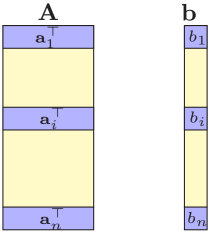
# Distributed Least squares in Small Space via sketching and Bias Reduction

Sachin Garg, Kevin Tan, Michał Dereziński

# Problem formulation

Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^{n}$, and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$.



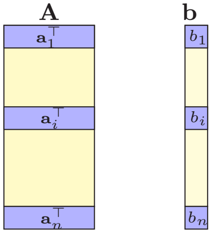

Aim: Find an "$\epsilon$-accurate" estimate to $\mathbf{x}^*$, but with space constraints. If

$$\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2,$$

then we say $\tilde{\mathbf{x}}$ is an $\epsilon$-accurate estimate to $\mathbf{x}^*$.

# Problem formulation

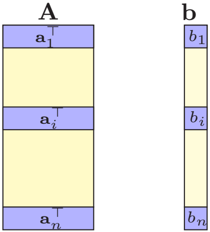
Let $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$.




Direct methods store $\mathbf{A}$, requiring $\tilde{O}(nd)$ space, rendering them unsuitable for memory-constrained computing systems when $n \gg d$.

# Problem formulation

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$.




Existing matrix sketching-based methods provide an $\epsilon$-accurate $\tilde{\mathbf{x}}$ but at least require $\tilde{O}(d^2/\epsilon)$ space [5].

# Our contribution

We propose an algorithm which provides an $\epsilon$-accurate $\tilde{\mathbf{x}}$ in $\tilde{O}(d^2)$ space in distributed computing environments.

- In distributed setup, our work provides an $\epsilon$-accurate $\tilde{\mathbf{x}}$ within 2 parallel data passes.
- Our work is based on debiasing techniques to recover *nearly unbiased* estimators of $\mathbf{x}^*$ using Leverage Score Sparsified (LESS) embeddings [6].
- Our theoretical analysis relies on proving higher moment-restricted *Bai-Silverstein inequalities*, which could be of independent interest to Random Matrix Theory (RMT) community [3].

# Matrix sketching for least squares

Let sketching matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, with $m \ll n$, and consider $\tilde{\mathbf{x}}$ as:

$$\tilde{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}} \|\mathbf{SAx} - \mathbf{Sb}\|^2.$$

Storing $\mathbf{SA}$ requires $\tilde{O}(md)$ space, potentially much lesser than $\tilde{O}(nd)$. Choices for $\mathbf{S}$ could be anything from

- Subgaussian matrices [1].
- Randomized Hadamard transforms [2].
- Sparse Matrices, e.g. Count Sketch [4].
- Subsampling e.g. approximate Leverage score subsampling [7],

and many others.

# Leverage score subsampling for least squares

The $i^{th}$ leverage score of $\mathbf{A}$ denoted by $\ell_i(\mathbf{A})$ is defined as $\ell_i(\mathbf{A}) = \mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{a}_i$.

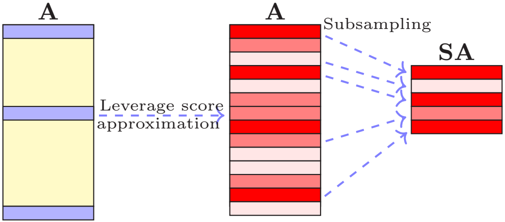



Figure: Visual illustration of Leverage score subsampling

Let $\tilde{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{SAx} - \mathbf{Sb}\|^2$. Then for $m = \tilde{O}(d/\epsilon)$:

$$\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2 \leq (1+\epsilon)\|\mathbf{Ax}^* - \mathbf{b}\|^2.$$

# Leverage score subsampling for least squares

The $i^{th}$ leverage score of $\mathbf{A}$ denoted by $\ell_i(\mathbf{A})$ is defined as $\ell_i(\mathbf{A}) = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{a}_i$.

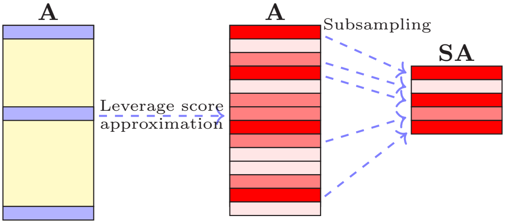



Figure: Visual illustration of Leverage Score subsampling

For $m = \tilde{O}(d/\epsilon)$ : $\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2$.

Space requirement: $\tilde{O}(d^2/\epsilon)$. For small $\epsilon$, this space requirement can be restrictive.

# Our approach

- Construct smaller sketches with a much smaller bias than the subsampling sketch.
- Leverage distributed averaging to recover an $\epsilon$-accurate estimate to $\mathbf{x}^*$.
- Turns out that in this distributed setup we can reduce the sketch size $m$ and recover a *nearly unbiased* $\tilde{\mathbf{x}}$.

Unfortunately, subsampled sketches still require $m = \tilde{O}(d/\sqrt{\epsilon})$.

We provide an algorithm that requires only $\tilde{O}(d^2)$ space.

# Find $\epsilon$-accurate $\tilde{\mathbf{x}}$ in distributed settings

Reducing $m$ and averaging multiple estimators for $\mathbf{x}^*$ leads to lesser space requirement.

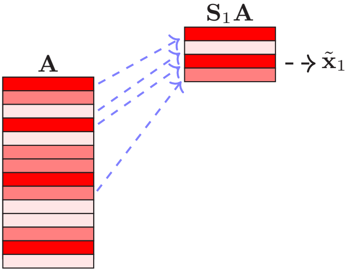



Figure: Averaging for Least squares via Leverage score subsampling

# Find $\epsilon$-accurate $\tilde{\mathbf{x}}$ in distributed settings

Reducing $m$ and averaging multiple estimators for $\mathbf{x}^*$ leads to lesser space requirement.

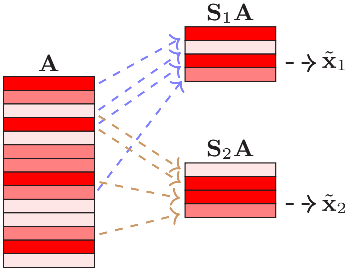



Figure: Averaging for Least squares via Leverage score subsampling

# Find $\epsilon$-accurate $\tilde{\mathbf{x}}$ in distributed settings

Reducing $m$ and averaging multiple estimators for $\mathbf{x}^*$ leads to lesser space requirement.

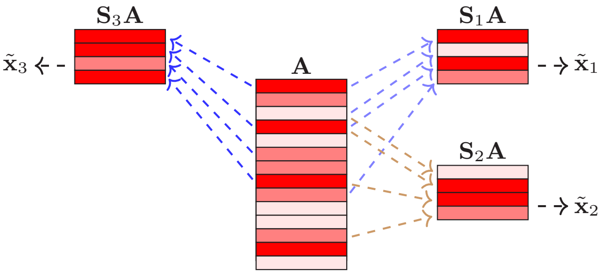



Figure: Averaging for Least squares via Leverage score subsampling

# Find $\epsilon$-accurate $\tilde{\mathbf{x}}$ in distributed settings

Reducing $m$ and averaging multiple estimators for $\mathbf{x}^*$ leads to lesser space requirement.

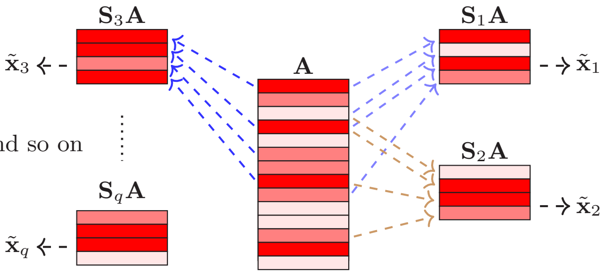
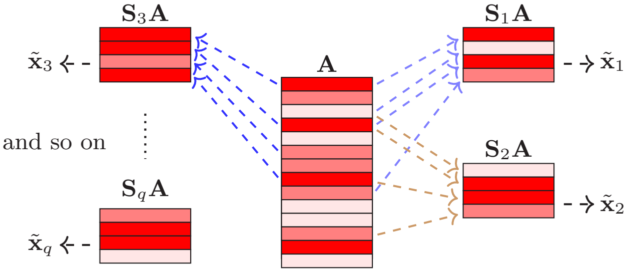



Figure: Averaging for Least squares via Leverage score subsampling

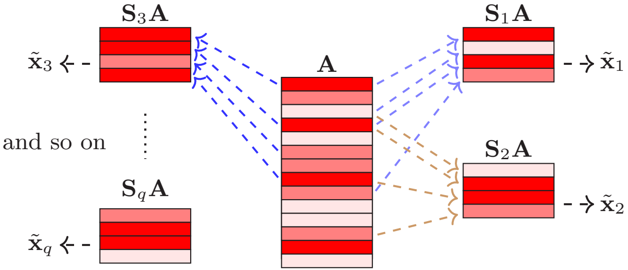
# Find $\epsilon$-accurate $\tilde{\mathbf{x}}$ in distributed settings




Let $\tilde{\mathbf{x}} = \frac{1}{q}\sum_{i=1}^{q} \tilde{\mathbf{x}}_i$ and let $q \to \infty$. Then,

$$\underbrace{\|\mathbf{A}\mathbb{E}[\tilde{\mathbf{x}}] - \mathbf{b}\|^2}_{\text{Bias}} \le (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2 \ll \underbrace{\mathbb{E}\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2}_{\text{Variance}}.$$

# Find $\epsilon$-accurate $\tilde{\mathbf{x}}$ in distributed settings




Let $\tilde{\mathbf{x}} = \frac{1}{q}\sum_{i=1}^{q} \tilde{\mathbf{x}}_i$ and let $q \to \infty$. Then,

$$\underbrace{\|\mathbf{A}\mathbb{E}[\tilde{\mathbf{x}}] - \mathbf{b}\|^2}_{\text{Bias}} \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2 \ll \underbrace{\mathbb{E}\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2}_{\text{Variance}}.$$

Our contribution: We propose an algorithm to recover an $\epsilon$-accurate $\tilde{\mathbf{x}}$ in $\tilde{O}(d^2)$ space in distributed settings.

# Our algorithm: Least squares using LESS embeddings

We construct $\mathbf{S} \in \mathbb{R}^{\tilde{O}(d)\times d}$ and every row of $\mathbf{S}$ is now formed by mixing (compressing) $\tilde{O}(1/\epsilon)$ rows from $\mathbf{A}$.

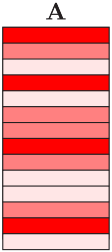



Figure: Sketching in small space via LESS embeddings

LESS: LEverage Score Sparsified.

# Our algorithm: Least squares using LESS embeddings

We construct $\mathbf{S} \in \mathbb{R}^{\tilde{O}(d) \times d}$ and every row of $\mathbf{S}$ is now formed by mixing (compressing) $\tilde{O}(1/\epsilon)$ rows from $\mathbf{A}$.

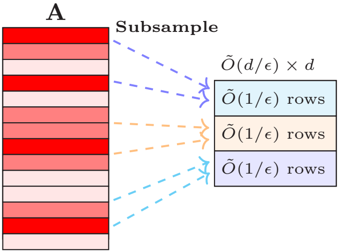



Figure: Sketching in small space via LESS embeddings

LESS: LEverage Score Sparsified.

# Our algorithm: Least squares using LESS embeddings

We construct $\mathbf{S} \in \mathbb{R}^{\tilde{O}(d) \times d}$ and every row of $\mathbf{S}$ is now formed by mixing (compressing) $\tilde{O}(1/\epsilon)$ rows from $\mathbf{A}$.

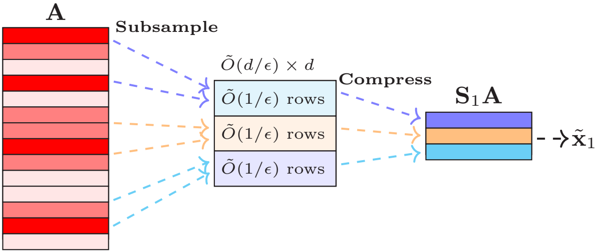



Figure: Sketching in small space via LESS embeddings

LESS: LEverage Score Sparsified.

# Our results

Let $\tilde{\mathbf{x}} = \frac{1}{q}\sum_{i=1}^{q} \tilde{\mathbf{x}}_i$. Then for $q = \frac{1}{\epsilon}$ we have,

$$\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2.$$

Importantly, $\mathbf{S}_i\mathbf{A}$ requires $\tilde{O}(d^2)$ space.

- We show that in streaming settings and distributed environments, an $\epsilon$-accurate estimate to $\mathbf{x}^*$ can be obtained in 2 data passes.
- We extend our results to distributed settings where data is uniformly partitioned across $q$ machines.

# Detailed technical result

**Theorem (Main result (informal) from the paper)**

*Given streaming access to* $\mathbf{A} \in \mathbb{R}^{n \times d}$ *and* $\mathbf{b} \in \mathbb{R}^n$, *within 2 passes over* $(\mathbf{A}, \mathbf{b})$, *in* $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + \epsilon^{-1} d^2)$ *time and* $\tilde{O}(d^2)$ *bits of space, we can construct a randomized estimator* $\tilde{\mathbf{x}}$ *for the least squares solution* $\mathbf{x}^*$ *such that:*

$$\begin{aligned} (\text{Bias}) \quad & \big\|\mathbf{A}\mathbb{E}[\tilde{\mathbf{x}}] - \mathbf{b}\big\|^2 \le (1+\epsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2, \\ (\text{Variance}) \quad & \mathbb{E}\big[\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2\big] \le 2\,\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|^2. \end{aligned}$$

Thank you!

# References I


[1] Dimitris Achlioptas. "Database-friendly random projections: Johnson-Lindenstrauss with binary coins". In: Journal of computer and System Sciences 66.4 (2003), pp. 671–687.

[2] Nir Ailon and Bernard Chazelle. "The fast Johnson–Lindenstrauss transform and approximate nearest neighbors". In: SIAM Journal on computing 39.1 (2009), pp. 302–322.

[3] Zhidong Bai and Jack W Silverstein. Spectral analysis of large dimensional random matrices. Vol. 20. Springer, 2010.

# References II


[4] Kenneth L Clarkson and David P Woodruff. "Low-rank approximation and regression in input sparsity time". In: Journal of the ACM (JACM) 63.6 (2017), pp. 1–45.

[5] Kenneth L Clarkson and David P Woodruff. "Numerical linear algebra in the streaming model". In: Proceedings of the forty-first annual ACM symposium on Theory 2009, pp. 205–214.

[6] Michal Derezinski et al. "Sparse sketches with small inversion bias". In: Conference on Learning Theory. PMLR. 2021, pp. 1467–1510.

# References III

[7] Shusen Wang, Alex Gittens, and Michael W Mahoney. "Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging". In: Journal of Machine Learning Research 18.218 (2018), pp. 1–50.