COMPSCI 614: Randomized Algorithms with Applications to Data Science

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2024.

Lecture 10

Logistics

- Problem Set 2 is due tonight at 11:59pm.
- One page project proposal due Tuesday 3/12.
- Quiz due Monday released after class.

Summary

Last Time:

- Count sketch for ℓ_2 heavy-hitters – estimate all entries of a vector x to error $\pm\epsilon\|x\|_2$ from a linear sketch of dimension $O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$.
- Analysis via linearity of expectation, variance, Chebyshev's inequality and median trick.

Summary

Last Time:

- Count sketch for ℓ_2 heavy-hitters – estimate all entries of a vector x to error $\pm\epsilon\|x\|_2$ from a linear sketch of dimension $O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$.
- Analysis via linearity of expectation, variance, Chebyshev's inequality and median trick.

Today:

- Approximate matrix multiplication via importance sampling.
- Application to fast low-rank approximation via sampling.

Approximate Matrix Multiplication

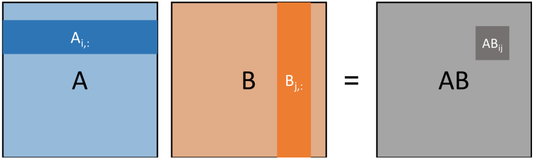
Matrix Multiplication Problem

Given $A, B \in \mathbb{R}^{n \times n}$ would like to compute $C = AB$. Requires n^ω time where $\omega \approx 2.373$ in theory.

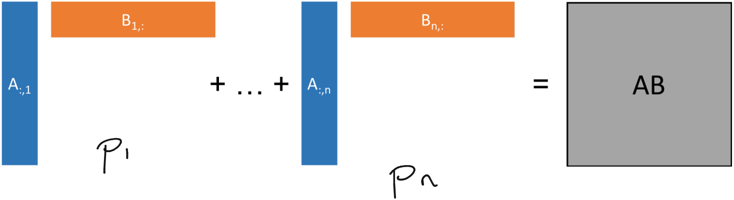
- We'll see how to compute an approximation in $O(n^2)$ time via a simple sampling approach.
- This is one of the fundamental building blocks of randomized numerical linear algebra.
- E.g. later in class we will use it to develop a fast algorithm for low-rank approximation.

Outer Product View of Matrix Multiplication

Inner Product View: $[AB]_{ij} = \langle A_{i,:}, B_{j,:} \rangle = \sum_{k=1}^{n} A_{ik} \cdot B_{kj}$.



Outer Product View: Observe that $C_k = A_{:,k} B_{k,:}$ is an $n \times n$ matrix with $[C_k]_{ij} = A_{jk} \cdot B_{kj}$. So $AB = \sum_{k=1}^{n} A_{:,k} B_{k,:}$



Basic Idea: Approximate **AB** by sampling terms of this sum.

Canonical AMM Algorithm

Approximate Matrix Multiplication (AMM):

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i \geq 0$ and $\sum_{[n]} p_i = 1$.
- Select $\mathbf{i_1}, \ldots, \mathbf{i_t} \in [n]$ independently, according to the distribution $\Pr[\mathbf{i_j} = k] = p_k$.
- Let $\overline{\mathbf{C}} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{p_{\mathbf{i_j}}} \cdot A_{:,\mathbf{i_j}} B_{\mathbf{i_j},:}$.

$\mathbb{E}\,\overline{C} = C = AB$

Canonical AMM Algorithm

Approximate Matrix Multiplication (AMM):

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i \geq 0$ and $\sum_{[n]} p_i = 1$.
- Select $\mathbf{i_1}, \ldots, \mathbf{i_t} \in [n]$ independently, according to the distribution $\Pr[\mathbf{i_j} = k] = p_k$.
- Let $\overline{\mathbf{C}} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{p_{\mathbf{i_j}}} \cdot A_{:,\mathbf{i_j}} B_{\mathbf{i_j},:}$.

Claim 1: $\mathbb{E}[\overline{\mathbf{C}}] = AB$

Canonical AMM Algorithm

Approximate Matrix Multiplication (AMM):

- Fix sampling probabilities $p_1, \dots, p_n$ with $p_i \geq 0$ and $\sum_{[n]} p_i = 1$.
- Select $\mathbf{i_1}, \dots, \mathbf{i_t} \in [n]$ independently, according to the distribution $\Pr[\mathbf{i_j} = k] = p_k$.
- Let $\overline{\mathsf{C}} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{p_{\mathbf{i_j}}} \cdot A_{:,\mathbf{i_j}} B_{\mathbf{i_j},:}$.

Claim 1: $\mathbb{E}[\overline{\mathsf{C}}] = AB$

$$\mathbb{E}[\overline{\mathsf{C}}] = \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{p_{\mathbf{i_j}}} \cdot A_{:,\mathbf{i_j}} B_{\mathbf{i_j},:}\right]$$

Canonical AMM Algorithm

Approximate Matrix Multiplication (AMM):

$p_1 = \dots = p_n = \frac{1}{n}$

- Fix sampling probabilities $p_1, \dots, p_n$ with $p_i \geq 0$ and $\sum_{[n]} p_i = 1$.
- Select $\mathbf{i}_1, \dots, \mathbf{i}_t \in [n]$ independently, according to the distribution $\Pr[\mathbf{i}_j = k] = p_k$.
- Let $\overline{\mathbf{C}} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{p_{\mathbf{i}_j}} \cdot A_{:,\mathbf{i}_j} B_{\mathbf{i}_j,:}$.

$\frac{n}{t} \cdot \dots \sum A_{:,i_j} B_{i_j,:}$

Claim 1: $\mathbb{E}[\overline{\mathbf{C}}] = AB$

$$\mathbb{E}[\overline{\mathbf{C}}] = \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{p_{\mathbf{i}_j}} \cdot A_{:,\mathbf{i}_j} B_{\mathbf{i}_j,:}\right] = \frac{1}{t} \sum_{j=1}^{t} \underbrace{\sum_{k=1}^{n} p_k \cdot \frac{1}{p_k} \cdot A_{:,k} B_{k,:}}$$

$$= \frac{1}{t} \sum_{j=1}^{t} AB = AB \quad \checkmark$$

Canonical AMM Algorithm

Approximate Matrix Multiplication (AMM):

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i \geq 0$ and $\sum_{[n]} p_i = 1$.
- Select $\mathsf{i_1}, \ldots, \mathsf{i_t} \in [n]$ independently, according to the distribution $\Pr[\mathsf{i_j} = k] = p_k$.
- Let $\overline{\mathsf{C}} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{p_{\mathsf{i_j}}} \cdot A_{:,\mathsf{i_j}} B_{\mathsf{i_j},:}$.

Claim 1: $\mathbb{E}[\overline{\mathsf{C}}] = AB$

$$\mathbb{E}[\overline{\mathsf{C}}] = \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{p_{\mathsf{i_j}}} \cdot A_{:,\mathsf{i_j}} B_{\mathsf{i_j},:}\right] = \frac{1}{t} \sum_{j=1}^{t} \sum_{k=1}^{n} p_k \cdot \frac{1}{p_k} \cdot A_{:,k} B_{k,:} = \frac{1}{t} \sum_{j=1}^{t} AB$$

Canonical AMM Algorithm

Approximate Matrix Multiplication (AMM):

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i \geq 0$ and $\sum_{[n]} p_i = 1$.
- Select $\mathbf{i_1}, \ldots, \mathbf{i_t} \in [n]$ independently, according to the distribution $\Pr[\mathbf{i_j} = k] = p_k$.
- Let $\overline{\mathbf{C}} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{p_{\mathbf{i_j}}} \cdot A_{:,\mathbf{i_j}} B_{\mathbf{i_j},:}$.

Claim 1: $\mathbb{E}[\overline{\mathbf{C}}] = AB$

$$\mathbb{E}[\overline{\mathbf{C}}] = \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{p_{\mathbf{i_j}}} \cdot A_{:,\mathbf{i_j}} B_{\mathbf{i_j},:}\right] = \frac{1}{t} \sum_{j=1}^{t} \sum_{k=1}^{n} p_k \cdot \frac{1}{p_k} \cdot A_{:,k} B_{k,:} = \frac{1}{t} \sum_{j=1}^{t} AB = AB$$

Canonical AMM Algorithm

Approximate Matrix Multiplication (AMM):

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i \geq 0$ and $\sum_{[n]} p_i = 1$.
- Select $\mathbf{i}_1, \ldots, \mathbf{i}_t \in [n]$ independently, according to the distribution $\Pr[\mathbf{i}_j = k] = p_k$.
- Let $\bar{\mathbf{C}} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{p_{\mathbf{i}_j}} \cdot A_{:,\mathbf{i}_j} B_{\mathbf{i}_j,:}$.

Claim 1: $\mathbb{E}[\bar{\mathbf{C}}] = AB$

$$\mathbb{E}[\bar{\mathbf{C}}] = \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{p_{\mathbf{i}_j}} \cdot A_{:,\mathbf{i}_j} B_{\mathbf{i}_j,:}\right] = \frac{1}{t} \sum_{j=1}^{t} \sum_{k=1}^{n} p_k \cdot \frac{1}{p_k} \cdot A_{:,k} B_{k,:} = \frac{1}{t} \sum_{j=1}^{t} AB = AB$$

Weighting by $\frac{1}{p_{\mathbf{i}_j}}$ keeps the expectation correct. Key idea behind **importance sampling** based methods.

Optimal Sampling Probabilities

Claim 2: $\mathbb{E}[\|AB - \overline{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n} \frac{\|A_{:,m}\|_2^2 \cdot \|B_{m,:}\|_2^2}{p_m}$.

Good exercsie – uses linearity of variance. I may ask you to prove it on the next problem set.

$$\sum_{ij}\left[(AB)_{ij} - \bar{C}_{ij}\right]^2 = \sum_{ij} \operatorname{Var}(\bar{C}_{ij})$$

Optimal Sampling Probabilities

Claim 2: $\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n} \frac{\|A_{:,m}\|_2^2 \cdot \|B_{m,:}\|_2^2}{p_m}$. $= V$ $\quad p_m \propto \|A_{:,m}\|_2^2 \, \|B_{m,:}\|_2^2$

Good exercsie – uses linearity of variance. I may ask you to prove it on the next problem set. $\sqrt{ab} = \sqrt{a} \cdot \sqrt{b}$ $\quad \sum p_i = 1$

Question: How should we set $p_1, \ldots, p_n$ to minimize this error?

$$\frac{\partial V}{\partial p_m} = \frac{-\|A_{:,m}\|_2^2 \cdot \|B_{m,:}\|_2^2}{p_m^2}$$

want:

$$\frac{\partial V}{p_1} = \frac{\partial V}{p_2} = \cdots = \frac{\partial V}{p_n}$$

so how should I set p_i?

$$p_m = \frac{\|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}{\sum_{j=1}^{n} \|A_{:,j}\|_2 \|B_{j,:}\|_2}$$

$$\frac{\partial V}{p_i} > \frac{\partial V}{p_j}$$

$$p_i' = p_i - \varepsilon$$

$$p_j' = p_j + \varepsilon$$

Optimal Sampling Probabilities

Claim 2: $\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n} \frac{\|A_{:,m}\|_2^{\cancel{2}} \cdot \|B_{m,:}\|_2^{\cancel{2}}}{p_m}$.

Good exercsie – uses linearity of variance. I may ask you to prove it on the next problem set.

Question: How should we set $p_1, \ldots, p_n$ to minimize this error?

Set $p_m = \frac{\|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}{\sum_{k=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2}$, giving:

$$\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n} \|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2 \cdot \left(\sum_{k=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2\right)$$

Claim 2: $\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n} \frac{\|A_{:,m}\|_2^2 \cdot \|B_{m,:}\|_2^2}{p_m}$.

$\mathrm{Var}(\bar{C}) \leq \mathbb{E}\bar{C}^2$

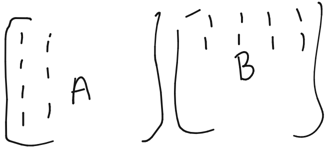
Good exercsie – uses linearity of variance. I may ask you to prove it on the next problem set.

Question: How should we set $p_1, \ldots, p_n$ to minimize this error?

Set $p_m = \frac{\|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}{\sum_{k=1}^{n}\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2}$, giving:

$$\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n}\|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2 \cdot \left(\sum_{k=1}^{n}\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2\right)$$

$$= \frac{1}{t}\left(\sum_{m=1}^{n}\|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2\right)^2$$



$\bar{C} = \frac{1}{t}\sum_{j=1}^{t}\frac{1}{p_{i_j}} A_{:i_j} B_{i_j:}$

$[\ldots] \circ \frac{1}{t}\sum_{j=1}^{t}\frac{1}{p_{i_j}} = n\,[\ldots] = AB$

Optimal Sampling Probabilities

Claim 2: $\mathbb{E}[\|AB - \overline{C}\|_F^2] \leq \frac{1}{t} \sum_{m=1}^{n} \frac{\|A_{:,m}\|_2^2 \cdot \|B_{m,:}\|_2^2}{p_m}$.

Good exercsie – uses linearity of variance. I may ask you to prove it on the next problem set.

Question: How should we set $p_1, \ldots, p_n$ to minimize this error?

Set $p_m = \frac{\|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}{\sum_{k=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2}$, giving:

$$\mathbb{E}[\|AB - \overline{C}\|_F^2] \leq \frac{1}{t} \sum_{m=1}^{n} \|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2 \cdot \left(\sum_{k=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2 \right)$$

$$= \frac{1}{t} \left(\sum_{m=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2 \right)^2$$

By the Cauchy-Schwarz inequality,

$$\sum_{m=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2 \leq \sqrt{\underbrace{\sum_{m=1}^{n} \|A_{:,k}\|_2^2}_{\|A\|_F^2}} \cdot \sqrt{\underbrace{\sum_{m=1}^{n} \|B_{k,:}\|_2^2}_{\|B\|_F^2}} = \|A\|_F \cdot \|B\|_F$$

Optimal Sampling Probabilities

Claim 2: $\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n} \frac{\|A_{:,m}\|_2^2 \cdot \|B_{m,:}\|_2^2}{p_m}$.

Good exercsie – uses linearity of variance. I may ask you to prove it on the next problem set.

Question: How should we set $p_1, \ldots, p_n$ to minimize this error?

Set $p_m = \frac{\|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}{\sum_{k=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2}$, giving:

$$\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{1}{t}\sum_{m=1}^{n} \|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2 \cdot \left(\sum_{k=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2\right)$$
$$= \frac{1}{t}\left(\sum_{m=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2\right)^2$$

By the Cauchy-Schwarz inequality,

$\sum_{m=1}^{n} \|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2 \leq \sqrt{\sum_{m=1}^{n} \|A_{:,k}\|_2^2} \cdot \sqrt{\sum_{m=1}^{n} \|B_{k,:}\|_2^2} = \|A\|_F \cdot \|B\|_F$

Overall: $\mathbb{E}[\|AB - \bar{C}\|_F^2] \leq \frac{\|A\|_F^2 \cdot \|B\|_F^2}{t}$.

Approximate Matrix Multiplication Variance

So far: With optimal sampling probabilities, approximate matrix multiplication satisfies $\mathbb{E}[\|AB - \overline{C}\|_F^2] \leq \frac{\|A\|_F^2 \cdot \|B\|_F^2}{t}$.

Approximate Matrix Multiplication Variance

So far: With optimal sampling probabilities, approximate matrix multiplication satisfies $\mathbb{E}[\|AB - \overline{\mathsf{C}}\|_F^2] \leq \frac{\|A\|_F^2 \cdot \|B\|_F^2}{t}$.

- Setting $t = \frac{1}{\epsilon^2 \sqrt{\delta}}$, by Markov's inequality:

$$\Pr[\|AB - \overline{\mathsf{C}}\|_F \geq \epsilon \cdot \|A\|_F \cdot \|B\|_F] \leq \delta.$$

$$= \Pr\left[\|AB - \overline{C}\|_F^2 \geq \epsilon^2 \|A\|_F^2 \cdot \|B\|_F^2\right] \leq \frac{\frac{1}{t}\cancel{\|A\|_F^2}\,\cancel{\|B\|_F^2}}{\epsilon^2 \cancel{\|A\|_F^2}\,\cancel{\|B\|_F^2}}$$

$$= \delta$$

Approximate Matrix Multiplication Variance

So far: With optimal sampling probabilities, approximate matrix multiplication satisfies $\mathbb{E}[\|AB - \overline{C}\|_F^2] \leq \frac{\|A\|_F^2 \cdot \|B\|_F^2}{t}$.

- Setting $t = \frac{1}{\epsilon^2 \sqrt{\delta}}$, by Markov's inequality:

$$\Pr[\|AB - \overline{C}\|_F \geq \epsilon \cdot \|A\|_F \cdot \|B\|_F] \leq \delta.$$

- **Note:** Its not so obvious how to improve the dependence on δ here, but it can be done using more advanced concentration inequalities. – Mahoney's book

$t = \frac{\log(1/\delta)}{\epsilon^2}$

AMM Upshot

Upshot: Sampling $t = O(1/\epsilon^2)$ columns/rows of A, B with probabilities proportional to $\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2$ yields, with good probability, an approximation $\overline{\mathsf{C}}$ with

$$\|AB - \overline{\mathsf{C}}\|_F \leq \epsilon \cdot \|A\|_F \cdot \|B\|_F.$$

AMM Upshot

Upshot: Sampling $t = O(1/\epsilon^2)$ columns/rows of A, B with probabilities proportional to $\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2$ yields, with good probability, an approximation $\overline{\mathsf{C}}$ with

$$\|AB - \overline{\mathsf{C}}\|_F \leq \epsilon \cdot \|A\|_F \cdot \|B\|_F.$$

- Probabilities take $O(n^2)$ time to compute. After sampling, $\overline{\mathsf{C}}$ takes $O(t \cdot n^2)$ time to compute.

AMM Upshot

Upshot: Sampling $t = O(1/\epsilon^2)$ columns/rows of A, B with probabilities proportional to $\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2$ yields, with good probability, an approximation $\overline{\mathsf{C}}$ with

$$\|AB - \overline{\mathsf{C}}\|_F \leq \epsilon \cdot \|A\|_F \cdot \|B\|_F.$$

- Probabilities take $O(n^2)$ time to compute. After sampling, $\overline{\mathsf{C}}$ takes $O(t \cdot n^2)$ time to compute.
- Can derive related bounds when probabilities are just approximate – i.e. $p_k \geq \beta \cdot \frac{\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2}{\sum_{m=1}^{n} \|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}$ for some $\beta > 0$.

($\beta = .5$

AMM Upshot

Upshot: Sampling $t = O(1/\epsilon^2)$ columns/rows of A, B with probabilities proportional to $\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2$ yields, with good probability, an approximation $\overline{\mathrm{C}}$ with

$$\|AB - \overline{\mathrm{C}}\|_F \leq \epsilon \cdot \|A\|_F \cdot \|B\|_F.$$

- Probabilities take $O(n^2)$ time to compute. After sampling, $\overline{\mathrm{C}}$ takes $O(t \cdot n^2)$ time to compute.
- Can derive related bounds when probabilities are just approximate – i.e. $p_k \geq \beta \cdot \frac{\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2}{\sum_{m=1}^n \|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}$ for some $\beta > 0$.
- Can also give bounds on $\|AB - \overline{\mathrm{C}}\|_2$, but analysis is much more complex. Will see tools in the coming weeks that let us do this.

matrix concentration

AMM Upshot

Upshot: Sampling $t = O(1/\epsilon^2)$ columns/rows of A, B with probabilities proportional to $\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2$ yields, with good probability, an approximation $\overline{\mathsf{C}}$ with

$$\|AB - \overline{\mathsf{C}}\|_F \leq \epsilon \cdot \|A\|_F \cdot \|B\|_F.$$

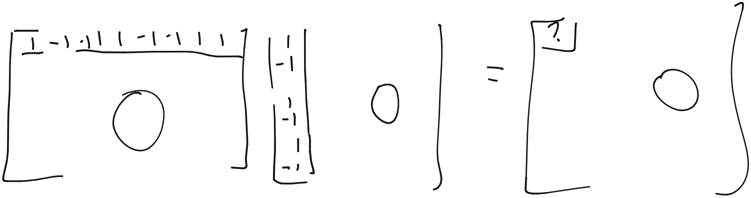
- Probabilities take $O(n^2)$ time to compute. After sampling, $\overline{\mathsf{C}}$ takes $O(t \cdot n^2)$ time to compute.
- Can derive related bounds when probabilities are just approximate – i.e. $p_k \geq \beta \cdot \frac{\|A_{:,k}\|_2 \cdot \|B_{k,:}\|_2}{\sum_{m=1}^{n} \|A_{:,m}\|_2 \cdot \|B_{m,:}\|_2}$ for some $\beta > 0$.
- Can also give bounds on $\|AB - \overline{\mathsf{C}}\|_2$, but analysis is much more complex. Will see tools in the coming weeks that let us do this.
- A classic example of using weighted importance sampling to decrease variance and in turn, sample complexity.

AMM Upshot

$\|AB\|_F \leq \|A\|_F \cdot \|B\|_F$

Think-Pair-Share 1: Ideally we would have *relative error*, $\|AB - \overline{\mathbf{C}}\|_F \leq \underbrace{\epsilon\|AB\|_F}_{=0}$. Could we get this via a tighter analysis or better sampling distribution?

$< \epsilon \|A\|_F \cdot \|B\|_F$



to achieve error $\epsilon \|AB\|_F$ I need to know if $AB = 0$ or not.

Randomized Low-Rank approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^T A\|_F = \min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F.$$

projection onto span Z

LSA

PCA

⋮

one of the main ways of approximating matrices

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|\overset{n \times d}{A} - \overset{n \times d}{ZZ^T A}\|_F = \min_{Z : Z^T Z = I} \|A - ZZ^T A\|_F.$$

$Z = \begin{bmatrix} | & & | \\ z_1 & \cdots & z_k \\ | & & | \end{bmatrix}$

$Z \in \mathbb{R}^{n \times k}$ s.t. $Z^T Z = I$

Why is $\mathrm{rank}(ZZ^T A) \leq k$?

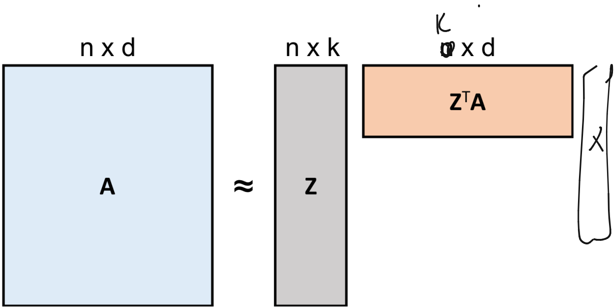
↳ columns are spanned by columns of Z of which there are k

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^TA\|_F = \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F.$$

Why is $\text{rank}(ZZ^TA) \leq k$?



Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - \underline{ZZ^T}A\|_F = \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F.$$

Why is $\text{rank}(ZZ^TA) \leq k$?

$$\min_{B:\text{rank}(B)\leq k} \|A - B\|_F$$

Why does it suffice to consider low-rank approximations of this form?

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n\times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n,d)$ we would like to find $Z \in \mathbb{R}^{n\times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^TA\|_F = \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F.$$

Why is $\operatorname{rank}(ZZ^TA) \leq k$?

Why does it suffice to consider low-rank approximations of this form? For any B with $\operatorname{rank}(B) = k$, let $Z \in \mathbb{R}^{n\times k}$ be an orthonormal basis for B's column span. Then $\|A - ZZ^TA\|_F \leq \|A - B\|_F$. So

$$\min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F = \min_{B:\operatorname{rank} B=k} \|A - B\|_F.$$

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^T A\|_F = \min_{Z : Z^T Z = I} \|A - ZZ^T A\|_F.$$

Why is $\operatorname{rank}(ZZ^T A) \leq k$?

Why does it suffice to consider low-rank approximations of this form? For any B with $\operatorname{rank}(B) = k$, let $Z \in \mathbb{R}^{n \times k}$ be an orthonormal basis for B's column span. Then $\|A - ZZ^T A\|_F \leq \|A - B\|_F$. So

$$\min_{Z : Z^T Z = I} \|A - ZZ^T A\|_F = \min_{B : \operatorname{rank} B = k} \|A - B\|_F.$$

How would one compute the optimal basis Z?

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^T A\|_F = \min_{Z:Z^TZ=I} \|A - ZZ^T A\|_F.$$

Why is $\text{rank}(ZZ^T A) \leq k$?

Why does it suffice to consider low-rank approximations of this form? For any B with $\text{rank}(B) = k$, let $Z \in \mathbb{R}^{n \times k}$ be an orthonormal basis for B's column span. Then $\|A - ZZ^T A\|_F \leq \|A - B\|_F$. So

eigenvectors of AA^T

$$\min_{Z:Z^TZ=I} \|A - ZZ^T A\|_F = \min_{B:\text{rank } B=k} \|A - B\|_F.$$

How would one compute the optimal basis Z? Compute the top k left singular vectors of A, which requires $O(nd^2)$ time, or $O(ndk)$ time for a high accuracy approximation with an iterative method.

$O(nd + nk^2)$

Sampling Based Algorithm

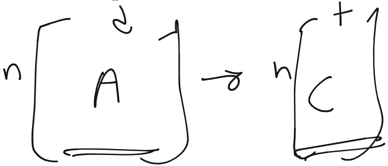
We will analysis a simple non-uniform sampling based algorithm for low-rank approximation, that gives a near optimal solution in $O(nd + nk^2)$ time.

Sampling Based Algorithm

We will analysis a simple non-uniform sampling based algorithm for low-rank approximation, that gives a near optimal solution in $O(nd + nk^2)$ time.

Linear Time Low-Rank Approximation:

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i = \frac{\|A_{:,i}\|_2^2}{\|A\|_F^2}$.
- Select $\mathbf{i}_1, \ldots, \mathbf{i}_t \in [n]$ independently, according to the distribution $\Pr[\mathbf{i}_j = k] = p_k$ for sample size $t \geq k$.
- Let $\mathbf{C} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{\sqrt{p_{\mathbf{i}_j}}} \cdot A_{:,\mathbf{i}_j}$.
- Let $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ consist of the top k left singular vectors of $\mathbf{C}$.



Sampling Based Algorithm

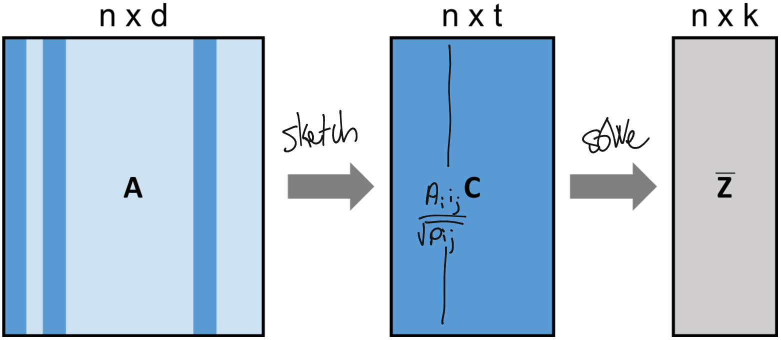
We will analysis a simple non-uniform sampling based algorithm for low-rank approximation, that gives a near optimal solution in $O(nd + nk^2)$ time.

Linear Time Low-Rank Approximation:

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i = \frac{\|A_{:,i}\|_2^2}{\|A\|_F^2}$.
- Select $\mathbf{i}_1, \ldots, \mathbf{i}_t \in [n]$ independently, according to the distribution $\Pr[\mathbf{i}_j = k] = p_k$ for sample size $t \geq k$.
- Let $\mathbf{C} = \cancel{\frac{1}{t}\sum_{j=1}^{t} \frac{1}{\sqrt{p_{\mathbf{i}_j}}} \cdot A_{:,\mathbf{i}_j}}$. $\frac{1}{t}\left[\frac{A_{i_1}}{\sqrt{p_{i_1}}}, \frac{A_{i_2}}{\sqrt{p_{i_2}}}, \ldots, \frac{A_{i_t}}{\sqrt{p_{i_t}}}\right]$
- Let $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ consist of the top k left singular vectors of $\mathbf{C}$.

Will use that $\mathbf{C}\mathbf{C}^T$ is a good approximation to the matrix product AA^T.

Sampling Based Algorithm



"sketch and solve"

Sampling Based Algorithm Approximation Bound

Theorem

The linear time low-rank approximation algorithm run with $t = \frac{k}{\epsilon^2 \cdot \sqrt{\delta}}$ samples outputs $\bar{\mathsf{Z}} \in \mathbb{R}^{n \times k}$ satisfying with probability at least $1 - \delta$:

$$\|A - \bar{\mathsf{Z}}\bar{\mathsf{Z}}^T A\|_F^2 \leq \min_{Z : Z^T Z = I} \|A - ZZ^T A\|_F^2 + 2\epsilon \|A\|_F^2.$$

Sampling Based Algorithm Approximation Bound

Theorem

The linear time low-rank approximation algorithm run with $t = \frac{k}{\epsilon^2 \cdot \sqrt{\delta}}$ samples outputs $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ satisfying with probability at least $1 - \delta$:

$$\|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T A\|_F^2 \leq \min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F^2 + 2\epsilon \|A\|_F^2.$$

Key Idea: By the approximate matrix multiplication result applied to the matrix product AA^T, with probability $\geq 1 - \delta$,

$$\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}} \cdot \|A\|_F \cdot \|A^T\|_F = \frac{\epsilon}{\sqrt{k}} \|A\|_F^2.$$

$$\mathbf{C}\mathbf{C}^T = \sum_{j=1}^{t} C_{:,j} \cdot C_{:,j}^T = \frac{1}{t} \sum_{j=1}^{t} \frac{A_{:,i_j}}{\sqrt{p_{i_j}}} \frac{A_{:,i_j}^T}{\sqrt{p_{i_j}}} = \frac{1}{t} \sum_j \frac{1}{p_{i_j}} A_{:,i_j} A_{:,i_j}^T = \text{AMM}$$

Sampling Based Algorithm Approximation Bound

Theorem

The linear time low-rank approximation algorithm run with $t = \frac{k}{\epsilon^2 \cdot \sqrt{\delta}}$ samples outputs $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ satisfying with probability at least $1 - \delta$:

$$\|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T A\|_F^2 \leq \min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F^2 + 2\epsilon \|A\|_F^2.$$

Key Idea: By the approximate matrix multiplication result applied to the matrix product AA^T, with probability $\geq 1 - \delta$,

$$\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}} \cdot \|A\|_F \cdot \|A^T\|_F = \frac{\epsilon}{\sqrt{k}} \|A\|_F^2.$$

Since $\mathbf{C}\mathbf{C}^T$ is close to AA^T, the top eigenvectors of these matrices (i.e. the top left singular vectors of A and $\mathbf{C}$ will not be too different.) So $\bar{\mathbf{Z}}$ can be used in place of the top left singular vectors of A to give a near optimal approximation.

Formal Analysis

Let $Z_* \in \mathbb{R}^{n \times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T \mathbf{C}\|_F^2$.

Formal Analysis

Let $Z_* \in \mathbb{R}^{n \times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T \mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n \times k}$, and any matrix B,

$$\|B - ZZ^T B\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^T BB^T Z).$$

$$\|A\|_F^2 = tr(AA^T)$$

$$\|B - ZZ^TB\|_F^2 \; tr\left((B - ZZ^TB)(B - ZZ^TB)^T\right)$$

$$= tr\left(BB^T - ZZ^TBB^T - BB^TZZ^T + ZZ^TBB^TZZ^T\right)$$

$$\downarrow$$

$$tr(BB^T) - tr(Z^TBB^TZ) \cancel{- tr(Z^TBBZ)} \cancel{+ tr(Z^TBB^TZ)}$$

$$tr(BB^T) - tr(Z^TBB^TZ)$$

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^T B\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^T BB^T Z).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \operatorname{arg\,min} \|A - ZZ^TA\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \operatorname{arg\,min} \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^TB\|_F^2 = \operatorname{tr}(BB^T) - \operatorname{tr}(Z^TBB^TZ).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\operatorname{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Proof from claims:

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^TA\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^TB\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^TBB^TZ).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Proof from claims:

$$\|\mathbf{C} - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T\mathbf{C}\|_F^2 \leq \|\mathbf{C} - Z_*Z_*^T\mathbf{C}\|_F^2 \implies \mathrm{tr}(\bar{\mathbf{Z}}^T\mathbf{C}\mathbf{C}^T\bar{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^T\mathbf{C}\mathbf{C}^TZ_*)$$

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^T B\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^T BB^T Z).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Proof from claims:

$$\|\mathbf{C} - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T\mathbf{C}\|_F^2 \leq \|\mathbf{C} - Z_*Z_*^T\mathbf{C}\|_F^2 \implies \mathrm{tr}(\bar{\mathbf{Z}}^T\mathbf{C}\mathbf{C}^T\bar{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^T\mathbf{C}\mathbf{C}^T Z_*)$$
$$\implies \mathrm{tr}(\bar{\mathbf{Z}}^T AA^T\bar{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^T AA^T Z_*) - 2\epsilon\|A\|_F^2$$

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^TA\|_F^2$. Similarly, $\overline{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^TB\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^TBB^TZ).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Proof from claims:

$$\begin{aligned}
\|\mathbf{C} - \overline{\mathbf{Z}}\overline{\mathbf{Z}}^T\mathbf{C}\|_F^2 \leq \|\mathbf{C} - Z_*Z_*^T\mathbf{C}\|_F^2 &\implies \mathrm{tr}(\overline{\mathbf{Z}}^T\mathbf{C}\mathbf{C}^T\overline{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^T\mathbf{C}\mathbf{C}^TZ_*) \\
&\implies \mathrm{tr}(\overline{\mathbf{Z}}^TAA^T\overline{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^TAA^TZ_*) - 2\epsilon\|A\|_F^2 \\
&\implies \|A - \overline{\mathbf{Z}}\overline{\mathbf{Z}}^TA\|_F^2 \leq \underbrace{\|A - Z_*Z_*^TA\|_F^2}_{OPT} + 2\epsilon\|A\|_F^2.
\end{aligned}$$