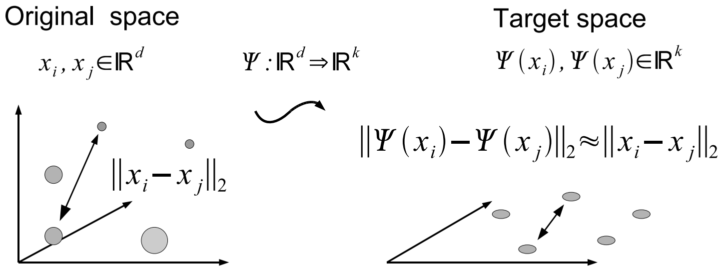
Accelerated Dense Random Projections

Edo Liberty[1]

Advisor: Steven Zucker

[1] Yale University, Department of Computer Science.

Dimensionality reduction



$$(1-\varepsilon)\|x_i - x_j\|_2 \leq \|\Psi(x_i) - \Psi(x_j)\|_2 \leq (1+\varepsilon)\|x_i - x_j\|_2$$

- $\binom{n}{2}$ distances are ε preserved
- Target dimension k smaller than original dimension d

What are they good for?

We will see that:

- The target dimension k can be significantly smaller than d.
- Ψ can be chosen independently of x_i.

This makes random projection very useful in:

- Approximate-nearest-neighbor algorithms
- Linear Embedding / Dimensionality reduction
- Rank k approximation
- ℓ_1 and ℓ_2 regression
- Compressed sensing
- Learning

...

Simple image search example

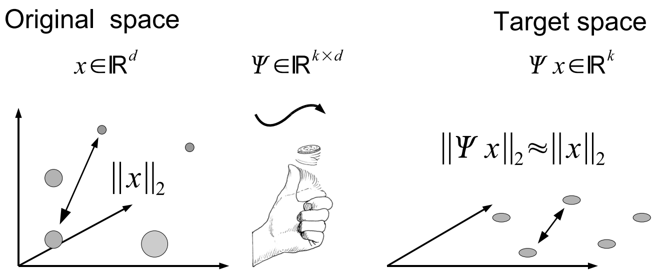
Simple task: search through your library of $10,000$ images for near duplicates (on your PC).

Problem: your images are 5 Mega-pixels each. Your library occupies 22 Gigabytes of disk space and does not fit in memory.

Possible solution: Project each image to a lower dimension (say 500). Then, search for close neighbors in the embedded points.

This can be done in memory on a moderately strong computer.

Random projections



A distribution $\mathbb{D}$ over $k \times d$ matrices Ψ s.t.

$$\forall_{x \in \mathbb{S}^{d-1}} \Pr_{\Psi \sim \mathbb{D}} [|\|\Psi x\|_2 - 1| > \varepsilon] \leq 1/n^2$$

All $\binom{n}{2}$ pairwise distances are preserved w.p. at least $1/2$.

Johnson Lindenstrauss Lemma

Lemma (Johnson Lindenstrauss (1984))

Let $\mathbb{D}$ denote the uniform distribution over all $k \times d$ projections

$$\forall\, x \in \mathbb{S}^{d-1} \Pr_{\Psi \sim \mathbb{D}} [|\|\Psi x\|_2 - 1| > \varepsilon] \leq c_1 e^{-c_2 \varepsilon^2 k}$$

This gives $\Pr \leq 1/n^2$ for $k = \Theta(log(n)/\varepsilon^2)$.

Definition

Such distributions are said to exhibit the JL property.

Johnson Lindenstrauss proof sketch

The distribution $\mathbb{D}$ is rotation invariant, thus:

$$\Pr_{\Psi \sim \mathbb{D}} [|\|\Psi x\|_2 - 1| > \varepsilon] = \Pr_{x \sim U(\mathbb{S}^{d-1})} [|\|I_k x\|_2 - 1| > \varepsilon]$$

Informally: projecting any *fixed* vector on a *random subspace* is equivalent to projecting a *random* vector on a fixed *subspace*.

The rest follows directly from the isoperimetric inequality on the sphere.

Gaussian i.i.d. distribution

Lemma (Frankl Meahara (1987))

Let $\mathbb{D}$ denote an i.i.d. Gaussian distribution for each entry of Ψ. Then, $\mathbb{D}$ exhibits the JL property.

Proof.

Due to the rotational invariance of $\mathbb{D}$

$$\Pr_{\Psi \sim \mathbb{D}} \left[\left| \|\Psi x\|_2 - 1 \right| > \varepsilon \right] = \Pr_{\Psi \sim \mathbb{D}} \left[\left| \|\Psi e_1\|_2 - 1 \right| > \varepsilon \right].$$

Also, $\|\Psi e_1\|_2 = \left\|\Psi^{(1)}\right\|_2$ which is distributed as χ^2 with k degrees of freedom. □

± 1 and Sub-Gaussian i.i.d. distributions

Lemma (Achlioptas (2003))

Let $\mathbb{D}$ denote an i.i.d. ± 1 distribution for each entry of Ψ. Then, $\mathbb{D}$ exhibits the JL property.

Proof.

$$\|\Psi x\|_2^2 = \sum_{i=1}^{k} \langle \Psi_{(i)}, x \rangle^2 = \sum_{i=1}^{k} y_i^2$$

The random variables y_i are i.i.d. and sub-Gaussian (Due to Hoeffding).

□

The proof above is due to Matousek (2006).

The need for speed

All of the above distributions are such that:

- Ψ requires $O(kd)$ space to store.
- Mapping $x \mapsto \Psi x$ requires $O(kd)$ operations.

Example: projecting a 5 Megapixel image to dimension 500:

- Ψ takes up roughly 10 Gigabytes of memory.
- It takes roughly 5 hours to compute $x \mapsto \Psi x$.
  (very optimistic estimate for a 2Ghz CPU)

Sparse i.i.d. distributions

Assume that $\mathbb{D}$ is such that $\Psi(i,j)$ is non-zero w.p. q.

Can $\mathbb{D}$ exhibit the JL property and $q = o(1)$?

We must have that

$$\Pr_{\Psi\sim\mathbb{D}}[|\|\Psi e_1\|_2 - 1| > \varepsilon] = \Pr_{\Psi\sim\mathbb{D}}\left[\left|\left\|\Psi^{(1)}\right\|_2 - 1\right| > \varepsilon\right] \leq 1/n^2$$

Thus, $\Psi^{(1)}$ must rely on $\Omega(\log(n))$ random bits.

This cannot be achieved!

Sparse i.i.d. distributions

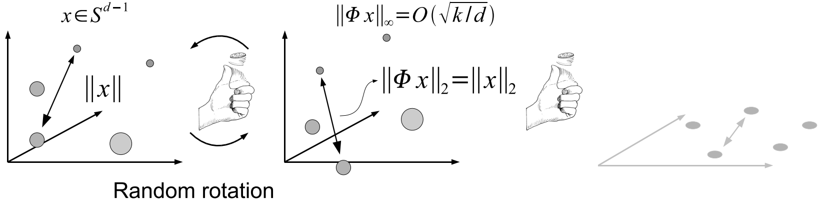
Lemma (Matousek (2006) Ailon Chazelle (2006))

Let $x \in \mathbb{S}^{d-1}$ be such that $\|x\|_\infty \leq \eta$. Let $\mathbb{D}$ be such that:

$$\Psi(i,j) = \begin{cases} 1/\sqrt{q} & w.p. \quad q/2 \\ -1/\sqrt{q} & w.p. \quad q/2 \\ 0 & w.p. \quad 1-q. \end{cases}$$

for some $q \in O(\eta^2 k)$,
$\mathbb{D}$ exhibits the JL property with respect to x.

FJLT, random rotation



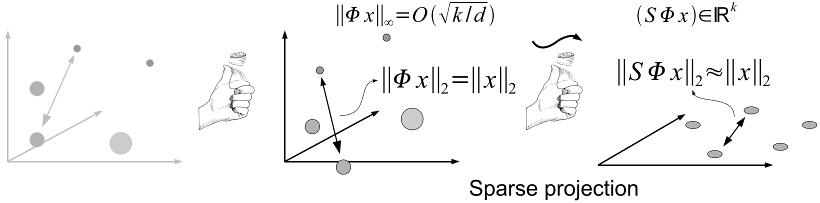
Lemma (Ailon, Chazelle (2006))

Let Φ be HD:

- *H is a Hadamard transform*
- *D is a random ± 1 diagonal matrix*

$$\forall x \in \mathbb{S}^{d-1} \quad w.h.p. \quad \|\Phi x\|_\infty \leq \sqrt{k/d}\}$$

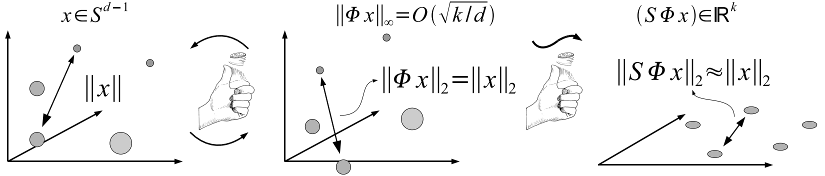
FJLT, sparse projection



Lemma (Ailon, Chazelle (2006))

After the rotation, an expected number of $O(k^3)$ nonzeros in S is sufficient for the JL property to hold.

FJLT algorithm, random rotation + sparse projection



Random rotation
Requires $d\,log(d)$ operations

Sparse projection
S contains $O(k^3)$ non zeros in expectation

Lemma (Ailon, Chazelle (2006))

Let $\mathbb{D}$ be the above distribution. $\mathbb{D}$ exhibits the JL property. Moreover, computing $x \mapsto S\Phi x$ requires $O(d \log(d) + k^3)$ operations in expectation.

Statement of results

Previous algorithms' application complexity:

	Naïve or Slower	Faster than naïve	$O(d\log(k))$	Optimal, $O(d)$
k in $O(\log d)$	JL, FJLT			
k in $\omega(\log d)$ and $o(\text{poly}(d))$	JL	FJLT		
k in $\Omega(\text{poly}(d))$ and $o((d\log(d))^{1/3})$	JL		FJLT	
k in $\omega((d\log d)^{1/3})$ and $O(d^{1/2-\delta})$	JL	FJLT		
k in $O(d^{1/2-\delta})$ and $k < d$	JL, FJLT			

Statement of results

Our contributions either match or outperform pervious results.

	Naïve or Slower	Faster than naïve	$O(d\log(k))$	Optimal, $O(d)$
k in $O(\log d)$	JL, FJLT, FWI		FJLTr	**JL + Mailman**
k in $\omega(\log d)$ and $o(\mathrm{poly}(d))$	JL	FJLT, FWI	**FJLTr**	
k in $\Omega(\mathrm{poly}(d))$ and $o((d\log(d))^{1/3})$	JL		FJLT, **FJLTr**, **FWI**	
k in $\omega((d\log d)^{1/3})$ and $O(d^{1/2-\delta})$	JL	FJLT, FJLTr	**FWI**	
k in $O(d^{1/2-\delta})$ and $k < d$	JL, FJLT, FJLTr	**JL concatenation**		

- Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes. SODA 08, best papers invitation to TALG, Discrete and Computational Geometry 08. with Nir Ailon.
- Dense Fast Random Projections and Lean Walsh Transforms. RANDOM 08. with Nir Ailon and Amit singer.
- The Mailman algorithm: a note on matrix vector multiplication. IPL 08. with Steven Zucker.

One dimensional Random walks

Consider the random walk distance:

$$Y = |\sum_{i=1}^{d} v(i)s(i)|$$

- $v(i) \in \mathbb{R}$ are scaler step sizes.
- $s(i)$ are ± 1 w.p $1/2$ each.

We have from Hoeffding's inequality that:

$$\Pr[Y - E[Y] \geq t] \leq e^{-t^2/2\|v\|_2^2}.$$

This can be slightly modified to obtain:

$$\Pr[Y - \|v\|_2 \geq \varepsilon] \leq c_1 e^{-c_2\varepsilon^2/\|v\|_2^2}$$

High dimensional Random walks

Now consider the walk:

$$Y = \left\| \sum_{i=1}^{d} M^{(i)} s(i) \right\|_2$$

- $M^{(i)} \in \mathbb{R}^k$ are *vector* valued steps.
- $s(i)$ are still ± 1 w.p $1/2$ each.

Lemma

$$\Pr\left[|Y - \|M\|_{Fro}| \geq \varepsilon \right] \leq c_1 e^{-c_2 \varepsilon^2 / \|M\|_2^2}$$

- *M is a matrix whose i'th column is $M^{(i)}$.*
- *$\|M\|_{Fro}$ and $\|M\|_2$ stand for the Frobenius and spectral norms of M.*

Measure concentration on the hypercube

Lemma (Ledoux Talagrand (1991))

Let $f : [0,1]^d \rightarrow \mathbb{R}$ *be a convex function.*
Let $\mathbb{D}$ *be a probability product space over* $[0,1]^d$*.*

$$\Pr_{s \sim \mathbb{D}} [|f(s) - \mu| > t] \leq 4e^{-t^2/8\|f\|_{Lip}^2}.$$

Here μ *is a median on* f *and* $\|f\|_{Lip}$ *is its Lipschitz constant.*

High dimensional Random walks

Setting $f(s) \leftarrow \left\| \sum_{i=1}^{d} M^{(i)} s(i) \right\|_2 = \|Ms\|_2$:

- $f(s)$ is convex, by convexity of the 2-norm.
- $\|f\|_{Lip} = \|M\|_2$, by definition.
- $|\mu - \|M\|_{Fro}| = O(\|M\|_2)$ (requires derivation).

Substituting into the hypercube concentration result we get

$$\Pr\left[|Y - \|M\|_{Fro}| \geq \varepsilon\right] \leq c_1 e^{-c_2 \varepsilon^2 / \|M\|_2^2}$$

as required.

From random walks to random projections

Consider the distribution $\Psi = AD$:

- A is a *fixed* $k \times d$ matrix.
- D is a diagonal matrix, $D(i,i) = s(i)$ (Rademacher).

We have that:

$$\|ADx\|_2 = \left\|\sum_{i=1}^{d} A^{(i)} D(i,i) x(i)\right\|_2 = \left\|\sum_{i=1}^{d} A^{(i)} x(i) s(i)\right\|_2 = \|Ms\|_2$$

where $M^{(i)} = A^{(i)} x(i)$.

From random walks to random projections

The random walk concentration result,

$$\Pr\left[\left|\|Ms\|_2 - \|M\|_{Fro}\right| \geq \varepsilon\right] \leq c_1 e^{-c_2\varepsilon^2/\|M\|_2^2},$$

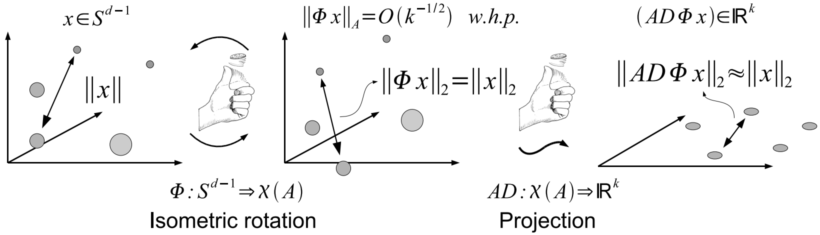
gives the JL property

$$\Pr\left[\left|\|ADx\|_2 - 1\right| \geq \varepsilon\right] \leq c_1 e^{-c_2\varepsilon^2 k}$$

If

- $\|M\|_{Fro} = 1$ true if A is column normalized.
- $\|M\|_2 = O(k^{-1/2})$.

Two stage projection process



Definition

$\|x\|_A \equiv \|M\|_2$, where $M^{(i)} = A^{(i)} x(i)$.

Definition

$\chi(A) \equiv \{x \in \mathbb{S}^{d-1} \mid \|x\|_A = O(k^{-1/2})\}$.

If $\|\Phi x\|_A = O(k^{-1/2})$ w.h.p., then $AD\Phi$ exhibits the *JL* property.

Four-wise independent projection matrix

Lemma

For a four-wise independent matrix, B:

$$\|x\|_4 = O(d^{-1/4}) \quad \rightarrow x \in \chi(B)$$

Lemma

If $k = O(d^{1/2})$, *there exists a* $k \times d$ *four-wise independent matrix* B *such that computing* $z \mapsto Bz$ *requires* $O(d \log(k))$ *operations.*

Four-wise independent projection matrix

Lemma

If $k = O(d^{1/2-\delta})$, *there exists a random rotation* Φ *such that* $\|\Phi x\|_4 = O(d^{-1/4})$ *w.p. at least* $1 - O(e^{-k})$.

Lemma

Computing $x \mapsto \Phi x$ *requires* $O(d \log(d))$ *operations.*

Thus $BD\Phi$ exhibits the JL property.

Improvement over the FJLT algorithm

- FJLT running time: $O(d\log(d) + k^3)$.
- FWI running time: $O(d\log(d))$ for $k \in O(d^{1/2-\delta})$.

	Naïve or Slower	Faster than naïve	$O(d\log(k))$	Optimal, $O(d)$
k in $O(\log d)$	JL, FJLT, FWI			
k in $\omega(\log d)$ and $o(\text{poly}(d))$	JL	FJLT, FWI		
k in $\Omega(\text{poly}(d))$ and $o((d\log(d)^{1/3})$	JL		FJLT, FWI	
k in $\omega((d\log d)^{1/3})$ and $O(d^{1/2-\delta})$	JL	FJLT	**FWI**	
k in $O(d^{1/2-\delta})$ and $k < d$	JL, FJLT			

The mailman algorithm

The running time lower bound for random projections is $O(d)$. Can this be achieved?

Claim

Any $k \times d$, ± 1 matrix, Ψ, can be applied to any vector $x \in \mathbb{R}^d$ in $O(kd/log(d))$ operation.

If $k = O(log(d))$, then a random i.i.d. ± 1 projection can be applied to vectors in optimal $O(d)$ time.

The mailman algorithm

For simplicity, assume Ψ is $k \times d$ and $d = 2^k$.

We have that $\Psi = UP$ if:

- U contains each possible column $\{+1, -1\}^k$.
- $P(i,j) = \delta(U^{(i)}, A^{(j)})$

Computing $x \mapsto Px$ requires $O(d)$ operations since P contains only d non-zeros.

The mailman algorithm

Applying U also requires only $O(d)$ operations.

$$U_2 = \left(1 \;\vdots\; -1 \right), \quad U_d = \left(\begin{array}{c:c} 1,\ldots,1 & -1,\ldots,-1 \\ \hdashline U_{d/2} & U_{d/2} \end{array} \right)$$

$$U_d z = \left(\begin{array}{c:c} 1,\ldots,1 & -1,\ldots,-1 \\ \hdashline U_{d/2} & U_{d/2} \end{array} \right) \left(\begin{array}{c} z_1 \\ \hdashline z_2 \end{array} \right) = \left(\begin{array}{c} \sum_{i=1}^{d/2} z_1(i) - z_2(i) \\ \hdashline U_{d/2}(z_1 + z_2) \end{array} \right)$$

This gives the following recursion:

$$T(d) = T(d/2) + O(d) \quad \Rightarrow \quad T(d) = O(d)$$

Remark

If $k \geq \log(d)$, Ψ *can be sectioned into* $\lceil k/\log(d) \rceil$ *submatrices of size at most* $\log(d) \times d$.

Mailman application speed

Running time for multiplying a $\log(d) \times d$ random ± 1 matrix to a double precision vector.

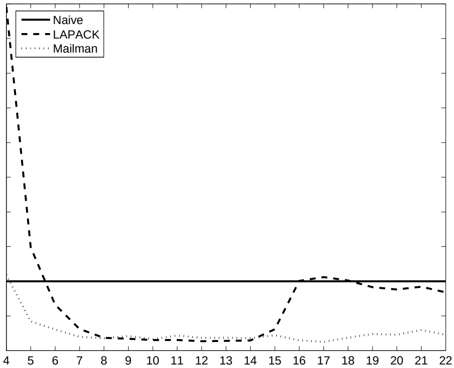


Figure: The experiments were run Xeon Quad core 2.33GHz machine running Linux Ubuntu with 8G of RAM and a Bus speed of 1333MHz.

Linear time projection

Using the Mailman algorithm gives the first $O(d)$ algorithm.

	Naïve or Slower	Faster than naïve	$O(d\log(k))$	Optimal, $O(d)$
k in $O(\log d)$	JL, FJLT, FWI			**JL + Mailman**
k in $\omega(\log d)$ and $o(\text{poly}(d))$	JL	FJLT, FWI		
k in $\Omega(\text{poly}(d))$ and $o((d\log(d))^{1/3})$	JL		FJLT, FWI	
k in $\omega((d\log d)^{1/3})$ and $O(d^{1/2-\delta})$	JL	FJLT	FWI	
k in $O(d^{1/2-\delta})$ and $k < d$	JL, FJLT			

Linear time projection

Can we achieve an $O(d)$ running time in general?

Proving the contrary will give a super-linear running time lower bound on performing Fourier transforms...

Look for a $k \times d$ matrix, A, which:

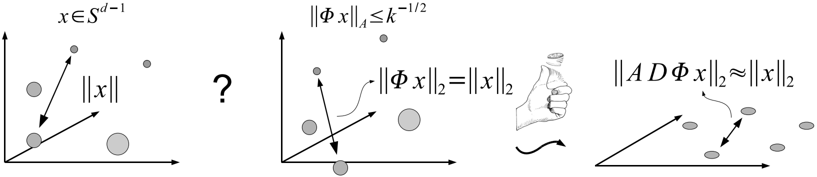
- is applicable in $O(d)$ operations
- exhibits the largest possible set $\chi(A)$.

Linear time projection

projection matrix	Application complexity	A is a good random projection for x if:
Any matrix		$\|x\|_A = O(k^{-1/2})$
four-wise independent	$O(d \log k)$	$\|x\|_4 = O(d^{-1/4})$
Lean Walsh	$O(d)$	$\|x\|_\infty = O(k^{-1/2} d^{-\delta})$
Identity copies	$O(d)$	$\|x\|_\infty = O((k \log k)^{-1/2})$

Table: Lean-Walsh matrices are dense ± 1 tensor product matrices. Identity-copies, is a horizontal concatenation of $\log(k)$ identity matrices.

Open questions



Random rotation
Applicable in $O(d)$ Operations ?

The matrix A requires
$O(d)$ Operations to apply

Can $O(d \log(d))$ running time be achieved for $k \in \omega(d^{1/2-\delta})$?

Open questions

	Naïve or Slower	Faster than naïve	$O(d\log(k))$	Optimal, $O(d)$
k in $O(\log d)$	JL, FJLT, FWI		FJLTr	**JL + Mailman**
k in $\omega(\log d)$ and $o(\text{poly}(d))$	JL	FJLT, FWI	**FJLTr**	**?**
k in $\Omega(\text{poly}(d))$ and $o((d\log(d)^{1/3})$	JL		FJLT, **FJLTr**, **FWI**	**?**
k in $\omega((d\log d)^{1/3})$ and $O(d^{1/2-\delta})$	JL	FJLT, FJLTr	**FWI**	**?**
k in $O(d^{1/2-\delta})$ and $k < d$	JL, FJLT, FJLTr	**JL concatenation**	**?**	**?**

Fin

Projection norm concentration

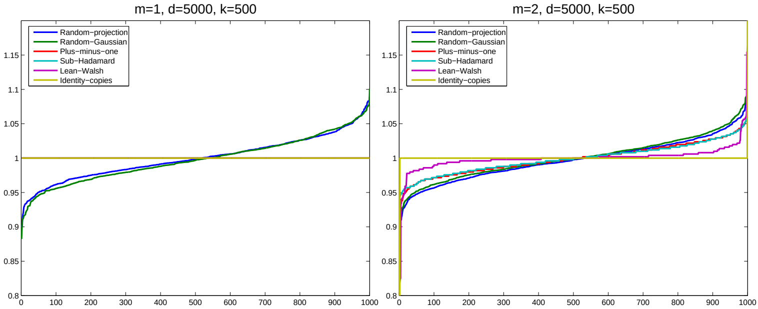


Figure: Accuracy of projection for six projection methods as a function of m, the number of non-zeros of value $1/\sqrt{m}$ in the input vectors. When $m = 1$ (left) all deterministic matrices exhibit zero distortion since their column norms are equal to 1. When $m = 2$ (right) all constructions might exhibit a distortion equal to their coherence.

Projection norm concentration

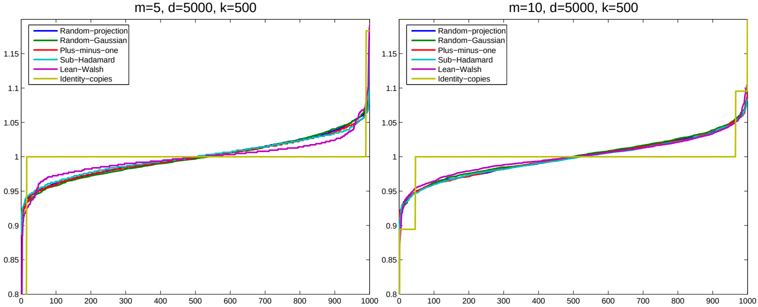


Figure: Small values of *m* give rise to better average behavior by deterministic matrices, but worse worst-case behavior. This stems from the fact that their average coherence is smaller but their maximum coherence is larger.

Projection norm concentration

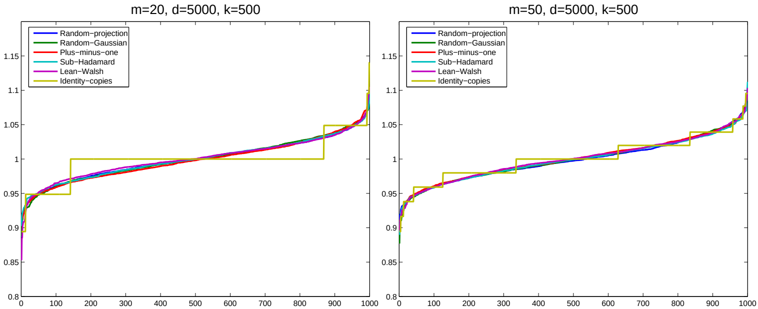


Figure: When *m* grows the behavior of deterministic matrices and dense random ones becomes indistinguishable, with the exception of Identity-copies.

Projection norm concentration

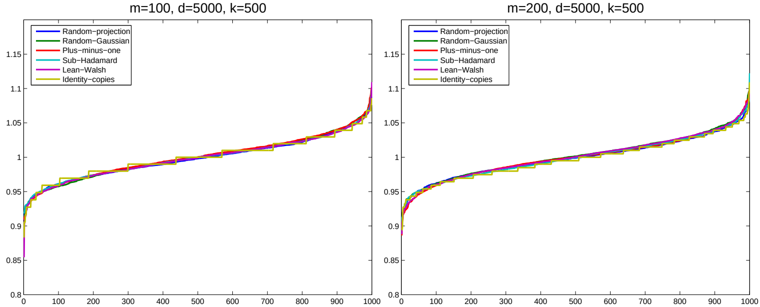


Figure: Large values of *m* allow all methods including Identity-copies to be used equally reliably.

Projection running time

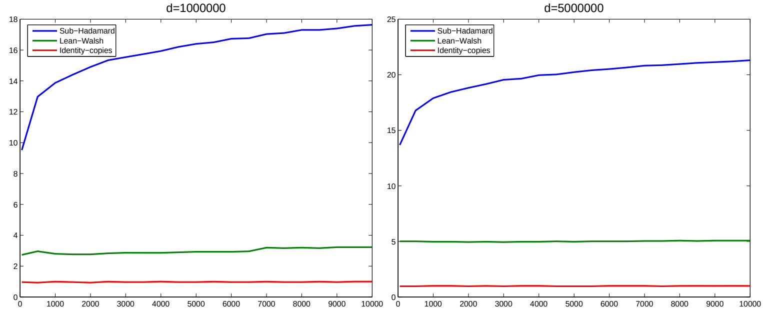


Figure: Running time of applying Sub-Hadamard, Lean-Walsh and Identity-copies $k \times d$ matrices. k ranges from 1 to 10^3 and $d = 10^5$ (left) $d = 5 \cdot 10^6$ (right).