Resources and Schedule

MathSci.ai

- **Data for Labs:** tgkolda/randalgslabs (github.com)
- **Schedule**
  - 11:00 – Part I
  - 11:35 – Labs for Part I
  - 11:50 – Part II
    - Randomized Matrix Multiplication
    - Randomized Least Squares
  - 12:25 – Labs for Part II
  - 12:40 – Adjourn

MathSci.ai

Part 2

Randomized Matrix Multiplication & Randomized Least Squares

Overview

MathSci.ai

- **Randomized Matrix Multiplication Goal:** Approximate the product of two matrices.
- **Randomized Least Squares Goal:** Approximate the solution to an overdetermined least squares problem.
- In both cases, we'll be considering row sampling in which we randomly sample and reweight rows of the matrices.
  - *We'll ask what measures of importance should we use for sampling to ensure key rows are more likely to be included?*
  - *What kind of guarantees can we get out of these methods?*
- We'll focus on results that can be proven with minimal tools from statistics and probability.

MathSci.ai

Randomized Matrix Multiplication

- Drineas, Kannan, and Mahoney (2006). **Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication**. SIAM Journal on Computing https://doi.org/10.1137/S0097539704442684

Randomized Matrix Multiplication

MathSci.ai

Given:

$$\mathbf{A} \in \mathbb{R}^{n \times m} \text{ and } \mathbf{B} \in \mathbb{R}^{n \times p}$$

We want to approximate:

$$\mathbf{A}^\intercal \mathbf{B} = \sum_{t=1}^{n} \mathbf{A}(t,:)^\intercal \mathbf{B}(t,:), \quad \text{(Sum of outer products)}$$

where $\mathbf{A}(t,:), \mathbf{B}(t,:)$ denotes row t of the matrix

Approach: Approximate this sum of rank-1 terms by sampling $s < n$ terms with replacement and reweighting appropriately.

Row Sampling Matrix

MathSci.ai

Definition (Distribution) We say a vector $\mathbf{p} \in [0,1]^n$ is a (discrete) probability distribution if $\sum_{i=1}^{n} p_i = 1$.

Definition (Multinomial) A multinomial discrete random variable with respect to a distribution $\mathbf{p}$ which we denote $\xi \in [n] \sim \text{MULTINOMIAL}(\mathbf{p})$ is defined such that $\mathbb{P}\{\xi = i\} = p_i$ for all $i \in [n]$.

Row Sampling Matrix

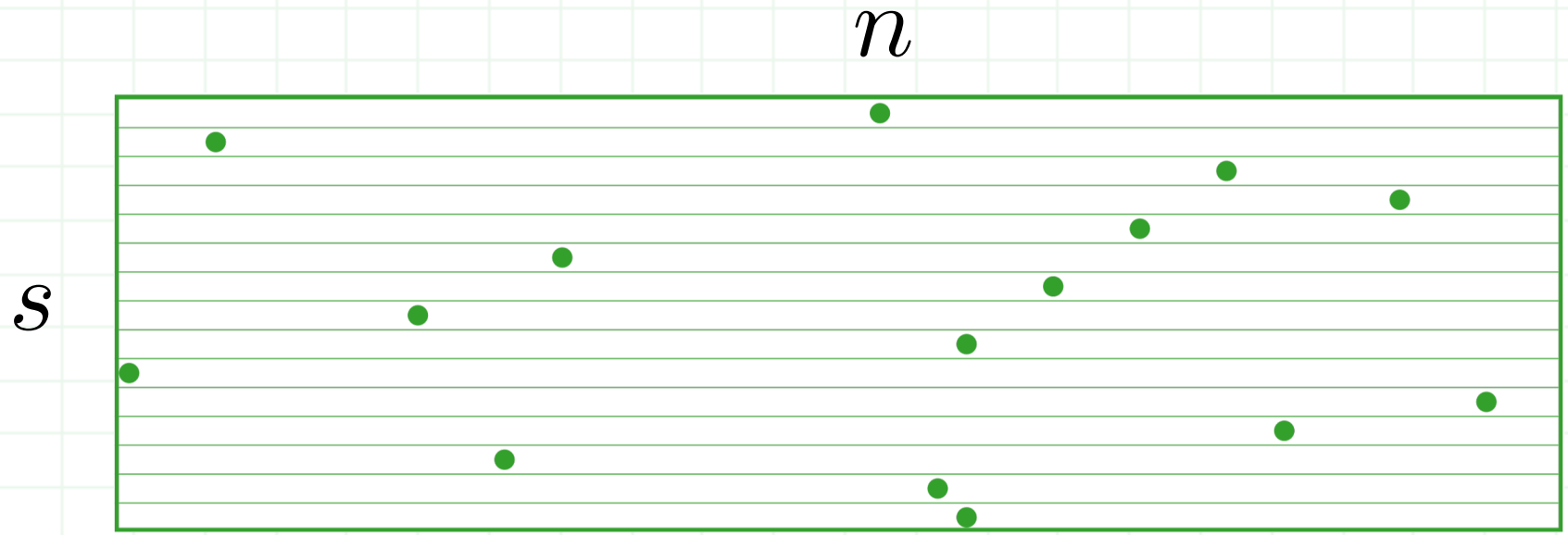
MathSci.ai

Definition (RandSample Matrix) We say $\boldsymbol{S} \in \mathbb{R}^{s \times n} \sim \textsc{randsample}(s, \mathbf{p})$ if $\boldsymbol{S}$ is defined as follows. For each $k \in [s]$, let the random variable $\xi_k = \textsc{multinomial}(\mathbf{p})$, and then

$$\boldsymbol{S}(k,i) = \begin{cases} \frac{1}{\sqrt{sp_k}} & \text{if } i = \xi_k \\ 0 & \text{otherwise} \end{cases} \quad \text{for all} \quad (k,i) \in [s] \otimes [n].$$

The matrix $\boldsymbol{S}$ has one nonzero per row, which means that the kth row of $\boldsymbol{SA}$ is row ξ_k of $\mathbf{A}$, reweighted by $\frac{1}{\sqrt{sp_{\xi_k}}}$.

Row Sampling Matrix



$$\boldsymbol{S} \in \mathbb{R}^{s \times n}$$

The matrix $\boldsymbol{S}$ has one nonzero per row, which means that the kth row of $\boldsymbol{SA}$ is row ξ_k of $\mathbf{A}$, weighted by $\frac{1}{\sqrt{sp_{\xi_k}}}$.

Randomized Matrix Multiplication

MathSci.ai

Definition (Approximate Product) Choose s rows, denoted $\{\xi^{(1)}, \ldots, \xi^{(s)}\}$, according to the probability distribution defined by $\mathbf{p} \in [0,1]^n$ Then form the approximate product

$$\frac{1}{s}\sum_{t=1}^{s}\frac{1}{p_{\xi^{(t)}}}\mathbf{A}(\xi^{(t)},:)^\intercal\mathbf{B}(\xi^{(t)},:) \triangleq (\boldsymbol{S}\mathbf{A})^\intercal\boldsymbol{S}\mathbf{B},$$

We want a guarantee on the quality of this approximation measured by:

$$\|\mathbf{A}^\intercal\mathbf{B} - (\boldsymbol{S}\mathbf{A})^\intercal\boldsymbol{S}\mathbf{B}\|_F^2$$

Randomized Matrix Multiplication

MathSci.ai

Definition (Approximate Product) Choose s rows, denoted $\{\xi^{(1)}, \ldots, \xi^{(s)}\}$, according to the probability distribution defined by $\mathbf{p} \in [0,1]^n$ Then form the approximate product

$$\frac{1}{s}\sum_{t=1}^{s}\frac{1}{p_{\xi^{(t)}}}\mathbf{A}(\xi^{(t)},:)^\intercal\mathbf{B}(\xi^{(t)},:) \triangleq (\mathbf{SA})^\intercal\mathbf{SB},$$

This approximation is for an arbitrary distribution $\mathbf{p}$. We want:

1. $\mathbf{p}$ to be easy to compute
2. $\mathbf{p}$ to bias towards important samples for better approximation

Importance Sampling

Recall that we are approximating a sum of rank-1 terms:

$$\mathbf{A}^\intercal\mathbf{B} = \sum_{t=1}^{n} \mathbf{A}(t,:)^\intercal\mathbf{B}(t,:)$$

Good measure of importance? The "size" of the term:

$$\|\mathbf{A}(t,:)^\intercal\mathbf{B}(t,:)\|_2$$

Because these are rank-1, the spectral norm has a simple form:

$$\|\mathbf{A}(t,:)^\intercal\mathbf{B}(t,:)\|_2 = \|\mathbf{A}(t,:)\|_2 \, \|\mathbf{B}(t,:)\|_2 \qquad (Check!)$$

So sampling proportional to the "size" of each term is equivalent to sampling proportional to the product of the row norms.

Importance Sampling

MathSci.ai

So normalizing our probabilities would look like:

$$p_k = \frac{\|\mathbf{A}(k,:)\|_2 \, \|\mathbf{B}(k,:)\|_2}{\sum_{t=1}^{n} \|\mathbf{A}(t,:)\|_2 \, \|\mathbf{B}(t,:)\|_2}$$

Because of how we will use RMM for least squares later, we're going to focus on the case where we only have access to $\mathbf{A}$:

$$p_k = \frac{\|\mathbf{A}(k,:)\|_2^2}{\sum_{t=1}^{n} \|\mathbf{A}(t,:)\|_2^2} = \frac{\|\mathbf{A}(k,:)\|_2^2}{\|\mathbf{A}\|_F^2}$$

Importance Sampling

MathSci.ai

So normalizing our probabilities would look like:

$$p_k = \frac{\|\mathbf{A}(k,:)\|_2 \, \|\mathbf{B}(k,:)\|_2}{\sum_{t=1}^{n} \|\mathbf{A}(t,:)\|_2 \, \|\mathbf{B}(t,:)\|_2}$$

Further relax this to *approximations* of the row norm squared:

$$p_k \geq \frac{\beta \|\mathbf{A}(k,:)\|_2^2}{\|\mathbf{A}\|_F^2}$$

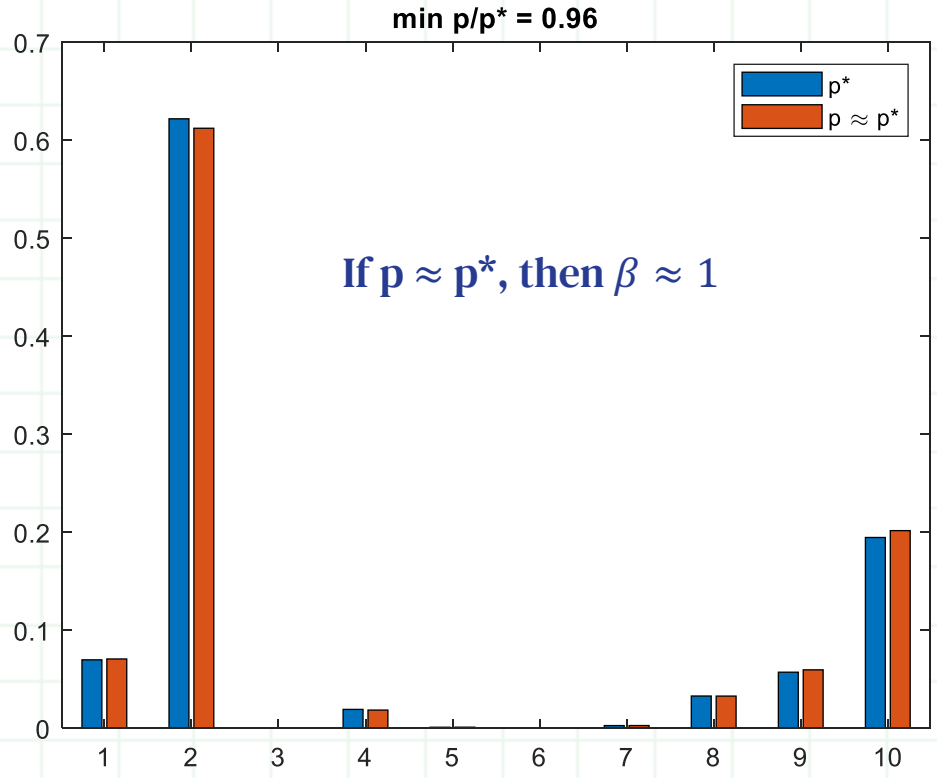
Recall that we can choose any probability distribution, but this will affect the accuracy of our approximation.

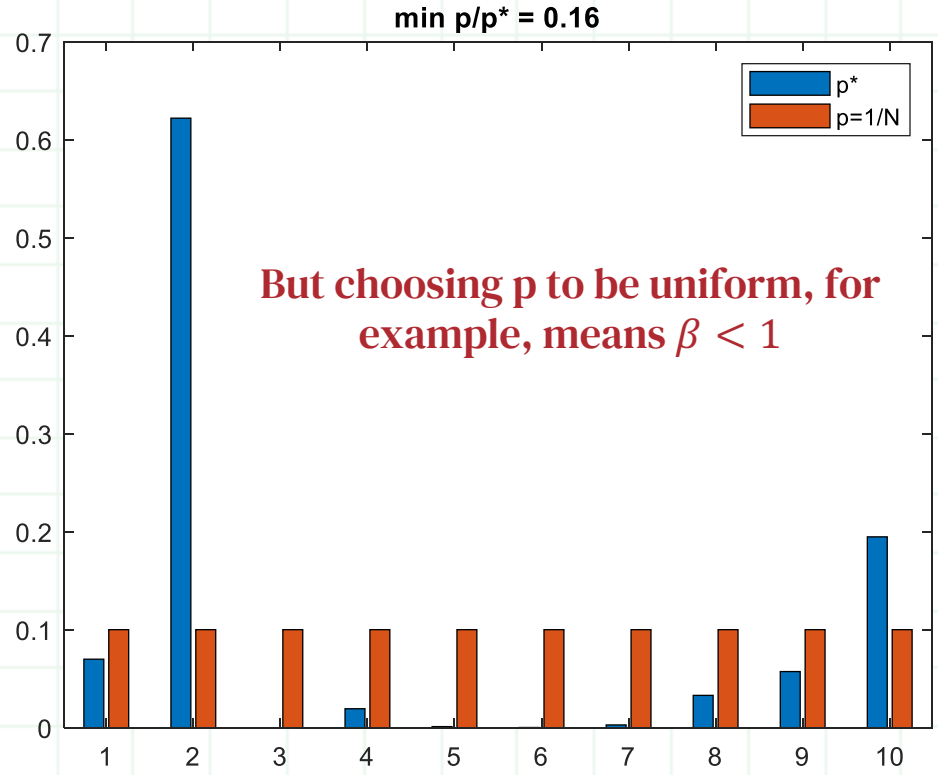
Misestimation Factor

Misestimation factor:

$$\beta \leq \min_k \frac{p_k \|\mathbf{A}\|_F^2}{\|\mathbf{A}(k,:)\|_2^2} \equiv \min_k \frac{p_k}{p_k^*} \in (0,1]$$

Can use any procedure to compute p_k but want β as close to one as possible.





Theorem – Randomized Matrix Multiplication

MathSci.ai

Theorem (Drineas, Kannan, and Mahoney, 2006) Choose s rows, denoted $\{\xi^{(1)}, \ldots, \xi^{(s)}\}$, according to the probability distribution $\mathbf{p} \in [0,1]^n$ with the property that there exists $\beta > 0$ such that

$$p_k \geq \beta \|\mathbf{A}(k,:)\|^2 / \|\mathbf{A}\|_F^2 \quad \text{for all} \quad k \in [n].$$

We then have the following guarantee on the quality of the approximate product:

$$\mathbb{E}\left[\|\mathbf{A}^\intercal\mathbf{B} - (\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\right] \leq \frac{1}{\beta s}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2.$$

Proof

Fix i, j to specify an element of the matrix product and let $\{\xi^{(1)}, \ldots, \xi^{(s)}\}$ be the indices of the sampled rows of $\mathbf{A}$ (and $\mathbf{B}$). We can examine:

$$[(\mathbf{SA})^\intercal \mathbf{SB}]_{ij}$$

} Scalar Random Variable

Proof

Fix i, j to specify an element of the matrix product and let $\{\xi^{(1)}, \ldots, \xi^{(s)}\}$ be the indices of the sampled rows of $\mathbf{A}$ (and $\mathbf{B}$). We can examine:

$$[(\mathbf{SA})^\intercal \mathbf{SB}]_{ij}$$

} Scalar Random Variable

This can be written as a sum of scalar random variables X_t for $t = 1, \ldots, s$ as follows:

$$X_t = \frac{\mathbf{A}(\xi^{(t)}, i)\mathbf{B}(\xi^{(t)}, j)}{sp_{\xi^{(t)}}} \quad \Longrightarrow \quad [(\mathbf{SA})^\intercal \mathbf{SB}]_{ij} = \sum_{t=1}^{s} X_t$$

Proof

Since X_t is $\mathbf{A}(k,i)\mathbf{B}(k,j)/\left(sp_k\right)$ with probability p_k:

$$\mathbb{E}[X_t] = \sum_{k=1}^{n} p_k \frac{\mathbf{A}_{ki}\mathbf{B}_{kj}}{sp_k} = \frac{1}{s}(\mathbf{A}^\intercal\mathbf{B})_{ij}$$

Proof

Since X_t is $\mathbf{A}(k,i)\mathbf{B}(k,j)/\left(sp_k\right)$ with probability p_k:

$$\mathbb{E}[X_t] = \sum_{k=1}^{n} p_k \frac{\mathbf{A}_{ki}\mathbf{B}_{kj}}{sp_k} = \frac{1}{s}(\mathbf{A}^\intercal\mathbf{B})_{ij}$$

Compute the expectation of an element of the sampled matrix product:

$$\mathbb{E}\left[(\mathbf{SA})^\intercal\mathbf{SB}\right]_{ij} = \mathbb{E}\left[\sum_{t=1}^{s} X_t\right] = \sum_{t=1}^{s}\mathbb{E}[X_t]$$

$$= \sum_{t=1}^{s}\frac{1}{s}(\mathbf{A}^\intercal\mathbf{B})_{ij} = (\mathbf{A}^\intercal\mathbf{B})_{ij}$$

Unbiased

Proof

(*Exercise*) Similarly, the variance can be computed using the fact that the $X_1, \ldots, X_s$ are indpendent.

$$\operatorname{Var}\left[(\mathbf{SA})^{\intercal}\mathbf{SB}\right]_{ij}] = \sum_{k=1}^{n} \frac{\mathbf{A}_{ki}^2 \mathbf{B}_{kj}^2}{sp_k} - \frac{1}{s}(\mathbf{A}^{\intercal}\mathbf{B})_{ij}$$

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^\intercal\mathbf{SB}\|_F^2\right]=\sum_{i=1}^{m}\sum_{j=1}^{p}\mathbb{E}\left[\left([(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}-(\mathbf{A}^\intercal\mathbf{B})_{ij}\right)^2\right],$$

Linearity of Expectation

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^\intercal\mathbf{SB}\|_F^2\right]=\sum_{i=1}^{m}\sum_{j=1}^{p}\mathbb{E}\left[\left([(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}-(\mathbf{A}^\intercal\mathbf{B})_{ij}\right)^2\right],$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{p}\mathbb{E}\left[\left([(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}-\mathbb{E}\left[[(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}\right]\right)^2\right],$$

Estimator is unbiased, i.e. $\mathbb{E}\left[[(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}\right]=(\mathbf{A}^\intercal\mathbf{B})_{ij}$

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^\intercal\mathbf{SB}\|_F^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{p}\mathbb{E}\left[\left([(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}-(\mathbf{A}^\intercal\mathbf{B})_{ij}\right)^2\right],$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{p}\mathbb{E}\left[\left([(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}-\mathbb{E}\left[[(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}\right]\right)^2\right],$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{p}\mathrm{Var}\left[[(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}\right],$$

Definition of variance, $\mathrm{Var}[X] = \mathbb{E}[(X-\mathbb{E}X)^2]$

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^\intercal\mathbf{SB}\|_F^2\right]=\sum_{i=1}^{m}\sum_{j=1}^{p}\mathbb{E}\left[\left([(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}-(\mathbf{A}^\intercal\mathbf{B})_{ij}\right)^2\right],$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{p}\mathbb{E}\left[\left([(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}-\mathbb{E}\left[[(\mathbf{SA})^\intercal\mathbf{SB}]_{ij}\right]\right)^2\right],$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{p}\mathrm{Var}\left[\left[(\mathbf{SA})^\intercal\mathbf{SB}\right]_{ij}\right],$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{p}\left(\sum_{k=1}^{n}\frac{\mathbf{A}_{ki}^2\mathbf{B}_{kj}^2}{sp_k}-\frac{1}{s}(\mathbf{A}^\intercal\mathbf{B})_{ij}\right).$$

Plug in expression for varaince

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^{\intercal}\mathbf{SB}\|_F^2\right]=\sum_{i=1}^{m}\sum_{j=1}^{p}\left(\sum_{k=1}^{n}\frac{\mathbf{A}_{ki}^2\mathbf{B}_{kj}^2}{sp_k}-\frac{1}{s}(\mathbf{A}^{\intercal}\mathbf{B})_{ij}\right),$$

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^\intercal\mathbf{SB}\|_F^2\right]=\sum_{i=1}^{m}\sum_{j=1}^{p}\left(\sum_{k=1}^{n}\frac{\mathbf{A}_{ki}^2\mathbf{B}_{kj}^2}{sp_k}-\frac{1}{s}(\mathbf{A}^\intercal\mathbf{B})_{ij}\right),$$

$$=\sum_{k=1}^{n}\frac{\left(\sum_{i=1}^{m}\mathbf{A}_{ki}^2\right)\left(\sum_{j=1}^{p}\mathbf{B}_{kj}^2\right)}{sp_k}-\frac{1}{s}\sum_{i=1}^{m}\sum_{j=1}^{p}(\mathbf{A}^\intercal\mathbf{B})_{ij},$$

Rearranging sums

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^\intercal\mathbf{SB}\|_F^2\right]=\sum_{i=1}^{m}\sum_{j=1}^{p}\left(\sum_{k=1}^{n}\frac{\mathbf{A}_{ki}^2\mathbf{B}_{kj}^2}{sp_k}-\frac{1}{s}(\mathbf{A}^\intercal\mathbf{B})_{ij}\right),$$

$$=\sum_{k=1}^{n}\frac{\left(\sum_{i=1}^{m}\mathbf{A}_{ki}^2\right)\left(\sum_{j=1}^{p}\mathbf{B}_{kj}^2\right)}{sp_k}-\frac{1}{s}\sum_{i=1}^{m}\sum_{j=1}^{p}(\mathbf{A}^\intercal\mathbf{B})_{ij},$$

$$=\frac{1}{s}\sum_{k=1}^{n}\frac{\|\mathbf{A}(k,:)\|_2^2\|\mathbf{B}(k,:)\|_2^2}{p_k}-\frac{1}{s}\|\mathbf{A}^\intercal\mathbf{B}\|_F^2,$$

Norm definitions

Proof

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^\intercal\mathbf{SB}\|_F^2\right]=\sum_{i=1}^{m}\sum_{j=1}^{p}\left(\sum_{k=1}^{n}\frac{\mathbf{A}_{ki}^2\mathbf{B}_{kj}^2}{sp_k}-\frac{1}{s}(\mathbf{A}^\intercal\mathbf{B})_{ij}\right),$$

$$=\sum_{k=1}^{n}\frac{\left(\sum_{i=1}^{m}\mathbf{A}_{ki}^2\right)\left(\sum_{j=1}^{p}\mathbf{B}_{kj}^2\right)}{sp_k}-\frac{1}{s}\sum_{i=1}^{m}\sum_{j=1}^{p}(\mathbf{A}^\intercal\mathbf{B})_{ij},$$

$$=\frac{1}{s}\sum_{k=1}^{n}\frac{\|\mathbf{A}(k,:)\|_2^2\|\mathbf{B}(k,:)\|_2^2}{p_k}-\frac{1}{s}\|\mathbf{A}^\intercal\mathbf{B}\|_F^2,$$

$$\leq\frac{1}{s}\sum_{k=1}^{n}\frac{\|\mathbf{A}(k,:)\|_2^2\|\mathbf{B}(k,:)\|_2^2}{p_k}$$

Frobenius norm is ≥ 0

Proof

MathSci.ai

So far we have not used any assumptions about p_k.

$$\mathbb{E}\left[\|\mathbf{AB} - (\mathbf{SA})^\intercal \mathbf{SB}\|_F^2\right] \leq \frac{1}{s}\sum_{k=1}^{n} \frac{\|\mathbf{A}(k,:)\|_2^2 \|\mathbf{B}(k,:)\|_2^2}{p_k}$$

Result from previous page.

Proof

So far we have not used any assumptions about p_k.

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^{\intercal}\mathbf{SB}\|_F^2\right] \leq \frac{1}{s}\sum_{k=1}^{n}\frac{\|\mathbf{A}(k,:)\|_2^2\|\mathbf{B}(k,:)\|_2^2}{p_k}$$

$$\leq \frac{1}{s}\sum_{k=1}^{n}\left(\|\mathbf{A}^{\intercal}\|_F^2\frac{\|\mathbf{A}(k,:)\|_2^2\|\mathbf{B}(k,:)\|_2^2}{\beta\|\mathbf{A}(k,:)\|_2^2}\right),$$

Theorem assumes $p_k \geq \beta\|\mathbf{A}(k,:)\|_2^2/\|\mathbf{A}^{\intercal}\|_F^2$

Proof

So far we have not used any assumptions about p_k.

$$\mathbb{E}\left[\|\mathbf{AB}-(\mathbf{SA})^{\intercal}\mathbf{SB}\|_F^2\right] \leq \frac{1}{s}\sum_{k=1}^{n}\frac{\|\mathbf{A}(k,:)\|_2^2\|\mathbf{B}(k,:)\|_2^2}{p_k}$$

$$\leq \frac{1}{s}\sum_{k=1}^{n}\left(\|\mathbf{A}^{\intercal}\|_F^2\frac{\|\mathbf{A}(k,:)\|_2^2\|\mathbf{B}(k,:)\|_2^2}{\beta\|\mathbf{A}(k,:)\|_2^2}\right),$$

$$= \frac{1}{\beta s}\|\mathbf{A}\|_F^2\sum_{k=1}^{n}\|\mathbf{B}(k,:)\|_2^2 = \frac{1}{\beta s}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2.$$

Simplifying

■

Markov's Inequality

MathSci.ai

Theorem gives us a bound on the expectation.

Proposition (Markov's inequality for positive random variables) For a random variable Z such that $Z \geq 0$ (a.e.),

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t} \quad \text{for any} \quad t > 0.$$

Markov's inequality enables us to change a bound on the expectation into a bound on the probability.

Proof

$$\text{Prob}_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\geq t\right]\leq\frac{\mathbb{E}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\right]}{t}$$

Frobenius norm non-negative, apply Markov

Proof

$$\text{Prob}_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\geq t\right]\leq\frac{\mathbb{E}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\right]}{t}$$

$$\leq\frac{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2}{\beta st}$$

Bound on expectation we just proved.

Proof

$$\text{Prob}_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\geq t\right]\leq\frac{\mathbb{E}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\right]}{t}$$
$$\leq\frac{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2}{\beta st}$$

Setting $\delta=\frac{1}{\beta st}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$ and solving we get:

$$t=\frac{1}{\beta s\delta}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$$

Proof

$$\begin{aligned}\text{Prob}_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\geq t\right] &\leq \frac{\mathbb{E}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\right]}{t}\\ &\leq \frac{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2}{\beta s t}\end{aligned}$$

Setting $\delta = \frac{1}{\beta s t}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$ and solving we get:

$$t = \frac{1}{\beta s \delta}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$$

Which gives:

$$\text{Prob}_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\geq \frac{1}{\beta s\delta}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\right]\leq\delta$$

Proof

$$\begin{aligned}\text{Prob}_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\geq t\right] &\leq \frac{\mathbb{E}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2\right]}{t}\\ &\leq \frac{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2}{\beta s t}\end{aligned}$$

Setting $\delta = \frac{1}{\beta s t}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$ and solving we get:

$$t = \frac{1}{\beta s \delta}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$$

Or for the complement event:

$$\text{Prob}_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{A}^\intercal\mathbf{B}-(\mathbf{S}\mathbf{A})^\intercal\mathbf{S}\mathbf{B}\|_F^2 \leq \frac{1}{\beta s\delta}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\right] \geq 1-\delta$$

Probabilistic Guarantee of Accuracy

MathSci.ai

Corollary (Drineas, Kannan, and Mahoney, 2006) Choose s rows, denoted $\{\xi^{(1)}, \ldots, \xi^{(s)}\}$, according to the probability distribution $\mathbf{p} \in [0,1]^n$ with the property that there exists $\beta > 0$ such that

$$p_k \geq \beta \|\mathbf{A}(k,:)\|^2 / \|\mathbf{A}\|_F^2 \quad \text{for all} \quad k \in [n].$$

Then with probability at least $1 - \delta$:

$$\|\mathbf{A}^\intercal \mathbf{B} - (\mathbf{S}\mathbf{A})^\intercal \mathbf{S}\mathbf{B}\|_F^2 \leq \frac{1}{\beta s \delta} \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2.$$

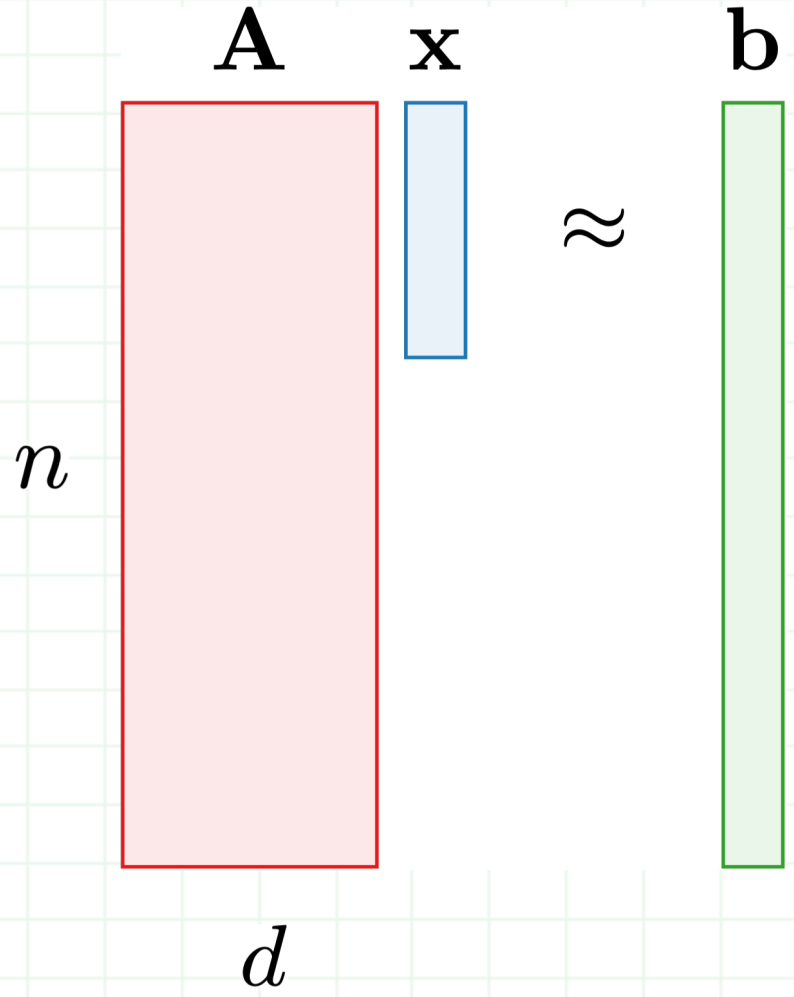
Proof only required Markov's inequality and a judicious choice of probabilities.

MathSci.ai

Randomized Least Squares

- Drineas, Mahoney, Muthukrishnan, and Sarlos (2011). **Faster Least Squares Approximation**. Numerische mathematic. https://doi.org/10.1007/s00211-010-0331-6
- Larsen, B. W. and T. G. Kolda (2022). **Sketching Matrix Least Squares via Leverage Score Estimates.** https://arxiv.org/abs/2201.10638

Overdetermined Least Squares Problem



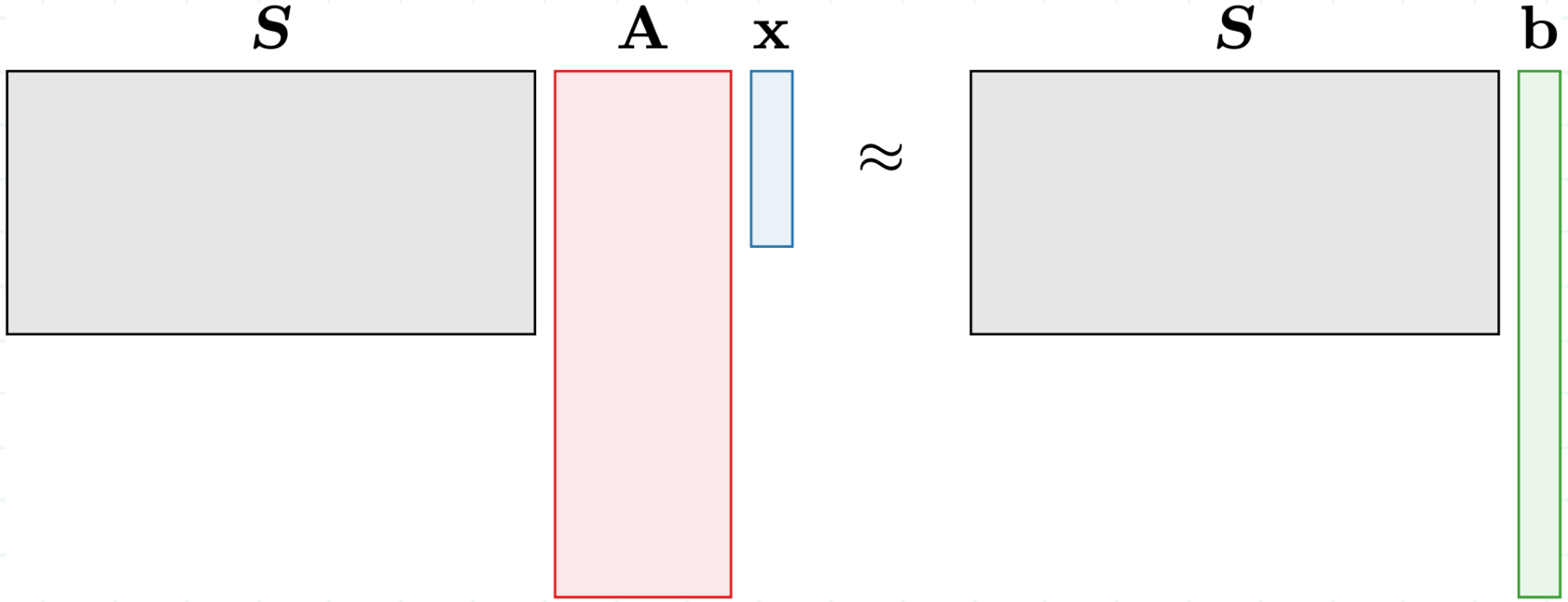
Given:

$$\mathbf{A} \in \mathbb{R}^{n \times d} \text{ with } n \gg d,$$
$$\mathbf{b} \in \mathbb{R}^{n}$$

Want:

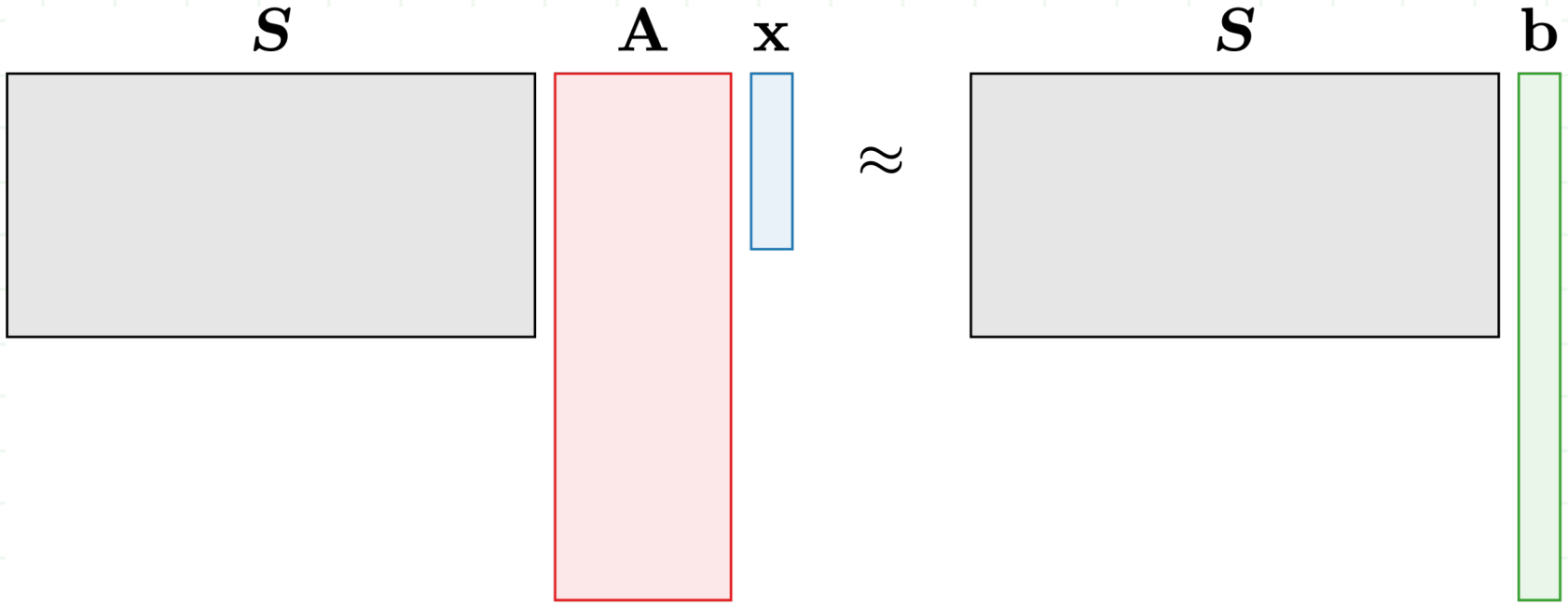
$$\mathbf{x}_{\text{opt}} = \underbrace{\arg\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2}_{\text{Minimizes Squared Residual}}$$

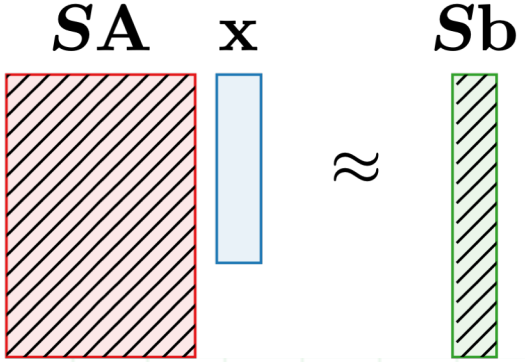
Sketching Least Squares



Random Sketching Matrix: $\boldsymbol{S} \in \mathbb{R}^{s \times n}, \boldsymbol{S} \sim \mathcal{D}, s < n$

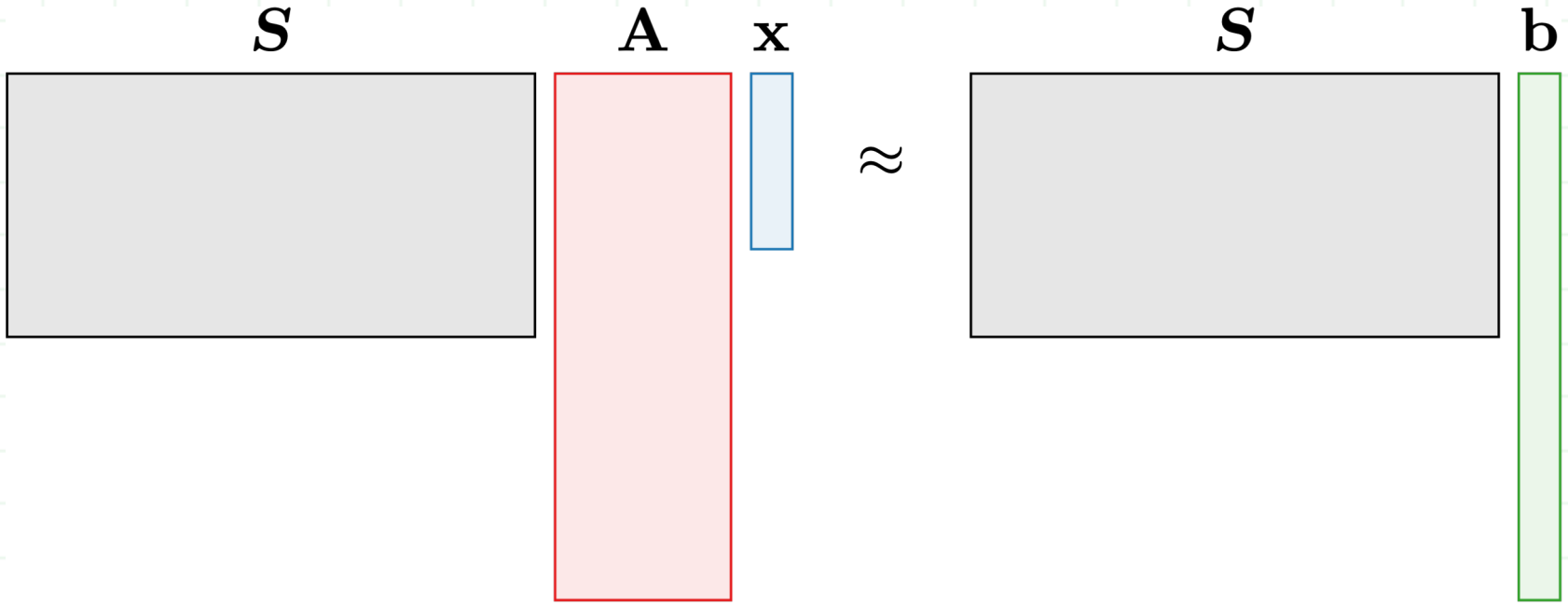
Sketching Least Squares

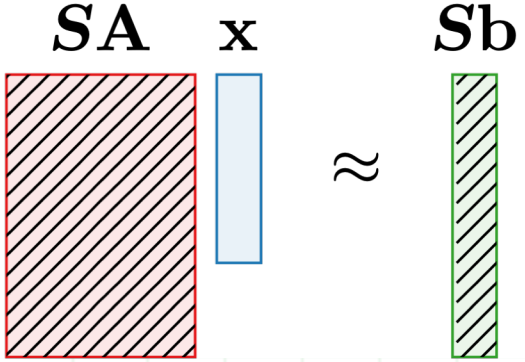




Smaller “Sketched”
Least Squares Problem

Sketching Least Squares



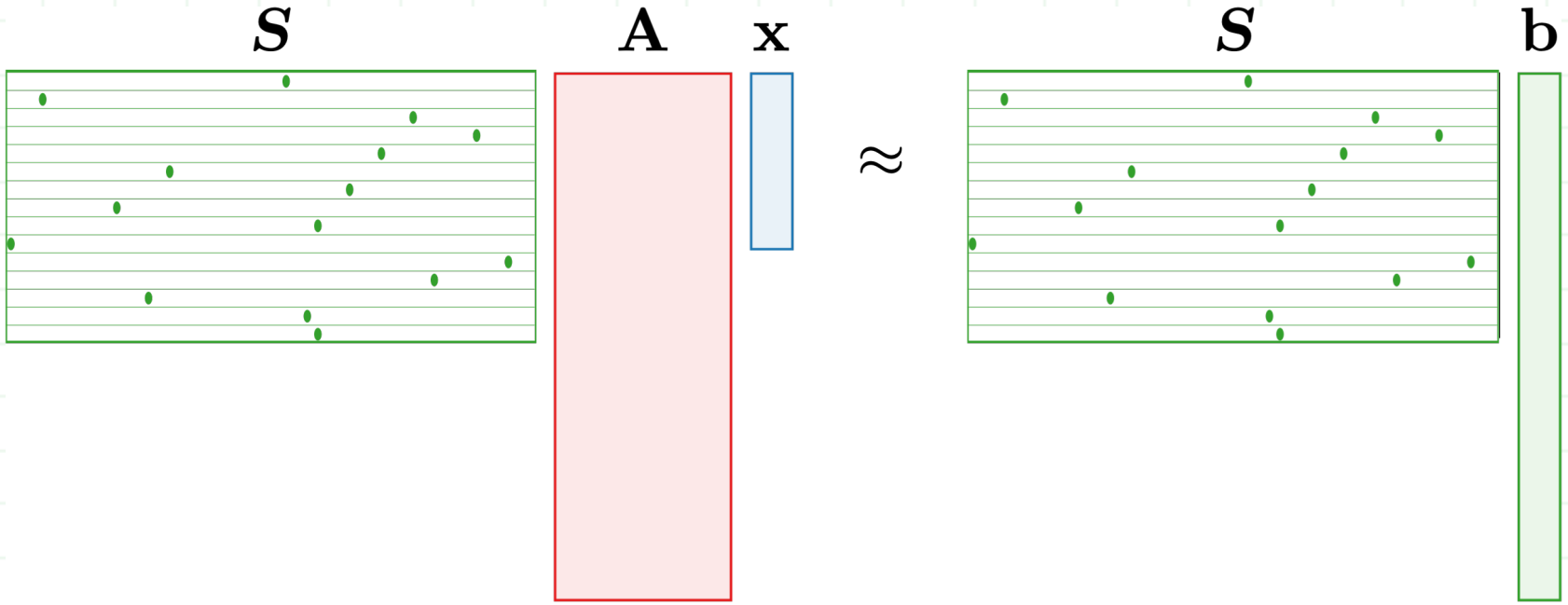


Smaller “Sketched” Least Squares Problem

Solve Smaller Problem

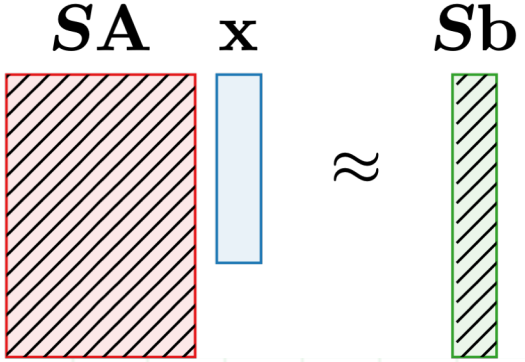
$$\widetilde{\mathbf{x}}_{\text{opt}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \|\boldsymbol{S}\mathbf{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_2^2$$

Sketching via Row Sampling



Need to chose sketching matrix. We're going to again use a row sampling matrix.

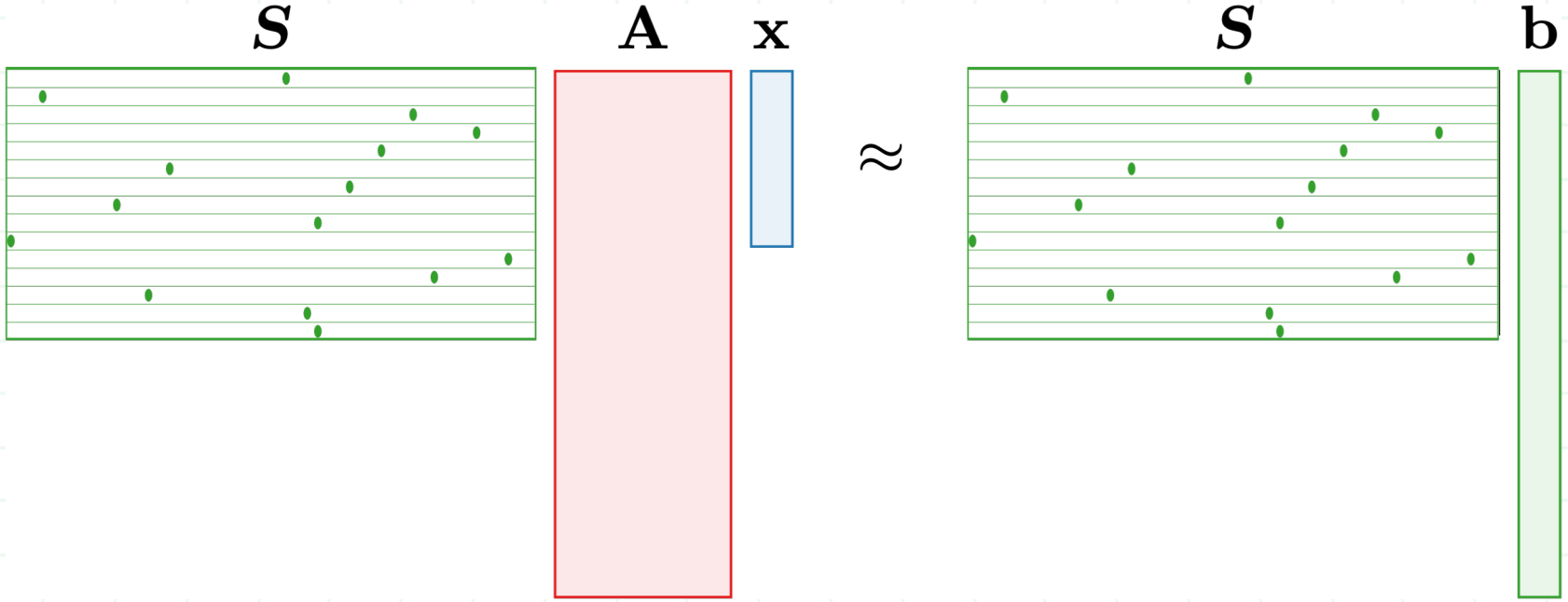
Can we get a good approximation with $s \ll n$?



Solve Smaller Problem

$$\widetilde{\mathbf{x}}_{\text{opt}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \|\boldsymbol{S}\mathbf{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_2^2$$

Sketching via Row Sampling



1. Select rows of $\mathbf{A}/\mathbf{b}$ with replacement
2. Probability p_i that row i is selected, with $\sum_{i=1}^{N} p_i = 1$
3. Row sampled is weighted by $1/\sqrt{sp_i}$ so that:

$$\mathbb{E}\left(\|\boldsymbol{S}\mathbf{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_2^2 \right) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

How should we chose p?

How do we measure success (ϵ-error)?

MathSci.ai

Original system with n rows

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

Sampled system with s rows

$$\widetilde{\mathbf{x}}_{\text{opt}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \|\boldsymbol{S}\mathbf{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_2^2$$

Relative residual error :

$$\frac{\|\mathbf{A}\widetilde{\mathbf{x}}_{\text{opt}} - \mathbf{b}\|_2^2 - \|\mathbf{A}\mathbf{x}_{\text{opt}} - \mathbf{b}\|_2^2}{\|\mathbf{A}\mathbf{x}_{\text{opt}} - \mathbf{b}\|_2^2} \leq \varepsilon$$

Equivalent expression:

$$\underbrace{\|\mathbf{A}\widetilde{\mathbf{x}}_{\text{opt}} - \mathbf{b}\|_2^2}_{\text{Residual with solution of sketched problem}} \leq (1+\varepsilon)\underbrace{\|\mathbf{A}\mathbf{x}_{\text{opt}} - \mathbf{b}\|_2^2}_{\text{Best possible residual}}$$

Some Notation/Definitions

MathSci.ai

SVD of the design matrix:

$$\mathbf{A} = \mathbf{U_A} \mathbf{\Sigma_A} \mathbf{V_A}^\intercal$$

Value of the optimal squared residual:

$$\mathcal{R}^2 \triangleq \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

Subspace orthogonal to the column space of $\mathbf{A}$:

- $\mathbf{U_A}$ is an orthonormal basis for the d-dim. column space of $\mathbf{A}$
- Define $\mathbf{U_A^\perp}$ to be the $(n-d)$-dim. orthgonal subspace
- Define $\mathbf{b}^\perp$ to be the projection of $\mathbf{b}$ onto this subpsace:

$$\mathbf{b}^\perp \triangleq \mathbf{U_A^\perp} \mathbf{U_A^\perp}^\intercal \mathbf{b}$$

Some Notation/Definitions

By definition:

$$\mathbf{U}_{\mathbf{A}}^{\intercal}\mathbf{b}^{\perp} = \mathbf{0}_n, \text{ i.e. } \mathbf{b}^{\perp} \text{ is orthogonal to the column space of } \mathbf{A}$$

Alternative way to write optimal residual:

$$\mathcal{R}^2 = \min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \|\mathbf{U}_{\mathbf{A}}^{\perp}\mathbf{U}_{\mathbf{A}}^{\perp\intercal}\mathbf{b}\|_2^2 = \|\mathbf{b}^{\perp}\|_2^2$$

This is because $\mathbf{x}$ can be chosen to exactly match the part of $\mathbf{b}$ in the column space of $\mathbf{A}$ but cannot match anything in orthogonal subspace. Thus the optimal $\mathbf{x}$ will yield the following decomposition:

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{opt}} + \mathbf{b}^{\perp}$$

Two conditions for an ϵ-accurate solution

MathSci.ai

Lemma For the overdetermined least squares problem, assume that the sketch matrix $\boldsymbol{S}$ satisfies the following two conditions for some $\epsilon \in (0, 1)$:

$$\sigma_{\min}^2(\mathbf{S}\mathbf{U}_\mathbf{A}) \geq 1/\sqrt{2}, \quad \text{and} \tag{1}$$

$$\|\mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2 \leq \epsilon \mathcal{R}^2/2. \tag{2}$$

Then the solution to the sketched problem, denoted $\widetilde{\mathbf{x}}_{\text{opt}}$ is ε-accurate, i.e.:

$$\|\mathbf{A}\widetilde{\mathbf{x}}_{\text{opt}} - \mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}_{\text{opt}} - \mathbf{b}\|_2^2$$

See Theorem 4.1 of Larsen and Kolda (2022) for a version of the proof.

Two conditions for an ϵ-accurate solution

MathSci.ai

The first property has to do with how well the subspace of $\mathbf{SU_A}$ approximates the subspace defined by $\mathbf{U_A}$, i.e. the column space of $\mathbf{A}$

$$\sigma^2_{\min}(\mathbf{SU_A}) \geq 1/\sqrt{2}$$

The second property is acutally a randomized matrix multiplication condition and thus by our results earlier we should sample by the row norms squared of $\mathbf{U_A}$.

$$\begin{aligned}\|\mathbf{U_A^T S^T S b^\perp}\|_2^2 = \|\mathbf{0}_n - \mathbf{U_A^T S^T S b^\perp}\|_2^2 &= \|\mathbf{U_A^T b^\perp} - \mathbf{U_A^T S^T S b^\perp}\|_2^2 \\ &\leq \epsilon \mathcal{R}^2/2.\end{aligned}$$

Both conditions point towards sampling by the *leverage scores* of $\mathbf{A}$.

Leverage Scores

MathSci.ai

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $n > d$

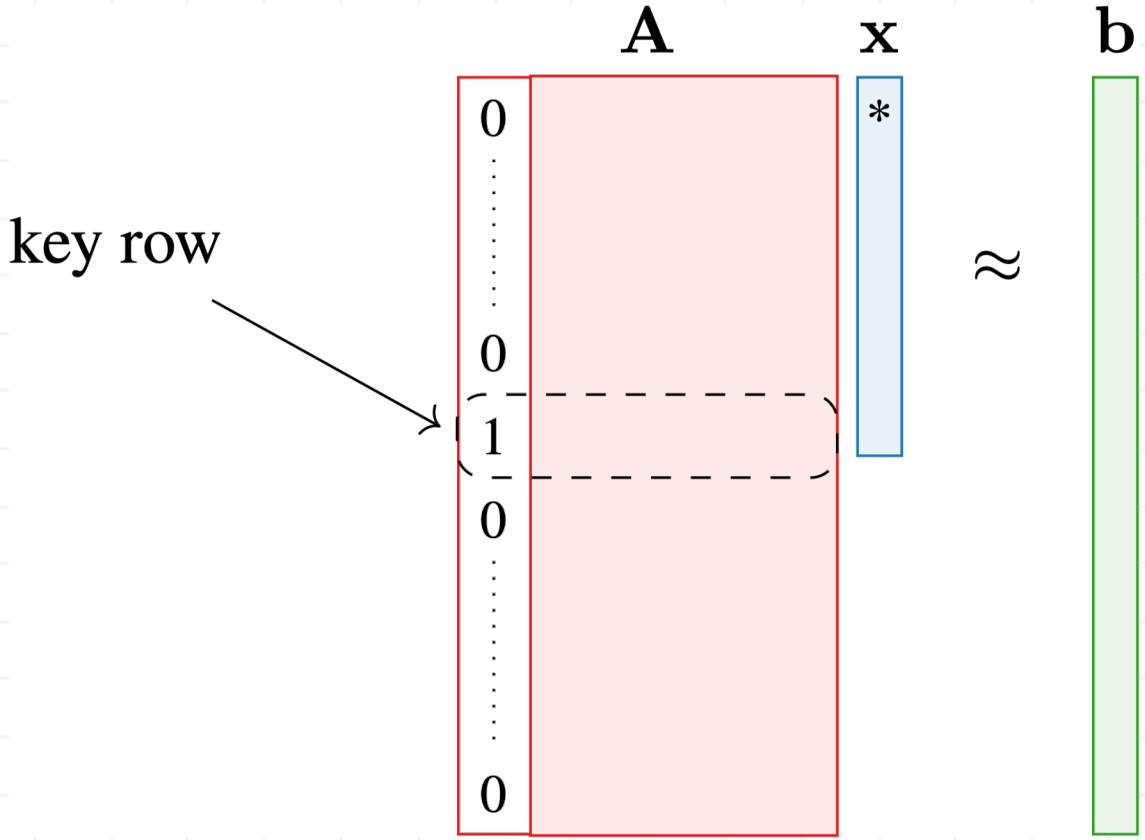
Let $\mathbf{Q} \in \mathbb{R}^{n \times d}$ be any orthogonal basis for column space of $\mathbf{A}$ (e.g. $\mathbf{U_A}$).

Leverage scores: $\ell_i(\mathbf{A}) = \|\mathbf{Q}(i,:)\|_2^2 \quad \text{for} \quad i \in \{1, 2, \ldots, n\}$

- $\sum_{i=1}^{n} \ell_i(\mathbf{A}) = d$
- $\max_i \ell_i(\mathbf{A}) \in [d/n, 1]$
- $p_i^* = \ell_i(\mathbf{A})/d$

Intuition: The leverage score of a row is its "fractional contribution" to the column space of $\mathbf{A}$.

Leverage Scores



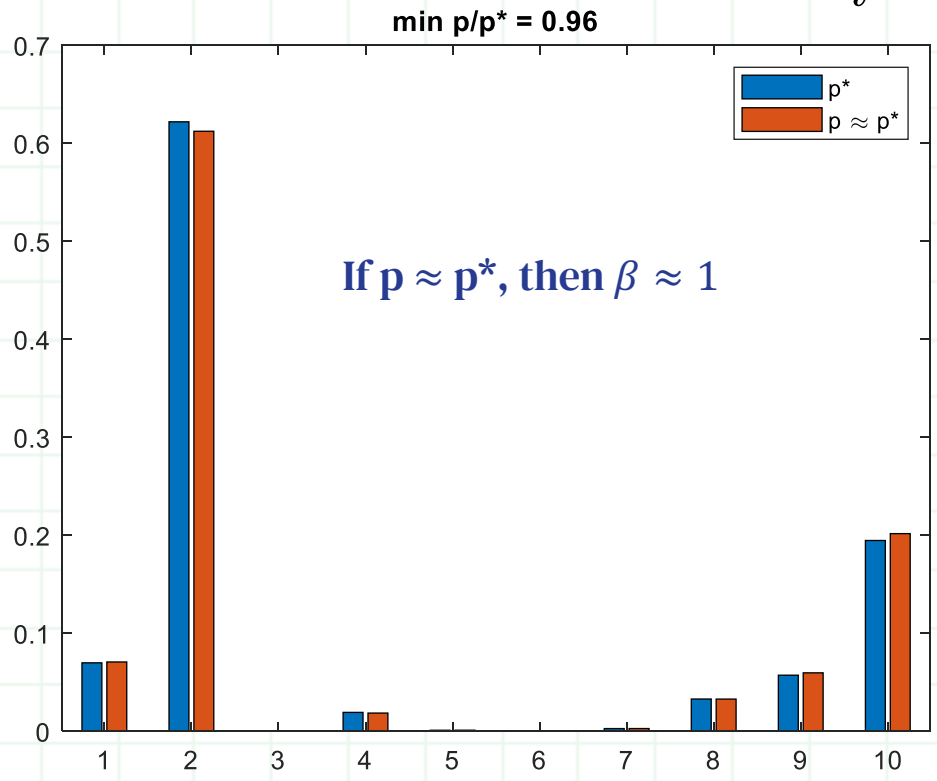
This row will have a leverage score of 1 because not including it will decrease the dimension of the column space by 1

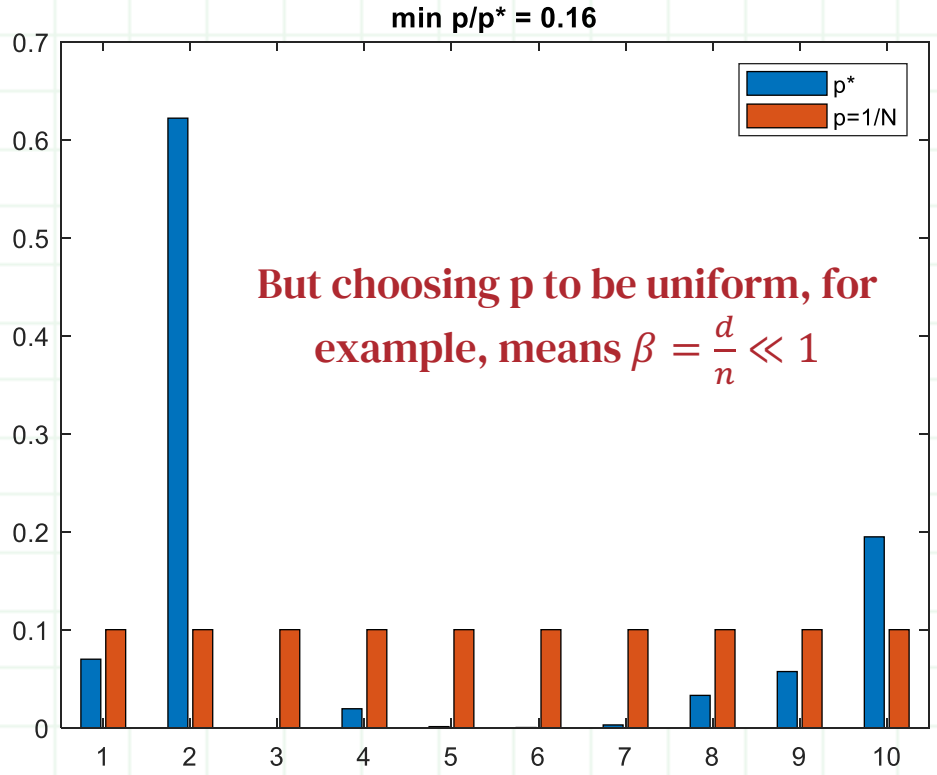
In general, exact leverage scores are prohibitively expensive to compute; require finding the SVD of $\mathbf{A}$.

Approximate Leverage Scores

Misestimate factor: $$\beta \leq \min_i \frac{p_i}{p_i^*} \equiv \min_i \frac{p_i d}{\ell_i(\mathbf{A})} \in (0, 1]$$

Can use any procedure to compute p_i but want β as close to one as possible.





Overview of Proof

$$p_i \geq \beta \ell_i(\mathbf{A})/r, \quad \text{for all} \quad i \in [n].$$

1. Show that first condition ($\sigma^2_{\min}(\mathbf{S}\mathbf{U}_\mathbf{A}) \geq 1/\sqrt{2}$) holds with probability at least $1 - \delta/2$ if (will not show, see references for proof):

$$s = Cr \log(r/\delta)/\beta \quad \text{where} \quad C = 144/(1 - 1/\sqrt{2})^2$$

2. Show that the second condition ($\|\mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2 \leq \epsilon \mathcal{R}^2/2$) holds with probability at least $1 - \delta/2$ if:

$$s = r/(\beta\delta\epsilon).$$

3. Use the union bound to show that both conditions hold simultaneously with probability at least $1 - \delta$ if both sample bounds are true.

Larsen & Kolda, arXiv:2201.10638 (2022); Drineas et al., *Numerische Mathematik* (2011), Woodruff, *FNT-TCS* (2014)

Lemma – Samples for Condition 2

MathSci.ai

Lemma Consider full rank $\mathbf{A} \in \mathbb{R}^{n \times d}$, its SVD $\mathbf{U_A \Sigma_A V_A^\intercal}$, and row leverage scores $\ell_i(\mathbf{A})$. Define the probability distribution $\mathbf{p} \in [0,1]^N$ and assume there exists $\beta \in (0,1]$ such that $p_i \geq \beta \ell_i(\mathbf{A})/r$ for all $i \in [n]$.

Construct row sampling and rescaling matrix $\mathbf{S} \in \mathbb{R}^{s \times n}$ by importance sampling by the leverage score overestimates.

Then provided $s \geq \frac{2d}{\beta\delta\epsilon}$, the property $\|\mathbf{U_A^\intercal S^\intercal S b}^\perp\|_F^2 \leq \epsilon \mathcal{R}^2/2$ holds with probability at least $1 - \delta$.

This is essentially the randomized matrix multiplication result from earlier!

Proof

$$\mathbb{E}\left[\|\mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2\right] = \mathbb{E}\left[\|\mathbf{0}_n - \mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2\right],$$

$$= \mathbb{E}\left[\|\mathbf{U}_\mathbf{A}\mathbf{b}^\perp - \mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2\right],$$

$$\mathbf{U}_\mathbf{A}\mathbf{b}^\perp = \mathbf{0}_n$$

Proof

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2\right] &= \mathbb{E}\left[\|\mathbf{0}_n - \mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2\right], \\ &= \mathbb{E}\left[\|\mathbf{U}_\mathbf{A}\mathbf{b}^\perp - \mathbf{U}_\mathbf{A}^\intercal \mathbf{S}^\intercal \mathbf{S}\mathbf{b}^\perp\|_2^2\right], \\ &\leq \frac{1}{\beta s}\|\mathbf{U}_\mathbf{A}\|_F^2 \|\mathbf{b}^\perp\|_2^2 = \frac{d}{\beta s}\|\mathbf{b}^\perp\|_2^2\end{aligned}$$

Randomized matrix multiplication result, $\|\mathbf{U}_\mathbf{A}\|_F^2 = d$

Proof

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{U}_{\mathbf{A}}^{\intercal}\mathbf{S}^{\intercal}\mathbf{S}\mathbf{b}^{\perp}\|_2^2\right] &= \mathbb{E}\left[\|\mathbf{0}_n - \mathbf{U}_{\mathbf{A}}^{\intercal}\mathbf{S}^{\intercal}\mathbf{S}\mathbf{b}^{\perp}\|_2^2\right], \\
&= \mathbb{E}\left[\|\mathbf{U}_{\mathbf{A}}\mathbf{b}^{\perp} - \mathbf{U}_{\mathbf{A}}^{\intercal}\mathbf{S}^{\intercal}\mathbf{S}\mathbf{b}^{\perp}\|_2^2\right], \\
&\leq \frac{1}{\beta s}\|\mathbf{U}_{\mathbf{A}}\|_F^2\|\mathbf{b}^{\perp}\|_2^2 = \frac{d}{\beta s}\|\mathbf{b}^{\perp}\|_2^2 \\
&= \frac{d}{\beta s}\mathcal{R}^2.
\end{aligned}
$$

Using $\|\mathbf{b}^{\perp}\|_2^2 = \mathcal{R}^2$

Proof

Apply Markov's inequality:

$$\Pr_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{U}_\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b}^\perp\|_2^2 \geq \frac{\epsilon\mathcal{R}^2}{2}\right] \leq \frac{2\mathbb{E}\left[\|\mathbf{U}_\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b}^\perp\|_2^2\right]}{\epsilon\mathcal{R}^2}$$
$$\leq \frac{2d}{\beta\epsilon s}$$

Proof

Apply Markov's inequality:

$$\Pr_{\mathbf{S}\sim\mathcal{D}}\left[\|\mathbf{U}_\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b}^\perp\|_2^2 \geq \frac{\epsilon\mathcal{R}^2}{2}\right] \leq \frac{2\mathbb{E}\left[\|\mathbf{U}_\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b}^\perp\|_2^2\right]}{\epsilon\mathcal{R}^2}$$
$$\leq \frac{2d}{\beta\epsilon s}$$

We thus need $\frac{2d}{\beta\epsilon s} \leq \delta$. Solving for s:

$$s \geq \frac{2d}{\beta\delta\epsilon}$$

■

Probabilistic Guarantee – Randomized Least Squares

MathSci.ai

Original system with n rows

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

Sampled system with s rows

$$\widetilde{\mathbf{x}}_{\text{opt}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \|\boldsymbol{S}\mathbf{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_2^2$$

$$\text{Prob}\left(\|\mathbf{A}\widetilde{\mathbf{x}}_{\text{opt}} - \mathbf{b}\|_2^2 \le (1+\varepsilon)\|\mathbf{A}\mathbf{x}_{\text{opt}} - \mathbf{b}\|_2^2 \right) > (1-\delta)$$

if $s = \frac{d}{\beta}\max\left\{C\log\left(\frac{d}{\delta}\right), \frac{1}{\delta\varepsilon}\right\}$ where $\beta \le \min_i \frac{p_i d}{\ell_i(\mathbf{A})} \in (0,1]$

Want this to be independent of n which it will be if β is.

Larsen & Kolda, arXiv:2201.10638 (2022); Drineas et al., *Numerische Mathematik* (2011), Woodruff, *FNT-TCS* (2014)

Leverage Score in Practice

For general matrices, computing the exact leverage scores is as expensive as solving the full least squares problem. How to deal with this?

Leverage Score in Practice

MathSci.ai

For general matrices, computing the exact leverage scores is as expensive as solving the full least squares problem. How to deal with this?

1. For arbitrary overdetermined least squares problems, random projections can be used to find approximate leverage scores with probabilistic guarantees on quality. See:

   Drineas, Magdon-Ismail, Mahoney, and Woodruff. Fast Approximation of Matrix Coherence and Statical Leverage. *JMLR* (2012). https://jmlr.org/papers/v13/drineas12a.html

Leverage Score in Practice

MathSci.ai

For general matrices, computing the exact leverage scores is as expensive as solving the full least squares problem. How to deal with this?

1. For arbitrary overdetermined least squares problems, random projections can be used to find approximate leverage scores with probabilistic guarantees on quality. See:

   Drineas, Magdon-Ismail, Mahoney, and Woodruff. Fast Approximation of Matrix Coherence and Statical Leverage. *JMLR* (2012). https://jmlr.org/papers/v13/drineas12a.html

2. Certain problems have special structure to the design problem that allows for faster computation of leverage scores. For example, the design matrix is the Khatri-Rao product of smaller matrices when solving for the CP tensor decomposition.

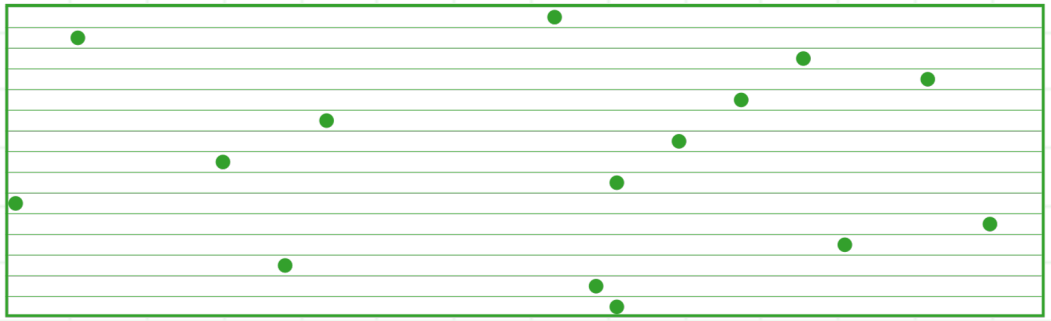
   Larsen and Kolda. Practical Leverage-Based Sampling for Low-Rank Tensor Decomposition. *SIMAX* (2022). https://doi.org/10.1137/21M1441754

Leverage Score in Practice

MathSci.ai

For general matrices, computing the exact leverage scores is as expensive as solving the full least squares problem. How to deal with this?

3. There is another family of algorithms for randomized least squares that first perform a "mixing" operation to create a new problem with all the leverage scores close to uniform (i.e. d/n). As a result, uniform sampling is now a good approximation to leverage score sampling.



Practical Considerations

- If a row is sampled multiple times, we can include the row once and reweight it appropriately to account for the multiple samples rather than repeating the row.

$$\sqrt{\frac{c}{sp_i}} \quad \text{where } c \text{ is the number of repeats for row } i$$

Practical Considerations

MathSci.ai

- If a row is sampled multiple times, we can include the row once and reweight it appropriately to account for the multiple samples rather than repeating the row.

$$\sqrt{\frac{c}{sp_i}} \quad \text{where } c \text{ is the number of repeats for row } i$$

- If there are some rows with very high leverage score approximations, they can be sampled quite frequently. We can include these row deterministically and then sample from the remaining rows. This again requires some careful reweighting.

Larsen and Kolda. Practical Leverage–Based Sampling for Low–Rank Tensor Decomposition. *SIMAX* (2022). https://doi.org/10.1137/21M1441754

Summary

MathSci.ai

- Showed how two linear algebra primitives can be randomized with probabilistic guarantees on accuracy.
- Were able to partially prove these results with basic techniques from probability and statistics (Markov's inequality, Union Bound).
- For a more complete version of the proof, see:

Larsen, B. W. and T. G. Kolda (2022). **Sketching Matrix Least Squares via Leverage Score Estimates.** https://arxiv.org/abs/2201.10638