Closing the Theory-Practice Gap in Oblivious Subspace Embeddings

Michał Dereziński

University of Michigan

Joint work with Shabarish Chenakkod, Xiaoyu Dong, and Mark Rudelson

Randomized Numerical Linear Algebra Workshop
ICERM, Providence, RI

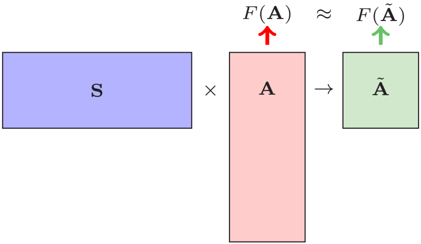
February 5, 2026

Outline

Sketching and subspace embeddings

Sketching

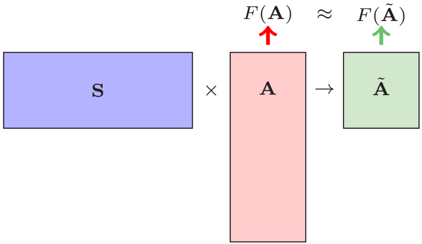
- Least squares
- Low-rank approximation
- Linear/Non-linear programming
- ...



Sketching and subspace embeddings

Sketching

- Least squares
- Low-rank approximation
- Linear/Non-linear programming
- ...



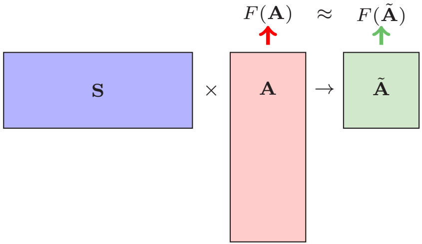
Intuition: We obtain a "good" sketch $\tilde{\mathbf{A}} = \mathbf{SA}$ of a tall matrix $\mathbf{A}$, if $\mathbf{S}$ is a "subspace embedding" for the range$(\mathbf{A}) = \{\mathbf{z} \mid \mathbf{z} = \mathbf{Ax}\}$,

$$(1-\epsilon)\|\mathbf{Ax}\| \le \|\mathbf{SAx}\| \le (1+\epsilon)\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

Sketching and subspace embeddings

Sketching

- Least squares
- Low-rank approximation
- Linear/Non-linear programming
- ...



Intuition: We obtain a "good" sketch $\tilde{\mathbf{A}} = \mathbf{SA}$ of a tall matrix $\mathbf{A}$, if $\mathbf{S}$ is a "subspace embedding" for the range$(\mathbf{A}) = \{\mathbf{z} \mid \mathbf{z} = \mathbf{Ax}\}$,

$$(1-\epsilon)\|\mathbf{Ax}\| \leq \|\mathbf{SAx}\| \leq (1+\epsilon)\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

Example: Approximating tall least squares, $\operatorname{argmin}_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|^2$, by

- Sketch-and-Solve: $\operatorname{argmin}_{\mathbf{x}} \|\mathbf{SAx} - \mathbf{Sb}\|^2$
- Sketch-and-Precondition: $\operatorname{argmin}_{\mathbf{z}} \|\mathbf{AR}^{-1}\mathbf{z} - \mathbf{b}\|^2$ for $\mathbf{SA} = \mathbf{QR}$.

*Of course, subspace embeddings are used not just for sketching tall matrices.

Oblivious Subspace Embeddings (OSE)

Definition (Oblivious Subspace Embedding)

Random $m \times n$ sketching matrix $\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$(1-\epsilon)\|\mathbf{Ax}\| \le \|\mathbf{SAx}\| \le (1+\epsilon)\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

Oblivious Subspace Embeddings (OSE)

Definition (Oblivious Subspace Embedding)

Random $m \times n$ sketching matrix $\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$(1-\epsilon)\|\mathbf{Ax}\| \leq \|\mathbf{SAx}\| \leq (1+\epsilon)\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

Popular example: Gaussian sketch

Random matrix with i.i.d. entries distributed as $\mathcal{N}(0, 1/m)$

Oblivious Subspace Embeddings (OSE)

Definition (Oblivious Subspace Embedding)

Random $m \times n$ sketching matrix $\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\| \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\| \qquad \text{for all } \mathbf{x}.$$

Popular example: Gaussian sketch
Random matrix with i.i.d. entries distributed as $\mathcal{N}(0, 1/m)$

Gaussian OSE is optimal

Gaussian sketching matrix is an (ϵ, δ, d)-OSE for $m = O(d/\epsilon^2)$.
Every (ϵ, δ, d)-OSE satisfies $m = \Omega(d/\epsilon^2)$.

Oblivious Subspace Embeddings (OSE)

Definition (Oblivious Subspace Embedding)

Random $m \times n$ sketching matrix $\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\| \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\| \qquad \text{for all } \mathbf{x}.$$

Popular example: Gaussian sketch
Random matrix with i.i.d. entries distributed as $\mathcal{N}(0, 1/m)$

Gaussian OSE is optimal

Gaussian sketching matrix is an (ϵ, δ, d)-OSE for $m = O(d/\epsilon^2)$.
Every (ϵ, δ, d)-OSE satisfies $m = \Omega(d/\epsilon^2)$.

Problem: Matrix multiplication $\mathbf{SA}$ is expensive!

Oblivious Subspace Embeddings (OSE)

Definition (Oblivious Subspace Embedding)

Random $m \times n$ sketching matrix $\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\| \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\| \qquad \text{for all } \mathbf{x}.$$

Popular example: Gaussian sketch

Random matrix with i.i.d. entries distributed as $\mathcal{N}(0, 1/m)$

Gaussian OSE is optimal

Gaussian sketching matrix is an (ϵ, δ, d)-OSE for $m = O(d/\epsilon^2)$.

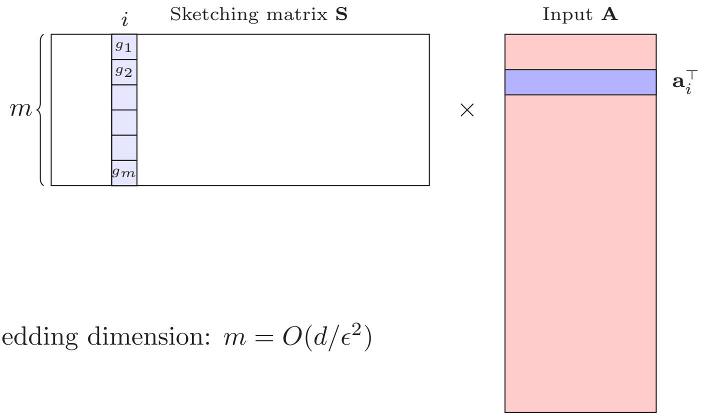
Every (ϵ, δ, d)-OSE satisfies $m = \Omega(d/\epsilon^2)$.

Problem: Matrix multiplication $\mathbf{SA}$ is expensive!

Solution: Sparse matrix $\mathbf{S}$ such as the CountSketch. [CW, STOC 2013]

Gaussian Sketch vs CountSketch

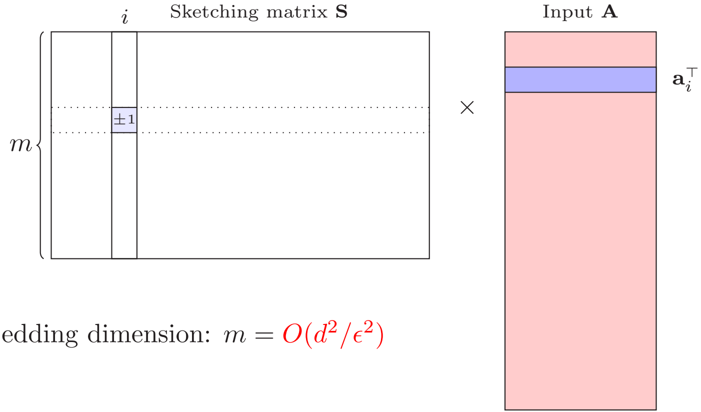
Gaussian sketch



Embedding dimension: $m = O(d/\epsilon^2)$

Gaussian Sketch vs CountSketch

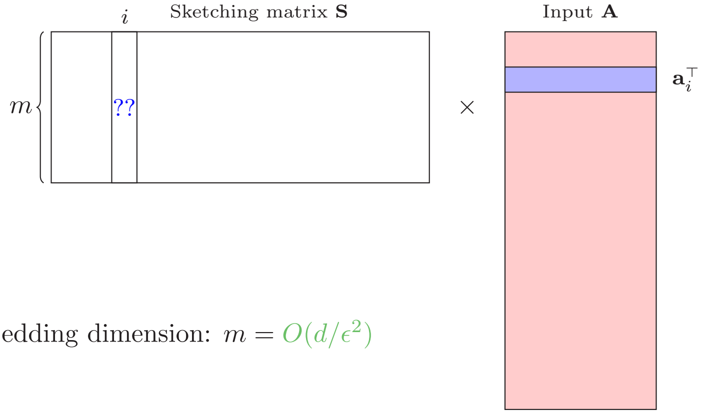
CountSketch



Embedding dimension: $m = O(d^2/\epsilon^2)$

Gaussian Sketch vs CountSketch

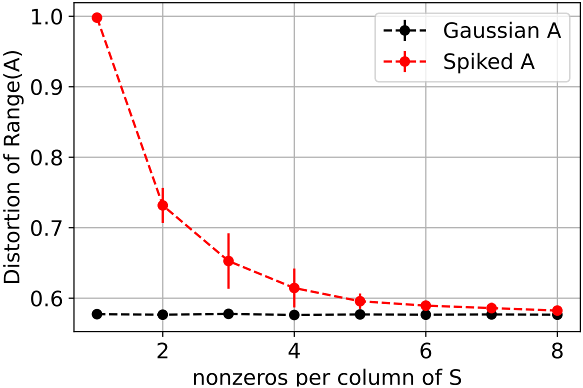
Optimal OSE?



Embedding dimension: $m = O(d/\epsilon^2)$

Question: What is the optimal sparsity for an (ϵ, δ, d)-OSE with optimal embedding dimension $m = O(d/\epsilon^2)$?

Effect of sparsity on embedding distortion



Embedding distortion ϵ of a sparse matrix $\mathbf{S}$ for range($\mathbf{A}$), where $\mathbf{A} \in \mathbb{R}^{10^5 \times 2 \cdot 10^3}$ and $m = 3d$. The "Spiked" $\mathbf{A}$ is formed by stacking copies of the identity matrix and randomly scaling the rows by 10^4.

Sparse Oblivious Subspace Embeddings

Conjecture [Nelson & Nguyen, FOCS 2013]

There is an (ϵ, δ, d)-OSE with optimal dimension $m = O(d/\epsilon^2)$ and $O(\log(d)/\epsilon)$ non-zero entries (nnz) per column.

	dimension m	nnz per column	reference
Gaussian	d/ϵ^2	m	folklore
CountSketch	d^2/ϵ^2	1	[CW, STOC 2013]
OSNAP	$d\log^8(d)/\epsilon^2$	$\log^3(d)/\epsilon$	[NN, FOCS 2013]
	$d\log^2(d)/\epsilon^2$	$\log^4(d)/\epsilon^2$	[BDN, STOC 2015]
$\vdots$	$d\log(d)/\epsilon^2$	$\log(d)/\epsilon$	[Cohen, SODA 2016]

Sparse Oblivious Subspace Embeddings

Conjecture [Nelson & Nguyen, FOCS 2013]

There is an (ϵ, δ, d)-OSE with optimal dimension $m = O(d/\epsilon^2)$ and $O(\log(d)/\epsilon)$ non-zero entries (nnz) per column.

	dimension m	nnz per column	reference
Gaussian	d/ϵ^2	m	folklore
CountSketch	d^2/ϵ^2	1	[CW, STOC 2013]
OSNAP	$d\log^8(d)/\epsilon^2$	$\log^3(d)/\epsilon$	[NN, FOCS 2013]
	$d\log^2(d)/\epsilon^2$	$\log^4(d)/\epsilon^2$	[BDN, STOC 2015]
$\vdots$	$d\log(d)/\epsilon^2$	$\log(d)/\epsilon$	[Cohen, SODA 2016]
	d/ϵ^2	$\log^4(d)/\epsilon^6$	[CDDR, STOC 2024]

Sparse Oblivious Subspace Embeddings

Conjecture [Nelson & Nguyen, FOCS 2013]

There is an (ϵ, δ, d)-OSE with optimal dimension $m = O(d/\epsilon^2)$ and $O(\log(d)/\epsilon)$ non-zero entries (nnz) per column.

	dimension m	nnz per column	reference
Gaussian	d/ϵ^2	m	folklore
CountSketch	d^2/ϵ^2	1	[CW, STOC 2013]
OSNAP	$d\log^8(d)/\epsilon^2$	$\log^3(d)/\epsilon$	[NN, FOCS 2013]
	$d\log^2(d)/\epsilon^2$	$\log^4(d)/\epsilon^2$	[BDN, STOC 2015]
$\vdots$	$d\log(d)/\epsilon^2$	$\log(d)/\epsilon$	[Cohen, SODA 2016]
	d/ϵ^2	$\log^4(d)/\epsilon^6$	[CDDR, STOC 2024]
	d/ϵ^2	$\log^3(d)/\epsilon$	[CDD, ICALP 2025]

Sparse Oblivious Subspace Embeddings

Conjecture [Nelson & Nguyen, FOCS 2013]

There is an (ϵ, δ, d)-OSE with optimal dimension $m = O(d/\epsilon^2)$ and $O(\log(d)/\epsilon)$ non-zero entries (nnz) per column.

	dimension m	nnz per column	reference
Gaussian	d/ϵ^2	m	folklore
CountSketch	d^2/ϵ^2	1	[CW, STOC 2013]
OSNAP	$d\log^8(d)/\epsilon^2$	$\log^3(d)/\epsilon$	[NN, FOCS 2013]
	$d\log^2(d)/\epsilon^2$	$\log^4(d)/\epsilon^2$	[BDN, STOC 2015]
$\vdots$	$d\log(d)/\epsilon^2$	$\log(d)/\epsilon$	[Cohen, SODA 2016]
	d/ϵ^2	$\log^4(d)/\epsilon^6$	[CDDR, STOC 2024]
	d/ϵ^2	$\log^3(d)/\epsilon$	[CDD, ICALP 2025]
	$d/\epsilon^2 \cdot \log^{o(1)}(d)$	$\log(d)/\epsilon \cdot \log^{o(1)}(d)$	[CDD, SODA 2026]

[CDD26] Chenakkod, Dereziński, and Dong. *"Optimal Subspace Embeddings: Resolving Nelson-Nguyen Conjecture Up to Sub-Polylogarithmic Factors,"* SODA 2026.

Outline

Main result

Theorem (Sparse OSE)

For $\varepsilon, \delta \in (0,1)$ such that $1/\epsilon\delta \le \mathrm{poly}(d)$, there is a (ε, δ, d)-OSE with

$$m = \theta_1 \frac{d}{\varepsilon^2} \quad \textit{and} \quad \text{nnz-per-col} = \theta_2 \frac{\log d}{\varepsilon},$$

where θ_1 and θ_2 are both bounded by $(\log d)^{O(\frac{1}{\log\log\log d})}$.

- First to match the conjecture of Nelson & Nguyen [FOCS 2013] up to sub-polylogarithmic factors.
- Simply stack nnz-per-col CountSketches on top of each other.
- Same general result also yields $\theta_1 = O(1)$ and $\theta_2 = (\log d)^{O(1)}$.

[CDD26] Chenakkod, Dereziński, and Dong. *"Optimal Subspace Embeddings: Resolving Nelson-Nguyen Conjecture Up to Sub-Polylogarithmic Factors,"* SODA 2026.

Analysis: Non-asymptotic Gaussian universality

The analysis of subspace embeddings require bounding the extreme singular values of the matrix $\mathbf{SU}$ for an orthogonal basis $\mathbf{U}$:

1. Covering nets. Limited to primarily dense $\mathbf{S}$ (e.g., Gaussian), but attains optimal embedding dimension $m = O(d/\epsilon^2)$
2. Matrix Chernoff. Applies to extremely sparse $\mathbf{S}$, but limited to sub-optimal embedding dimension $m = \Omega(d \log(d)/\epsilon^2)$.

Our approach: Non-asymptotic Gaussian universality theory

New trace moment universality bounds via iterative decoupling.

Are Oblivious Subspace Embeddings the right notion?

Classical OSE definition

$\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with prob. $1 - \delta$

$$(1-\epsilon)\|\mathbf{Ax}\| \le \|\mathbf{SAx}\| \le (1+\epsilon)\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

[CDDR24] Chenakkod, Dereziński, Dong, and Rudelson. "Optimal embedding dimension for sparse subspace embeddings," STOC 2024.
[CEMT25] Camaño, Epperly, Meyer, and Tropp. "Faster linear algebra algorithms with structured random matrices," arXiv:2508.21189, 2025.

Are Oblivious Subspace Embeddings the right notion?

Classical OSE definition

$\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with prob. $1-\delta$

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}\| \le \|\mathbf{S}\mathbf{A}\mathbf{x}\| \le (1+\epsilon)\|\mathbf{A}\mathbf{x}\| \qquad \text{for all } \mathbf{x}.$$

- What about $\epsilon > 1$? [CDDR24]

[CDDR24] Chenakkod, Dereziński, Dong, and Rudelson. "Optimal embedding dimension for sparse subspace embeddings," STOC 2024.

[CEMT25] Camaño, Epperly, Meyer, and Tropp. "Faster linear algebra algorithms with structured random matrices," arXiv:2508.21189, 2025.

Are Oblivious Subspace Embeddings the right notion?

Classical OSE definition

$\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with prob. $1-\delta$

$$(1-\epsilon)\|\mathbf{Ax}\| \le \|\mathbf{SAx}\| \le (1+\epsilon)\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

- What about $\epsilon > 1$? [CDDR24]
- Do we need both inequalities? [CEMT25]

[CDDR24] Chenakkod, Dereziński, Dong, and Rudelson. "Optimal embedding dimension for sparse subspace embeddings," STOC 2024.
[CEMT25] Camaño, Epperly, Meyer, and Tropp. "Faster linear algebra algorithms with structured random matrices," arXiv:2508.21189, 2025.

Are Oblivious Subspace Embeddings the right notion?

Classical OSE definition

$\mathbf{S}$ is an (ϵ, δ, d)-OSE if for any $n \times d$ matrix $\mathbf{A}$, with prob. $1-\delta$

$$(1-\epsilon)\|\mathbf{Ax}\| \le \|\mathbf{SAx}\| \le (1+\epsilon)\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

- What about $\epsilon > 1$? [CDDR24]
- Do we need both inequalities? [CEMT25]

Motivation: Is a Gaussian sketch of dimension $m = 1.1d$ an OSE?

[CDDR24] Chenakkod, Dereziński, Dong, and Rudelson. "Optimal embedding dimension for sparse subspace embeddings," STOC 2024.

[CEMT25] Camaño, Epperly, Meyer, and Tropp. "Faster linear algebra algorithms with structured random matrices," arXiv:2508.21189, 2025.

Subspace embeddings and condition numbers

Generalized OSE: For any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$\alpha\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\| \leq \beta\|\mathbf{A}\mathbf{x}\| \qquad \text{for all } \mathbf{x}.$$

[MDM$^+$23] Murray et al. *"Randomized Numerical Linear Algebra: A Perspective on the Field with an Eye to Software,"* arXiv:2302.11474, 2023.

Subspace embeddings and condition numbers

Generalized OSE: For any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$\alpha\|\mathbf{Ax}\| \leq \|\mathbf{SAx}\| \leq \beta\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

Restricted condition number: Smallest $\kappa = \beta/\alpha$ such that this holds.

[MDM+23] Murray et al. *"Randomized Numerical Linear Algebra: A Perspective on the Field with an Eye to Software,"* arXiv:2302.11474, 2023.

Subspace embeddings and condition numbers

Generalized OSE: For any $n \times d$ matrix $\mathbf{A}$, with probability $1 - \delta$

$$\alpha\|\mathbf{Ax}\| \leq \|\mathbf{SAx}\| \leq \beta\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

Restricted condition number: Smallest $\kappa = \beta/\alpha$ such that this holds.

- The quantity of interest when using subspace embeddings for preconditioning (e.g., sketch-and-precondition least squares)
- Gaussian sketch of dimension $m = 1.1d$ satisfies $\kappa = O(1)$

[MDM+23] Murray et al. *"Randomized Numerical Linear Algebra: A Perspective on the Field with an Eye to Software,"* arXiv:2302.11474, 2023.

Subspace embeddings and condition numbers

Generalized OSE: For any $n \times d$ matrix $\mathbf{A}$, with probability $1-\delta$

$$\alpha\|\mathbf{Ax}\| \le \|\mathbf{SAx}\| \le \beta\|\mathbf{Ax}\| \qquad \text{for all } \mathbf{x}.$$

Restricted condition number: Smallest $\kappa = \beta/\alpha$ such that this holds.

- The quantity of interest when using subspace embeddings for preconditioning (e.g., sketch-and-precondition least squares)
- Gaussian sketch of dimension $m = 1.1d$ satisfies $\kappa = O(1)$

Question: What about a *sparse* sketch with dimension $m = 1.1d$?

[MDM+23] Murray et al. *"Randomized Numerical Linear Algebra: A Perspective on the Field with an Eye to Software,"* arXiv:2302.11474, 2023.

Result: Restricted condition number of sparse OSEs

Theorem (Restricted condition number of sparse sketch)

For any fixed $\theta > 0$ and large enough d, an embedding matrix $\mathbf{S}$ with

$$m = (1+\theta)d \quad \textit{and} \quad \text{nnz-per-col} = (\log(d)/\theta)^{O(1)},$$

with high probability has restricted condition number $O(1/\theta)$.

- Matches the restricted condition number of the Gaussian sketch.
- Conclusion: Even very small sparse sketches can be effective for preconditioning iterative solvers.

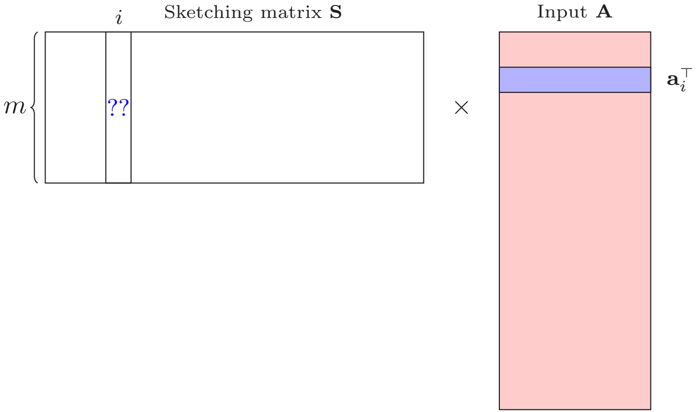
[CDDR24] Chenakkod, Dereziński, Dong, and Rudelson. *"Optimal embedding dimension for sparse subspace embeddings,"* STOC 2024.

Outline

Data-aware sparse embeddings

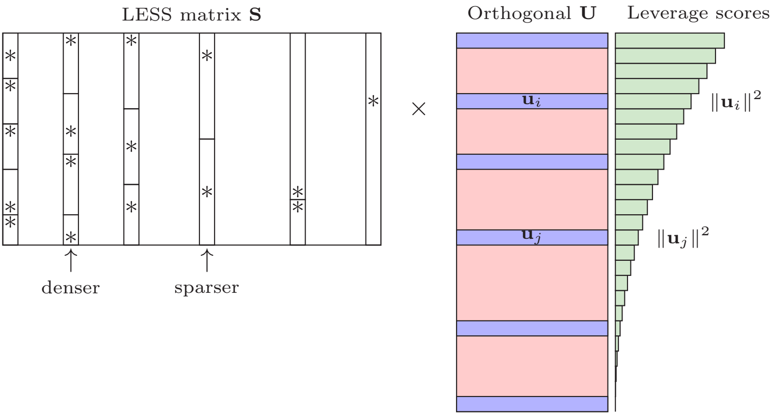
What if we want a subspace embedding for a specific subspace U?

Example: $U = \mathrm{range}(\mathbf{A})$ for a given matrix $\mathbf{A}$ (as in least squares)



Question: Can we further reduce the sparsity of $\mathbf{S}$ by knowing $\mathbf{A}$?

Leverage Score Sparsification (LESS)



Definition: Let $\mathbf{U}$ be the orthonormal basis matrix for range($\mathbf{A}$). The ith leverage score of $\mathbf{A}$ is the squared norm of the ith row of $\mathbf{U}$.

[DLDM21] Dereziński, Liao, Dobriban, and Mahoney. *"Sparse sketches with small inversion bias,"* COLT 2021.

Leverage score sparsified subspace embedding

Theorem ([CDD25])

Let $\mathbf{U}$ *be an* $n \times d$ *orthogonal matrix with rows* $\mathbf{u}_1, \ldots, \mathbf{u}_n$*, and let* $\mathbf{S}$ *be an* $m \times n$ *sketching matrix with* nnz_i *non-zeros in the* i*th column. If*

$$(LESS) \qquad m \geq C\frac{d}{\epsilon^2} \quad \textit{and} \quad \mathrm{nnz}_i \geq \max\left\{1,\ C\|\mathbf{u}_i\|^2 \frac{\log^3(d)}{\epsilon}\right\}$$

then with high probability $\mathbf{S}$ *is an embedding for* $\mathrm{range}(\mathbf{U})$*, i.e.:*

$$(1-\epsilon)\|\mathbf{x}\| \leq \|\mathbf{S}\mathbf{U}\mathbf{x}\| \leq (1+\epsilon)\|\mathbf{x}\| \qquad \textit{for all } \mathbf{x}.$$

[CDD25] Chenakkod, Dereziński, and Dong. *"Optimal oblivious subspace embeddings with near-optimal sparsity,"* ICALP 2025.

Leverage score sparsified subspace embedding

Theorem ([CDD25])

Let $\mathbf{U}$ *be an* $n \times d$ *orthogonal matrix with rows* $\mathbf{u}_1, ..., \mathbf{u}_n$, *and let* $\mathbf{S}$ *be an* $m \times n$ *sketching matrix with* nnz_i *non-zeros in the* i*th column. If*

$$(LESS) \qquad m \geq C\frac{d}{\epsilon^2} \quad \textit{and} \quad \mathrm{nnz}_i \geq \max\left\{1,\ C\|\mathbf{u}_i\|^2\frac{\log^3(d)}{\epsilon}\right\}$$

then with high probability $\mathbf{S}$ *is an embedding for* $\mathrm{range}(\mathbf{U})$, *i.e.:*

$$(1-\epsilon)\|\mathbf{x}\| \leq \|\mathbf{SUx}\| \leq (1+\epsilon)\|\mathbf{x}\| \qquad \textit{for all } \mathbf{x}.$$

- Since $\|\mathbf{u}_i\|^2 \leq 1$, maximum per-column sparsity is near-optimal

[CDD25] Chenakkod, Dereziński, and Dong. *"Optimal oblivious subspace embeddings with near-optimal sparsity,"* ICALP 2025.

Leverage score sparsified subspace embedding

Theorem ([CDD25])

Let $\mathbf{U}$ *be an* $n \times d$ *orthogonal matrix with rows* $\mathbf{u}_1, \ldots, \mathbf{u}_n$*, and let* $\mathbf{S}$ *be an* $m \times n$ *sketching matrix with* nnz_i *non-zeros in the ith column. If*

$$(LESS) \qquad m \geq C\frac{d}{\epsilon^2} \quad \textit{and} \quad \mathrm{nnz}_i \geq \max\left\{1,\ C\|\mathbf{u}_i\|^2 \frac{\log^3(d)}{\epsilon}\right\}$$

then with high probability $\mathbf{S}$ *is an embedding for* $\mathrm{range}(\mathbf{U})$*, i.e.:*

$$(1-\epsilon)\|\mathbf{x}\| \leq \|\mathbf{SUx}\| \leq (1+\epsilon)\|\mathbf{x}\| \qquad \textit{for all } \mathbf{x}.$$

- Since $\|\mathbf{u}_i\|^2 \leq 1$, maximum per-column sparsity is near-optimal
- But $\sum_i \|\mathbf{u}_i\|^2 = d$, so the total number of non-zeros in $\mathbf{S}$ is only

$$\underbrace{n + O\Big(d\log^3(d)/\epsilon\Big)}_{\text{LESS}} \quad \ll \quad \underbrace{O\Big(n\log(d)/\epsilon\Big)}_{\text{Nelson-Nguyen conj.}}$$

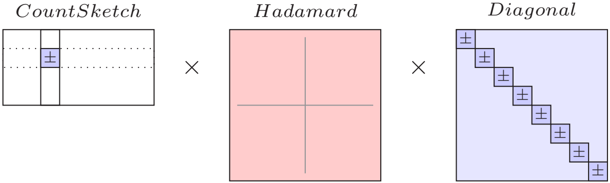
[CDD25] Chenakkod, Dereziński, and Dong. *"Optimal oblivious subspace embeddings with near-optimal sparsity,"* ICALP 2025.

Implication: CountSketch $\times$ RHT just works!

Corollary (CountSketch $\times$ RHT)

As long as the input is sufficiently tall, i.e., $n \geq O(d \log^3(d)/\epsilon)$, then "CountSketch $\times$ RHT" with $m = O(d/\epsilon^2)$ is an (ϵ, δ, d)-OSE.

Note: Improves on previously known condition $n \geq O(d^3)$ [CFS21]



Proof. RHT flattens the leverage scores so that $\|\mathbf{u}_i\|^2 = O(d/n)$, and

$$\text{nnz-per-col} \;=\; \max\left\{1,\; C\|\mathbf{u}_i\|^2 \frac{\log^3(d)}{\epsilon}\right\} \;=\; 1 \quad \text{for all } i.$$

[CFS21] Cartis, Fiala, and Shao. *"Hashing embeddings of optimal dimension, with applications to linear least squares"*, arXiv:2105.11815, 2021.

Outline

Conclusions

- Nelson-Nguyen Conjecture: There is an (ϵ, δ, d)-OSE with dimension $O(d/\epsilon^2)$ and $O(\log(d)/\epsilon)$ nnz per column. [NN, FOCS'13]
- Our main result: Resolving the conjecture up to sub-polylogarithmic factors [CDD, SODA'26]
- Other takeaways:
  - Even small $m = 1.1d$ sparse sketches can be subspace embeddings! *(if you define it right)* [CDDR, STOC'24]
  - CountSketch $\times$ RHT just works! [CDD, ICALP'25]

References I

Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson.
Toward a unified theory of sparse dimensionality reduction in euclidean space.
In Proceedings of the forty-seventh annual ACM symposium on Theory of Computing, pages 499–508, 2015.

Tatiana Brailovskaya and Ramon van Handel.
Universality and sharp matrix concentration inequalities.
Geometric and Functional Analysis, pages 1–105, 2024.

Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong.
Optimal oblivious subspace embeddings with near-optimal sparsity.
In 52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025), pages 55–1, 2025.

Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong.
Optimal subspace embeddings: Resolving nelson-nguyen conjecture up to sub-polylogarithmic factors.
ACM-SIAM Symposium on Discrete Algorithms, 2026.
arXiv:2508.14234.

Shabarish Chenakkod, Michał Dereziński, Xiaoyu Dong, and Mark Rudelson.
Optimal embedding dimension for sparse subspace embeddings.
In Proceedings of the 56th Annual ACM Symposium on Theory of Computing, pages 1106–1117, 2024.

Chris Camaño, Ethan N Epperly, Raphael A Meyer, and Joel A Tropp.
Faster linear algebra algorithms with structured random matrices.
arXiv preprint arXiv:2508.21189, 2025.

References II

Coralia Cartis, Jan Fiala, and Zhen Shao.
Hashing embeddings of optimal dimension, with applications to linear least squares.
arXiv preprint arXiv:2105.11815, 2021.

Michael B Cohen.
Nearly tight oblivious subspace embeddings by trace inequalities.
In Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms, pages 278–287. SIAM, 2016.

Kenneth L. Clarkson and David P. Woodruff.
Low-rank approximation and regression in input sparsity time.
J. ACM, 63(6):54:1–54:45, January 2017.

Michał Dereziński, Zhenyu Liao, Edgar Dobriban, and Michael W Mahoney.
Sparse sketches with small inversion bias.
In Proceedings of the 34th Conference on Learning Theory, 2021.

Riley Murray, James Demmel, Michael W Mahoney, N Benjamin Erichson, Maksim Melnichenko, Osman Asif Malik, Laura Grigori, Piotr Luszczek, Michał Dereziński, Miles E Lopes, Tianyu Liang, Hengrui Luo, and Jack Dongarra.
Randomized numerical linear algebra: A perspective on the field with an eye to software.
arXiv preprint arXiv:2302.11474, 2023.

Jelani Nelson and Huy L. Nguyên.
OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings.
In Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, FOCS '13, pages 117–126. IEEE Computer Society, 2013.

Outline

Approach: Trace moment universality

back to slide

Fact

$\mathbf{S}$ is an (ϵ, δ, d)-OSE iff for any $n \times d$ $\mathbf{U}$ with orthonormal columns,

$$\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\| \le \epsilon \quad \text{with prob. } 1 - \delta.$$

Suffices to show a trace moment condition:

$$\underbrace{\mathbb{E}\left[\operatorname{tr}\left((\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right)^{2q}\right]^{\frac{1}{2q}}}_{\|(\mathbf{SU})^\top(\mathbf{SU})-\mathbf{I}_d\|_{2q}} \le \epsilon \quad \text{for} \quad q = \log(d/\delta)$$

If $\mathbf{S}$ is Gaussian, then $\mathbf{SU} = \mathbf{G}$ is also Gaussian (rotation invariance). Conclusion follows from standard guarantees for Gaussian ensembles.

If $\mathbf{S}$ is *not* Gaussian, then we separate out the universality error:

$$\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\|_{2q} = \underbrace{\|\mathbf{G}^\top\mathbf{G} - \mathbf{I}_d\|_{2q}}_{\text{Gaussian}} + \underbrace{\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\|_{2q} - \|\mathbf{G}^\top\mathbf{G} - \mathbf{I}_d\|_{2q}}_{\text{Universality error}}$$

Approach: Trace moment universality

Theorem ([BvH24])

Let $\mathbf{M} = \sum_i \mathbf{Z}_i$ *be a sum of independent symmetric random matrices. Let* $\mathbf{G}$ *be Gaussian with the same mean and covariance as* $\mathbf{M}$*. Then:*

$$\Big| \|\mathbf{M}\|_{2q} - \|\mathbf{G}\|_{2q} \Big| \lesssim R_{2q}(\mathbf{M}) q^2 \quad \textit{where} \quad R_{2q}(\mathbf{M}) = \Big(\sum_i \|\mathbf{Z}_i\|_{2q}^{2q} \Big)^{\frac{1}{2q}}$$

[BvH24] Brailovskaya and van Handel. "Universality and sharp matrix concentration inequalities," *Geometric and Functional Analysis*, 2024.

Approach: Trace moment universality

Theorem ([BvH24])

Let $\mathbf{M} = \sum_i \mathbf{Z}_i$ *be a sum of independent symmetric random matrices. Let* $\mathbf{G}$ *be Gaussian with the same mean and covariance as* $\mathbf{M}$*. Then:*

$$\Big| \|\mathbf{M}\|_{2q} - \|\mathbf{G}\|_{2q} \Big| \lesssim R_{2q}(\mathbf{M}) q^2 \quad \textit{where} \quad R_{2q}(\mathbf{M}) = \Big(\sum_i \|\mathbf{Z}_i\|_{2q}^{2q} \Big)^{\frac{1}{2q}}$$

Bounding moments via Gaussian universality:

$$\|\mathbf{M}\|_{2q} \lesssim \underbrace{\|\mathbf{G}\|_{2q}}_{\text{Gaussian}} + \underbrace{R_{2q}(\mathbf{M}) q^2}_{\text{Universality error}}$$

In many cases (e.g., sparse subspace embeddings), universality error dominates Gaussian moments.

[BvH24] Brailovskaya and van Handel. "Universality and sharp matrix concentration inequalities," *Geometric and Functional Analysis*, 2024.

Key refinement: Iterative decoupling

$$\text{[BvH24]:} \quad \|\mathbf{M}\|_{2q} \lesssim \underbrace{\|\mathbf{G}\|_{2q}}_{\text{Gaussian}} + \underbrace{R_{2q}(\mathbf{M})q^2}_{\text{Universality error}}$$

Goal: Reduce the universality error bound when $R_{2q}(\mathbf{M})q^2 \gg \|\mathbf{G}\|_{2q}$

Key refinement: Iterative decoupling

$$[\text{BvH24}]: \quad \|\mathbf{M}\|_{2q} \lesssim \underbrace{\|\mathbf{G}\|_{2q}}_{\text{Gaussian}} + \underbrace{R_{2q}(\mathbf{M})q^2}_{\text{Universality error}}$$

Goal: Reduce the universality error bound when $R_{2q}(\mathbf{M})q^2 \gg \|\mathbf{G}\|_{2q}$

$$\|\mathbf{M}\|_{2q} = \left\| \left(\sum_i \mathbf{Z}_i \right)^2 \right\|_q^{1/2} \leq \underbrace{\left\| \sum_i \mathbf{Z}_i^2 \right\|_q^{1/2}}_{\text{diagonal}} + \underbrace{\left\| \sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2} \right\|_q^{1/2}}_{\text{off-diagonal}}$$

Key refinement: Iterative decoupling

$$\text{[BvH24]:} \quad \|\mathbf{M}\|_{2q} \lesssim \underbrace{\|\mathbf{G}\|_{2q}}_{\text{Gaussian}} + \underbrace{R_{2q}(\mathbf{M})q^2}_{\text{Universality error}}$$

Goal: Reduce the universality error bound when $R_{2q}(\mathbf{M})q^2 \gg \|\mathbf{G}\|_{2q}$

$$\|\mathbf{M}\|_{2q} = \left\| \Big(\sum_i \mathbf{Z}_i \Big)^2 \right\|_q^{1/2} \leq \Bigg\| \underbrace{\sum_i \mathbf{Z}_i^2}_{\text{diagonal}} \Bigg\|_q^{1/2} + \Bigg\| \underbrace{\sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2}}_{\text{off-diagonal}} \Bigg\|_q^{1/2}$$

Here, the off-diagonal term dominates, but we can use decoupling:

$$\Bigg\| \sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2} \Bigg\|_q \lesssim \underbrace{\| \mathbf{M} \cdot \mathbf{M}' \|_q}_{\ll \|\mathbf{M}^2\|_q}$$

where $\mathbf{M}'$ is an independent identically distributed copy of $\mathbf{M}$.

Key refinement: Iterative decoupling

$$\text{[BvH24]:}\quad \|\mathbf{M}\|_{2q} \lesssim \underbrace{\|\mathbf{G}\|_{2q}}_{\text{Gaussian}} + \underbrace{R_{2q}(\mathbf{M})q^2}_{\text{Universality error}}$$

<u>Goal</u>: Reduce the universality error bound when $R_{2q}(\mathbf{M})q^2 \gg \|\mathbf{G}\|_{2q}$

$$\|\mathbf{M}\|_{2q} = \left\| \Big(\sum_i \mathbf{Z}_i \Big)^2 \right\|_q^{1/2} \leq \Big\| \underbrace{\sum_i \mathbf{Z}_i^2}_{\text{diagonal}} \Big\|_q^{1/2} + \Big\| \underbrace{\sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2}}_{\text{off-diagonal}} \Big\|_q^{1/2}$$

Here, the off-diagonal term dominates, but we can use decoupling:

$$\Big\| \sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2} \Big\|_q \lesssim \underbrace{\| \mathbf{M} \cdot \mathbf{M}' \|_q}_{\ll \|\mathbf{M}^2\|_q} \underbrace{\lesssim^* \|\mathbf{G}\|_{2q} \cdot R_{2q}(\mathbf{M})q^2}_{^*\text{idealized bound}}$$

where $\mathbf{M}'$ is an independent identically distributed copy of $\mathbf{M}$.

Key refinement: Iterative decoupling

decoupling + [BvH24]: $\|\mathbf{M}\|_{2q} \lesssim^* \underbrace{\|\mathbf{G}\|_{2q}}_{\text{Gaussian}} + \underbrace{(\|\mathbf{G}\|_{2q} \cdot R_{2q}(\mathbf{M}) q^2)^{1/2}}_{\text{Universality error}}$

Goal: Reduce the universality error bound when $R_{2q}(\mathbf{M}) q^2 \gg \|\mathbf{G}\|_{2q}$

$$\|\mathbf{M}\|_{2q} = \left\| \Big(\sum_i \mathbf{Z}_i \Big)^2 \right\|_q^{1/2} \le \Bigg\| \underbrace{\sum_i \mathbf{Z}_i^2}_{\text{diagonal}} \Bigg\|_q^{1/2} + \Bigg\| \underbrace{\sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2}}_{\text{off-diagonal}} \Bigg\|_q^{1/2}$$

Here, the off-diagonal term dominates, but we can use decoupling:

$$\Bigg\| \sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2} \Bigg\|_q \lesssim \underbrace{\| \mathbf{M} \cdot \mathbf{M}' \|_q}_{\ll \|\mathbf{M}^2\|_q} \underbrace{\lesssim^* \|\mathbf{G}\|_{2q} \cdot R_{2q}(\mathbf{M}) q^2}_{^*\text{idealized bound}}$$

where $\mathbf{M}'$ is an independent identically distributed copy of $\mathbf{M}$.

Key refinement: Iterative decoupling

decoupling + [BvH24]: $\|\mathbf{M}\|_{2q} \lesssim^* \underbrace{\|\mathbf{G}\|_{2q}}_{\text{Gaussian}} + \underbrace{(\|\mathbf{G}\|_{2q} \cdot R_{2q}(\mathbf{M}) q^2)^{1/2}}_{\text{Universality error}}$

Goal: Reduce the universality error bound when $R_{2q}(\mathbf{M}) q^2 \gg \|\mathbf{G}\|_{2q}$

$$\|\mathbf{M}\|_{2q} = \Big\| \Big(\sum_i \mathbf{Z}_i \Big)^2 \Big\|_q^{1/2} \le \Big\| \underbrace{\sum_i \mathbf{Z}_i^2}_{\text{diagonal}} \Big\|_q^{1/2} + \Big\| \underbrace{\sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2}}_{\text{off-diagonal}} \Big\|_q^{1/2}$$

Here, the off-diagonal term dominates, but we can use decoupling:

$$\Big\| \sum_{i_1 \neq i_2} \mathbf{Z}_{i_1} \mathbf{Z}_{i_2} \Big\|_q \lesssim \underbrace{\| \mathbf{M} \cdot \mathbf{M}' \|_q}_{\ll \|\mathbf{M}^2\|_q} \underbrace{\lesssim^* \|\mathbf{G}\|_{2q} \cdot R_{2q}(\mathbf{M}) q^2}_{\textit{Why not decouple again?}}$$

where $\mathbf{M}'$ is an independent identically distributed copy of $\mathbf{M}$.

Iterative decoupling for subspace embeddings

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\big(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\big), \qquad \gamma_k \lesssim 1/k$$

Iterative decoupling for subspace embeddings

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\big\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\big\|_{2q} \lesssim C_k\big(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\big), \qquad \gamma_k \lesssim 1/k$$

$$\text{where} \qquad \mathcal{B} \lesssim \sqrt{\frac{d}{m}} + \frac{q}{\mathrm{nnz}} \qquad \text{(Benign term)}$$

Iterative decoupling for subspace embeddings

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\big(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\big), \qquad \gamma_k \lesssim 1/k$$

$$\text{where} \qquad \mathcal{B} \lesssim \sqrt{\frac{d}{m}} + \frac{q}{\mathrm{nnz}} \qquad \text{(Benign term)}$$

$$\text{and} \qquad \mathcal{U} \lesssim \frac{q^{2.5}}{\mathrm{nnz}} + \frac{q^2}{\sqrt{\mathrm{nnz}}}\sqrt{\frac{d}{m}} \qquad \text{(Universality error)}$$

Iterative decoupling for subspace embeddings

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\big\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\big\|_{2q} \lesssim C_k\big(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\big), \qquad \gamma_k \lesssim 1/k$$

where $\mathcal{B} \lesssim \sqrt{\frac{d}{m}} + \frac{q}{\mathrm{nnz}}$ (Benign term)

and $\mathcal{U} \lesssim \frac{q^{2.5}}{\mathrm{nnz}} + \frac{q^2}{\sqrt{\mathrm{nnz}}}\sqrt{\frac{d}{m}}$ (Universality error)

Reminder: We need $\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\|_{2q} \le \epsilon$ for $q \gtrsim \log d$

Iterative decoupling for subspace embeddings

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\left(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\right), \qquad \gamma_k \lesssim 1/k$$

$$\text{where} \quad \mathcal{B} \lesssim \sqrt{\frac{d}{m}} + \frac{q}{\mathrm{nnz}} \qquad \text{(Benign term)}$$

$$\text{and} \quad \mathcal{U} \lesssim \frac{q^{2.5}}{\mathrm{nnz}} + \frac{q^2}{\sqrt{\mathrm{nnz}}}\sqrt{\frac{d}{m}} \qquad \text{(Universality error)}$$

Reminder: We need $\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\|_{2q} \le \epsilon$ for $q \gtrsim \log d$

- The condition $\mathcal{B} \lesssim \epsilon$ precisely yields Nelson-Nguyen conjecture:

$$m \gtrsim \frac{d}{\epsilon^2} \qquad \text{and} \qquad \mathrm{nnz} \gtrsim \frac{\log d}{\epsilon}$$

Iterative decoupling for subspace embeddings

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\left(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\right), \qquad \gamma_k \lesssim 1/k$$

$$\text{where} \quad \mathcal{B} \lesssim \sqrt{\frac{d}{m}} + \frac{q}{\mathrm{nnz}} \qquad \text{(Benign term)}$$

$$\text{and} \quad \mathcal{U} \lesssim \frac{q^{2.5}}{\mathrm{nnz}} + \frac{q^2}{\sqrt{\mathrm{nnz}}}\sqrt{\frac{d}{m}} \qquad \text{(Universality error)}$$

Reminder: We need $\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\|_{2q} \le \epsilon$ for $q \gtrsim \log d$

- The condition $\mathcal{B} \lesssim \epsilon$ precisely yields Nelson-Nguyen conjecture:

$$m \gtrsim \frac{d}{\epsilon^2} \qquad \text{and} \qquad \mathrm{nnz} \gtrsim \frac{\log d}{\epsilon}$$

- But in order to bound the universality error, $\mathcal{U} \lesssim \epsilon$, we need:

$$\mathrm{nnz} \gtrsim \frac{(\log d)^{2.5}}{\epsilon} + (\log d)^4$$

Iterative decoupling for subspace embeddings

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\big(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\big), \qquad \gamma_k \lesssim 1/k$$

$$\text{where} \qquad \mathcal{B} \lesssim \sqrt{\frac{d}{m}} + \frac{q}{\mathrm{nnz}} \qquad \text{(Benign term)}$$

$$\text{and} \qquad \mathcal{U} \lesssim \frac{q^{2.5}}{\mathrm{nnz}} + \frac{q^2}{\sqrt{\mathrm{nnz}}}\sqrt{\frac{d}{m}} \qquad \text{(Universality error)}$$

<u>Reminder</u>: We need $\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\|_{2q} \le \epsilon$ for $q \gtrsim \log d$

- The condition $\mathcal{B} \lesssim \epsilon$ precisely yields Nelson-Nguyen conjecture:

$$m \gtrsim \frac{d}{\epsilon^2} \qquad \text{and} \qquad \mathrm{nnz} \gtrsim \frac{\log d}{\epsilon}$$

- But in order to bound the universality error, $\mathcal{U} \lesssim \epsilon$, we need:

$$\mathrm{nnz} \gtrsim \frac{(\log d)^{2.5}}{\epsilon} + (\log d)^4$$

<u>Question</u>: How does γ_k (iterative decoupling) help here?

Optimizing the sparsity-dimension trade-off

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\big\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\big\|_{2q} \lesssim C_k\big(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\big), \qquad \gamma_k \lesssim 1/k$$

Strategy: Make $\mathcal{B}$ smaller and $\mathcal{U}$ larger while keeping $\mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \epsilon$

Optimizing the sparsity-dimension trade-off

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\left(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\right), \qquad \gamma_k \lesssim 1/k$$

Strategy: Make $\mathcal{B}$ smaller and $\mathcal{U}$ larger while keeping $\mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \epsilon$

$$\text{Suppose} \qquad m \gtrsim (\log d)^{O(\gamma_k)}\frac{d}{\epsilon^2} \quad \text{and} \quad \text{nnz} \gtrsim (\log d)^{O(\gamma_k)}\frac{\log d}{\epsilon}$$

Optimizing the sparsity-dimension trade-off

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\left(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\right), \qquad \gamma_k \lesssim 1/k$$

Strategy: Make $\mathcal{B}$ smaller and $\mathcal{U}$ larger while keeping $\mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \epsilon$

$$\text{Suppose} \qquad m \gtrsim (\log d)^{O(\gamma_k)}\frac{d}{\epsilon^2} \quad \text{and} \quad \text{nnz} \gtrsim (\log d)^{O(\gamma_k)}\frac{\log d}{\epsilon}$$

$$\Rightarrow \quad \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \Big(\underbrace{\epsilon \cdot (\log d)^{-O(\gamma_k)}}_{\ll \epsilon}\Big)^{1-\gamma_k}\Big(\underbrace{\epsilon \cdot (\log d)^{O(1)}}_{\gg \epsilon}\Big)^{\gamma_k} \lesssim \epsilon.$$

Optimizing the sparsity-dimension trade-off

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\left(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\right), \qquad \gamma_k \lesssim 1/k$$

Strategy: Make $\mathcal{B}$ smaller and $\mathcal{U}$ larger while keeping $\mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \epsilon$

$$\text{Suppose} \quad m \gtrsim (\log d)^{O(\gamma_k)}\frac{d}{\epsilon^2} \quad \text{and} \quad \text{nnz} \gtrsim (\log d)^{O(\gamma_k)}\frac{\log d}{\epsilon}$$

$$\Rightarrow \quad \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \Big(\underbrace{\epsilon \cdot (\log d)^{-O(\gamma_k)}}_{\ll \epsilon}\Big)^{1-\gamma_k}\Big(\underbrace{\epsilon \cdot (\log d)^{O(1)}}_{\gg \epsilon}\Big)^{\gamma_k} \lesssim \epsilon.$$

Question: Why can't we just take $k \to \infty$ and $\gamma_k = O(1/k) \to 0$?

Optimizing the sparsity-dimension trade-off

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\big(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\big), \qquad \gamma_k \lesssim 1/k$$

Strategy: Make $\mathcal{B}$ smaller and $\mathcal{U}$ larger while keeping $\mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \epsilon$

$$\text{Suppose} \qquad m \gtrsim (\log d)^{O(\gamma_k)}\frac{d}{\epsilon^2} \quad \text{and} \quad \text{nnz} \gtrsim (\log d)^{O(\gamma_k)}\frac{\log d}{\epsilon}$$

$$\Rightarrow \quad \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \Big(\underbrace{\epsilon \cdot (\log d)^{-O(\gamma_k)}}_{\ll \epsilon}\Big)^{1-\gamma_k}\Big(\underbrace{\epsilon \cdot (\log d)^{O(1)}}_{\gg \epsilon}\Big)^{\gamma_k} \lesssim \epsilon.$$

Question: Why can't we just take $k \to \infty$ and $\gamma_k = O(1/k) \to 0$?

$$\text{Decoupling constant:} \qquad C_k = \exp\big(\exp\big(O(k)\big)\big)$$

Optimizing the sparsity-dimension trade-off

After k steps of *iterative decoupling* for $\mathbf{S}$ with nnz nonzeros per col,

$$\left\|(\mathbf{SU})^\top(\mathbf{SU}) - \mathbf{I}_d\right\|_{2q} \lesssim C_k\left(\mathcal{B} + \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k}\right), \qquad \gamma_k \lesssim 1/k$$

Strategy: Make $\mathcal{B}$ smaller and $\mathcal{U}$ larger while keeping $\mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \epsilon$

$$\text{Suppose} \qquad m \gtrsim (\log d)^{O(\gamma_k)}\frac{d}{\epsilon^2} \quad \text{and} \quad \text{nnz} \gtrsim (\log d)^{O(\gamma_k)}\frac{\log d}{\epsilon}$$

$$\Rightarrow \quad \mathcal{B}^{1-\gamma_k}\mathcal{U}^{\gamma_k} \lesssim \Big(\underbrace{\epsilon \cdot (\log d)^{-O(\gamma_k)}}_{\ll \epsilon}\Big)^{1-\gamma_k}\Big(\underbrace{\epsilon \cdot (\log d)^{O(1)}}_{\gg \epsilon}\Big)^{\gamma_k} \lesssim \epsilon.$$

Question: Why can't we just take $k \to \infty$ and $\gamma_k = O(1/k) \to 0$?

$$\text{Decoupling constant:} \qquad C_k = \exp\big(\exp\big(O(k)\big)\big)$$

To ensure that $C_k \lesssim (\log d)^{O(1/k)}$, we let $k = O(\log\log\log d)$, recovering the main result we claimed.

Final random matrix theory result

back to slide

Theorem

The $m \times n$ *stacked CountSketch matrix* $\mathbf{S}$ *with*

$$m \geq (\log d)^{O(\frac{1}{\log\log\log d})} \frac{d}{\epsilon^2} \quad \textit{and} \quad \mathrm{nnz} \geq (\log d)^{O(\frac{1}{\log\log\log d})} \frac{\log d}{\epsilon},$$

for any $n \times d$ *orthogonal matrix* $\mathbf{U}$ *with high probability satisfies:*

$$1 - \epsilon \leq \sigma_{\min}(\mathbf{SU}) \leq \sigma_{\max}(\mathbf{SU}) \leq 1 + \epsilon.$$

Note: The result holds independently of the choice of $\mathbf{U}$, so the matrix $\mathbf{S}$ is an *oblivious* subspace embedding.

Recall: Nelson-Nguyen conjecture proposes the following condition:

$$m \geq C \frac{d}{\epsilon^2} \quad \text{and} \quad \mathrm{nnz} \geq C \frac{\log d}{\epsilon}$$