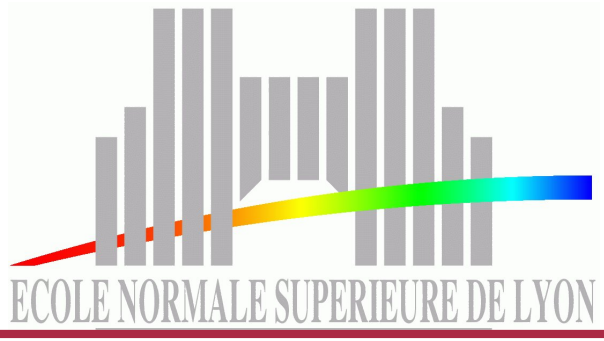






UMPA

Random Projections: on sub-Gaussianity and sparsity

https://hal.science/hal-04690731

Aurélien Garivier, Emmanuel Pilliat
Unité de Mathématiques Pures et Appliquées
École Normale Supérieure de Lyon

Outline

Dimensionality Reduction

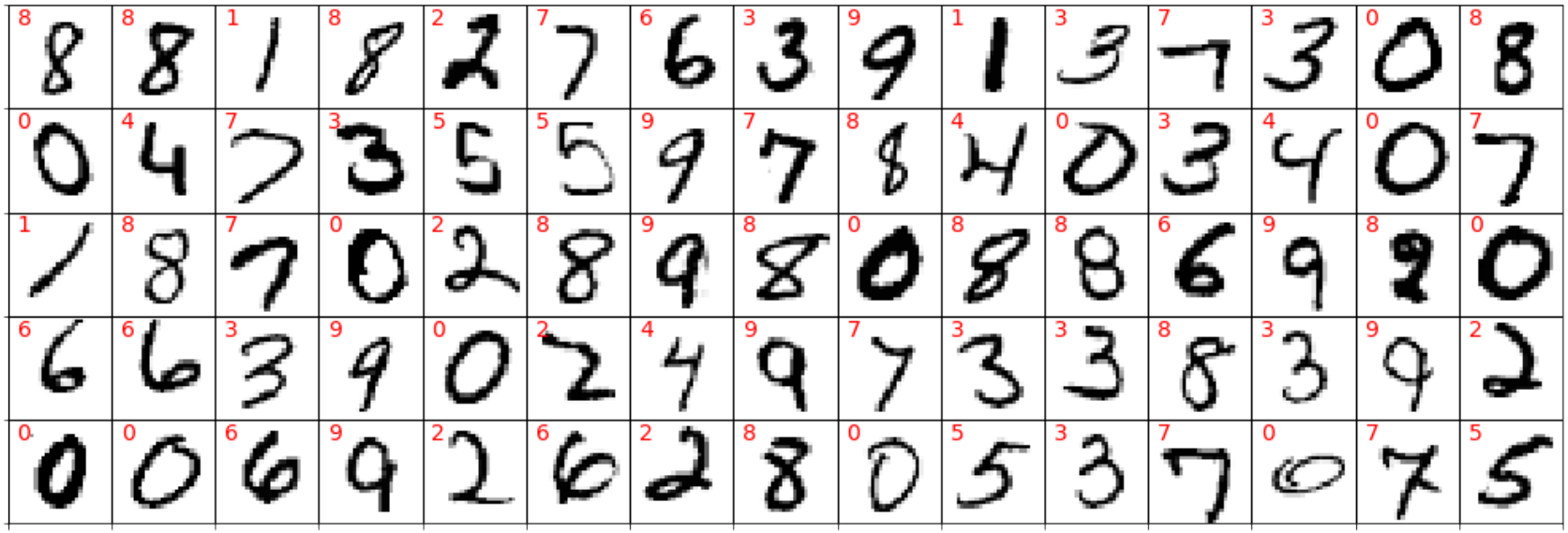
Random Projections

The Sub-Gaussian Case

Sparsity Conditions

MNIST experiment

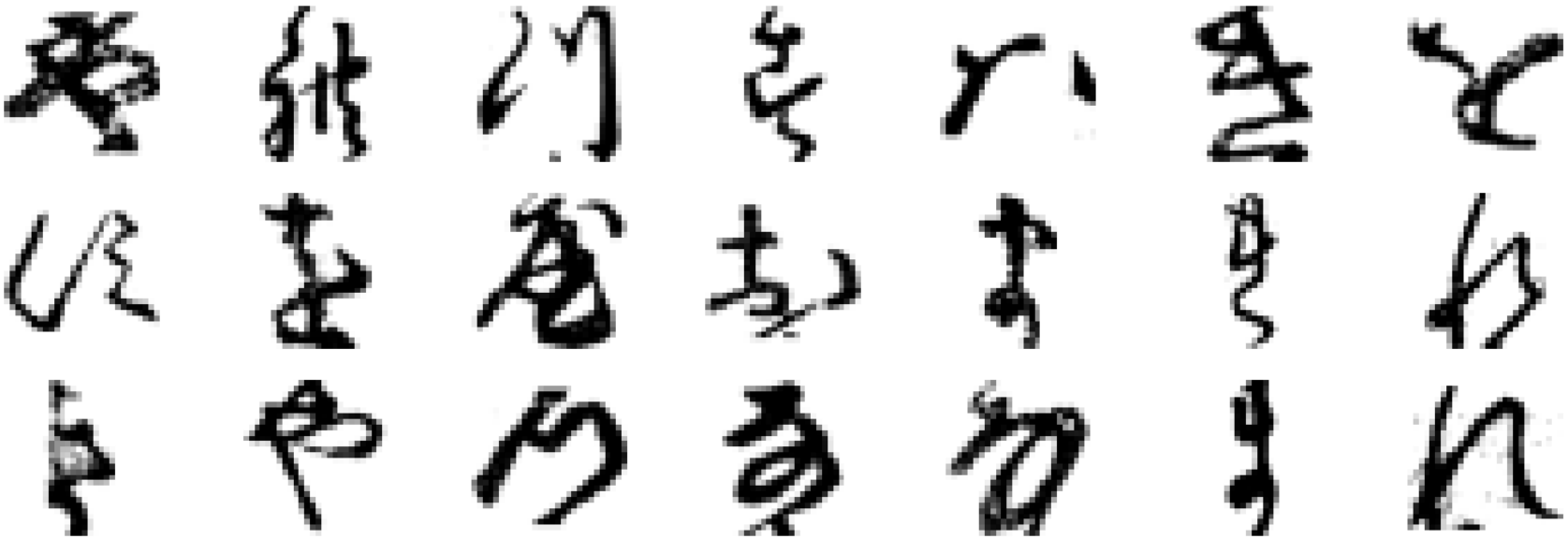
Dataset of $n = 70000$ images 28×28 of handwritten digits with their labels



Goal: build a *classifier* $[0,1]^{28\times 28} \to \{0,1,\ldots,9\}$

K-MNIST experiment

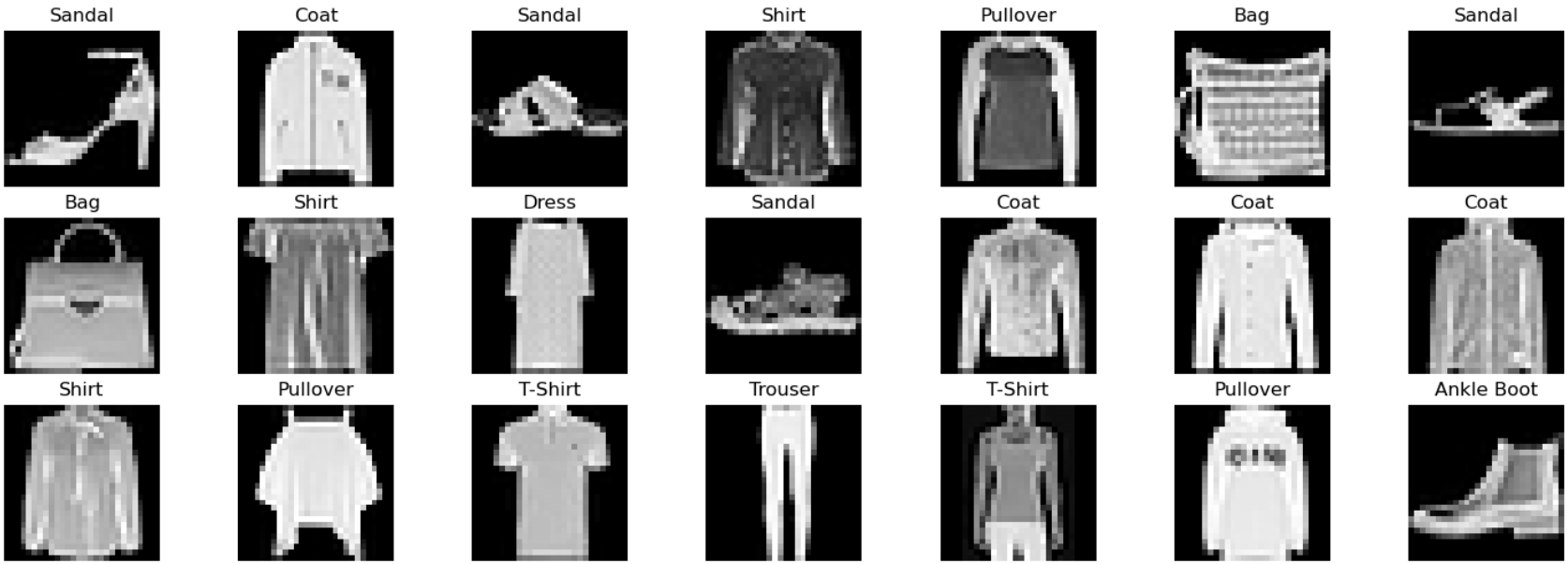
Dataset of $n = 70000$ images 28×28 of Kuzushiji with their labels



Goal: build a *classifier* $[0,1]^{28\times 28} \to \{0,1,\ldots,9\}$

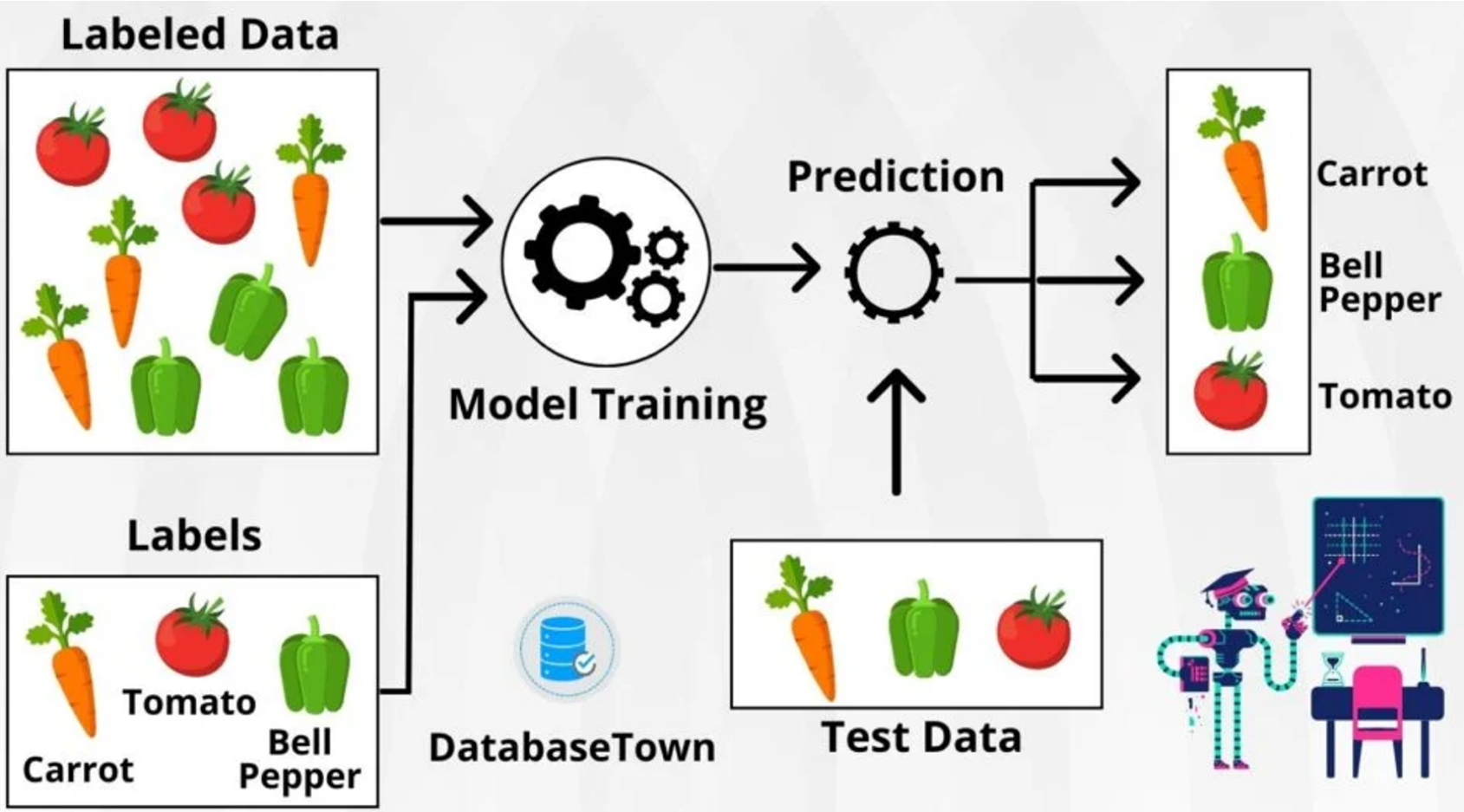
fashion-MNIST experiment

Dataset of $n = 70000$ images 28×28 of clothes digits with their labels



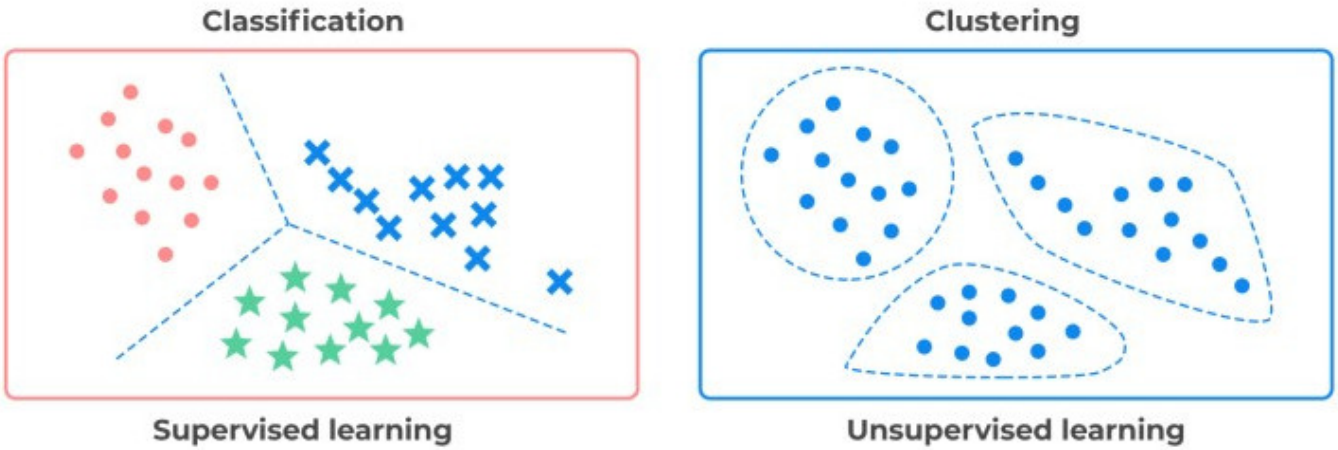
Goal: build a *classifier* $[0,1]^{28\times 28} \to \{0,1,\ldots,9\}$

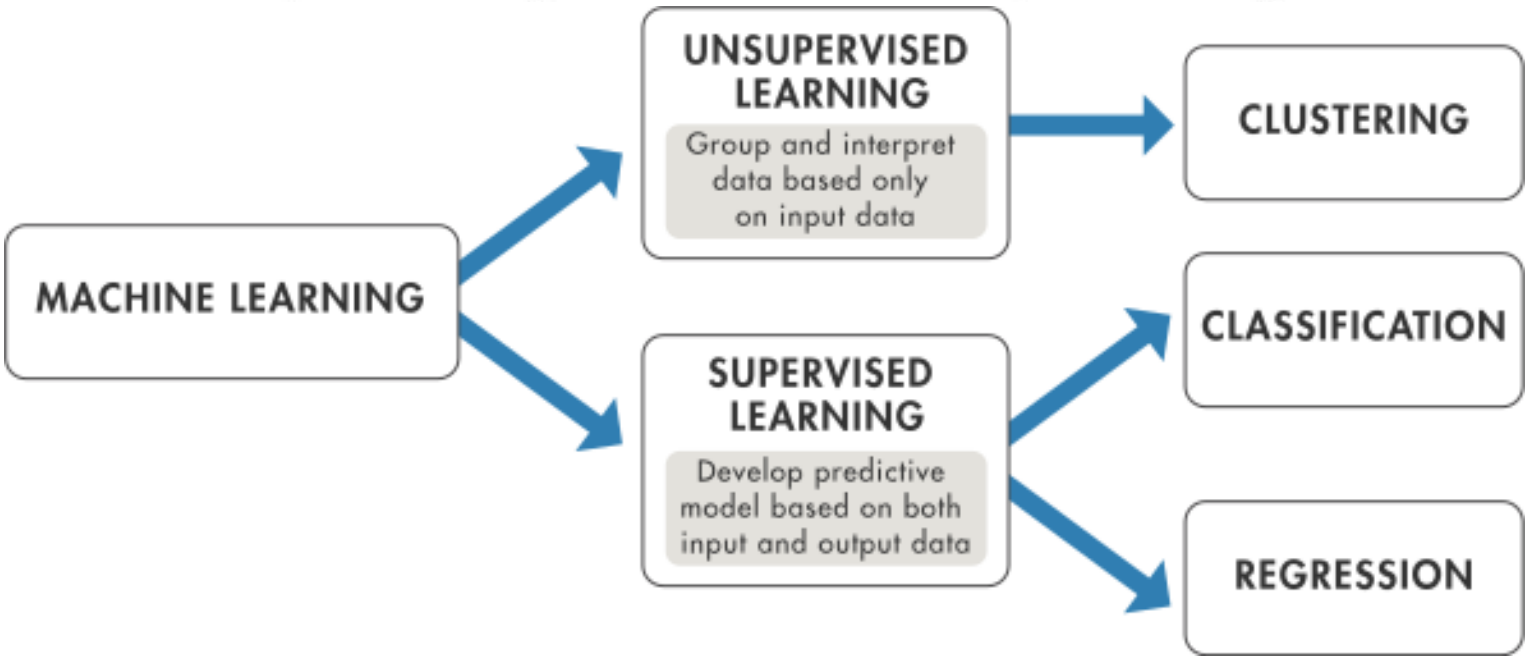
Supervised Learning



Src: `https://medium.com`

Machine Learning





Src: `https://medium.com`

Spectacular successes

- Computer vision

Natural Language Processing

. . . and their combination

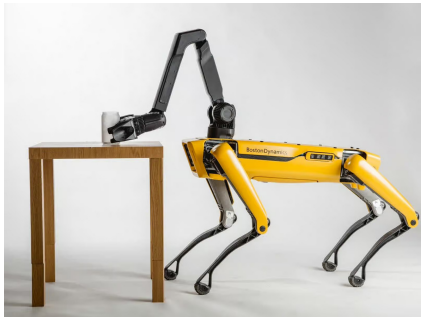


https://link.springer.com/article/10.1007

- Game solving (strategies)

AlphaGo

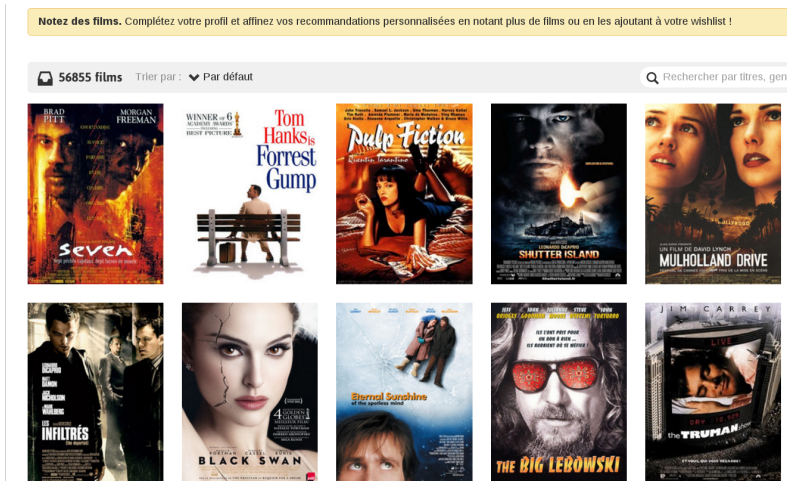
- Autonomous agents



https://bostondynamics.com/

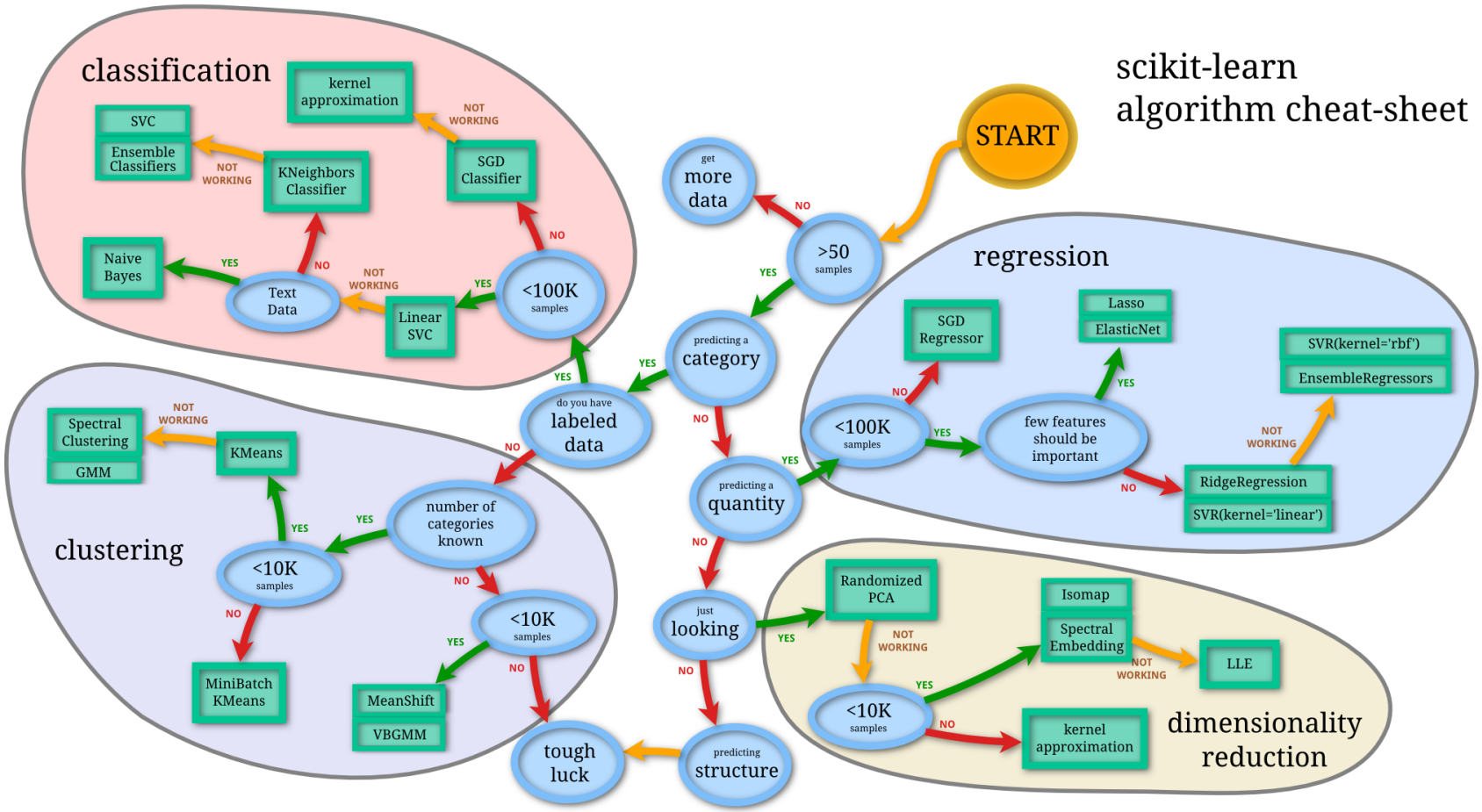
- Recommender systems

(press, ads, movies, etc.)



http://www.vodkaster.com/

Scikit-learn ML cheat-sheet



Src: `https://scikit-learn.org`

Why dimensionality reduction?

ENS DE LYON

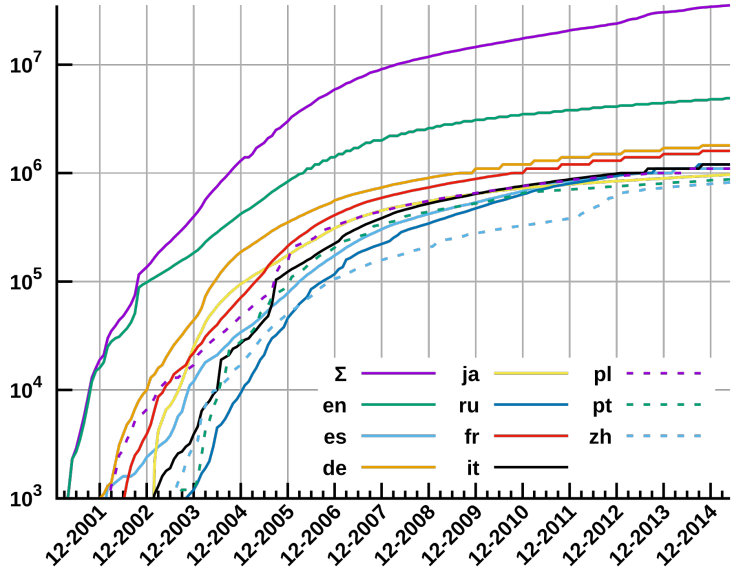
- Image: ImageNet

$n = 14 \times 10^6$, $p = 469 \times 387$



- Text: Wikipedia

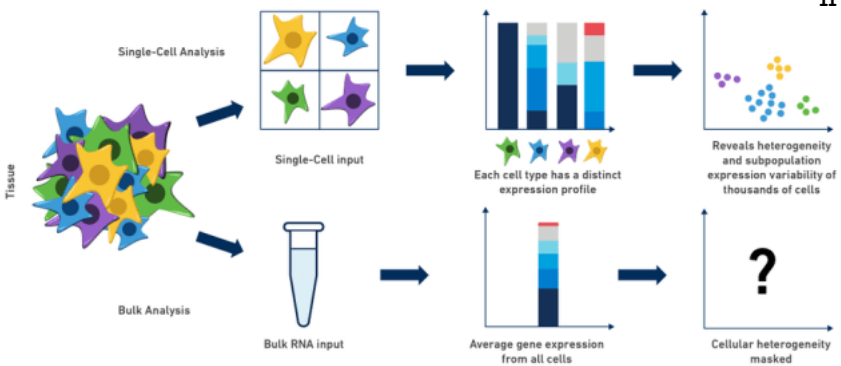
$n = 4.7 \times 10^9$, $p > 10^5$



Src: `https://wikipedia.com/`
`https://dmpi.duke.edu/`

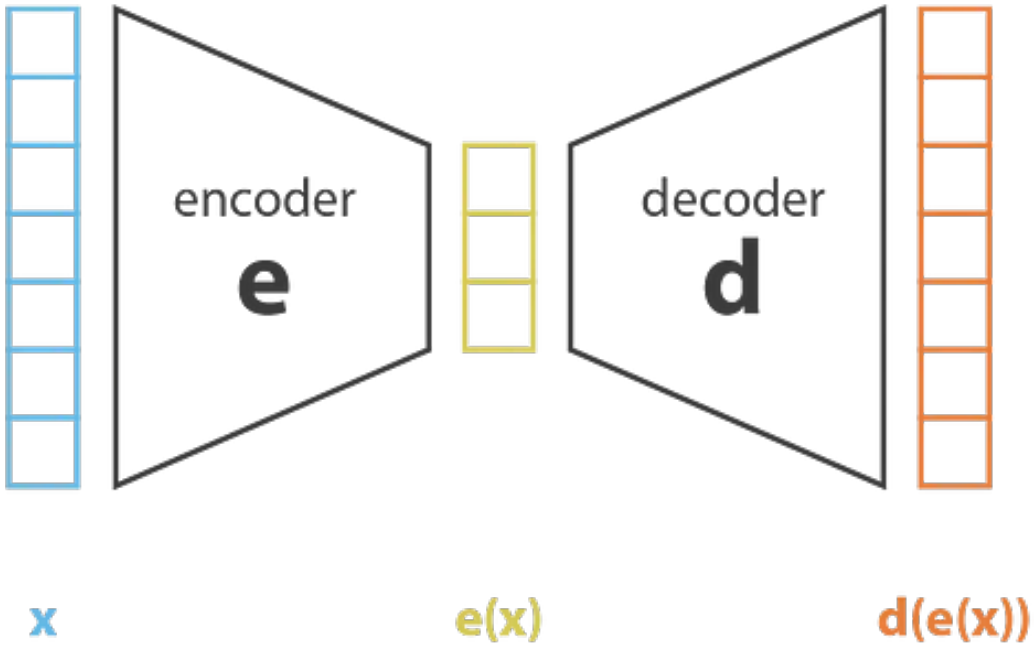
- Biology: single cell

$n > 10^7$, $p > 10^6$



Dimensionality reduction

- Data: $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathcal{M}_{n,p}(\mathbb{R})$, $p \gg 1$
- Replace x_i with $y_i = \mathrm{e}(x_i)$, where $\mathrm{e} : \mathbb{R}^p \to \mathbb{R}^d$, $d \ll p$
- Hopefully $\exists \mathrm{d} : \mathbb{R}^d \to \mathbb{R}^p$ s.t. $\forall i \in \{1, \ldots, n\}$, $\mathrm{d}(\mathrm{e}(x_i)) \approx x_i$



Src: `https://towardsdatascience.com/`

Linear solution: PCA

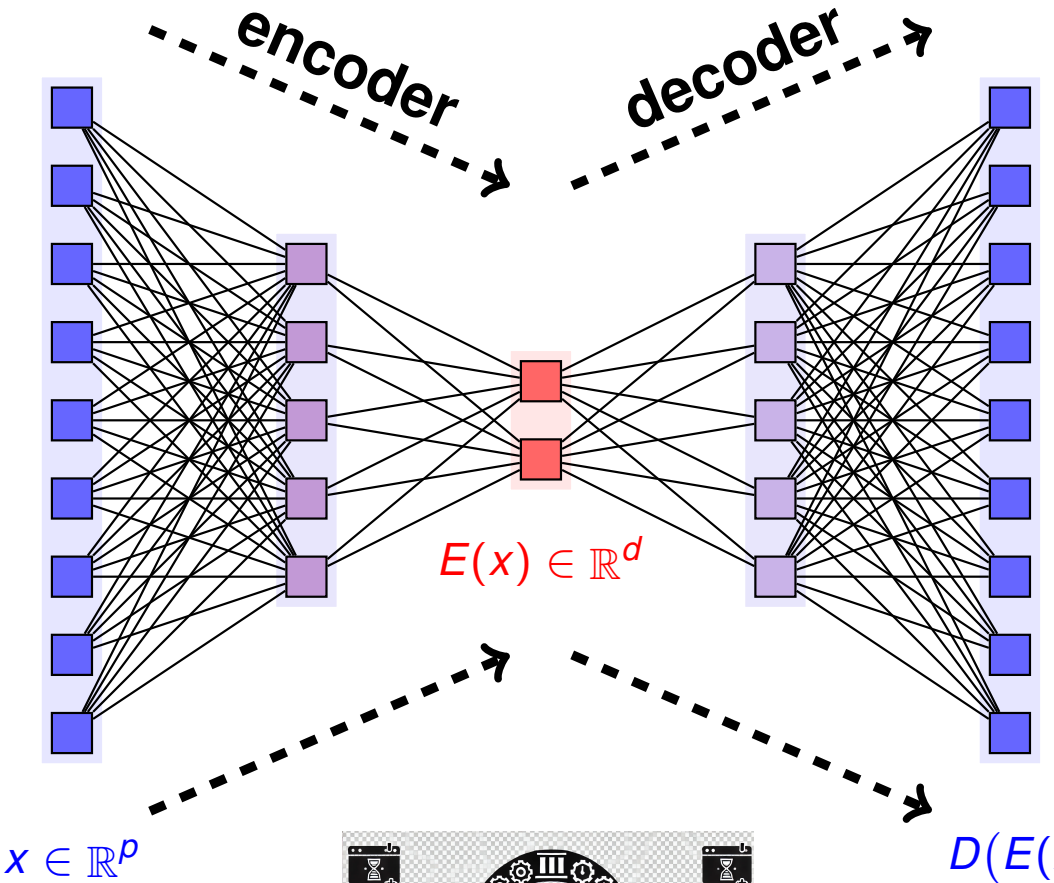
PCA aims at finding the compression matrix A and the recovering matrix B such that the total squared distance between the original and the recovered vectors is minimal:

$$\underset{A \in \mathcal{M}_{d,p}(\mathbb{R}), B \in \mathcal{M}_{p,d}(\mathbb{R})}{\arg\min} \sum_{i=1}^{n} \left\| x_i - B A x_i \right\|_2^2$$

Theorem

Let $G = \sum_{i=1}^{n} x_i x_i^T$, and let $u_1, \ldots, u_d$ be the eigenvectors of G corresponding to the d largest eigenvalues of G.
Then a solution to the PCA optimization problem is $B = \left(\begin{array}{c|c|c} u_1 & \cdots & u_d \end{array} \right)$, and $A = B^T$.

Beyond linearity: auto-encoders



$$x_0 = x$$
$$x_{i+1} = \sigma\left(W_i\, x_i\right)$$
$$\theta = (W_0, W_1, \dots)$$

Minimize reconstruction error

$$\min_{D,E} \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - D\left(E(x_i)\right) \right\|_2^2$$



Outline

Dimensionality Reduction

Random Projections

The Sub-Gaussian Case

Sparsity Conditions

Johnson-Lindenstrauss Lemma

ENS DE LYON

Theorem [*Johnson & Lindenstrauss '84*]

Let $x_1, \dots, x_n \in \mathbb{R}^p$, and let $\varepsilon > 0$. Then, for every $d \geq \dfrac{4\log(n)}{\varepsilon - \log(1+\varepsilon)}$, there exists a matrix $A \in \mathcal{M}_{d,p}(\mathbb{R})$ such that

$$\forall i, j \in [n], \quad (1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Ax_i - Ax_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$

- d does not depend on p (!)
- dependence on ε: $\dfrac{4\log(n)}{\varepsilon - \log(1+\varepsilon)} \leq \dfrac{8\log(n)}{\varepsilon^2 - \varepsilon^3}$
- how to find such a quasi-isometry A?
- the construction of A does not depend on x (!)

[Extensions of Lipschitz maps into a Hilbert space, *Johnson & Lindenstrauss '84*] Choose $A_{k,\ell} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{d}\right)$

If $y = \frac{x_i - x_j}{\|x_i - x_j\|_2} \in \mathbb{R}^p$ and $Y = Ay$, we hope that $\|Y\|_2^2 \approx \|y\|_2^2 = 1$

- $\forall k,\ Y_k = \sum_{\ell=1}^{p} A_{k,\ell} y_\ell$, so $\mathbb{E}[Y_k] = 0$ and $\mathbb{E}\left[Y_k^2\right] = \sum_{\ell=1}^{p} \frac{y_\ell^2}{d} = \frac{1}{d}$
- Hence $\mathbb{E}\left[\|Y\|_2^2\right] = \sum_{k=1}^{d} \mathbb{E}\left[Y_k^2\right] = 1$ $\rightarrow$ OK in expectation

$\Longrightarrow$ if we prove $\mathbb{P}\left((1-\varepsilon) \leq \|Y\|_2^2 \leq (1+\varepsilon)\right) \geq 1 - \frac{2}{n^2}$,

by the union bound for all $y = \frac{x_i - x_j}{\|x_i - x_j\|_2}$ $(i < j)$ we conclude:

$$\mathbb{P}\left(\bigcap_{1 \leq i < j \leq n} \left\{ (1-\varepsilon)\|x_i - x_j\|_2^2 \leq \left\|A(x_i - x_j)\right\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2 \right\} \right)$$

$$\geq 1 - \frac{n(n-1)}{2} \frac{2}{n^2} > 0$$

The Gaussian Case

- Gaussian stability: $Y_k = \sum_{\ell=1}^{p} A_{k,\ell} y_\ell \sim \mathcal{N}\left(0, \frac{1}{d}\right)$

- $$\mathbb{P}\left(\|Y\|_2^2 \geq 1 + \varepsilon \right) = \mathbb{P}\left(\frac{1}{d} \sum_{k=1}^{d} \left(\underbrace{\sqrt{d} Y_k}_{\sim \mathcal{N}(0,1)} \right)^2 \geq 1 + \varepsilon \right)$$

$$= \mathbb{P}\left(\frac{1}{d} \sum_{k=1}^{d} \underbrace{Z_k^2 - 1}_{U_k} \geq \varepsilon \right)$$

where $Z_k \sim \mathcal{N}(0,1)$ and $U_k = Z_k^2 - 1$

Chernoff's method

For every $\lambda > 0$,

$$\mathbb{P}\left(\frac{1}{d}\sum_{k=1}^{d} U_k \geq \varepsilon\right) = \mathbb{P}\left(e^{\lambda \sum_{k=1}^{d} U_k} \geq e^{\lambda d \varepsilon}\right)$$

$$\leq \frac{\mathbb{E}\left[e^{\lambda \sum_{k=1}^{d} U_k}\right]}{e^{d\lambda\varepsilon}} = \frac{\mathbb{E}\left[\prod_{k=1}^{d} \exp\left(\lambda U_k\right)\right]}{e^{d\lambda\varepsilon}}$$

$$= \frac{\mathbb{E}\left[e^{\lambda U_k}\right]^d}{e^{d\lambda\varepsilon}}$$

$$= \exp\left(-d\Big(\underbrace{\lambda\varepsilon - \ln \mathbb{E}\left[e^{\lambda U_k}\right]}_{\text{choose } \lambda \text{ to maximize}}\Big)\right)$$

Chernoff's bound for χ^2

ENS DE LYON

Lemma

If $Z \sim \mathcal{N}(0,1)$ and $U = Z^2 - 1$, then

$$\varphi^*(\varepsilon) = \sup_\lambda \lambda\varepsilon - \log \mathbb{E}\left[e^{\lambda U}\right] = \frac{\varepsilon - \log(1+\varepsilon)}{2} \,.$$

Proof: For every $\lambda < 1/2$,

$$\mathbb{E}\left[e^{\lambda U}\right] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda(z^2-1)} e^{-\frac{z^2}{2}} dz = \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(1-2\lambda)z^2}{2}} dz = e^{-\lambda} \frac{1}{\sqrt{1-2\lambda}} \,.$$

Hence $\varphi(\lambda) = \log \mathbb{E}\left[e^{\lambda U}\right] = -\frac{1}{2}\log(1-2\lambda) - \lambda$. The concave function $\lambda \mapsto \lambda\varepsilon - \varphi(\lambda)$ is maximized at λ^* s.t. $x = \varphi'(\lambda^*) = \frac{1}{1-2\lambda^*} - 1$, that is at $\lambda^* = \frac{1}{2}\left(1 - \frac{1}{1+\varepsilon}\right) = \frac{\varepsilon}{2(1+\varepsilon)}$. Hence

$$\varphi^*(\varepsilon) = \lambda^*\varepsilon - \varphi(\lambda^*) = \frac{\varepsilon - \log(1+\varepsilon)}{2} \,.$$

Wrapping up

- $$\mathbb{P}\left(\|Y\|_2^2 \geq 1+\varepsilon \right) = \mathbb{P}\left(\frac{1}{d}\sum_{k=1}^{d} \left(\sqrt{d}Y_k\right)^2 - 1 \geq \varepsilon \right)$$
$$\leq e^{-d\frac{\varepsilon-\log(1+\varepsilon)}{2}} \leq \frac{1}{n^2}$$

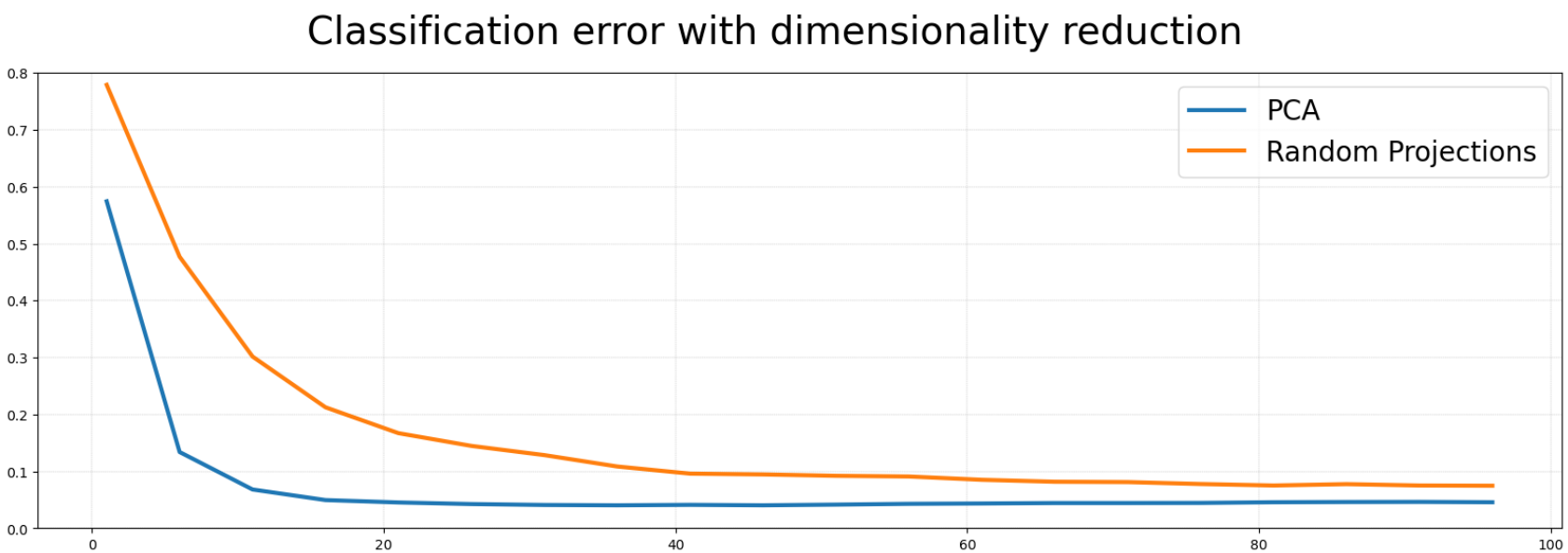
as soon as $d \geq \frac{4\log(n)}{\varepsilon-\log(1+\varepsilon)}$. Similarly,

$$\mathbb{P}\left(\|Y\|_2^2 \leq (1-\varepsilon)\|y\|_2^2 \right) = \mathbb{P}\left(\frac{1}{d}\sum_{k=1}^{d} \left(\sqrt{d}Y_k\right)^2 - 1 \leq -\varepsilon \right)$$
$$\leq e^{-d\frac{\varepsilon-\log(1+\varepsilon)}{2}} \leq \frac{1}{n^2}$$

- Observe that if $d \geq \frac{4\log(n)+2\log\frac{1}{\delta}}{\varepsilon-\log(1+\varepsilon)}$,

$$\mathbb{P}\left(1-\varepsilon \leq \|Y\|_2^2 \leq 1+\varepsilon \right) \geq 1 - \frac{2\delta}{n^2}$$

Experiment on MNIST



Only with Gaussians?

$$\sqrt{d}A_{k,\ell} \overset{iid}{\sim} \mathcal{N}(0,1)$$

Gaussian stability

$$\sqrt{d}Y_k = \sum_{\ell=1}^{p} \sqrt{d}A_{k,\ell}y_\ell \sim \mathcal{N}(0,1)$$

Chernoff deviation of χ^2

$$\|Y\|_2^2 = \frac{1}{d}\sum_{k=1}^{d}\left(\sqrt{d}Y_k\right)^2 \approx 1$$

Outline

ENS DE LYON

Dimensionality Reduction

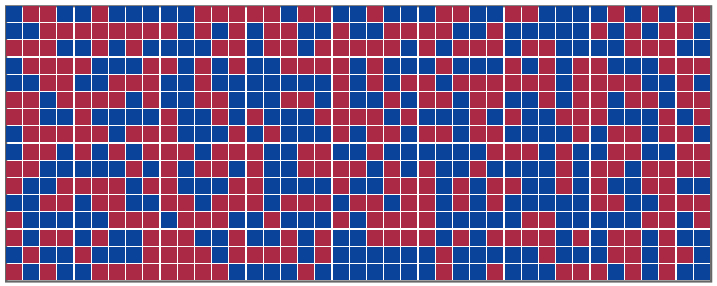
Random Projections

The Sub-Gaussian Case

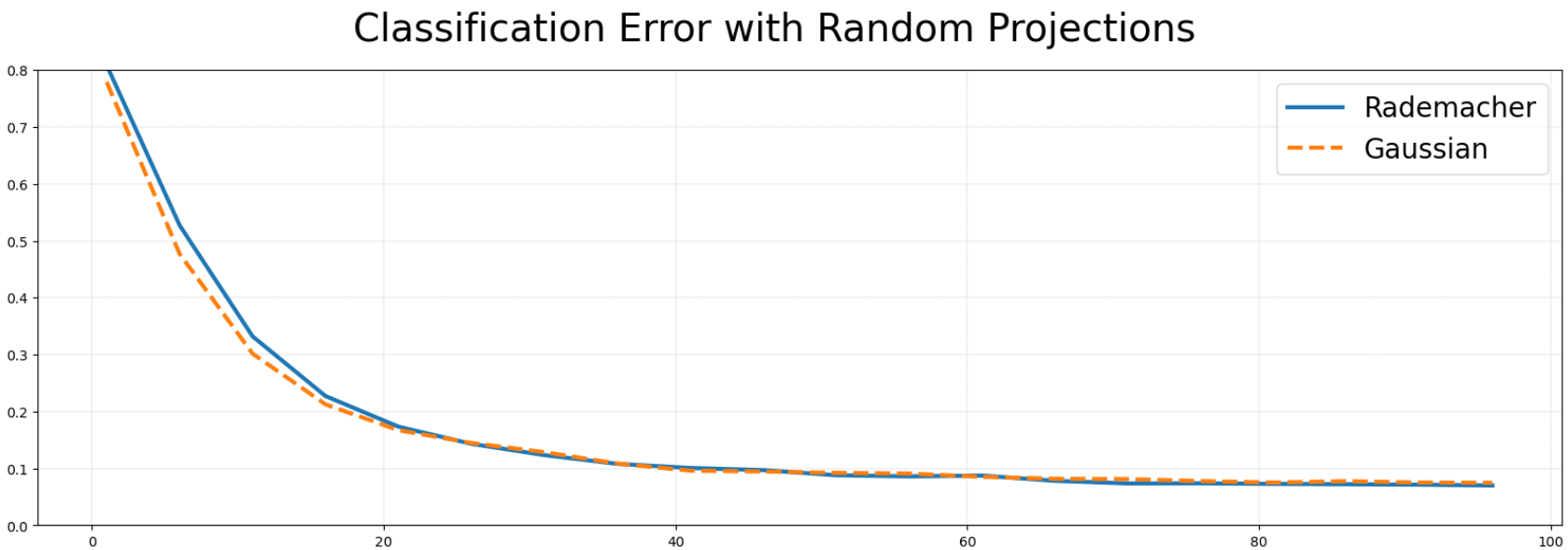
Sparsity Conditions

Using other variables

- Can we replace the Gaussian entries with other (simpler) independent variables?
- $\mathbb{E}\left[\|Y\|_2^2\right] = 1$ as soon as $\mathbb{E}[A_{k,\ell}] = 0$ and $\mathbb{V}ar[A_{k,\ell}] = \frac{1}{d}$
- Example: random signs $A_{k,\ell} \sim \mathcal{U}\left(\left\{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right\}\right)$
- Experimentally: yes!



Experiment on MNIST



Sub-Gaussian Variables

ENS DE LYON

Definition

Z is said to be σ^2-sub-Gaussian ($Z \sim \sigma^2$-sG) if $\forall \lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda Z}\right] \leq e^{\frac{\sigma^2 \lambda^2}{2}} = \left[e^{\lambda G}\right]$$

where $G \sim \mathcal{N}(0, \sigma^2)$.

Equivalent definitions:

- $\exists c > 0, \forall t > 0, \mathbb{P}(|Z| > t) \leq c\mathbb{P}(|G| > t)$
- $\exists c > 0, \forall p > 0, \mathbb{E}\left[|Z|^p\right] \leq c\mathbb{E}\left[|G|^p\right]$

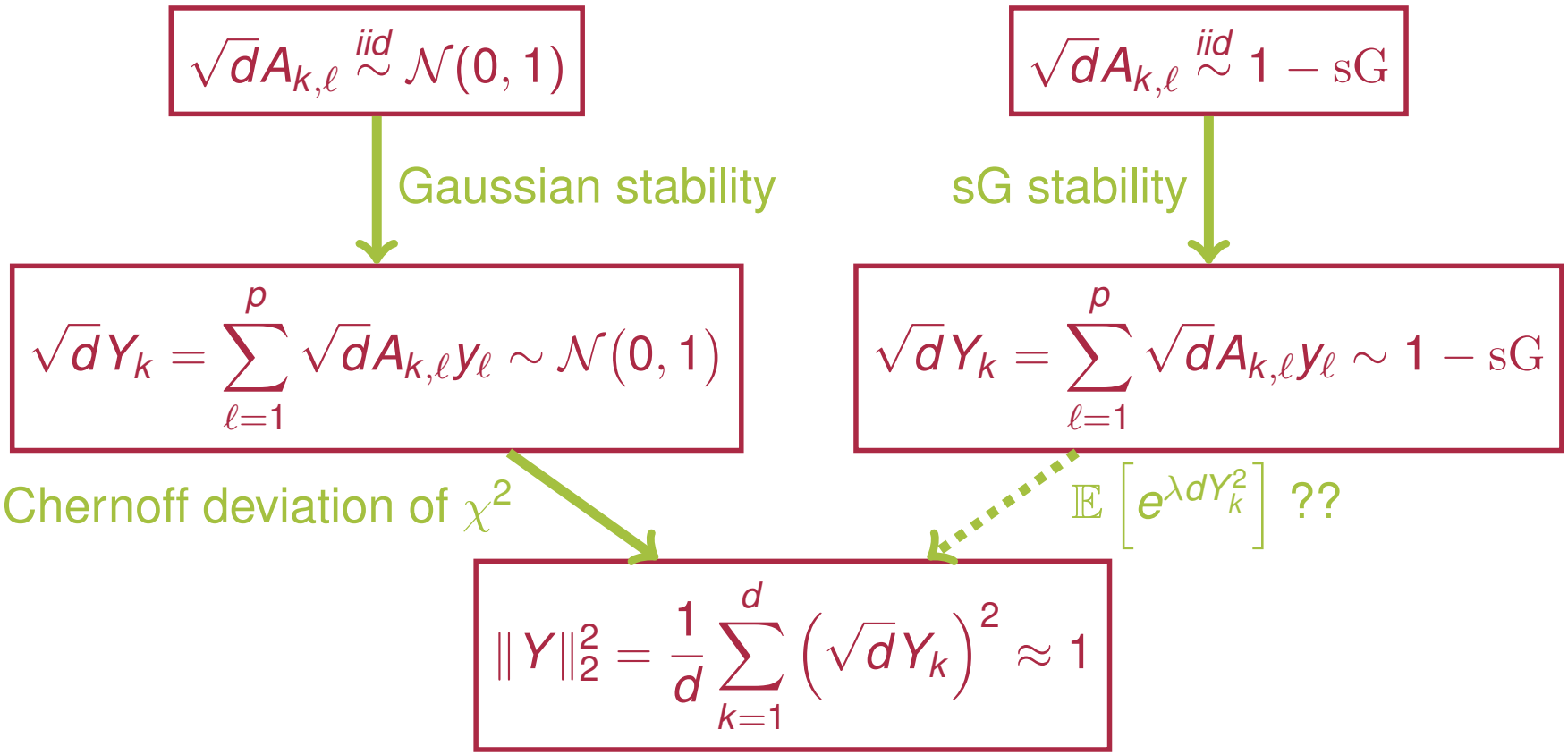
Stability: if $Z_1 \sim \sigma_1^2 - \mathrm{sG} \perp\!\!\!\perp Z_2 \sim \sigma_2^2 - \mathrm{sG}$ and $\gamma > 0$ then

$$Z_1 + \gamma Z_2 \sim (\sigma_1^2 + \gamma^2 \sigma_2^2) - \mathrm{sG}$$

Examples

- If $Z \sim \mathcal{N}(0, \sigma^2)$, then $Z \sim \sigma^2 - \mathrm{sG}$ 
- If $\mathbb{P}(a \leq Z \leq b) = 1$, then $Z \sim \left(\frac{b-a}{2}\right)^2 - \mathrm{sG}$
- A mixture of Gaussians is sub-Gaussian
- etc. . .

JL for sub-Gaussian variables?



The sub-Gaussian Trick

Sub-Gaussian Trick

Let Z be a 1-sub-Gaussian random variable. For all $\lambda \in \left[0, \frac{1}{2}\right)$,

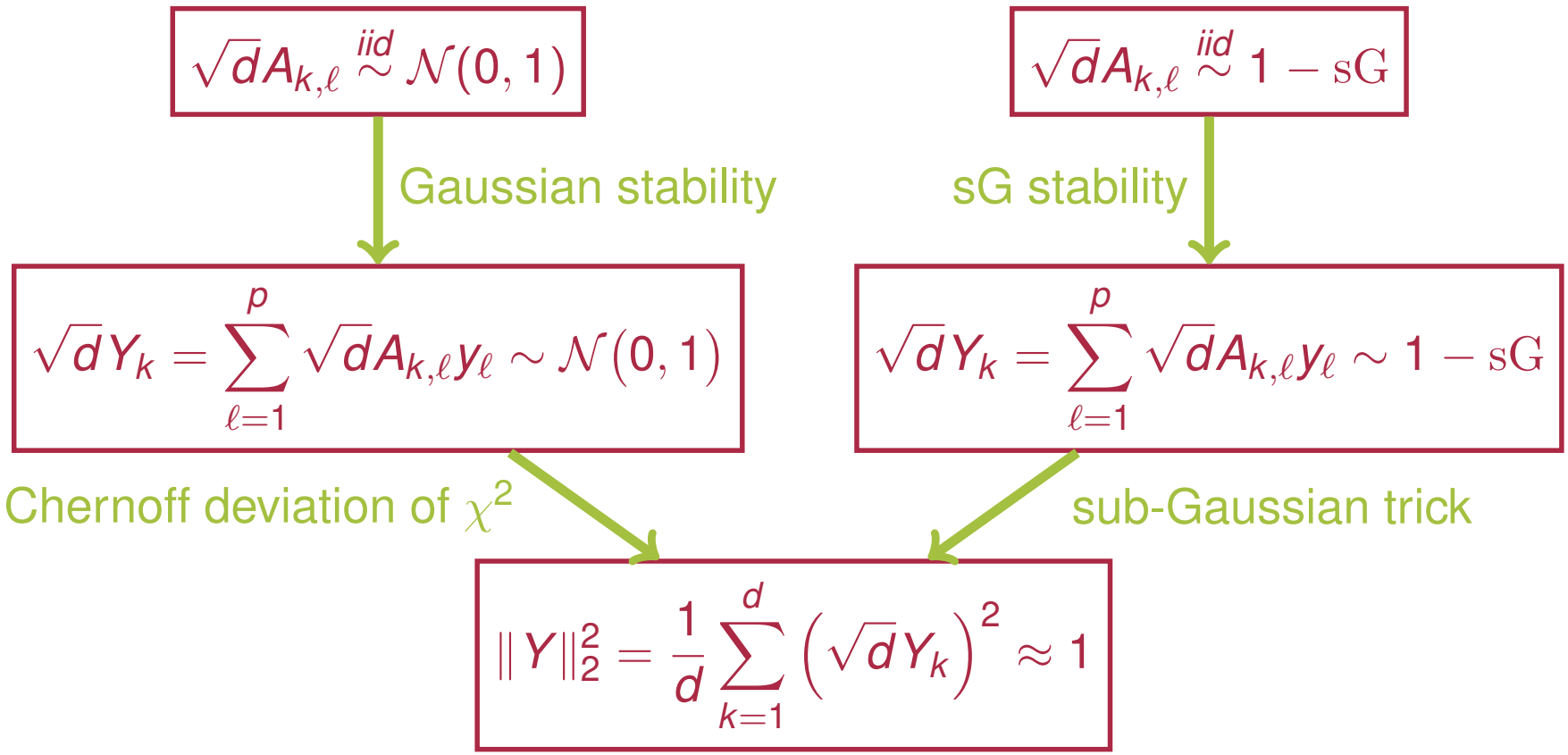
$$\mathbb{E}\left[e^{\lambda Z^2}\right] \leq \mathbb{E}\left[e^{\lambda G^2}\right] \quad \text{where } G \sim \mathcal{N}(0,1)\ .$$

Proof: Since $\forall x > 0, e^{x^2} = \mathbb{E}_G\left[e^{\sqrt{2x}G}\right]$, by Fubini's theorem,

$$\mathbb{E}_Z\left[e^{\lambda Z^2}\right] = \mathbb{E}_Z\left[\mathbb{E}_G\left[e^{\sqrt{2\lambda} Z\, G}\right]\right] = \mathbb{E}_G\left[\mathbb{E}_Z\left[e^{\sqrt{2\lambda} G\, Z}\right]\right] \leq \mathbb{E}_G\left[e^{\lambda G^2}\right]$$

with equality if and only if $Z \sim \mathcal{N}(0,1)$.

JL for sub-Gaussian variables



NB: for $\lambda < 0$ (left deviations), another argument is needed

Acceptable entries

$\rightarrow$ What are the variables $Z = \sqrt{d} A_{k,\ell}$ that

- have variance 1
- and are 1-sub-Gaussian?

✓ $Z \sim \mathcal{U}(\{-1, 1\})$ works since

$$\mathbb{E}\left[e^{\lambda Z}\right] = \cosh(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k k!} = e^{\frac{\lambda^2}{2}}$$

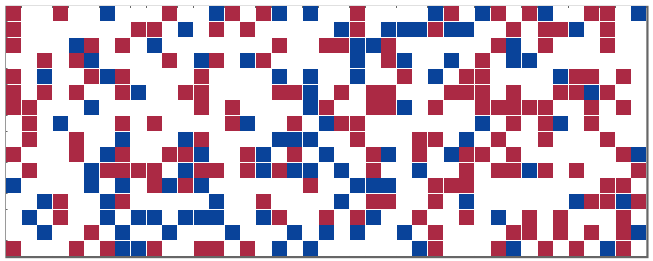
✓ $Z \sim \mathcal{U}\left([-\sqrt{3}, \sqrt{3}]\right)$ similarly

? Even more simple entries? with 0s?

Sparse entries?

[Database-friendly random projections *Achlioptas '01*] suggests

$$Z = \begin{cases} -\frac{1}{\sqrt{q}} & \text{with probability } \frac{q}{2} \\ 0 & \text{with probability } 1-q \\ \frac{1}{\sqrt{q}} & \text{with probability } \frac{q}{2} \end{cases}$$

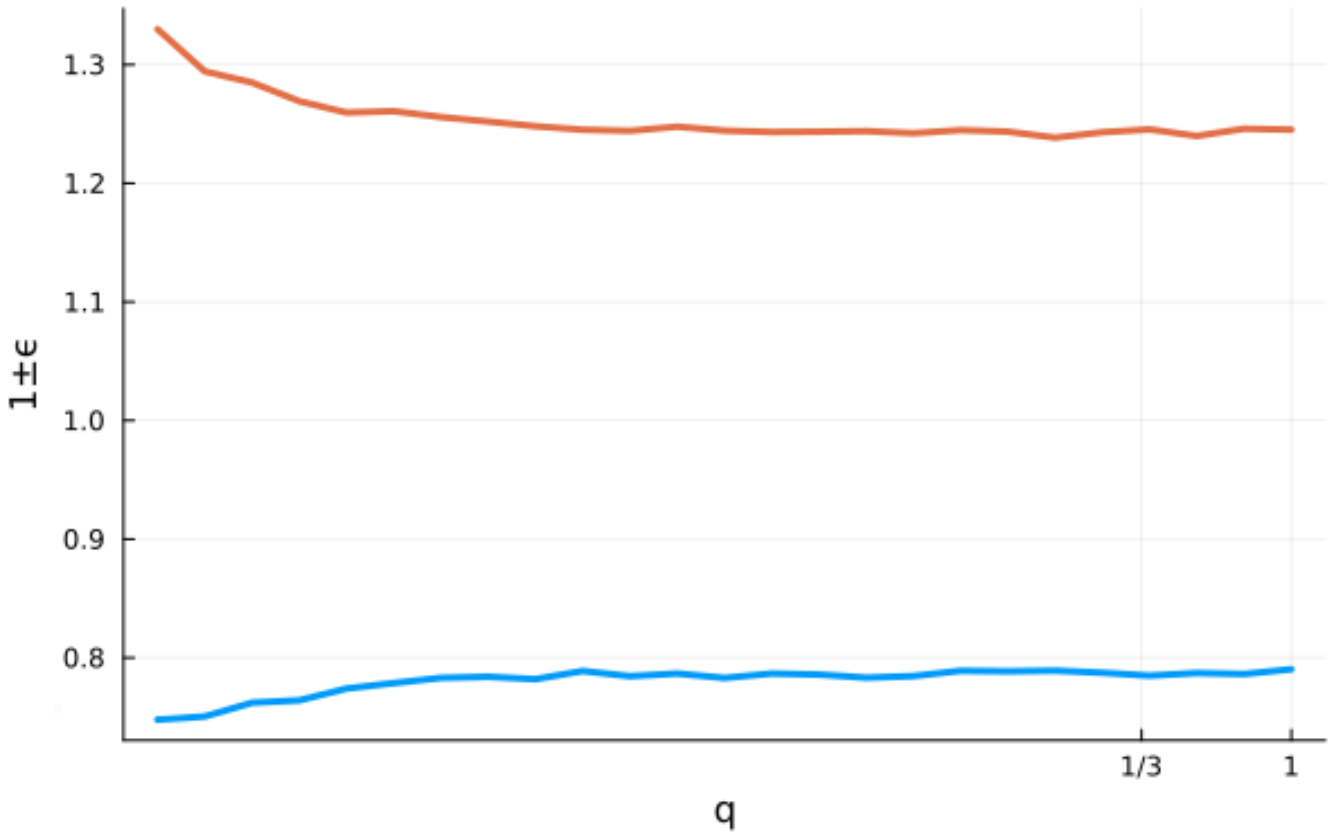


with $q = 1/3$

Proposition

If X is a 1-sub-Gaussian random variable of variance 1, then $P(X = 0) \leq 2/3$, with equality if and only $\mathbb{P}(X = -\sqrt{3}) = \mathbb{P}(X = \sqrt{3}) = \mathbb{P}(X \neq 0)/2$

What if $q < 1/3$?

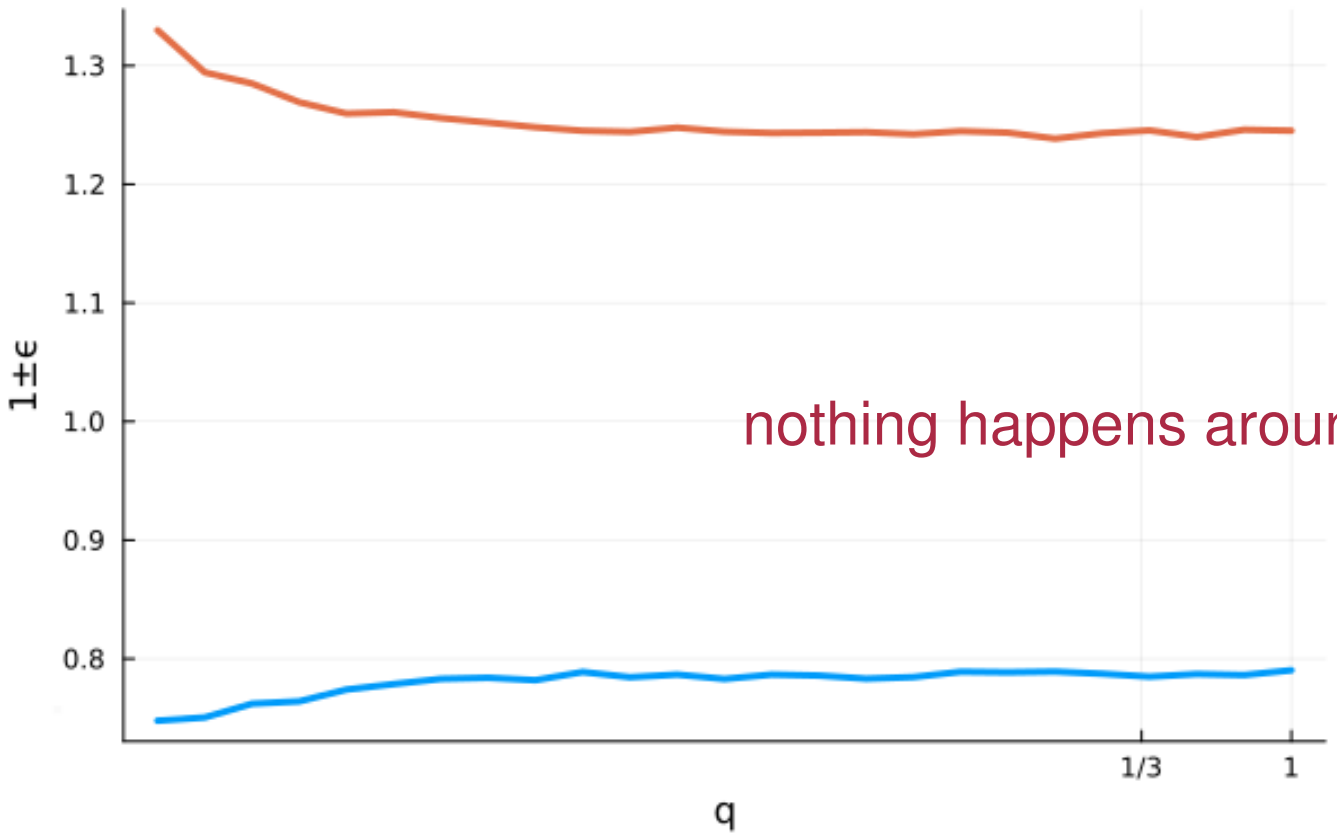


$\max_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$\min_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$(x_i)_{1 \le i \le n=100}$ with indep. Gaussian entries $\quad p = 10^5 \quad d = 500$

What if $q < 1/3$?



$\max_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

nothing happens around $q = 1/3$...

$\min_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$(x_i)_{1 \le i \le n=100}$ with indep. Gaussian entries $\quad p = 10^5 \quad d = 500$

What is too conservative?

$$\sqrt{d}A_{k,\ell} \overset{iid}{\sim} 1 - \mathrm{sG}$$

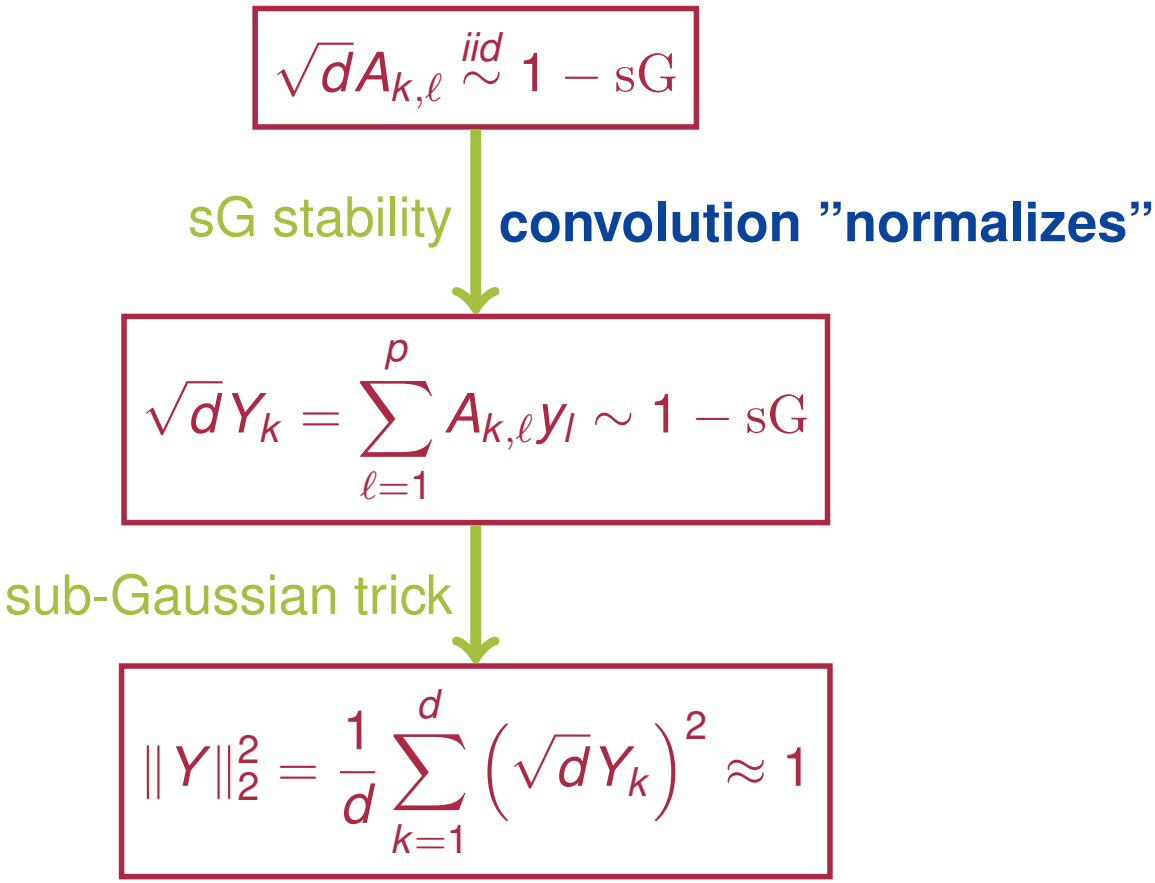
sG stability

$$\sqrt{d}Y_k = \sum_{\ell=1}^{p} A_{k,\ell} y_l \sim 1 - \mathrm{sG}$$

sub-Gaussian trick

$$\|Y\|_2^2 = \frac{1}{d}\sum_{k=1}^{d}\left(\sqrt{d}Y_k\right)^2 \approx 1$$

What is too conservative?



Outline

Dimensionality Reduction

Random Projections

The Sub-Gaussian Case

Sparsity Conditions

Ensuring $\sqrt{d}\,Y_k \sim 1 - \mathrm{sG}$

- coefficient matrix $U \in \mathbb{R}^{d \times p}$ of iid $1 - \mathrm{sG}$ with variance 1
- mask matrix $\zeta \in \mathbb{R}^{d \times p}$ of iid $\mathcal{B}(q)$, $\perp\!\!\!\perp U$
- $A_{k\ell} = \frac{1}{\sqrt{dq}} \zeta_{k\ell} U_{k\ell}$

Theorem

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ and $\varepsilon > 0$. If $d \geq \dfrac{36 \log\left(2n^2\right)}{\varepsilon^2 - \varepsilon^3}$ and

$$q \geq \frac{\log(dn^2)}{\varepsilon^2} \max_{i \neq j} \left(18 \left\| \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_4^4 + 2 \left\| \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_\infty^2 \right),$$

A is an ε-quasi-isometry with positive probability

Sparsity condition

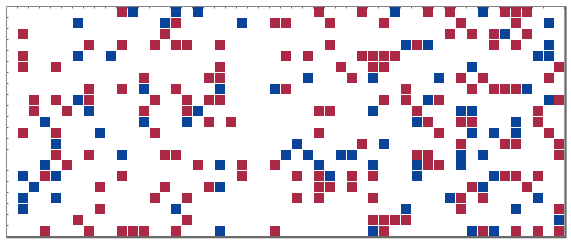
$$q \geq \frac{\log(dn^2)}{\varepsilon^2} \max_{i \neq j} \left(18 \left\| \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_4^4 + 2 \left\| \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_\infty^2 \right)$$

- Full data: if $\forall i, j, \ell$, $\frac{c}{\sqrt{p}} \leq \left(\frac{x_i - x_j}{\|x_i - x_j\|_2} \right)_\ell \leq \frac{C}{\sqrt{p}}$,

$$18 \left\| \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_4^4 + 2 \left\| \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_\infty^2 \leq 20 \left(\frac{C}{c} \right)^4 \frac{1}{p}$$

$\rightarrow$ ok for $q \gtrsim \frac{\log(dn^2)}{\varepsilon^2 p}$

- Sparse data: $\left(\frac{x_i}{\|x_i\|} \right)_\ell \in \{0\} \cup \left[\frac{c}{\sqrt{s}}, \frac{C}{\sqrt{s}} \right]$



$\rightarrow$ ok for $q \gtrsim \frac{\log(dn^2)}{\varepsilon^2 s}$

Proof sketch

Let $y \in \mathbb{R}^p$ s.t. $\|y\|_2 = 1$ and for $k \in [d]$, $G_k = \left\{ 1 - \frac{\varepsilon}{3} \leq \sum_{\ell=1}^{p} \frac{y_k^2}{q} \zeta_{k\ell} \leq 1 + \frac{\varepsilon}{3} \right\}$.

By Bernstein's ineq., $\mathbb{P}(\bar{G}_k) \leq \frac{1}{dn^2}$ as soon as $q \geq \frac{18\|y\|_4^4 + 2\|y\|_\infty^2}{\varepsilon^2} \log(dn^2)$

On the event G_k, Y_k is almost $\frac{\|y\|_2^2}{d}$-sub-Gaussian:

$$\mathbb{E}\left[\exp\left(\lambda \frac{\sqrt{d}}{\|y\|_2} Y_k \right) \mathbb{1}_{G_k} \right] \leq \exp\left(\frac{\lambda^2}{2} \left(1 + \frac{\varepsilon}{3} \right) \right)$$

so that $\frac{\sqrt{d} Y_k}{\sqrt{1+\varepsilon/3}} \mathbb{1}_{G_k}$ is 1-sub-Gaussian and hence $\mathbb{E}\left[e^{\frac{\lambda d Y_k^2}{1+\varepsilon/3}} \mathbb{1}_{G_k} \right] \leq \frac{1}{\sqrt{1-2\lambda}}$

It follows that

$$\mathbb{P}\left(\bigcap_{k=1}^{d} G_k \cap \left\{ \|Ay\|_2^2 > 1 + \varepsilon \right\} \right) \leq \mathbb{P}\left(G \cap \left\{ \frac{1}{d} \sum_{k=1}^{d} \frac{d Y_k^2}{1 + \frac{\varepsilon}{3}} > \underbrace{\frac{1+\varepsilon}{1+\frac{\varepsilon}{3}}}_{\geq 1 + \frac{\varepsilon}{3}} \right\} \right)$$

$$\leq e^{-d\left(\frac{\varepsilon^2 - \varepsilon^3}{36} \right)}$$

More sparsity against larger d

- coefficient matrix $U \in \mathbb{R}^{d \times p}$ of iid $1 - \mathrm{sG}$ with variance 1
- mask matrix $\zeta \in \mathbb{R}^{d \times p}$ of iid $\mathcal{B}(q)$, $\perp\!\!\!\perp$ U
- $A_{k\ell} = \frac{1}{\sqrt{dq}} \zeta_{k\ell} U_{k\ell}$

Theorem

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ and $\varepsilon > 0$. If d such that

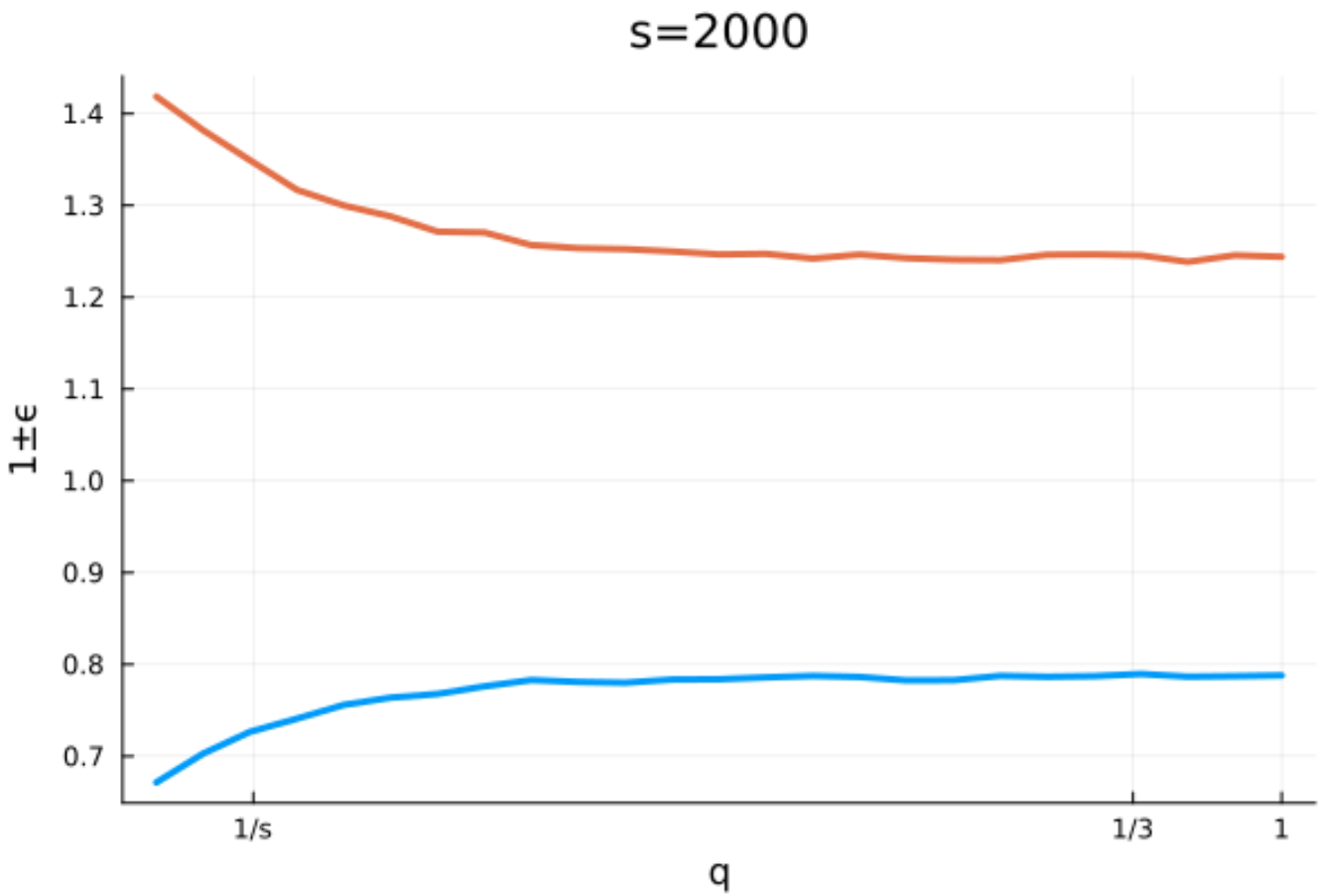
$$d \geq \frac{12}{\varepsilon^2} \log(3n/\delta) \left(1 + \sqrt{4 \log(nd/\delta)} + 2 \log(nd/\delta)\right)^2 ,$$

and if

$$q \geq \max_{i \neq j} \frac{\|x_i - x_j\|_\infty^2}{\|x_i - x_j\|_2^2}$$

then A is an ε-quasi-isometry with probability at least $1 - \delta$.

Is it what we can see?

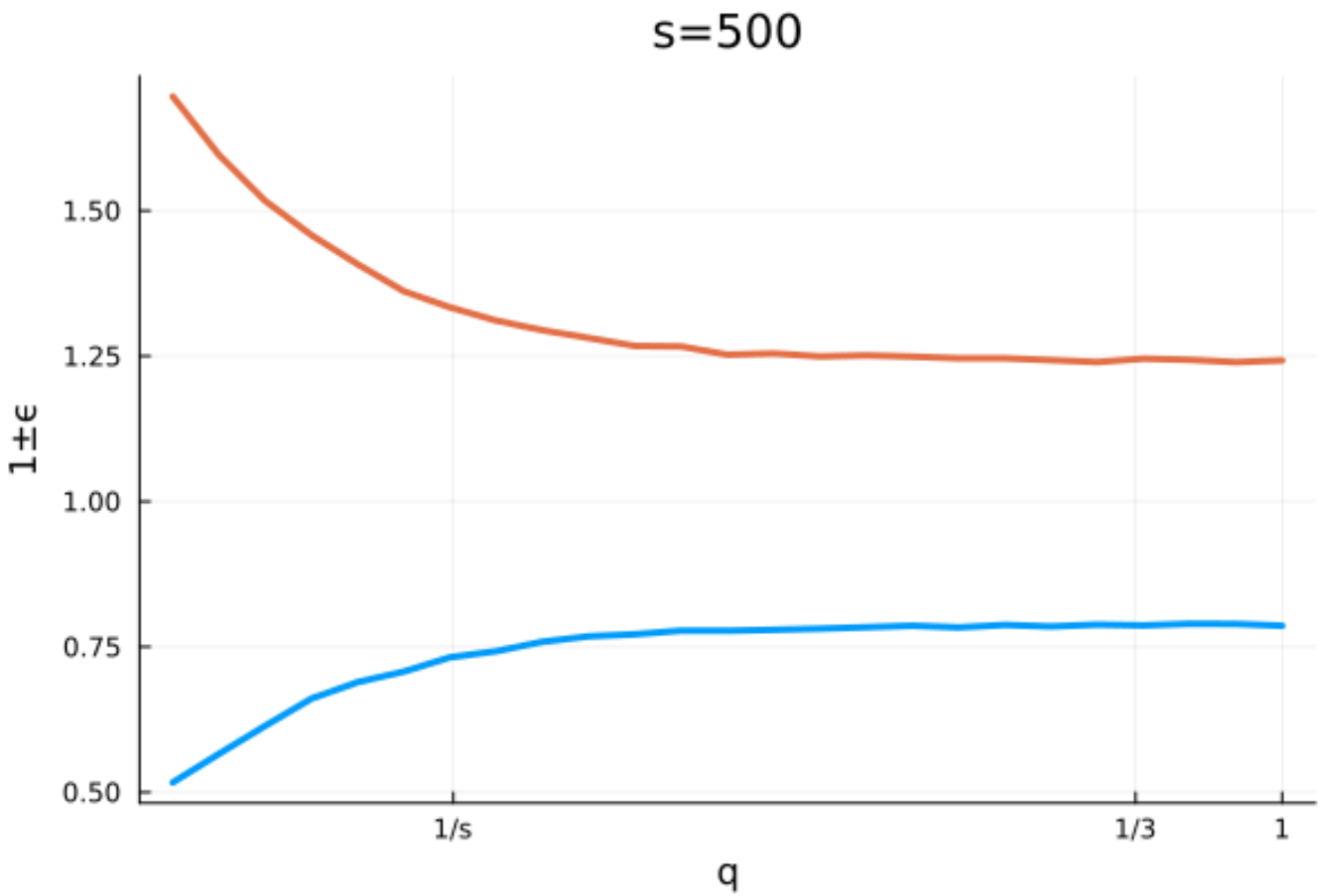


$\max_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$\min_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$(x_i)_{1 \leq i \leq n=100}$ with s non-zero indep $\mathcal{N}(0,1)$ $\quad p = 10^5 \quad d = 500$

Is it what we can see?

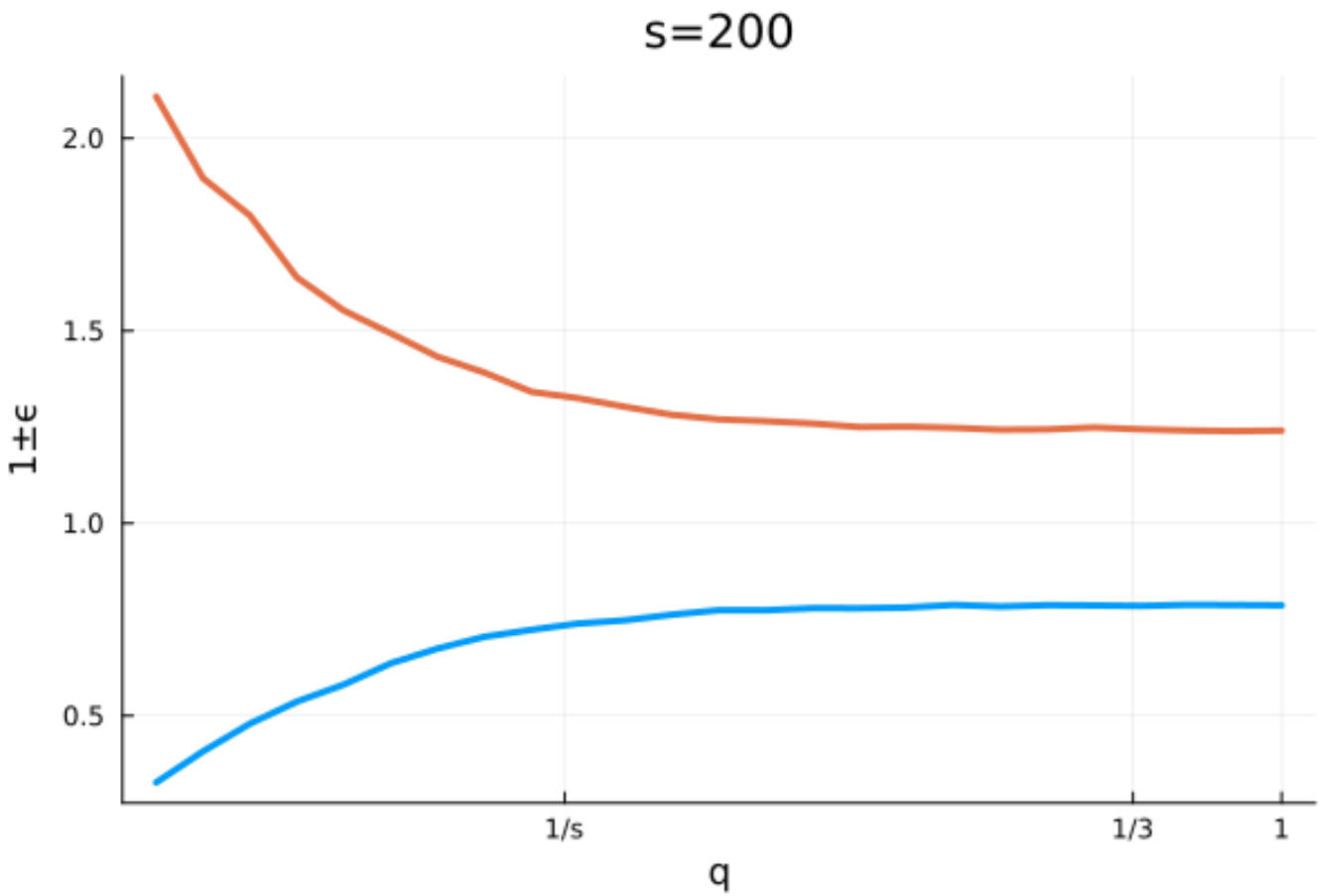


$\max_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$\min_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$(x_i)_{1 \le i \le n=100}$ with s non-zero indep $\mathcal{N}(0,1)$ $\quad p = 10^5 \quad d = 500$

Is it what we can see?

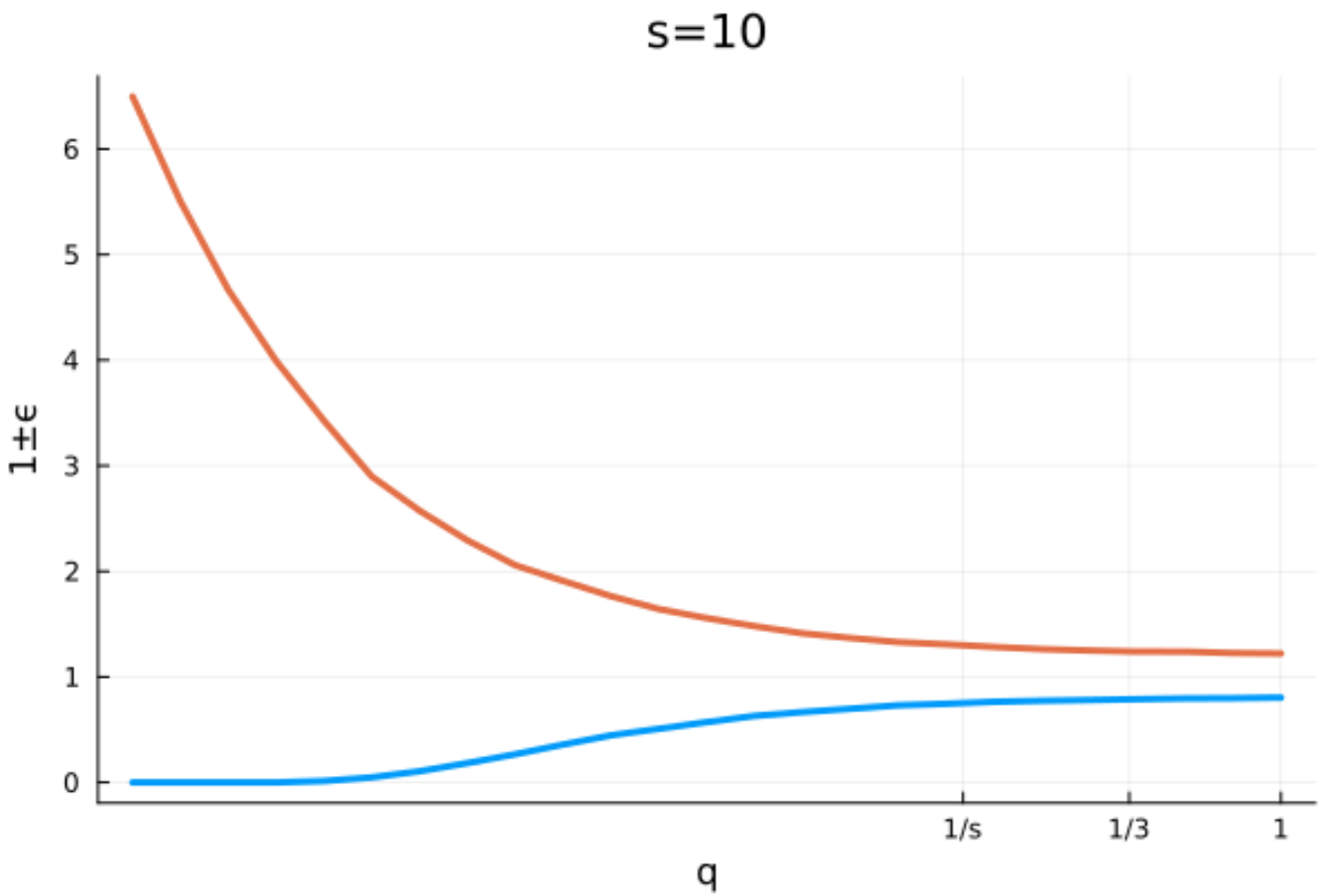


$\max_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$\min_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$(x_i)_{1 \le i \le n=100}$ with s non-zero indep $\mathcal{N}(0,1)$ $\quad p = 10^5 \quad d = 500$

Is it what we can see?

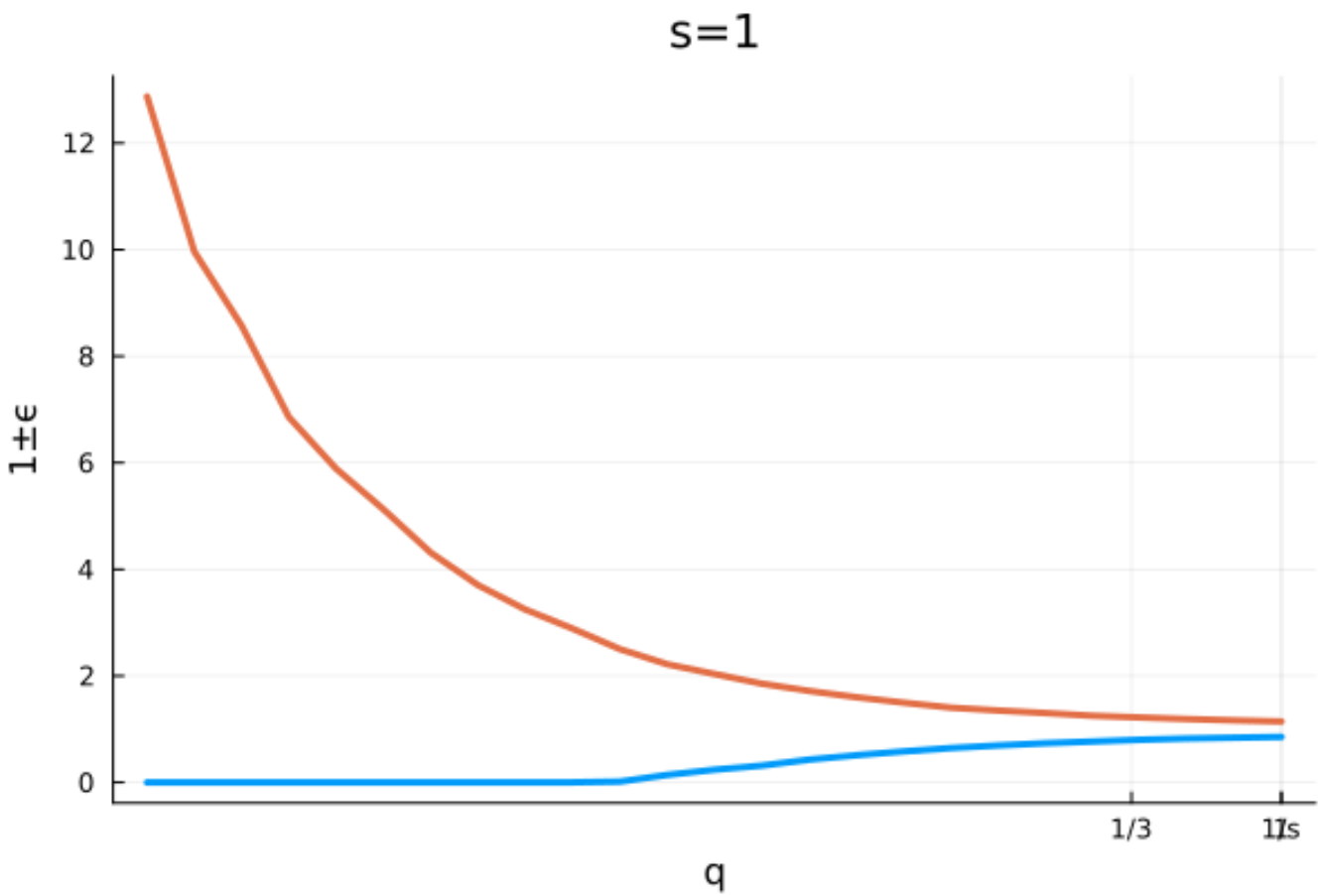


$\max_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$\min_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$(x_i)_{1 \le i \le n=100}$ with s non-zero indep $\mathcal{N}(0,1)$ $\quad p = 10^5 \quad d = 500$

Is it what we can see?



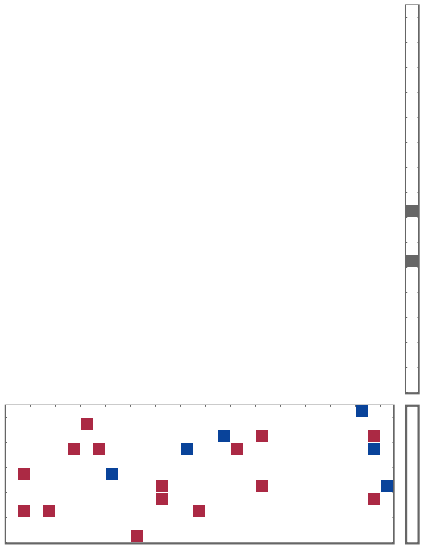
$\max_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$\min_{i,j} \frac{\|A(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2}$

$(x_i)_{1 \le i \le n=100}$ with s non-zero indep $\mathcal{N}(0,1)$ $\quad p = 10^5 \quad d = 500$

A lower bound

Think of a s-sparse vector:
$\mathbb{P}(Y = 0) \geq (1-q)^{ds} \approx e^{-qds}$ – the quasi-isometry property can hold only if $q \gg \frac{1}{ds}$



More precisely, if $d \sim 1/\varepsilon^2$ we need $q \gg 1/s$:

Theorem

Assume that $A \in \mathbb{R}^{d \times p}$ has iid entries $A_{k\ell} = \frac{1}{\sqrt{dq}} \zeta_{k\ell} U_{k\ell}$ as with $\zeta_{k\ell} \overset{iid}{\sim} \mathcal{B}(q) \perp\!\!\!\perp U_{k\ell} \overset{iid}{\sim} \mathcal{U}(\{-1,1\}$. Let $y \in \mathbb{R}^p$ be a unit vector with coordinates in $\{-1/\sqrt{s}, 1/\sqrt{s}, 0\}$.
If $q \leq \frac{1}{\big(2(d\varepsilon^2) \vee 240\big)s}$, $\mathbb{P}\Big(\|Ay\|^2 \in [1-\varepsilon, 1+\varepsilon]\Big) \leq 1 - e^{-5000}$.

Thank you!

ENS DE LYON

