Lecture 22 — November 16, 2017

Guest Lecture by Cameron Musco *Scribe: Jacob Klegar*

1 Low-Rank Approximation and Clustering Via Sketching

We approach solving the problems of low-rank approximation and clustering via a generalization of subspace embeddings called "Projection-Cost Preserving Sketches", or PCP's for short. The material for this lecture comes from a paper by Cohen et al. [1].

1.1 Low-Rank Approximation

Definition 1 (Low-Rank Approximation). *Given $A \in \mathbb{R}^{n\times d}$, find $\underset{\text{rank } k \text{ matrix } B}{\operatorname{argmin}} \|A - B\|_F^2$. Or equivalently, find $\underset{\text{rank } k \text{ projection matrix } P}{\operatorname{argmin}} \|A - PA\|_F^2$.*

The output P^* of Low-Rank Approximation is a projection onto A's top k singular vectors, i.e.,

$$P^*A = U_k U_k^T A = A_k$$

where U_k is the matrix with the top k singular vectors of A.

This takes $O(nd^2)$ time, and approximate iterative methods take $\tilde{O}(nnz(A)k)$ time, where $nnz(A)$ is the number of nonzero values of A. We want to do better.

Definition 2 (PCP). *$\tilde{A} \in \mathbb{R}^{n\times m}$ is an (ε, k)-PCP for A if $\forall$ rank k projections P:*

$$(1-\varepsilon)\|A - PA\|_F^2 \le \|\tilde{A} - P\tilde{A}\|_F^2 \le (1+\varepsilon)\|A - PA\|_F^2$$

Ideally we have $m \ll d$.

Now assuming we have $\tilde{A}$ an (ε, k)-PCP for A, let $\tilde{P}^* = \underset{\text{rank } k \text{ projections } P}{\operatorname{argmin}} \|\tilde{A} - P\tilde{A}\|_F^2$. Then

$$\begin{aligned}
\|A - \tilde{P}^*A\|_F^2 &\le \frac{1}{1-\varepsilon}\|\tilde{A} - \tilde{P}^*\tilde{A}\|_F^2 \\
&\le \frac{1}{1-\varepsilon}\|\tilde{A} - P^*\tilde{A}\|_F^2 \\
&\le \frac{1+\varepsilon}{1-\varepsilon}\|A - P^*A\|_F^2
\end{aligned}$$

For small ε we have $\frac{1+\varepsilon}{1-\varepsilon} = 1 + O(\varepsilon)$.

The run time using PCP is now $O(nm^2)$.

Theorem 3. *We can compute an (ε, k)-PCP for A in $nnz(A) + \tilde{O}(nk^2/\varepsilon^2)$ time with $m \approx k/\varepsilon^2$.*

This would make the total run time $O(nnz(A)) + \tilde{O}(nk^2/\varepsilon^4)$. Before proving this theorem, we show another application of PCPs.

1.2 Constrained Low-Rank Approximation Problem

Definition 4 (Constrained Low-Rank Approximation Problem)**.** *Let $T \subseteq$ all rank k projection matrices. Then find $\underset{P \in T}{\operatorname{argmin}} \|A - PA\|_F^2$.*

Claim 5. *If $\tilde{A}$ is an (ε, k)-PCP for A, and $\tilde{P} \leq \gamma \min_{P \in T} \|\tilde{A} - P\tilde{A}\|_F^2$, then*

$$\|A - \tilde{P}A\|_F^2 \leq (1 + O(\varepsilon))\gamma \min_{P \in T} \|A - PA\|_F^2$$

This follows from the same reasoning as before.

Definition 6 (k-means clustering)**.** *Given $a_1, ..., a_n \in \mathbb{R}^d$, which we can represent as the rows of $A \in \mathbb{R}^{n \times d}$,*

$$\min_{\text{partitions into } k \text{ sets } C = \{C_1, \ldots, C_k\}} \sum_{i=1}^{k} \sum_{j \in C_i} \|a_j - \mu(C_i)\|_2^2$$

where $\mu(C_i) = \frac{1}{|C_i|} \sum_{j \in C_i} a_j$ is the centroid.

We now show that k-means is constrained low-rank approximation. Let

$$f(C, A) = \sum_{j \in C_i} \|a_j - \mu(C_i)\|_2^2.$$

Then we will show

$$f(C, A) = \|A - P_C A\|_F^2$$

for some rank k projection matrix P_C.

We have $P_C = Z_C^T Z_C$, where Z_C is a cluster indicator matrix, i.e., $Z_C \in \mathbb{R}^{k \times n}$ and

$$(Z_C)_{ij} = \begin{cases} \frac{1}{\sqrt{|C_i|}} & \text{if } a_j \in C_i \\ 0 & \text{otherwise.} \end{cases}$$

Note Z_C is an orthogonal matrix and $Z_C Z_C^T = I$, which implies P_C is a projection.

So we get $\|A - Z_C^T Z_C A\|_F^2$.

After showing these applications, we now show how to get a PCP sketch.

2 Projection-Cost Preserving Sketches

2.1 Subspace Embeddings

Definition 7 (Subspace Embedding)**.** *Given $A \in \mathbb{R}^{n\times d}$, S is an ε-subspace embedding if $\forall x \in \mathbb{R}^n, \|x^T AS\|_2^2 \in (1\pm\varepsilon)\|x^T A\|_2^2$.*

Observation 8. *If S is a subspace embedding, it is an (ε, k)-PCP for any k.*

Let $Y \in \mathbb{R}^{n\times n}, Y = I - P$.

Then PCP is equivalent to

$$\|Y\tilde{A}\|_F^2 = \sum_{i=1}^n \|y_i^T \tilde{A}\|_2^2 \in (1\pm\varepsilon)\|YA\|_F^2.$$

Then set $\tilde{A} = AS$. We get

$$\|YAS\|_F^2 = \sum_{i=1}^n \|y_i^T AS\|_2^2 \in (1\pm\varepsilon)\sum_{i=1}^n \|y_i^T A\|_2^2 = (1\pm\varepsilon)\|YA\|_F^2$$

S works but is too expensive. Typically $S \in \mathbb{R}^{d\times m}$ where $m = \Theta(d/\varepsilon^2)$. We want $m = \Theta(k/\varepsilon^2)$.

2.2 Smaller m

Theorem 9. *$S \in \mathbb{R}^{d\times m}$, where $S_{ij} = \pm\frac{1}{\sqrt{m}}$ independently at random. Then if $m = O(k\log(1/\delta)\varepsilon^{-2})$, then $\tilde{A} = AS$ is (ε, k)-PCP with probability $1-\delta$.*

That is, letting $Y = I - P$ for any rank k projection P, we want the PCP guarantee:

$$|\|YA\|_F^2 - \|YAS\|_F^2| \le \varepsilon\|YA\|_F^2$$

Write $A = A_k + A_{\overline{k}}$, where A_k is A projected onto its top k singular vectors (what we care about) and $A_{\overline{k}}$ is the rest (noise).

Then our expression becomes

$$|\|Y(A_k + A_{\overline{k}})\|_F^2 - \|Y(A_k + A_{\overline{k}})S\|_F^2|$$

Now using the fact that $\|M\|_F^2 = \operatorname{tr}(MM^T)$:

$$\operatorname{tr}(YA_kA_k^TY) + \operatorname{tr}(YA_{\overline{k}}A_{\overline{k}}^TY) + 2\operatorname{tr}(YA_kA_{\overline{k}}^TY) - \operatorname{tr}(YA_kSS^TA_k^TY) - \operatorname{tr}(YA_{\overline{k}}SS^TA_{\overline{k}}^TY) - 2\operatorname{tr}(YA_kSS^TA_{\overline{k}}^TY)$$

Note that $\operatorname{tr}(YA_kA_{\overline{k}}^TY) = 0$ since $A_k, A_{\overline{k}}$ are orthogonal.

2.3 Head Terms (Subspace Embedding)

We show

$$|\operatorname{tr}(YA_kA_k^TY) - \operatorname{tr}(YA_kSS^TA_k^TY)| \le \varepsilon\|YA\|_F^2$$

Note the left hand side is

$$|\|YA_k\|_F^2 - \|YA_kS\|_F^2| \le \varepsilon\|YA\|_F^2$$

A_k is rank k, so since $m \approx k/\varepsilon^2$, S is an ε-subspace embedding for A_k, i.e., $\forall x, \|x^TA_kS\|_2^2 \in (1 \pm \varepsilon)\|x^TA_k\|_2^2$.

2.4 Tail Term (Approximate Matrix Multiplication)

We bound

$$|\operatorname{tr}(YA_{\overline{k}}A_{\overline{k}}^TY) - \operatorname{tr}(YA_{\overline{k}}SS^TA_{\overline{k}}^TY)|$$

Recall $Y = I - P$.

$$\|(I-P)A_{\overline{k}}\|_F^2 = \|A_{\overline{k}}\|_F^2 - \|PA_{\overline{k}}\|_F^2$$

So we get

$$\|A_{\overline{k}}\|_F^2 - \|PA_{\overline{k}}\|_F^2 - \|A_{\overline{k}}S\|_F^2 + \|PA_{\overline{k}}S\|_F^2$$

If $m > \log(1/\delta)\varepsilon^{-2}$, then $|\|A_{\overline{k}}\|_F^2 - \|A_{\overline{k}}S\|_F^2| \le \varepsilon\|A_{\overline{k}}\|_F^2 \le \varepsilon\|(I-P)A\|_F^2$ for any P.

Now

$$\begin{aligned}&|\|PA_{\overline{k}}\|_F^2 - \|PA_{\overline{k}}S\|_F^2| \\ &= |\operatorname{tr}(P[A_{\overline{k}}A_{\overline{k}}^T - A_{\overline{k}}SS^TA_{\overline{k}}^T]P)|\end{aligned}$$

Let $M = A_{\overline{k}}A_{\overline{k}}^T - A_{\overline{k}}SS^TA_{\overline{k}}^T$ and let $\lambda_1 > ... > \lambda_k > 0$ be its first k eigenvalues. Then we get

$$
\begin{aligned}
&= \sum_{i=1}^{k} \lambda_i(M)| \\
&\leq \sum_{i=1}^{k} |\lambda_i(M)| \\
&\leq \sqrt{k}\sqrt{\sum_{i=1}^{k} \lambda_i^2(M)} \\
&\leq \sqrt{k}\|PMP\|_F \\
&\leq \sqrt{k}\|M\|_F
\end{aligned}
$$

Recall (Approximate Matrix Multiplication) that for any C, D,

$$
\|CD - CSS^T D\| \leq \frac{1}{\sqrt{m}}\|C\|_F\|D\|_F
$$

where m is the number of columns in S.

Here, we take $C = D = A_{\overline{k}}$. Then we get

$$
\begin{aligned}
&\leq \sqrt{k}\frac{\varepsilon}{\sqrt{k}}\|A_{\overline{k}}\|_F^2 \\
&\leq \varepsilon\|A_{\overline{k}}\|_F^2 \\
&\leq \varepsilon\|(I-P)A\|_F^2
\end{aligned}
$$

for any P.

2.5 Cross Term

We show

$$
|\operatorname{tr}(YA_kSS^TA_{\overline{k}}^TY)|
$$

is small. Set $C = AA^T$ and let C^+ be the pseudoinverse of C. Then this becomes

$$
\begin{aligned}
&|\operatorname{tr}(YCC^+A_kSS^TA_{\overline{k}}^TY)| \\
&= |\operatorname{tr}(Y^2CC^+A_kSS^TA_{\overline{k}}^T)| \\
&= |\operatorname{tr}(YCC^+A_kSS^TA_{\overline{k}}^T)| \\
&= |\operatorname{tr}((YCC^{+/2})(C^{+/2}A_kSS^TA_{\overline{k}}^T)| \\
&\leq \sqrt{\operatorname{tr}(YCC^{+/2}C^{+/2}CY)}\sqrt{\operatorname{tr}(A_{\overline{k}}SS^TA_k^TC^{+/2}C^{+/2}A_kSS^TA_{\overline{k}}^T)}
\end{aligned}
$$

where the last inequality comes from Cauchy-Schwarz. Then the first part can be bounded by

$$\begin{aligned}
&\sqrt{\operatorname{tr}(YCC^{+/2}C^{+/2}CY)} \\
&= \sqrt{\operatorname{tr}(YCY)} \\
&= \sqrt{\operatorname{tr}(YAA^TY)} \\
&= \|YA\|_F
\end{aligned}$$

Using SVD, we get $A_k = V_k\Sigma_k V_k^T$, so

$$\begin{aligned}
&\sqrt{\operatorname{tr}(A_{\bar{k}}SS^TA_k^TC^{+/2}C^{+/2}A_kSS^TA_{\bar{k}}^T)} \\
&= \sqrt{\operatorname{tr}(A_{\bar{k}}SS^TV_k\Sigma_kU_k^TU\Sigma^{-2}U^TU_k\Sigma_kV_k^TSS^TA_{\bar{k}}^T)} \\
&= \sqrt{\operatorname{tr}(A_{\bar{k}}SS^TV_kV_k^TSS^TA_{\bar{k}}^T)} \\
&= \|A_{\bar{k}}SS^TV_k\|_F \\
&\le \frac{1}{\sqrt{m}}\|A_{\bar{k}}\|_F\|V_k\|_F \\
&\le \frac{\varepsilon}{\sqrt{k}}\|(I-P)A\|_F\sqrt{k} \\
&= \varepsilon\|(I-P)A\|_F
\end{aligned}$$

again for any P.

References

[1] Michael Cohen, Sam Elder, Cameron Musco, Christopher Musco, Madalina Persu. Dimensionality Reduction for k-Means Clustering and Low Rank Approximation. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 163–172. ACM, 2015.