COMS E6998-9: Algorithmic Techniques for Massive Data	Oct 22, 2015

Lecture 11 – Applications of Dimension Reduction

Instructor: *Alex Andoni* — Scribes: *Marshall Ball*

Today we looked at two applications of dimension reduction for improving the time complexity of two classical problems in P:

(a) Matrix Multiplication.

(b) Least Square Regression

As usual, we will make our lives easier by considering approximate variants (of the optimization versions) of the above problems.

1 Matrix Multiplicaiton

Definition 1. *(Exact) Matrix Multiplication* is the following problem:

- Given $A, B \in \mathbb{R}^{n\times d}$,
- Compute: $C = A^\top B \in \mathbb{R}^{d\times d}$.

In general, you may consider the problem for arbitrary fields $\mathcal{K}$, but we will restrict our attention to $\mathbb{R}$. (One may also consider matrices of arbitrary dimension.)

Naively, we can solve the above problem in time $O(nd^2)$. The state of the art for $n \times n$ matrices is time $O(n^\omega)$ for $\omega \approx 2.36\ldots$. This will yield an algorithm for our problem with complexity $O(d^2 n^{\omega-2})$.

However as usual, we are interested in a near linear time, $\sim O(nd)$, algorithm. To do this exactly is hard, so we will relax the problem to an approximate version.

First, we define the following norm to characterize our approximation guarrantee;

Definition 2. For a matrix $Z \in \mathbb{R}^{m\times n}$, the *(squared) frobenius norm* is defined as follows:

$$\|Z\|_F^2 := \sum_{i,j} Z_{i,j}^2.$$

Definition 3. *(Approximate) Matrix Multiplication* is the following problem:

- Given $A, B \in \mathbb{R}^{n\times d}$,
- Compute: $C' \in \mathbb{R}^d$ such that the following holds with high probility,

$$\|C' - A^\top B\|_F \leq \varepsilon \|A\|_F \times \|B\|_F.$$

Some notation for what follows:

$$A = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \qquad B = \begin{bmatrix} y_1^\top \\ \vdots \\ y_n^\top \end{bmatrix}$$

1.1 A First Algorithm: Sampling via a Horovitz-Thompson Estimator

We begin by noting the following:

Claim 1. $A^\top B = \sum_{k=1}^n x_k y_k^\top$ *($xy^\top$ is the "outer-product" of vectors x and y).*

Proof.

$$C_{ij} = \left(\sum_{k=1}^n x_k y_k^t \right)_{ij} = \sum_{k=1}^n x_{ki} y_{kj}.$$

□

From this, we derive the following algorithm (we will fix parameters in the analysis):

- Sample m coordinates k_t from $[n]$ ($2m$ vectors: x_{k_t}, y_{k_t}, $t \in [m]$) where the probability of sampling coordinate k is $p_k \propto \|x\|_k \|y\|_k$.
- Then simply output,

$$C' = \sum_{t=1}^m \frac{x_{k_t} y_{k_t}}{p_{k_t}}.$$

Theorem 2.

$$\Pr\left[\|C' - C\|_F > \varepsilon \|A\|_F \|B\|_F \right] < \frac{1}{\varepsilon^2 m}.$$

Notice that this means we can take $m = \Omega(1/\varepsilon^2)$.

Proof.
- *Expectation*

$$\begin{aligned} \mathbb{E}[C'] &= \frac{1}{m} \mathbb{E}\left[\sum_{t=1}^m \frac{x_{k_t} y_{k_t}^\top}{p_{k_t}} \right] \\ &= \frac{1}{m} \sum_{t=1}^m \sum_{k=1}^n \frac{p_k x_k y_k^\top}{p_k} \\ &= \sum_{k=1}^n x_k y_k^\top = C. \end{aligned}$$

- *Variance*

$$
\begin{aligned}
V &= \mathbb{E}\left[\|C' - C\|_F^2\right] \\
&= \mathbb{E}\left[\sum_{i,j}(C'_{ij} - C_{ij})^2\right] \\
&= \sum_{i,j} \mathrm{Var}[C'_{ij}] \\
&\leq \sum_{i,j} \mathrm{Var}\left[\tfrac{1}{m}\sum_{t=1}^{m} \underbrace{\frac{x_{k_t i} y_{k_t j}}{p_{k_t}}}_{\text{id. dist. var.}}\right] \\
&= \sum_{i,j} \tfrac{1}{m}\mathrm{Var}\left[\frac{x_{ki} y_{kj}}{p_k}\right] \qquad \text{(randomness over } k\text{)} \\
&\leq \tfrac{1}{m}\sum_{i,j} \mathbb{E}\left[\left(\frac{x_{ki} y_{kj}}{p_k}\right)^2\right] \\
&= \tfrac{1}{m}\sum_{i,j}\sum_{k=1}^{n} p_k \left(\frac{x_{ki} y_{kj}}{p_k}\right)^2 \\
&= \tfrac{1}{m}\sum_{k=1}^{n} \frac{1}{p_k} \sum_{i,j} x_{ki}^2 y_{kj}^2 \\
&= \tfrac{1}{m}\sum_{k=1}^{n} \frac{1}{p_k} \|x_k\|_F^2 \|y_k\|_F^2
\end{aligned}
$$

So, take

$$
p_k := \frac{\|x_k\|_F \|y_k\|_F}{\sum_{i=1}^{n} \|x_i\|_F \|y_i\|_F}.
$$

Then (via Cauchy-Schwartz),

$$
V \leq \frac{\left(\sum_{k=1}^{n} \|x_k\|_F \|y_k\|_F\right)^2}{m} \leq \frac{\left(\sum_{k=1}^{n} \|x_k\|_F^2\right)\left(\sum_{k=1}^{n} \|y_k\|_F^2\right)}{m} = \frac{\|A\|_F^2 \|B\|_F^2}{m}
$$

- So applying Chebyshev to the above,

$$
\Pr\left[\|C' - C\|_F^2 > \varepsilon^2 \|A\|_F^2 \|B\|_F^2\right] \leq \frac{\mathbb{E}\left[\|C' - C\|_F\right]}{\varepsilon^2 \|A\|_F^2 \|B\|_F^2} \leq \frac{1}{m\varepsilon^2}
$$

□

1.2 A Second Algorithm: Using Dimension Reduction

Note 1. We can view the above algorithm as the following:

- Choose a random $\Pi \in \mathbb{R}^{m \times n}$ where

$$\Pi_{i,j} := \begin{cases} \frac{1}{\sqrt{mp_k}} & \text{if } (i,j) = (t, k_t) \\ 0 & \text{otherwise} \end{cases}$$

- Compute:

$$C' = (\Pi A)^\top (\Pi B).$$

Observe that the above algorithm requires two passes over the data, one to sample Π (compute the p_k's) and one to compute the "reduced" matrix product (or the sum in our previous formulation).

Given this "randomized-projection/embedding" formulation of our approximation algorithm, it seems an appropriate place to invoke the magic of Johnson-Lindenstrauss. Consider the following definition:

Definition 4. $\Pi \in \mathbb{R}^{m \times n}$ is an (ε, δ)-*dimension reducing matrix*, (ε, δ)-DR, if

$$\forall x \in \mathbb{R}^n, \Pr\left[\left| \|\Pi x\|_2^2 - \|x\|_2^2 \right| > \varepsilon \|x\|_2^2 \right] \leq \delta.$$

Given some (ε, δ)-DR matrix Π, our algorithm is to simply compute:

$$C' = (\Pi A)^\top (\Pi B).$$

Theorem 3. *Π is (ε, δ)-DR $\implies \Pr[\|C' - C\|_F > 3\varepsilon \|A\|_F \|B\|_F] \leq 3d^2\delta$.*

Remark 1. With a more precise version of the Johnson-Lindenstrauss lemma we can remove the d^2 factor from the above.

Corollary 4. *If we choose $m = O(1/\varepsilon^2 \log(1/\delta)), \delta = \frac{1}{10d^2}$, then (naively) we can compute C' in time $O(mnd) + O(dmd) = O(\frac{nd+d^2}{\varepsilon^2} \log d)$.*

By the above remark, the $\log d$ factor is simply an artifact of our analysis.

To prove the theorem, we will show $C'_{ij} \approx C_{ij}$ with probability $\geq 1 - 3\delta$ and then take a union bound (hence the d^2).

Proof. First some notation:

$$A = \begin{bmatrix} A_1 & \cdots & A_d \end{bmatrix} \qquad B = \begin{bmatrix} B_1 & \cdots & B_d \end{bmatrix}$$

$$a_i := \frac{A_i}{\|A_i\|_2} \qquad b_i := \frac{B_i}{\|B_i\|_2}$$

Note:

- $C_{ij} = A_i^\top B_j = \|A_i\| \|B_j\| a_i^\top b_j$.

- With probability $\geq 1 - 3\delta$,

$$\begin{aligned} C'_{ij} = (\Pi A_i)^\top (\Pi B_j) &= \|A_i\| \|B_j\| (\Pi a_i)^\top (\Pi b_j) \\ &= \|A_i\| \|B_j\| \left[\|\Pi a_i\|^2 + \|\Pi b_j\|^2 - \tfrac{1}{2} \|\Pi a_i - \Pi b_J\|^2 \right] \\ &= \|A_i\| \|B_j\| \left[\|a_i\|^2 + \|b_j\|^2 - \|a_i - b_j\|^2 \pm 3\varepsilon \right] \qquad (\Pi \text{ is } (\varepsilon, \delta) - \text{DR}) \\ &= \|A_i\| \|B_j\| \left[a_i b_j \pm 3\varepsilon \right] \end{aligned}$$

So with probability $\geq 1 - 3\delta$, $(C'_{ij} - C_{ij})^2 \leq \|A_i\|_2^2 \|B_j\|_2^2 (3\varepsilon)^2$. This implies (via union bound) that with probability $\geq 1 - 3\delta d^2$,

$$\|C' - C\|_F \leq \sum_{ij} 9\varepsilon^2 \|A_i\|^2 \|B_j\|^2 = 9\varepsilon^2 \|A\|_F^2 \|B\|_F^2.$$

□

2 Least Squares Regression

Definition 5. *(Exact) Least Squares Regression* is the following problem:

- Given $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$,
- find $x^* = \operatorname{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$.

We can consider least squares regression as a simple learning problem where the i-th row of A, $a^{(i)}$, is labeled with b_i according to some approximately linear function.

Definition 6. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *linear* if

$$\exists y \in \mathbb{R}^d : f(x) = \langle x, y \rangle.$$

If $\exists x : Ax = b$ then the problem is easy. In general, we are only assume $\exists x : Ax \approx b$.

In general, we can do least squares regression via Singular Value Decomposition (in time $\tilde{O}(nd^{\omega-1})$), but perhaps we can speed things up by loosening the approximation.

Definition 7. *(Approximate) Least Squares Regression* is the following problem:

- Given $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$,
- Let $x^* = \operatorname{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$. Find $x \in \mathbb{R}^d$ such that

$$\|Ax - b\|_2 \leq (1 + \varepsilon) \|Ax^* - b\|_2.$$

To solve this problem we will use dimension reduction, as promised.

First, we define a special kind of dimension reducing matrix:

Definition 8. $\Pi \in \mathbb{R}^{m \times n}$ is a (d, ε, δ)*-subspace embedding*, (d, ε, δ)-SE, if $\forall P \subset \mathbb{R}^n$ such that P is a d-dimensional subspace,

$$\Pr[\forall p \in P : |\|\Pi p\| - \|p\|| \leq \varepsilon \|p\|] \geq 1 - \delta.$$

Then given some such SE Π, our alorithm is simply: find $\text{argmin}_x \|\Pi Ax - \Pi b\|$ (via SVD).

Naively, the time to reduce dimension is $O(mnd)$. The time to perform SVD on the result is $O(md^{\omega-1}$.

So if we take $m = O(d/\varepsilon^2)$, then the resulting algorithm has time complexity

$$O(\frac{nd^2}{\varepsilon} + md^{\omega-1}).$$

If we use a faster version of Johnson-Lindenstrauss, we can acheive $O_\varepsilon\left((n \log n + m^3)d\right)$ time complexity.

Unfortunately, at this point we ran out of time. We will finish up this application of dimension reduction next lecture.