5.5 Matrix sketches

Wednesday, February 5, 2020 2:42 PM

Matrix sampling & sketches

Let $A \in \mathbb{R}^{m \times n}$
$B \in \mathbb{R}^{n \times p}$. We want to approximate AB.
Naive approach $O(mnp)$ time.

Let $A(:,k)$ be the kth column of A
$B(k,:)$ be the kth row of B.

Then $$AB = \sum_{k=1}^{n} A(:,k)\, B(k,:)$$ (outer product)

Let's try to sample AB by taking components with prob. p_k.

i.e. Let $z = k$ w.p. p_k for $k \in [n]$, a random variable

Define $X = \frac{1}{p_z} A(:,z)\, B(z,:)$, a matrix r.v.

Then the entry-wise expectation

$$\mathbb{E} X = \sum_{k=1}^{n} \mathbb{P}(z=k) \frac{1}{p_k} A(:,k) B(k,:) = \sum_{k=1}^{n} A(:,k) B(k,:) = AB.$$

(this cancellation is the reason we scale by $\frac{1}{p_k}$)

Define $\mathrm{Var}(X) = \mathbb{E}(\|AB - X\|_F^2)$, the entry-wise variance.

Then $$\mathrm{Var}(X) = \sum_{i=1}^{m} \sum_{j=1}^{p} \mathrm{Var}(x_{ij}) = \sum_{ij} \mathbb{E}(x_{ij}^2) - \mathbb{E}(x_{ij})^2 = \left(\sum_{ij} \sum_{k=1}^{n} p_k \cdot \frac{1}{p_k^2} a_{ik}^2 b_{kj}^2 \right) - \underbrace{\|AB\|_F^2}_{\text{doesn't matter for minimizing } p_k}.$$

We want to minimize variance, by choosing appropriate p_k.

$$\sum_{ij} \sum_{k} p_k \cdot \frac{1}{p_k^2} a_{ik}^2 b_{kj}^2 = \sum_k \frac{1}{p_k} \left(\sum_i a_{ik}^2 \right) \left(\sum_j b_{kj}^2 \right) = \sum_k \frac{1}{p_k} |A(:,k)|^2\, |B(k,:)|^2$$

Note that for any $c_k \ge 0$, $\sum_k \frac{c_k}{p_k}$ is minimized by $p_k \sim \sqrt{c_k}$.

(proof by taking derivatives $p_1 + \cdots + p_n = 1$

$$\frac{\partial f}{\partial p_1} = \frac{\partial}{\partial p_1} \left(\frac{c_1}{(1-(p_2+\cdots+p_n))^2} + \frac{c_k}{p_k} \right)$$

$$\frac{\partial f}{\partial p_k} = \frac{\partial}{\partial p_k}\left(\frac{c_1}{(1-(p_2+\dots+p_n))^2} + \frac{c_k}{p_k} \right)$$

$$= \frac{c_1}{(1-(p_2+\dots+p_n))^2} - \frac{c_n}{p_k^2} = 0$$

$$\frac{p_k}{1-(p_2+\dots+p_n)} = \sqrt{\frac{c_k}{c_1}}$$

$$p_k = \sqrt{c_k} \cdot \frac{1-(p_2+\dots+p_n)}{\sqrt{c_1}} \quad \forall\, k \neq 1. \quad)$$

Thus, we want to pick $p_k \sim |A(:,k)|\,|B(k,:)|$.

Note, when $B=A^T$, $p_k \sim |A(:,k)|^2$, the squared length of the columns

Even if $B \neq A^T$, we can still use that as an easy to analyze upper bound.

Use $p_k = \dfrac{|A(:,k)|^2}{\|A\|_F^2}$,

$$\Rightarrow \mathbb{E}\left(\|AB - X\|_F^2\right) = \mathrm{Var}(X) \leq \|A\|_F^2 \sum_k |B(k,:)|^2 = \|A\|_F^2 \|B\|_F^2 .$$

Repeat with s independent trials, getting $X_1, \dots, X_s$.

Then $\mathrm{Var}(\bar{X}) = \frac{1}{s}\sum_{i=1}^{s} X_i = \frac{1}{s}\mathrm{Var}(X) \leq \frac{1}{s}\|A\|_F^2 \|B\|_F^2$.

$$\begin{bmatrix} A \\ m \times n \end{bmatrix} \begin{bmatrix} B \\ n \times p \end{bmatrix} \approx \begin{bmatrix} \text{Sampled scaled columns of } A \\ m \times s \end{bmatrix} \begin{bmatrix} \text{Corresponding scaled rows of } B \\ s \times p \end{bmatrix}$$

Note $\quad \frac{1}{s}\sum_{i=1}^{s} X_i = \frac{1}{s}\left(\frac{A(:,k_1)B(k_1,:)}{p_{k_1}} + \dots + \frac{A(:,k_s)B(k_s,:)}{p_{k_s}} \right)$

$= CR$, where

C has columns $\dfrac{A(:, k_i)}{\sqrt{s p_{k_i}}}$. Note $\mathbb{E}(CC^T) = AA^T$

R has rows $\dfrac{B(k_i, :)}{\sqrt{s p_{k_i}}}$ $\mathbb{E}(R^T R) = B^T B$.

Theorem 6.5 Suppose $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$. The product AB can be estimated by CR as given above, and the error is bounded by

$$\mathbb{E}\left(\|AB - CR\|_F^2\right) \le \frac{\|A\|_F^2 \|B\|_F^2}{s}.$$

Thus, to ensure $\mathbb{E}\left(\|AB - CR\|_F^2\right) \le \varepsilon^2 \|A\|_F^2 \|B\|_F^2$, it suffices to make $s \ge \frac{1}{\varepsilon^2}$. If $\varepsilon = \Omega(1)$, $s = O(1)$, so CR can be computed in $O(mp)$ time.

When is this a good estimate?

Consider case $B = A^T$ for simplicity.

Then if $A = I$, $\|II^T\|_F^2 = n$, but $\dfrac{\|I\|_F^2 \|I\|_F^2}{s} = \dfrac{n^2}{s}$,

so need $s > n$ for bound to be useful.

Trivial estimate of 0-matrix gives error $\|AA^T\|_F^2$, so need our bound to be at least as good.

Let's analyze using SVD.

When is SVD approximation good? When the top p singular values take up a large constant fraction of the Frobenius mass.

Suppose $\exists\, 0 < c < 1$ and a small integer p s.t. for a matrix A,

$$\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 \ge c\left(\sigma_1^2 + \cdots + \sigma_r^2\right), \text{ where } r = \operatorname{rank}(A).$$

Note $\|AA^T\|_F^2 = \sum_{t=1}^{r} \sigma_t^4$, and $\|A\|_F^2 = \sum_{t=1}^{r} \sigma_t^2$. $\quad (\sigma_1^2 + \cdots + \sigma_r^2) \le \dfrac{\sigma_1^2 + \cdots + \sigma_p^2}{c}$

Then $\mathbb{E}\left(\|AA^T - CR\|_F^2\right) \le \dfrac{\|A\|_F^2 \|A^T\|_F^2}{s}$

In order for the approximation to be good, we want

In order for the approximation to be good, we want

$$\frac{\|A\|_F^2 \|A^T\|_F^2}{s} \le \|AA^T\|_F^2$$

$$(\sigma_1^2 + \cdots + \sigma_r^2)^2 \le s(\sigma_1^4 + \cdots + \sigma_r^4)$$

$$\Rightarrow \quad s \ge \frac{(\sigma_1^2 + \cdots + \sigma_r^2)^2}{(\sigma_1^4 + \cdots + \sigma_r^4)}$$

$$\max_{\sigma_1, \ldots, \sigma_r} \frac{(\sigma_1^2 + \cdots + \sigma_r^2)}{(\sigma_1^4 + \cdots + \sigma_r^4)} = r.$$ So not a good bound.

But $$\frac{(\sigma_1^2 + \cdots + \sigma_r^2)^2}{(\sigma_1^4 + \cdots + \sigma_r^4)} \le \frac{(\sigma_1^2 + \cdots + \sigma_p^2)}{c^2(\sigma_1^4 + \cdots + \sigma_r^4)} \le \frac{(\sigma_1^2 + \cdots + \sigma_p^2)}{c^2(\sigma_1^4 + \cdots + \sigma_p^4)} \le \frac{p}{c^2}$$

Thus if $s \ge \frac{p}{c^2}$, then the approximation is better than the 0 matrix

Thus, we don't need to sample that many columns if the mass is contained in a few singular vectors.

Intuition is that we are sampling according to squared column length, so we will probably pick out columns with large singular components.

Matrix sketch

Let's motivate with a biological example.

An RNA microarray chip measures expression levels of different snippets of mRNA in a cell, often corresponding to gene expression

For a sample, get vector $\vec{x} = (x_1, \ldots, x_n)$, where each $x_i \in \mathbb{R}^+$ is the expression of a particular sequence corresponding to an ASO-probe.

Suppose we want to find a low-dimensional representation that is predictive of a condition. The SVD is a natural way to do this, but the singular vectors might be mixes of the original sample vectors, which is hard to interpret. Also, the SVD takes $O(n^3)$ computation, and destroys sparsity. Another idea is to sketch the matrix by sampling columns.

Thm 6.9 Let $A \in \mathbb{R}^{m\times n}$ and $r, s \in \mathbb{Z}^+$.
Let $C \in \mathbb{R}^{m\times s}$ of s columns of A picked via length squared sampling.

Let $C \in \mathbb{R}^{m \times s}$ of s columns of A picked via length squared sampling.
Let $R \in \mathbb{R}^{r \times n}$ of r rows of A " " " " "

Then we can find from C & R an $s \times r$ matrix U s.t.

$$\mathbb{E}\left(\|A - CUR\|_2^2\right) \leq \|A\|_F^2 \left(\frac{2}{\sqrt{r}} + \frac{2r}{s}\right).$$

If we fix s, we minimize error with $r = s^{2/3}$.
Choose $s = \frac{1}{\varepsilon^3}$ and $r = \frac{1}{\varepsilon^2}$. Then $\mathbb{E}\left(\|A - CUR\|_2^2\right) = O(\varepsilon)\|A\|_F^2$.

Looks like the bound you got for SVD $\|A - A_k\| \leq \frac{\|A\|_F^2}{\sqrt{k}}$, $k = \frac{1}{\varepsilon^2}$.

So when the first several singular values are large, sampling works well because columns are near a low-dimensional subspace.

$$\underbrace{\begin{bmatrix} A \\ \\ n \times m \end{bmatrix}}_{} = \underbrace{\begin{bmatrix} \text{Sample} \\ \text{columns} \\ \\ n \times s \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} \text{Multiplier} \end{bmatrix}_{s \times r}}_{U} \underbrace{\begin{bmatrix} \text{Sample rows} \\ r \times m \end{bmatrix}}_{R}$$

Lemma 6.6 If RR^T is invertible, then $P = \underbrace{R^T (RR^T)^{-1}}_{\text{Moore-Penrose pseudo-inverse } R^+} R$

orthogonal projection operator
- (i) $P\vec{x} = \vec{x}$ for every vector $\vec{x} = R^T \vec{y}$ (if x in row space of R)
- (ii) If $\vec{x} \perp R^T \vec{y}$ for all $\vec{y}$, then $P\vec{x} = 0$.

If RR^T is not invertible, let $\operatorname{rank}(RR^+) = r$ and $RR^T = \sum_{t=1}^{r} \sigma_t \vec{u}_t \vec{v}_t^T$ the $SVD(RR^T)$.
Then $P = \underbrace{R^T \left(\sum_{t=1}^{r} \frac{1}{\sigma_t^2} \vec{u}_t \vec{v}_t^T \right)}_{R^+ \in \mathbb{R}^{m \times r}} \underbrace{R}_{\in \mathbb{R}^{r \times m}}$ satisfies those properties.

Prop 6.7 $A \approx AP$ and $\mathbb{E}\left(\|A - AP\|_2^2\right) \leq \frac{1}{\sqrt{r}} \|A\|_F^2$.

So sample s columns of A to form C, and choose corresponding
[illegible] s rows of P to form a $s \times m$ matrix.

So sample s columns of A to form C, and choose corresponding sampled s rows of P to form a $s \times m$ matrix, which we can decompose into s rows of R^+, multiplied by R.

We will not take the time to prove this, but the matrix sketch largely follows from sampled matrix multiplication on AP.