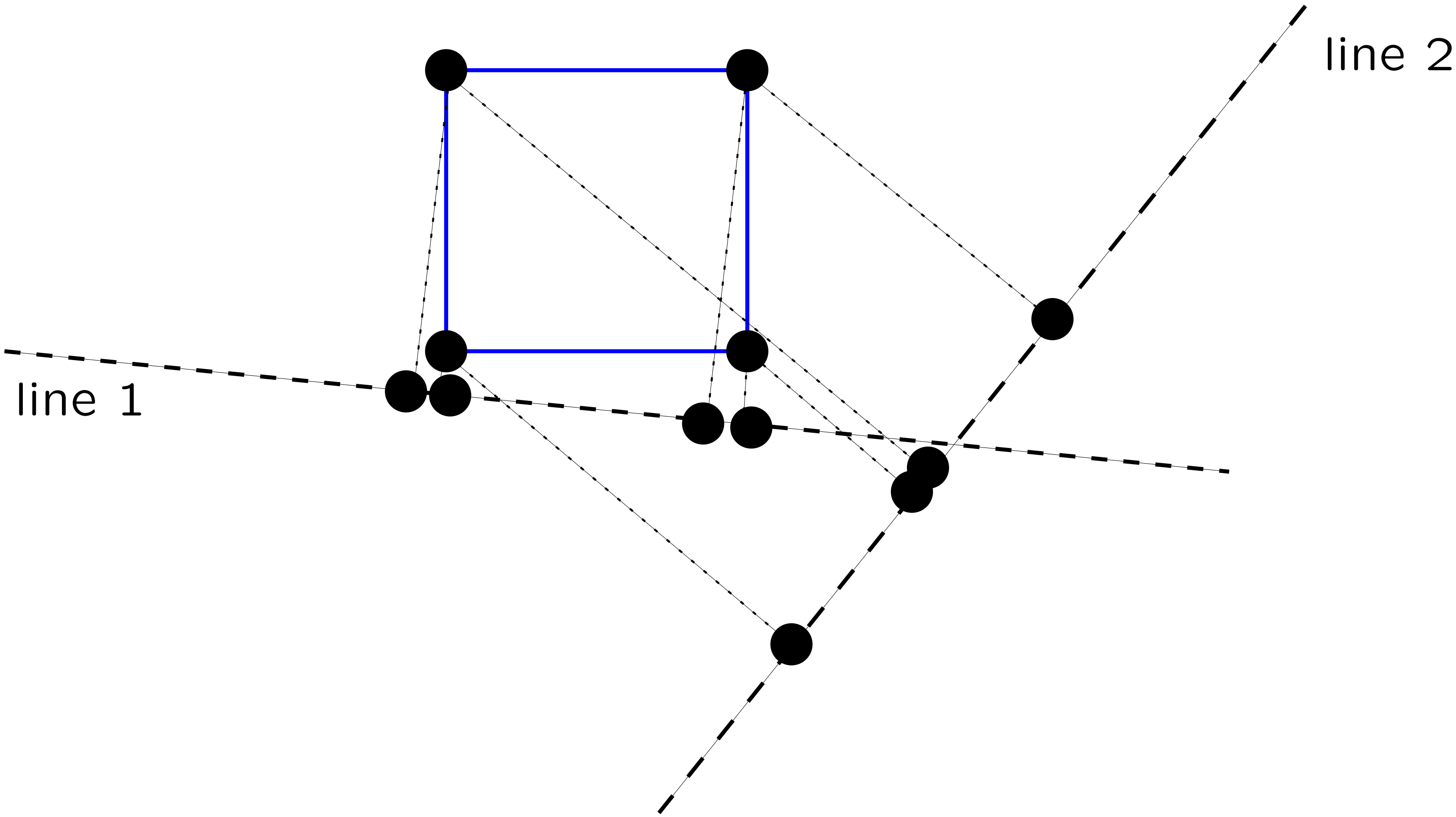
with high probability

- If this is possible, then solve $\mathsf{RLM}_X(b', A')$
- Since $k \ll m$, solving $\mathsf{RLM}_X(b', A')$ should be faster
- $\mathsf{RLM}_X(b', A') = \mathsf{RLM}_X(b, A)$ with high probability

Nothing short of a miracle!

Losing dimensions = "projection"

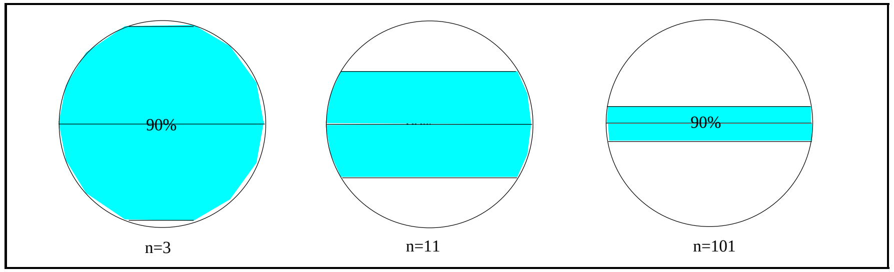
In the plane, hopeless



In 3D: no better

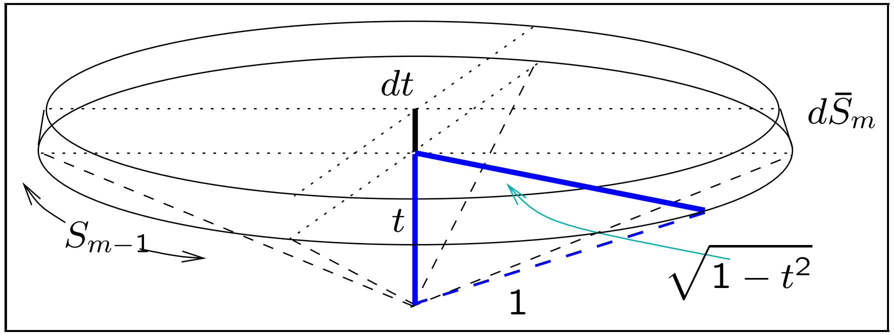
Concentration of measure and the Johnson-Lindenstrauss Lemma

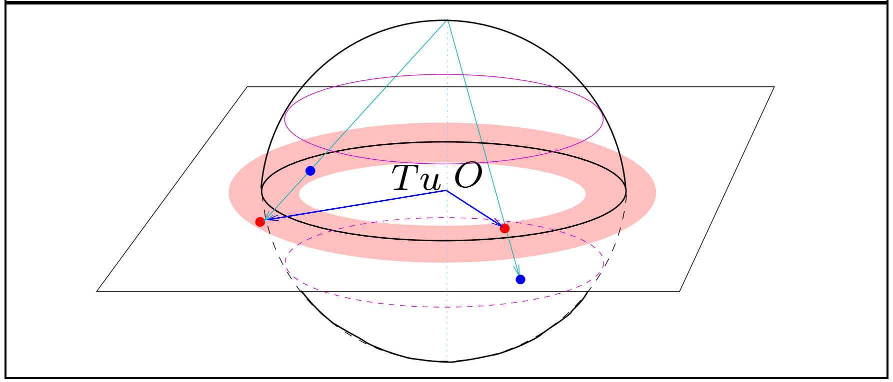
Summary

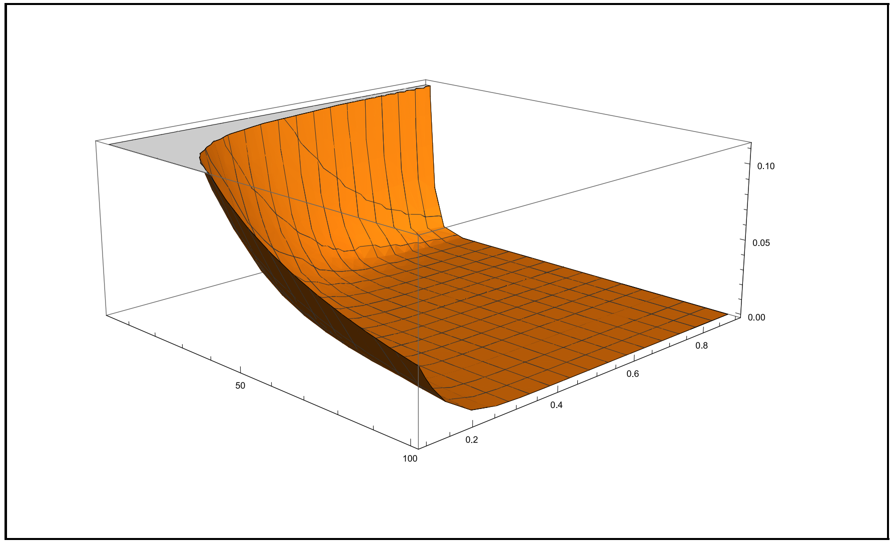


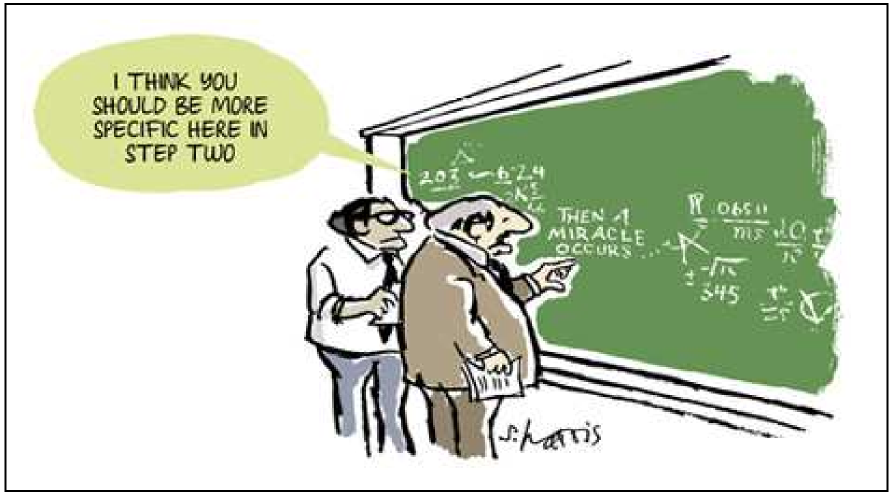
Thm.

Let T be a $k \times m$ rectangular matrix with each component sampled from $N(0, \frac{1}{\sqrt{k}})$, and $u \in \mathbb{R}^m$ s.t. $\|u\| = 1$. Then $\mathsf{E}(\|Tu\|^2) = 1$









Controlling the distortion of $\|Tu\|$

- $B =$ set of unit vectors; by concentration of measure $\forall u \in B\ \exists \mathcal{D}$, const > 0

$$\mathbf{P}(1 - t \leq \|Tu\| \leq 1 + t) \geq 1 - \mathcal{D}e^{-\text{const}\ t^2}$$

- Union bound on B:

$$\mathbf{P}(\forall u \in B\ (1 - t \leq \|Tu\| \leq 1 + t)) \geq 1 - |B|\mathcal{D}e^{-\text{const}\ t^2}$$

- We want nonzero probability: $\Rightarrow |B|\mathcal{D}e^{-\text{const}\ t^2} < 1$

- Set $\sqrt{\text{const}} \times t = \varepsilon\sqrt{k}$: $\Rightarrow |B|\mathcal{D}e^{-\varepsilon^2 k} < 1$ *hiding lotta details here*

- $\Rightarrow k > \varepsilon^{-2} \ln(|B|) +$ (other const)

- $\Rightarrow \exists$ const $\mathcal{C}$ s.t. $\boxed{k > \mathcal{C}\varepsilon^{-2} \ln(|B|)}$

Application to Euclidean distances

- Let $A \subset \mathbb{R}^m$ with $|A| = n$
- $\forall x \neq y \in A$ we have

$$\|Tx - Ty\|^2 = \|T(x-y)\|^2 = \|x-y\|^2 \left\| T\frac{x-y}{\|x-y\|} \right\|^2 = \|x-y\|^2 \|Tu\|^2$$

- $\|u\| = 1$, so by previous results $\boxed{\mathsf{E}(\|Tx - Ty\|^2) = \|x-y\|^2}$
- **Also, T has low distortion with high probability**
- Since $|\{x - y \mid x, y \in A\}|$ is $O(n^2)$, can choose $k > \mathcal{C}\varepsilon^{-2}\ln(n)$

Thm.[*Johnson-Lindenstrauss lemma*]

$\forall \varepsilon \in (0,1)\ \exists$ a $k \times m$ matrix T such that

$$\forall x, y \in A \quad (1-\varepsilon)\|x-y\|^2 \leq \|Tx - Ty\|^2 \leq (1+\varepsilon)\|x-y\|^2 \quad (*)$$

Proof

We showed T has nonzero probability of existing, so $\exists T$ such that $(*)$ holds

JLL properties

- $\forall\, \varepsilon > 0, k \approx \frac{1.8}{\varepsilon^2} \ln n$ [Venkatasubramanian & Wang 2011]
- Can be shown that T must be sampled at worst $O(n)$ times to ensure low distortion "$\forall x, y \in A$"
- But on average $\|Tx - Ty\| = \|x - y\|$, and distortion decreases exponentially with n
- **In most cases, you need only sample once**

We only need a logarithmic number of dimensions
in function of the number of points

Surprising fact:

k is independent of the original number of dimensions m

Many possible random projectors for JLL

- orthogonal proj. on a random k-dim. linear subspace of $\mathbb{R}^m$ *(used in the original proof of JLL)* [Johnson & Lindenstrauss, 1984]

- random $k \times m$ matrices with entries drawn from $N(0, \frac{1}{\sqrt{k}})$ *(modern treatments of JLL)* E.g. [Dasgupta & Gupta, 2003]

- random $k \times m$ matrices with entries in $\{-1, 1\}$ with probability $\frac{1}{2}$; and random $k \times m$ matrices with entries $= -1$ and $= 1$ with probability $\frac{1}{6}$, and 0 with probability $\frac{2}{3}$ *(used for sparse data sets)* [Achlioptas 2003]

- sparser projectors are possible [Dasgupta et al., 2010; Kane & Nelson 2010; Allen-Zhu et al., 2014]

- also faster but *dense*! [Liberty, 2009]

Applying the JLL to the RLM

The main theorem

Thm.

Let $T : \mathbb{R}^m \to \mathbb{R}^k$ be a JLL random projection, and $b, A_1, \ldots, A_n \in \mathbb{R}^m, X$ be a RLM instance. For any given vector $x \in \mathbb{R}^n$, we have:

(i) If $b = \sum_{i=1}^{n} x_i A_i$ then $Tb = \sum_{i=1}^{n} x_i T A_i$;

(ii) If $b \neq \sum_{i=1}^{n} x_i A_i$ then $\mathbf{P}\left(Tb \neq \sum_{i=1}^{n} x_i T A_i\right) \geq 1 - 2e^{-\mathcal{C}k}$;

(iii) If $b \neq \sum_{i=1}^{n} y_i A_i$ for all $y \in X \subseteq \mathbb{R}^n$, where $|X|$ is finite, then

$$\mathbf{P}\left(\forall y \in X \; Tb \neq \sum_{i=1}^{n} y_i T A_i\right) \geq 1 - 2|X|e^{-\mathcal{C}k};$$

for some constant $\mathcal{C} > 0$ (independent of n, k).

[VPL, arXiv:1507.00990v1/math.OC]

Proof (i)

Since T is a matrix, (i) follows by linearity

Proof (ii)

Cor.

$\forall \varepsilon \in (0,1)$ and $z \in \mathbb{R}^m$, there is a constant $\mathcal{C}$ such that

$$\mathbf{P}((1-\varepsilon)\|z\| \leq \|Tz\| \leq (1+\varepsilon)\|z\|) \geq 1 - 2e^{-\mathcal{C}\varepsilon^2 k}$$

Proof

By the JLL

Lemma

If $z \neq 0$, there is a constant $\mathcal{C}$ such that $\mathbf{P}(Tz \neq 0) \geq 1 - 2e^{-\mathcal{C}k}$

Proof

Consider events $\mathscr{A} : Tz \neq 0$ and $\mathscr{B} : (1-\varepsilon)\|z\| \leq \|Tz\| \leq (1+\varepsilon)\|z\|$
$\Rightarrow \mathscr{A}^{\mathsf{c}} \cap \mathscr{B} = \varnothing$, othw $Tz = 0 \Rightarrow (1-\varepsilon)\|z\| \leq \|Tz\| = 0 \Rightarrow z = 0$, contradiction
$\Rightarrow \mathscr{B} \subseteq \mathscr{A} \Rightarrow \mathbf{P}(\mathscr{A}) \geq \mathbf{P}(\mathscr{B}) \geq 1 - e^{-\mathcal{C}\varepsilon^2 k}$ by Corollary
Holds $\forall \varepsilon \in (0,1)$ hence result

Now it suffices to apply the Lemma to $Ax - b$

Proof (iii)

By the union bound on (ii)

$$\mathbf{P}\left(\forall y \in X \ Tb \neq \sum_{i=1}^{n} y_i T A_i\right) = \mathbf{P}\left(\bigcap_{y \in X} \{Tb \neq \sum_{i=1}^{n} y_i T A_i\}\right)$$

$$= 1 - \mathbf{P}\left(\bigcup_{y \in X} \{Tb \neq \sum_{i=1}^{n} y_i T A_i\}^{\mathsf{C}}\right) \geq 1 - \sum_{y \in X} \mathbf{P}\left(\{Tb \neq \sum_{i=1}^{n} y_i T A_i\}^{\mathsf{C}}\right)$$

$$\text{[by (ii)]} \qquad \geq 1 - \sum_{y \in X} 2e^{-\mathcal{C}k} = 1 - 2|X|e^{-\mathcal{C}k}$$

Consequences of the main theorem

- (i) and (ii) allow us to lose dimensions when checking certificates

  given x, with high probability $b = \sum_i x_i A_i \Leftrightarrow Tb = \sum_i x_i T A_i$

- (iii) allows us to lose dimensions when solving IFP whenever $|X|$ is polynomially bounded

  e.g. knapsack set $\{x \in \{0,1\}^n \mid \sum_{i \leq n} \alpha_i x_i \leq d\}$ for a fixed d with $\alpha > 0$

- (iii) also hints as to why LFP is not so easy:
  - LFP $\Leftrightarrow$ Cone Membership
  - $Tb \notin \text{cone}(TA_1, \ldots, TA_n)$ can be written as $\bigcap_{x \in \mathbb{R}^n} \{Tb \neq \sum_{i \leq n} x_i T A_i\}$
  - where $|X| = |\mathbb{R}^n| = 2^{\aleph_0}$, certainly *not* polynomially bounded!

What to do when $|X|$ is superpolynomial

Separating hyperplanes

Project separating hyperplanes instead

- **Convex $C \subseteq \mathbb{R}^m$, $x \notin C$: then $\exists$ hyperplane c separating x, C**

- In particular, true if $C = \text{cone}(A_1, \ldots, A_n)$ for $A \subseteq \mathbb{R}^m$

- **We aim to show $x \in C \Leftrightarrow Tx \in TC$ with high probability**

- As above, if $x \in C$ then $Tx \in TC$ by linearity of T
  real issue is proving the converse

Projecting the separation

Thm.

Given $c, b, A_1, \ldots, A_n \in \mathbb{R}^m$ of unit norm s.t. $b \notin \text{cone}\{A_1, \ldots, A_n\}$ pointed, $\varepsilon > 0$, $c \in \mathbb{R}^m$ s.t. $c^\top b < -\varepsilon$, $c^\top A_i \geq \varepsilon$ $(i \leq n)$, and T a random projector:

$$\mathbf{P}\left[Tb \notin \text{cone}\{TA_1, \ldots, TA_n\}\right] \geq 1 - 4(n+1)e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k}$$

for some constant $\mathcal{C}$.

Proof

Let $\mathscr{A}$ be the event that T approximately preserves $\|c - \chi\|^2$ and $\|c + \chi\|^2$ for all $\chi \in \{b, A_1, \ldots, A_n\}$. Since $\mathscr{A}$ consists of $2(n+1)$ events, by the JLL Corollary (squared version) and the union bound, we get

$$\mathbf{P}(\mathscr{A}) \geq 1 - 4(n+1)e^{-\mathcal{C}(\varepsilon^2 - \varepsilon^3)k}$$

Now consider $\chi = b$

$$\begin{aligned}
\langle Tc, Tb \rangle &= \frac{1}{4}(\|T(c+b)\|^2 - \|T(c-b)\|^2) \\
\text{by JLL} \quad &\leq \frac{1}{4}(\|c+b\|^2 - \|c-b\|^2) + \frac{\varepsilon}{4}(\|c+b\|^2 + \|c-b\|^2) \\
&= c^\top b + \varepsilon < 0
\end{aligned}$$

and similarly $\langle Tc, TA_i \rangle \geq 0$

[VPL, arXiv:1507.00990v1/math.OC]

Consequences of projecting separations

- **Result allows solving LFP by random projection**
  hence, by bisection, also LP
- Probability of being correct depends on ε (the larger the better)
- Largest ε given by $\max\{\varepsilon \geq 0 \mid c^\top b \leq -\varepsilon \wedge \forall i \leq n\ (c^\top A_i \geq \varepsilon)\}$
  an LP, so this does not help us much
- **If** $\mathrm{cone}(A_1, \ldots, A_n)$ **is almost non-pointed,** ε **is very small at best**

Improving ε

- We can slightly worsen the decrease rate of $\mathbf{P}(Tb \notin TA)$ in exchange for a better requirement on ε
- Proof is based on showing that the minimum distance to a cone is also approximately projected by T
- **This version also works with non-pointed cones**

When X is large, infinite or uncountable

A surprising result

Thm.

Given $p \in \mathbb{R}^m$, $X \subseteq \mathbb{R}^m$ at most countable s.t. $p \notin X$, $T : \mathbb{R}^m \to \mathbb{R}^k$ a random projector for **any** $k \geq 1$, then $\mathbf{P}(Tp \notin TX) = 1$.

Proof

We have:

$$\mathbf{P}(Tp \notin TX) = \mathbf{P}\Big(\bigcap_{x \in X} \{Tp \neq Tx\} \Big) = \mathbf{P}\Big(\bigcap_{x \in X} \{T(p - x) \neq 0\} \Big)$$

Note that:

(a) $p \notin X \Rightarrow \forall x \in X \ (p - x \neq 0)$

(b) T has full rank k with prob. 1

Hence $\forall x \in X \ \mathbf{P}(T(p - x) \neq 0) = 1$

Intersection of countably many almost sure events is almost sure

Surprising since $k = 1$ suffices: solve IFP on a line!

[VPL, arXiv:1509.00630v1/math.OC]

Consequences

- In theory, solve any IFP by random projection to **only one** constraint!
- **In practice, it doesn't work**
- Assume $X = \{x \in \mathbb{Z}^n_+ \mid Ax = b\}$
  If TAx is extremely close to Tb, a floating point approximation might yield $TA = Tb$ even though $A \neq b$
- You would need to compute using actual real numbers, rather than floating point

Enforcing a projected minimal distance

- Instead of $Tp \notin TX$, require $\min\limits_{x\in X} \|Tp - Tx\| > \tau$ for some $\tau \geq 0$

Thm.

Given $\tau, \delta > 0$ and $p \notin X \subseteq \mathbb{R}^m$ where $|X|$ finite,

$$d = \min_{x\in X} \|p - x\| > 0,$$

and $T : \mathbb{R}^m \to \mathbb{R}^k$ a Gaussian random projector with $k \geq \dfrac{\log(\frac{|X|}{\delta})}{\log(\frac{d}{\tau})}$,

$$\mathbf{P}\Big(\min_{x\in X} \|T(p) - T(x)\| > \tau\Big) > 1 - \delta.$$

- If p, X not integral, d can get very small and yield more floating point approximation errors; else, $d \geq 1$

What if $|X|$ is infinite?

- Let $X = \{Ax \mid x \in \mathbb{Z}^n_+\}$, where $A \in \mathbb{Z}^{mn}$ has at least one positive row
- For any $b \in \mathbb{Z}^m$ the problem $b \in X$ is equivalent, with high probability, to its projection to a $O(\log n)$-dimensional space
- The idea is to separate one positive row and apply random projection to the others
- **Not many real problems have a strictly positive row**

What if there is no positive row?

λ_X =doubling dimension of X:

"The doubling dimension of a metric space X is the smallest positive integer λ_X such that every ball in X can be covered by 2^{λ_X} balls of half the radius"

My own interpretation: attach a notion of dimension to any metric space

Thm.

Given $p \notin X \subseteq \mathbb{R}^m$, let $T : \mathbb{R}^m \to \mathbb{R}^k$ be a random projector with $k \geq \mathcal{C} \log_2(\lambda_X)$, where $\mathcal{C}$ is a constant. Then $\mathbf{P}(T(p) \notin T(X))$ can be made arbitrarily close to 1.

Proof

Sketch: cover X by balls, argue for each $X \cap$ ball along the lines of "projecting minimal distances" above, and use the union bound on the ball cover.

The doubling dimension λ_X turns up as a coefficient of a decreasing exponential in a probability bound, which makes it possible to derive the high probability guarantee

[VPL, arXiv:1509.00630v1/math.OC]

Computational results, a.k.a. the bad news

The dream

Project m rows $Ax = b$ to $k \ll m$ rows $TAx = Tb$ and solve projected LP/IP

1. **Assumption**: solving $TAx = Tb \wedge x \geq 0$ yields x' s.t. $Ax' = b$

2. **Trust the method**: solve `netlib` and `miplib` (IP)
   make sure results are good

3. **Showdown!** Pick huge, bad-ass LP/IP encoding some world-saving features and solve it!

Solving the `netlib`

- **Assumption verified on almost every instance!**
- However, need $k \approx O(m)$, say $k = m - 10$ or so
- **Most** `netlib` **instances are too small**
- Now TA is same size but denser than A: **slow**

Some results on uniform dense LP

- Matrix product TA takes too long
  Not counting it: can be streamlined and parallelized
  Also: faster JL transforms can be used

- **Infeasible instances** *(sizes from* 1000×1500 *to* 2000×2400*)*:

Uniform	ϵ	$k \approx$	CPU saving	accuracy
$(-1, 1)$	0.1	$0.5m$	30%	50%
$(-1, 1)$	0.15	$0.25m$	92%	0%
$(-1, 1)$	0.2	$0.12m$	99.2%	0%
$(0, 1)$	0.1	$0.5m$	10%	100%
$(0, 1)$	0.15	$0.25m$	90%	100%
$(0, 1)$	0.2	$0.12m$	97%	100%

- **Feasible instances**: similar CPU, 100% accuracy
  - **No solution** x' **of** $TAx = Tb \wedge x \geq 0$ **satisfies** $Ax = b$
  - Proved this happens almost surely

Some results on IP

- Similar story as for LPs
- **But: every solution** x' **of** $TAx = Tb \wedge x \in \mathbb{Z}^n_+$ **satisfies** $Ax = b$
- Proved this <u>also</u> happens almost surely

Issues

- LFP projection *only ever preserves feasibility*, not the certificate
- IFP projection preserves feasibility and certificate
- **Paradox**: IP solved by BB, which only solves LP relaxations
- Orthants occupy less volume than whole spaces
  $\Rightarrow$ *projection works better for* Uniform$(0,1)$
- Low CPU savings and accuracy:
  we are getting killed by the constant $\mathcal{C}$ in $k \approx \frac{\mathcal{C}}{\varepsilon^2} \ln n$

Bet: this method will start paying off at really large m, n

Some references

- K. Vu, P.-L. Poirion, L. Liberti, *Using the Johnson-Lindenstrauss lemma in linear and integer programming*, arXiv report 1507.00990v1/math.OC, 2015
- K. Vu, P.-L. Poirion, L. Liberti, *Gaussian random projections for Euclidean membership problems*, arXiv report 1509.00630v1/math.OC, 2015
- W. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, in Contemporary Mathematics, Vol. 26, 189-206, 1984
- S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures and Algorithms, **22**(1):60-65, 2003
- S. Venkatasubramanian, Q. Wang, *The Johnson-Lindenstrauss transform: an empirical study*, in Proceedings of ALENEX, 2011
- J. Nash, C^1 *Isometric embeddings*, in Annals of Mathematics, **60**(3):383-396, 1954