COMPSCI 614: Randomized Algorithms with Applications to Data Science

Prof. Cameron Musco
University of Massachusetts Amherst. Spring 2024.
Lecture 10
• 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 $\ell_2$ 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 $A, B \in \mathbb{R}^{n \times n}$ we would like to compute $C = AB$. Requires $n^\omega$ 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.
Approximate Matrix Multiplication (AMM):

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

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

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

Weighting by $\frac{1}{p_{i_j}}$ keeps the expectation correct. Key idea behind importance sampling based methods.
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 exercise – 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:

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

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

By the Cauchy-Schwarz inequality,

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

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

- 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: It's not so obvious how to improve the dependence on $\delta$ here, but it can be done using more advanced concentration inequalities.
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 \( \bar{C} \) with

\[
\|AB - \bar{C}\|_F \leq \epsilon \cdot \|A\|_F \cdot \|B\|_F.
\]

- Probabilities take \( O(n^2) \) time to compute. After sampling, \( \bar{C} \) 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 \( \|AB - \bar{C}\|_2 \), 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.
Think-Pair-Share 1: Ideally we would have relative error, $\|AB - \bar{C}\|_F \leq \epsilon \|AB\|_F$. Could we get this via a tighter analysis or better sampling distribution?
Randomized Low-Rank approximation
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 \( \text{rank}(ZZ^T A) \leq k \)?

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

\[
\min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F = \min_{\text{rank}(B) = k} \|A - B\|_F.
\]

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.
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 $i_1, \ldots, i_t \in [n]$ independently, according to the distribution $Pr[i_j = k] = p_k$ for sample size $t \geq k$.
- Let $C = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{\sqrt{p_{i_j}}} \cdot A_{:,i_j}$.
- Let $\bar{Z} \in \mathbb{R}^{n \times k}$ consist of the top $k$ left singular vectors of $C$.

Will use that $CC^T$ is a good approximation to the matrix product $AA^T$. 
Sampling Based Algorithm

A \rightarrow C \rightarrow \bar{Z}
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 \( \overline{Z} \in \mathbb{R}^{n \times k} \) satisfying with probability at least \( 1 - \delta \):

\[
\|A - \overline{Z}\overline{Z}^T A\|_F \leq \min_{Z:Z^T Z = I} \|A - ZZ^T A\|_F^2 + 2\epsilon \|A\|_F^2.
\]

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

\[
\|AA^T - CC^T\|_F \leq \frac{\epsilon}{\sqrt{k}} \cdot \|A\|_F \cdot \|A^T\|_F = \frac{\epsilon}{\sqrt{k}} \|A\|_F^2.
\]

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

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

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

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

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

**Proof from claims:**

$$\|C - ZZ^T C\|_F^2 \leq \|C - Z_\ast Z_\ast^T C\|_F^2 \implies \text{tr}(\bar{Z}^T CC^T \bar{Z}) \geq \text{tr}(Z_\ast^T CC^T Z_\ast)$$

$$\implies \text{tr}(\bar{Z}^T AA^T \bar{Z}) \geq \text{tr}(Z_\ast^T AA^T Z_\ast) - 2\epsilon \|A\|_F^2$$

$$\implies \|A - \bar{Z}Z^T A\|_F^2 \leq \|A - Z_\ast Z_\ast^T A\|_F^2 + 2\epsilon \|A\|_F^2.$$