CSE 392: Matrix and Tensor Algorithms for Data Spring 2024

Lecture 12 — Feb 26

Instructor: Shashanka Ubaru *Scribe: Jeil Oh*

Sketching and Iterative Methods

Overview of Sketching Methods

Sketching methods are crucial for processing data that is too large to fit in memory, especially in streaming settings. These methods enable the creation of low-rank approximations of matrices efficiently.

Sketch Size:

- Dense matrices: For a rank-k approximation, the Gaussian method requires $O(k)$ space.
- SRFT/SRHT: For the same, these methods need $O(k \log(k/\epsilon))$ space, with ϵ being the precision parameter.
- Sparse matrices: CountSketch method requires $O(k^2)$ space.

Overview of Iterative Methods

Iterative methods, known for their improved numerical results, require multiple passes over the data. They are pivotal across numerous fields such as system solvers, optimization, control systems, PDE solvers, scientific computing, NLP, and various industries including oil refineries, automotive modeling, electronics, and major tech firms like Google and Twitter.

Computing the partial SVD to obtain the top k singular vectors/values.

- Subspace iteration or block power method.
- Krylov subspace method.

Subspace iteration methods

Power Method

- Let us start with $k = 1$ (finding the top singular vector/value).
- Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, with SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, find a vector $\mathbf{z} \approx \mathbf{v}_1$.

Power Method Algorithm

```
1. Choose a random vector $\mathbf{z}_0$. E.g., $\mathbf{z}_0 \sim \mathcal{N}(0,1)$.
2. $\mathbf{z}_0 = \mathbf{z}_0 / \|\mathbf{z}_0\|_2$
3. For $l = 1, \ldots, q$
    • $\mathbf{z}_l = \mathbf{A}^T(\mathbf{A}\mathbf{z}_{l-1})$
    • $\mathbf{z}_l = \mathbf{z}_l / \|\mathbf{z}_l\|_2$
4. Return $\mathbf{z}_q$
```

Runtime $= O(\mathrm{nnz}(A) \cdot q) = O(\mathrm{nnz}(A) \cdot \frac{\log d}{\gamma})$

Convergence

Theorem 1 (Power Method Convergence)**.** *Let $\gamma = \frac{\sigma_1 - \sigma_2}{\sigma_1}$ be the parameter capturing the gap between the first and second largest singular values. If the Power Method is initialized with a random Gaussian vector with $A \in \mathbb{R}^{n \times d}$, then, with high probability, after $q = O\left(\frac{\log(1/\epsilon)}{\gamma}\right)$ steps, we have:*

$$\|v_1 - z_q\|_2 \leq \epsilon.$$

Total runtime:

$$O(\mathrm{nnz}(A)q) = O\left(\mathrm{nnz}(A) \cdot \frac{\log(1/\epsilon)}{\gamma}\right).$$

Above also implies,

$$\|A^T z_q\|_F^2 \geq (1 - \epsilon^2)\|A^T v_1\|_F^2.$$

Proof

Let us write $z_0 = \sum_{i=1}^d \mu_i v_i$ in terms of the right singular vector basis. If $\mu = [\mu_1, \ldots, \mu_d]$, we have $\mu = \frac{V^T g}{\|g\|_2}$ for random Gaussian g. Since V is orthogonal, we have $\|\mu\|_2 = 1$.

With high probability,

$$\frac{1}{\mathrm{poly}(d)} \leq |\mu_i| \leq 1 \quad \text{for } i = 1, \ldots, d.$$

Note that μ is Gaussian. We can show that $\mathrm{poly}(d) \approx d^3$ with high probability.

After q steps, we have $z_q = c(A^T A)^q z_0$ for some scaling c. If we write $z_q = \sum_{i=1}^d \rho_i v_i$, we have $\rho_i = c\mu_i^{2q}\mu_i$.

Since $A^T A = V\Sigma^2 V^T$, if the gap parameter is $\gamma = \frac{\sigma_1 - \sigma_2}{\sigma_1}$, we can show that, for all $j \geq 2$:

$$\frac{\sigma_j}{\sigma_1} \leq (1 - \gamma).$$

For all $j \geq 2$,

$$\frac{|\rho_j|}{|\rho_1|} \leq (1-\gamma)^{2q} \frac{|\mu_j|}{|\mu_1|} \leq (1-\gamma)^{2q} \text{poly}(d).$$

For any $0 < x \leq 1$, we can show that $(1-x)^q \leq e^{-q}$.

Proof:

The Taylor series expansion of $\log(1-x)$ around 0 is given by:

$$\log(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}.$$

For $0 < x \leq 1$, this series converges, and we have:

$$\log(1-x) \leq -x.$$

Taking exponentials on both sides, we get:

$$1 - x \leq e^{-x}.$$

Raising both sides to the power of q, we obtain:

$$(1-x)^q \leq e^{-qx}.$$

Hence proved.

If we set $q = \frac{\log(\text{poly}(d)/\varepsilon)}{\gamma} = O\left(\frac{\log(d/\varepsilon)}{\gamma}\right)$ and then we get $\frac{|\rho_j|}{|\rho_1|} \leq \sqrt{\frac{\varepsilon}{d}}$.

Since z_q is a unit vector, we have $\sum_i \rho_i^2 = 1$, and $|\rho_1| \leq 1$, hence $\rho_1^2 \geq 1 - d(\sqrt{\varepsilon/d})^2 \implies |\rho_1| \geq 1 - \varepsilon$.

Therefore,

$$\|v_1 - z_q\|_2 = 2 - 2\langle v_1, z_q \rangle \leq 2\varepsilon.$$

Analysis without Gap

Theorem 2 (Gapless Power Method Convergence)**.** *If Power Method is initialized with a random Gaussian vector, then, with high probability, after $q = O\left(\frac{\log(d/\varepsilon)}{\varepsilon}\right)$ steps, we obtain a z_q satisfying:*

$$\|A - Az_q z_q^T\|_F^2 \leq (1+\varepsilon)\|A - Av_1 v_1^T\|_F^2.$$

Gap γ might be too small. Then, we do not care to find v_1. Say $\sigma_1 = \sigma_2$, then v_2 is as good as v_1.

Proof:

We know that $\|A - Az_q z_q^T\|_F^2 = \|A\|_F^2 - \|Az_q z_q^T\|_F^2$. So, to prove the above, we need to show $\|Az_q\|_2^2 \geq (1-\varepsilon)^2 \sigma_1^2$.

We have,

$$\|Az_q\|_2^2 = z_q^T A^T A z_q = \sum_{i=1}^{d} \rho_i^2 \sigma_i^2,$$

where $\rho_i = v_i^T z_q$. For $q = O\left(\frac{\log(d/\varepsilon)}{\varepsilon}\right)$, from our previous analysis we have $\rho_1 \geq (1-\varepsilon)$. Hence,

$$\|Az_q\|_2^2 \geq \sum_{i=1}^{d} \rho_i^2 \sigma_i^2 \geq \rho_1^2 \sigma_1^2 \geq (1-\varepsilon)^2 \sigma_1^2.$$

Subspace Iteration

For larger $k \geq 1$ (finding the top-k singular vectors/values):

- Block Power Method aka Simultaneous Iteration aka Subspace Iteration aka Orthogonal Iteration.

Block Power Method

1. Choose $S \in \mathbb{R}^{d \times k}$ a random Gaussian matrix.
2. $Z_0 = \mathrm{orth}(S)$
3. For $l = 1, \ldots, q$
    - $Z_l = A^T(AZ_{l-1})$
    - $Z_l = \mathrm{orth}(Z_l)$
4. Return Z_q

Total runtime: $O(\mathrm{nnz}(A)kq)$.

Equivalent to sketching with input $(A^T A)^q$.

With $q = O\left(\frac{\log(d/\varepsilon)}{\varepsilon}\right)$, we obtain a nearly optimal low-rank approximation:

$$\|A - AZ_q Z_q^T\|_F^2 \leq (1+\varepsilon)\|A - AV_k V_k^T\|_F^2.$$

For $q = O\left(\frac{\log(nd)}{\varepsilon}\right)$, we have:

$$\|A - AZ_q Z_q^T\|_2 \leq (1+\varepsilon)\|A - A_k\|_2.$$