600.464 Randomized and Big Data Algorithms
Homework #4 Answers

Ravindra Gaddipati

November 29, 2016

Problem 1 (4 points)

With the distribution given, pairwise distances are not preserved within some fixed ϵ as n grows. To preserve the norm through the multiplication by A, we normalize A through factor α such that $E[||\alpha Ax||_2^2] = 1$. With the given probability distribution, we choose $\alpha = \sqrt{1/20}$. We define the probability of success as

$$P(||x||(1-\epsilon) \leq ||Ax|| \leq ||x||(1+\epsilon))$$

Letting $x = v/||v||$ such that $||x|| = 1$ for vector v,

$$P(1-\epsilon \leq ||Ax|| \leq 1+\epsilon)$$

Let A_i be the column vectors of A such that $A = [A_0| \dots |A_n]$.

We examine the case $x = (1, 0, \dots, 0)$. Then,

$$||Ax||_2^2 = ||A_1||_2^2$$

For this vector, we compute the probability the norm is perfectly preserved,

$$P(||Ax|| = 1) = P(||A_i||_2^2) = \binom{r}{20}\left(\frac{20}{r}\right)^{20}\left(1 - \frac{20}{r}\right)^{r-20}$$

Where this represents the probability A_1 contains exactly 20 ones. Taking the limit,

$$\lim_{r\to\infty}\left(\binom{r}{20}\left(\frac{20}{r}\right)^{20}\left(1 - \frac{20}{r}\right)^{r-20}\right) = 0$$

Thus $P(||Ax|| \neq 1) = 1$ as $r \to \infty$. With an input set of n points, there exists $O(n^2)$ vectors, where the error for a single vector is greater than $O(1/n^2)$. Therefore as our input set of n points grows, the probability of failing grows faster. As a result, we cannot bound the error as n continues to grow for a fixed ϵ. This distribution to sparsify matrix A does not allow for the preservation of pairwise distances.

Problem 2 (4 points)

If C_1 is a coreset of C_2, and C_2 is a coreset of C_3, then by the definition of a (k, ϵ) coreset we have:

$$
\begin{aligned}
|\nu(Z, C_2) - \nu(Z, C_1)| \leq \epsilon\nu(Z, C_2) \\
|\nu(Z, C_3) - \nu(Z, C_2)| \leq \epsilon\nu(Z, C_3)
\end{aligned}
$$

Adding the two and using the triangle inequality,

$$
\begin{aligned}
|\nu(Z, C_2) - \nu_w(Z, C_1)| + |\nu(Y, C_3) - \nu_w(Y, C_2)| &\leq \epsilon\nu(Z, C + 2) + \delta\nu(Y, C_3) \\
|\nu(Z, C_2) - \nu_w(Z, C_1) + \nu(Y, C_3) - \nu_w(Y, C_2)| &= |\nu(Y, C_3) - \nu_w(Z, C_1) + \nu(Z, C_2) - \nu_w(Y, C_2)| \\
&= |\nu(Y, C_3) - \nu_w(Z, C_1) - \nu_w(Y, C_2) - \nu(Z, C_2)| \\
|\nu(Y, C_3) - \nu_w(Z, C_1) - \nu_w(Y, C_2) - \nu(Z, C_2)| &\geq |\nu(Y, C_3) - \nu_w(Z, C_1)| - |\nu_w(Y, C_2) - \nu(Z, C_2)|
\end{aligned}
$$

Where we obtain the last expression by examination of the RHS. Using the initial expression,

$$
\begin{aligned}
|\nu(Y, C_3) - \nu_w(Z, C_1)| - |\nu_w(Y, C_2) - \nu(Z, C_2)| &\leq |\nu(Z, C_2) - \nu_w(Z, C_1)| + |\nu(Y, C_3) - \nu_w(Y, C_2)| \\
|\nu(Y, C_3) - \nu_w(Z, C_1)| - |\nu_w(Y, C_2)| &\leq |\nu(Z, C_2) - \nu_w(Z, C_1)| + |\nu(Y, C_3) - \nu_w(Y, C_2)| \\
&\leq \epsilon\nu(Z, C_2) + \delta\nu(Y, C_3)
\end{aligned}
$$

Since Z is always valid for Y, we can show that the following inequality holds,

$$
\begin{aligned}
|\nu(Y, C_3) - \nu_w(Z, C_1)| &\leq \epsilon\nu(Z, C_2) + \delta\nu(Y, C_3) + |\nu_w(Y, C_2)| \\
&\leq \epsilon\nu(Y, C_2) + \delta\nu(Y, C_3) + |\nu_w(Y, C_2)| \\
&\leq (\beta + \epsilon)\nu(Y, C_3) + \delta\nu(Y, C_3) \\
&\leq O(\delta + \epsilon)\nu(Y, C_3)
\end{aligned}
$$

Thus C_1 is a $(k, O(\epsilon + \delta))$ coreset for C_3.

Problem 3 (3 points)

During the coreset construction, A partitions P. With the $[\alpha, \beta]$ bicriteria, we are able to generate disjoint ring sets where the size in controlled by β. However by choosing random αk centers as our partitioning set, there is no restriction on the location of the centers. To be a valid coreset, then

$$|\nu(X, P) - \nu(C, S)| \leq \epsilon\nu(C, P)$$

With fixed ϵ, and arbitrary partitioning of P can result in a arbitrarily large $\nu(C, S)$- for example we can place them all very far from all $x \in P$. For any ϵ it is possible to place our αk centers such that the coreset is not valid since the centers are unbounded. Thus there is no way to guarantee that a random partitioning will result in a valid coreset.

Problem 4 (4 points)

We demonstrate the proof with the normalized vector,

$$x = \frac{v}{||v||}$$

And we want to show that, for $||x|| = 1$,

$$(1-\epsilon)||x|| \leq \frac{1}{\sqrt{k}}||Mx|| \leq (1+\epsilon)||x||$$

Let Y_i be a single element of the product, $Y_i = \sum_{j=1}^{n} M_{ij}x_i$.

$$\begin{aligned}
||Mx||^2 - 1 &= \left(\frac{1}{\sqrt{k}}\sqrt{\sum_i Y_i^2}\right) - 1 \\
&= \frac{1}{k}\sum_i Y_i^2 - 1 \\
&= \frac{1}{k}\left(\sum_i Y_i^2 - k\right)
\end{aligned}$$

Using lemma 4 from the JL-1 notes, we note that $||Mx||^2 - 1$ has the same distribution as $\frac{1}{k}Z = \frac{1}{k}(Y_1^2 + \cdots + Y_k^2 - k$ with subgaussian tails. As a result,

$$\begin{aligned}
P(||Mx|| \geq 1+\epsilon) &= P\left(||Mx||^2 \geq (1+\epsilon)^2\right) \\
&\leq P(||Mx||^2 \geq 1+\epsilon^2) \\
&= P\left(\frac{1}{\sqrt{k}}Z \geq \epsilon^2\right) \\
&= P(Z \geq \epsilon^2 k) \\
P(Z \geq \epsilon^2 k) &\leq e^{-a\epsilon^4 k} \\
&= e^{-\epsilon^4 k}
\end{aligned}$$

When choosing $a = 1$. Bounding for an error of 0.05 on both sides,

$$\begin{aligned}
P(Z \geq \epsilon^2 k) \leq e^{-\epsilon^4 k} &\leq 0.05 \\
-\epsilon^4 k &\leq \ln(0.05) \\
k &\geq \frac{\ln(0.05)}{\epsilon^4}
\end{aligned}$$

k is $O(1/\epsilon^4)$ as required, and the error is bounded by 0.1 by choosing k by above.

Collaborators

I worked with Matthew Ige, Emily Wagner, and Phillipe Piantonne on these problems. The following sources were used.
`people.csail.mit.edu/dannyf/coresets.pdf`
`ttic.uchicago.edu/ gregory/courses/LargeScaleLearning/lectures/jl.pdf`