23. Matrix sketching

Tuesday, November 2, 2021 9:18 PM

Recall: SVDs are one way to approximate a matrix (MAT1850)

Today: Sampling for matrix products.

Let $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$. We want to approximate $AB \in \mathbb{R}^{m \times p}$

Naive approach: Matrix multiplication $O(mnp)$ time. (maybe slightly faster with Strassen, etc.)

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 prod.)

Let's sample AB by taking components with prob. $p_k > 0$. ← arbitrary for now.

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

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

Then the entry-wise expectation

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

cancellation is reason we scaled by $\frac{1}{p_k}$.

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

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

Want to choose p_k to minimize variance.

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

Euclidean norm

Lemma: Let $f(p_1, \dots, p_K) = \sum_{k=1}^{n} \frac{c_k}{p_k}$, where $c_k \geq 0$ and $p_k > 0$.

Then subject to the condition $p_1 + \cdots + p_n = 1$, the minimum of f is achieved by $p_k \sim \sqrt{c_k}$.

[illegible] $\frac{c_1}{\ \ } + \sum^{n} c_k$ (an unconstrained optimization) is achieved by $p_k \sim \sqrt{c_k}$.

proof. WLOG say $c_1 > 0$ and remove p_1 by $f(p_2, \dots, p_n) = \dfrac{c_1}{1-(p_2+\dots+p_n)} + \sum_{k=2}^{n} \dfrac{c_k}{p_k}$. (an unconstrained optimization)

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

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

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

Note: When $B = A^T$, $p_k \sim |A(:,k)|^2$ ← squared length of cols

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

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

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

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

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

$$\begin{bmatrix} \\ A \\ \\ m \times n \\ \\ \end{bmatrix} \begin{bmatrix} \\ B \\ \\ n \times p \\ \\ \end{bmatrix} \approx \underset{m \times s}{\overset{C}{\begin{bmatrix} \text{Scaled} \\ \text{sampled} \\ \text{cols of} \\ A \end{bmatrix}}} \underset{s \times p}{\overset{R}{\begin{bmatrix} \text{Corresponding} \\ \text{scaled rows of} \\ B \end{bmatrix}}}$$

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

C has cols $\dfrac{A(:,k_i)}{\sqrt{s\,p_{k_i}}}$

R has rows $\dfrac{B(k_i,:)}{\sqrt{s\,p_{k_i}}}$

Note: $\mathbb{E}(CC^T) = AA^T$

$\mathbb{E}(R^TR) = B^TB$.

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

$$\mathbb{E}(\|AB - CR\|_F^2) \leq \frac{\|A\|_F^2\|B\|_F^2}{s}$$

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

To ensure $\mathbb{E}\left(\|AB-CR\|_F^2\right) \le \varepsilon^2 \|A\|_F^2 \|B\|_F^2$, suffices to choose $s \ge \frac{1}{\varepsilon^2}$.

$\Rightarrow$ CR can be computed in $O\big(\underbrace{\tfrac{1}{\varepsilon^2} mp}_{\text{compute } CR} + \underbrace{mn}_{\text{sample cols of } A} + \underbrace{np}_{\text{sample rows of } B}\big)$ time.

When is this a good estimate?

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

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

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

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

Analysis via SVD

When is SVD approximation good? $(A = U\Sigma V^T)$

When top p singular values take up a large constant fraction of Frobenius mass.

Recall: $\|A\|_F^2 = \sigma_1^2 + \dots + \sigma_r^2$, where $r = \mathrm{rank}(A)$ and $\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$.

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

$$\sigma_1^2 + \dots + \sigma_p^2 \ge c\left(\sigma_1^2 + \dots + \sigma_r^2\right).$$

Note $\|AA^T\|_F^2 = \sum_{t=1}^r \sigma_t^4$ and $\|A\|_F^2 = \sum_{t=1}^r \sigma_t^2$.

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

For approx. to be good, we want $\frac{\|A\|_F^2 \|A^T\|_F^2}{s} \le \|AA^T\|_F^2$

$$\Longleftrightarrow \left(\sigma_1^2 + \dots + \sigma_r^2\right)^2 \le s\left(\sigma_1^4 + \dots + \sigma_r^4\right)$$

$$\Longleftrightarrow s \ge \frac{\left(\sigma_1^2 + \dots + \sigma_r^2\right)^2}{\left(\sigma_1^4 + \dots + \sigma_r^4\right)}.$$

$$\max_{\sigma_1, \dots, \sigma_r} \frac{\left(\sigma_1^2 + \dots + \sigma_r^2\right)^2}{\left(\sigma_1^4 + \dots + \sigma_r^4\right)} = r$$

because max is when $\sigma_1 = \sigma_2 = \dots = \sigma_r$, $\frac{r^2}{r} = r$.

So not good bound.

But
$$\frac{\left(\sigma_1^2 + \dots + \sigma_r^2\right)^2}{\left(\sigma_1^4 + \dots + \sigma_r^4\right)} \le \frac{\left(\sigma_1^2 + \dots + \sigma_p^2\right)^2}{c^2\left(\sigma_1^4 + \dots + \sigma_r^4\right)} \le \frac{\left(\sigma_1^2 + \dots + \sigma_p^2\right)^2}{c^2\left(\sigma_1^4 + \dots + \sigma_p^4\right)} \le \frac{p}{c^2}.$$

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

Thus, we don't need to sample that many columns if the mass is concentrated

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

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

Matrix sketch

If $A \in \mathbb{R}^{m \times n}$ is a data matrix $\begin{array}{cc} & \text{feature} \\ \text{Sample} & \begin{bmatrix} & & \end{bmatrix} \end{array}$ (e.g. expression of mRNA in a cell),

we may want a low-dimensional representation.

SVD $A = U \Sigma V^T$ is a natural solution, but takes $O(mn \cdot \min(n,m)) \simeq O(n^3)$

└ also, destroys sparsity and interpretability because singular vectors are mixes of samples

What about sketching by sampling cols?

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 cols of A picked via length square sampling

let $R \in \mathbb{R}^{r \times n}$ of r rows " " " " " " "

Then we can find from C & R a matrix $U \in \mathbb{R}^{s \times r}$ s.t.

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

If we fix s, we can minimize error with $r = s^{2/3}$

Choose $s = \frac{1}{\varepsilon^3}$ and $r = \frac{1}{\varepsilon^2}$. Then $\mathbb{E}\left(\|A - CUR\|_2^2\right) = O(\varepsilon)\|A\|_F^2$.

Recall: $\|A - A_k\|_2^2 = \sigma_{k+1}^2 \le \frac{\sigma_1^2 + \dots + \sigma_k^2}{k} \le \frac{\|A\|_F^2}{k}$. (similar form)

best rank-k approx via SVD

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

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

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

Orthogonal projection operator

$$\begin{cases} \text{(i)} \quad P\vec{x} = \vec{x} \text{ for every vector } \vec{x} = R^T\vec{y} \\ \text{(ii)} \quad \text{If } \vec{x} \perp R^T y \ \forall \vec{y}, \text{ then } P\vec{x} = 0 \end{cases}$$

(if x is row space of R)

If RR^T not invertible, let $\operatorname{rank}(RR^T) = r$ and $RR^T = \sum_{t=1}^{r} \sigma_t \vec{u}_t \vec{v}_t^T$ the SVD.

Then $P = \underbrace{R^T \left(\sum_{t=1}^{r} \frac{1}{\sigma_t^2} \vec{u}_t \vec{v}_t^T \right)}_{R^+ \in \mathbb{R}^{m \times r}} R$ satisfies these properties.

$\in \mathbb{R}^{r \times m}$

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

(proof omitted)

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

(Use Thm 6.5)

Matrix sketch follows from sampled matrix multiplication on AP.