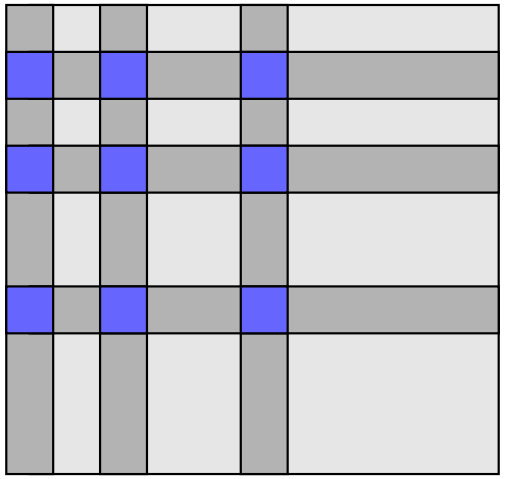
Low Rank Approximation
Lecture 5

Daniel Kressner
Chair for Numerical Algorithms and HPC
Institute of Mathematics, EPFL
daniel.kressner@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Randomized column/row sampling

Aim: Obtain rank-r approximation from randomly selected rows and columns of A.



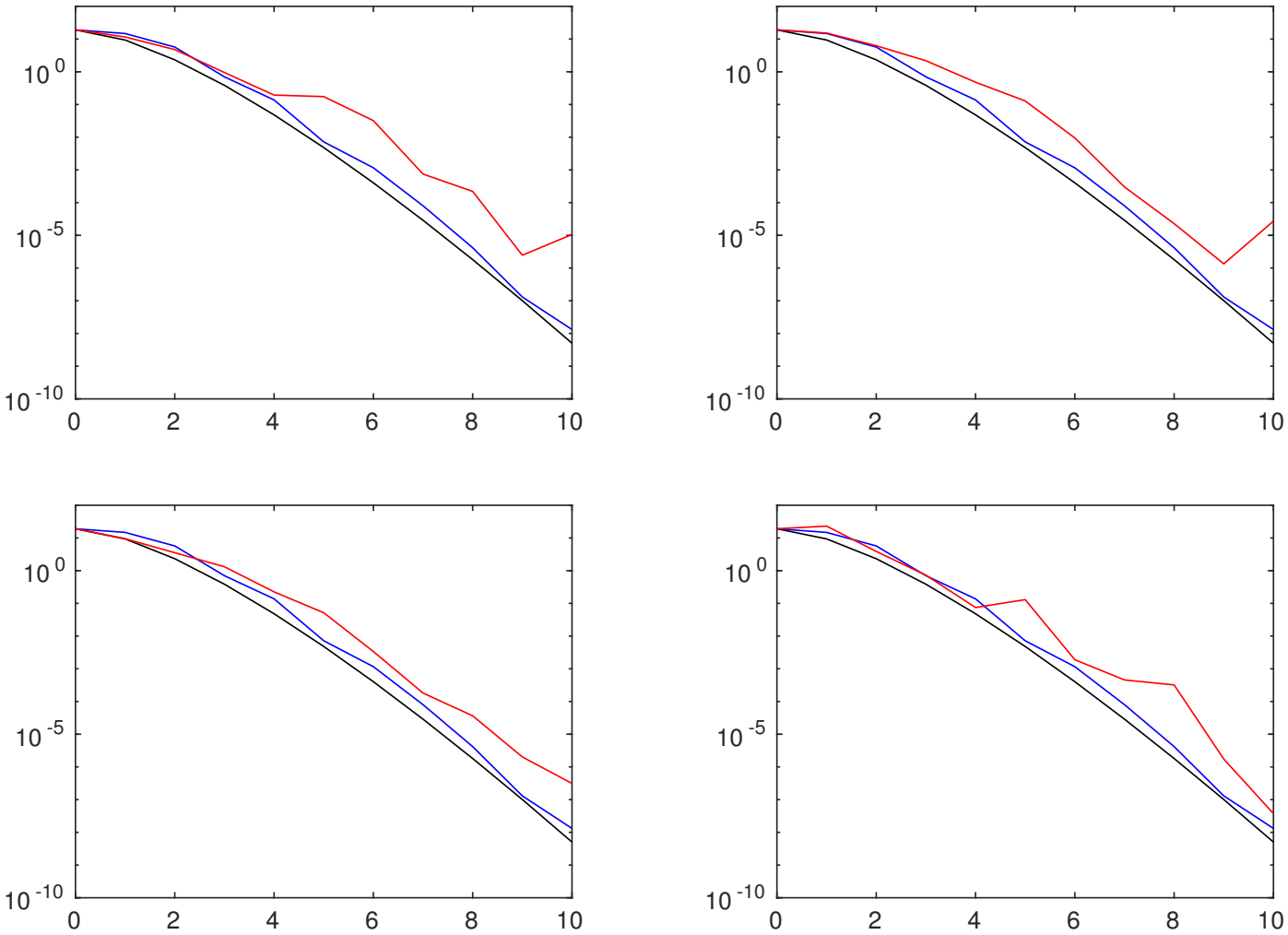
Popular sampling strategies:

- Uniform sampling.
- Sampling based on row/column norms.
- Sampling based on more complicated quantities.

Preliminaries on randomized sampling

Exponential function example from Lecture 4 (Slide 14).

Comparison between best approximation, greedy approximation, approximation obtained by randomly selecting rows.



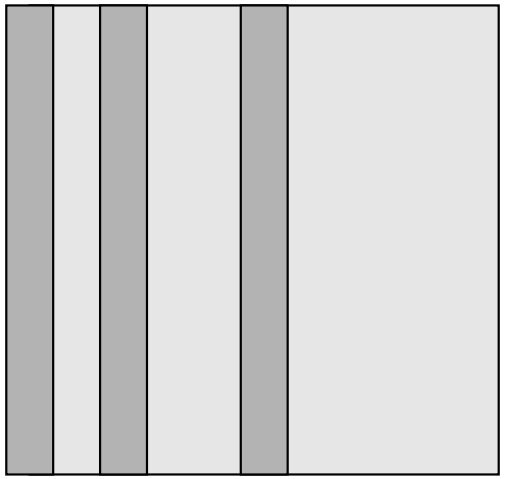
Preliminaries on randomized sampling

A simple way to fool uniformly random row selection:

$$U = \begin{pmatrix} 0_{(n-r)\times r} \\ I_r \end{pmatrix}$$

for n very large and $r \ll n$.

Column sampling



Basic algorithm aiming at rank-r approximation:

1. Sample (and possibly rescale) $k > r$ columns of A $\rightsquigarrow m \times k$ matrix C.
2. Compute SVD $C = U\Sigma V^T$ and set $Q = U_r \in \mathbb{R}^{m \times r}$.
3. Return low-rank approximation $QQ^T A$.

- Can be combined with streaming algorithm [Liberty'2007] to limit memory/cost of working with C.
- Quality of approximation crucially depends on sampling strategy.

Column sampling

Lemma

For any matrix $C \in \mathbb{R}^{m \times r}$, let Q be the matrix computed above. Then

$$\|A - QQ^T A\|_2^2 \leq \sigma_{r+1}(A)^2 + 2\|AA^T - CC^T\|_2.$$

Proof. We have

$$\begin{aligned} & (A - QQ^T A)(A - QQ^T A)^T \\ = \; & (I - QQ^T)CC^T(I - QQ^T) + (I - QQ^T)(AA^T - CC^T)(I - QQ^T) \end{aligned}$$

Hence,

$$\begin{aligned} \|A - QQ^T A\|_2^2 &= \lambda_{\max}\big((A - QQ^T A)(A - QQ^T A)^T\big) \\ &\leq \lambda_{\max}\big((I - QQ^T)CC^T(I - QQ^T)\big) + \|AA^T - CC^T\|_2 \\ &= \sigma_{r+1}(C)^2 + \|AA^T - CC^T\|_2. \end{aligned}$$

The proof is completed by applying Weyl's inequality:

$$\sigma_{r+1}(C)^2 = \lambda_{r+1}(CC^T) \leq \lambda_{r+1}(AA^T) + \|AA^T - CC^T\|_2.$$

Random column sampling

Using the lemma, the goal now becomes to approximate the matrix product AA^T using column samples of A.

Notation:

$$A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}, \quad C = \begin{bmatrix} c_1 & \cdots & c_k \end{bmatrix}$$

General sampling method:

Input: $A \in \mathbb{R}^{m \times n}$, probabilities $p_1, \ldots, p_n \neq 0$, integer k.
Output: $C \in \mathbb{R}^{m \times k}$ containing selected columns of A.

```
for t = 1, ..., k do
    Pick j_t ∈ {1, ..., n} with P[j_t = ℓ] = p_ℓ, ℓ = 1, ..., n,
    independently and with replacement.
    Set c_t = a_{j_t} / sqrt(k p_{j_t}).
end for
```

Random column sampling

Lemma

For the matrix C returned by algorithm, it holds that

$$\mathbb{E}[CC^T] = AA^T, \quad \mathrm{Var}[(CC^T)_{ij}] = \frac{1}{k}\sum_{\ell=1}^{n} \frac{a_{i\ell}^2 a_{j\ell}^2}{p_\ell} - \frac{1}{k}(AA^T)_{ij}^2.$$

Proof. For fixed i, j, consider $X_t = \left(c_t c_t^T\right)_{ij} = \frac{1}{kp_{j_t}} a_{i,j_t} a_{j,j_t}$, for which

$$\mathbb{E}[X_t] = \sum_{\ell=1}^{n} p_\ell \frac{1}{kp_\ell} a_{i,\ell} a_{j,\ell} = \frac{1}{k}(AA^T)_{ij}.$$

Analogously,

$$\mathrm{Var}(X_t) = \mathbb{E}[(X_t - \mathbb{E}[X_t])^2] = \mathbb{E}[X_t^2] - \mathbb{E}[X_t]^2 = \frac{1}{k^2}\sum_{\ell=1}^{n} \frac{a_{i\ell}^2 a_{j\ell}^2}{p_\ell} - \frac{1}{k^2}(AA^T)_{ij}^2.$$

Because of independence, it follows $\mathbb{E}[\sum_t X_t] = k \cdot \mathbb{E}[X_t] = (AA^T)_{ij}$, and analogously for variance.

Random column sampling

As a consequence of the lemma,

$$\begin{aligned}
\mathbb{E}[\|AA^T - CC^T\|_F^2] &= \sum_{ij} \mathbb{E}[(AA^T - CC^T)_{ij}^2] \\
&= \sum_{ij} \mathrm{Var}[(CC^T)_{ij}] \\
&= \frac{1}{k} \sum_{ij} \Big(\sum_{\ell=1}^{n} \frac{a_{i\ell}^2 a_{j\ell}^2}{p_\ell} - \frac{1}{k}(AA^T)_{ij}^2 \Big) \\
&= \frac{1}{k} \left[\sum_{\ell=1}^{n} \frac{1}{p_\ell} \|a_\ell\|_2^4 - \|AA^T\|_F^2 \right].
\end{aligned}$$

Lemma

The choice $p_\ell = \|a_\ell\|_2^2 / \|A\|_F^2$ *minimizes* $\mathbb{E}[\|AA^T - CC^T\|_F^2]$ *and yields*

$$\mathbb{E}[\|AA^T - CC^T\|_F^2] = \frac{1}{k} \left[\|A\|_F^4 - \|AA^T\|_F^2 \right]$$

Proof. Established by showing that this choice of p_ℓ satisfies first-order conditions of constrained optimization problem.

Random column sampling

Norm based sampling:

Input: $A \in \mathbb{R}^{m \times n}$, integer k.

Output: $C \in \mathbb{R}^{m \times k}$ containing selected columns of A.

Set $p_\ell = \|a_\ell\|_2^2 / \|A\|_F^2$ for $\ell = 1, \ldots, n$.
for $t = 1, \ldots, k$ **do**
    Pick $j_t \in \{1, \ldots, n\}$ with $\mathbb{P}[j_t = \ell] = p_\ell$, $\ell = 1, \ldots, n$, independently and with replacement.
    Set $c_t = a_{j_t} / \sqrt{k p_{j_t}}$.
end for
Compute SVD $C = U \Sigma V^T$ and set $Q = U_r \in \mathbb{R}^{m \times r}$.
Return low-rank approximation $QQ^T A$.

Random column sampling

Lemma

For the matrix C returned by algorithm, it holds with probability $1-\delta$ that

$$\|AA^T - CC^T\|_F \leq \frac{\eta}{\sqrt{k}}\|A\|_F,$$

where $\eta = 1 + \sqrt{8 \cdot \log(1/\delta)}$.

Proof. Aim at applying Azuma-Hoeffding inequality. Define

$$F(i_1, i_2, \ldots, i_k) = \|AA^T - CC^T\|_F,$$

with $C = \begin{bmatrix} a_{i_1} & \cdots & a_{i_k} \end{bmatrix}$. Quantify the effect of varying an index (w.l.o.g. the first one) on F:

$$\begin{aligned}
& |F(i_1, i_2, \ldots, i_k) - F(i_1', i_2, \ldots, i_k)| \\
= \quad & \left| \|AA^T - CC^T\|_F - \|AA^T - C'C'^T\|_F \right| \\
\leq \quad & \|CC^T - C'C'^T\|_F \leq \frac{1}{kp_{i_1}}\|a_{i_1}\|_2^2 + \frac{1}{kp_{i_1'}}\|a_{i_1'}\|_2^2 \\
\leq \quad & \frac{2}{k}\|A\|_F^2 := \Delta.
\end{aligned}$$

Random column sampling

This implies that Doob martingales $g_\ell = \mathbb{E}[f(i_1, \ldots, i_k)|i_1, \ldots, i_\ell]$ for $1 \leq \ell \leq k$ satisfy

$$|g_{\ell+1} - g_\ell| \leq \Delta.$$

Note that $g_k = \mathbf{E}[\|AA^T - CC^T\|_F]$. By lemma and Jensen's inequality we know that $g_k \leq \|A\|_F^2/\sqrt{k}$. Applying Azuma-Hoeffding inequality to g_k yields

$$P\big[\|AA^T - CC^T\|_F \geq \|A\|_F^2/\sqrt{k} + \gamma\big] \leq \exp(-\gamma^2/2k\Delta^2) =: \delta.$$

Setting $\gamma = \sqrt{8 \cdot \log(1/\delta)}$ completes the proof.

Random column sampling

Theorem (Drineas/Kannan/Mahoney'2006)

For the matrix Q returned by the algorithm above it holds that

$$\mathbf{E}\left[\|A - QQ^T A\|_2^2\right] \leq \sigma_{r+1}^2(A) + \varepsilon\|A\|_F^2 \text{ for } k \geq 4/\varepsilon^2.$$

With probability at least $1 - \delta$,

$$\|A - QQ^T A\|_2^2 \leq \sigma_{r+1}^2(A) + \varepsilon\|A\|_F^2 \text{ for } k \geq 4(1 + \sqrt{8 \cdot \log(1/\delta)})^2/\varepsilon^2.$$

Proof. Follows from combining very first lemma with last two lemmas.

Remarks:

- Dependence of k on ε pretty bad. Unlikely to achieve something significantly better without assuming further properties of A (e.g., incoherence of singular vectors) with sampling based on row norms only.
- Simple "counter example":

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} e_1 & \frac{1}{\sqrt{n}} e_1 & \cdots & \frac{1}{\sqrt{n}} e_1 & \frac{1}{\sqrt{n}} e_2 \end{pmatrix} \in \mathbb{R}^{n \times (n+1)}.$$

Random column sampling

[Drineas/Mahoney/Muthukrishnan'2007]: Let V_k contain k dominant right singular vectors of A. Setting

$$p_\ell = \| V_k(\ell, :)\|_2^2 / k, \qquad \ell = 1, \ldots, n$$

and sampling $\mathcal{O}(k^2(\log 1/\delta)/\varepsilon^2)$ columns[1] yields

$$\|A - QQ^T A\|_F \leq (1 + \varepsilon)\|A - \mathcal{T}_k(A)\|_F$$

with probability $1 - \delta$.

Relative error bound!

CUR decomposition can be obtained by applying ideas to rows and columns (yielding R and C, respectively) and choosing U appropriately.

[1] There are variants that improve this to $\mathcal{O}(k \log k \log(1/\delta)/\varepsilon^2)$.