15-859 Algorithms for Big Data — Fall 2022

Solutions for Problem Set 2

Problem 1: ℓ_1-Median Subspace Embedding

As seen in the class, if S is an $m \times d$ matrix with independent Cauchy random variables scaled by m, then with probability $\geq 9/10$, for all x,

$$\Omega(1)\|Ax\|_1 \leq \|SAx\|_1 \leq O(d \log d)\|Ax\|_1$$

if $m = O(d \log d)$. With a proof similar to that of in class, it is not hard to see that the statement continues to hold true even if $m = O(d \log(d/\epsilon)/\epsilon^2)$ if $\epsilon \geq 1/\text{poly}(d)$. We also assume without loss of generality that A is a well-conditioned basis satisfying for all x,

$$\|x\|_1 \leq \|Ax\|_1 \leq d\|x\|_1.$$

Fix a vector x. Now using the 1-stability property of Cauchy random variables, we have that

$$(SAx)_i = \frac{1}{m}\sum_j C_{ij}(Ax)_j \sim \frac{1}{m}\|Ax\|_1 C_i$$

where the last notation means that $(SAx)_i$ is distributed as $C_i \cdot (\|Ax\|_1/m)$ with C_i being a Cauchy random variable. Thus, we obtain that $|(SAx)_i| \sim (\|Ax\|_1/m)|C_i|$. Importantly, as the rows of S are independent we obtain that $|(SAx)_1|, \ldots, |(SAx)_m|$ are independent as well. From now on we just consider the independent standard Cauchy random variables $C_1, \ldots, C_m$. We have

$$\begin{aligned}
\mathbf{Pr}[|C_i| \geq (1+\epsilon)] &= \mathbf{Pr}[C_i \geq 1+\epsilon] + \mathbf{Pr}[C_i \leq -1-\epsilon] \\
&= \frac{1}{\pi}\left[\int_{1+\epsilon}^{\infty} \frac{1}{1+x^2}\mathrm{d}x + \int_{-\infty}^{-(1+\epsilon)} \frac{1}{1+x^2}\mathrm{d}x\right] \\
&= \frac{1}{\pi}[\tan^{-1}(\infty) - \tan^{-1}(1+\epsilon) + \tan^{-1}(-(1+\epsilon)) + \tan^{-1}(-\infty)] \\
&= \frac{1}{\pi}[\pi - 2\tan^{-1}(1+\epsilon)].
\end{aligned}$$

We now use the fact that $\tan^{-1}(1+\epsilon) \geq \pi/4 + \epsilon/6$ for $0 < \epsilon < 1$. Hence, $\mathbf{Pr}[|C_i| \geq (1+\epsilon)] \leq (1/\pi)(\pi/2 - \epsilon/3) \leq 1/2 - \epsilon/(3\pi)$. Let X_i be the indicator random variable that denotes if $|C_i| \geq 1+\epsilon$. Thus, we have $\mathbf{E}[X_i] \leq (1/2 - \epsilon/(3\pi))$. By Chernoff bound, we have $\mathbf{Pr}[\sum_i X_i \geq (1+\delta)m\mu] \leq \exp(-\delta^2 m\mu/3)$ where $\mu = \mathbf{E}[X_i]$. We can also see that $\mu \geq 1/4$. Picking $\delta = \epsilon/10$, we conclude that with probability $\geq 1 - \exp(-\epsilon^2 m\mu/300)$,

$$\sum_i X_i \leq (1+\epsilon/10)m\mu \leq m(1+\epsilon/10)(1/2 - \epsilon/3\pi) \leq m(1/2 - \epsilon/100)$$

for small enough ϵ. Using the fact that $\mu \geq 1/4$, and $m = O(d\log(d/\epsilon)/\epsilon^2)$, we obtain that with probability $1 - \exp(-O(d\log(d/\epsilon)))$,

$$\sum_i X_i \leq m(1/2 - \epsilon/100).$$

Thus we showed that number of indices i such that $|(SAx)_i| \geq (1/m)\|Ax\|_1(1+\epsilon)$ is at most $(m/2)(1-\epsilon/50)$ with probability $\geq 1-\exp(-O(d\log(d/\epsilon)))$. Similarly, we can show that number of indices i such that $|(SAx)_i| \leq (1/m)\|Ax\|_1(1-\epsilon)$ is at most $(m/2)(1-\epsilon/50)$ with probability $\geq 1-\exp(-O(d\log(d/\epsilon)))$. Thus, by a union bound, with probability $\geq 1-2\exp(-O(d\log(d/\epsilon)))$,

$$\|SAx\|_{\text{med}} \in [(1-\epsilon)\|Ax\|_1/m, (1+\epsilon)\|Ax\|_1/m].$$

Now consider the set $M = \{Ax \mid \|Ax\|_1 = 1\}$ and an $\epsilon/\text{poly}(d)$-net $N \subseteq M$ for M. So, for every $Ax \in M$, there is an $Ay \in N$ such that $\|Ax - Ay\|_1 \leq \epsilon/\text{poly}(d)$. As given the size of N can be taken as $\exp(O(d\log(\text{poly}(d)/\epsilon))) = \exp(O(d\log(d/\epsilon)))$. Now, by a union bound, we can assume that for all $Ay \in N$,

$$\|SAy\|_{\text{med}} \in [(1-\epsilon)\|Ay\|_1/m, (1+\epsilon)\|Ay\|_1/m] = [(1-\epsilon)/m, (1+\epsilon)/m].$$

Now consider an arbitrary $Ax \in M$ and let $Ay \in N$ be such that $\|Ax - Ay\|_1 \leq \epsilon/\text{poly}(d)$. We now have for all i,

$$|(SAx)_i| \leq |(SAy)_i| + |(SA(x-y))_i| \leq |(SAy)_i| + \|SA(x-y)\|_\infty.$$

Similarly, $|(SAx)_i| \geq |(SAy)_i| - \|SA(x-y)\|_\infty$ for all i. This then implies that

$$\|SAy\|_{\text{med}} - \|SA(x-y)\|_\infty \leq \|SAx\|_{\text{med}} \leq \|SAy\|_{\text{med}} + \|SA(x-y)\|_\infty$$

Now, using the fact that $\|SA(x-y)\|_\infty \leq \|SA(x-y)\|_1 \leq O(d\log d)\|Ax-Ay\| \leq O(d\log d)\epsilon/\text{poly}(d) \leq \epsilon/m$, we obtain that

$$\|SAx\|_{\text{med}} \in [(1-2\epsilon)/m, (1+2\epsilon)/m]$$

for all $Ax \in M$. By scaling we obtain that for all x,

$$\|SAx\|_{\text{med}} \in [(1-2\epsilon)\|Ax\|_1/m, (1+2\epsilon)\|Ax\|_1/m].$$

Problem 2:

The main high level idea is as follows: Suppose you want to approximate a quantity $a + b$ with both $a, b > 0$. Suppose you know that $b \leq \gamma a$ and you know a quantity $a' \in (1 \pm \epsilon/2)a$. Now, even if $b' \leq (\epsilon/2)(b/\gamma) + (1+\epsilon)b \leq (\epsilon/2)a + (1+\epsilon)b$, we have $a' + b' \leq (1+\epsilon)(a+b)$. Which essentially means that we do not have to approximate the smaller quantity very well to approximate the whole sum well.

We use the countsketch matrix with r rows to estimate the norm. We process until we see the vector u to obtain the sketch Su. Then instead of updating the sketch Su, we start afresh (still using the same CountSketch matrix) and compute a sketch Sv. Using the AMM property of CountSketch matrices, we have that with probability $\geq 9/10$, the following properties hold,

$$|\|Sv\|_2^2 - \|v\|_2^2| \leq \frac{10}{\sqrt{r}}\|v\|_2^2$$

and

$$|\langle Su, Sv\rangle - \langle u, v\rangle| \leq \frac{10}{\sqrt{r}}\|u\|_2\|v\|_2.$$

We now have

$$\|u+v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 + 2\langle u, v\rangle$$

and let $\alpha = (1 \pm \epsilon/2)\|u\|_2^2$. Conditioned on above events,

$$\begin{aligned}
\alpha + \|Sv\|_2^2 + 2\langle Su, Sv\rangle &\leq (1+\epsilon/2)\|u\|_2^2 + (1+10/\sqrt{r})\|v\|_2^2 + 2\langle u, v\rangle + \frac{10}{\sqrt{r}}\|u\|_2\|v\|_2 \\
&\leq (1+\epsilon/2)(\|u+v\|_2^2) + \|u+v\|_2^2 \frac{(10/\sqrt{r})\|v\|_2^2 + (10/\sqrt{r})\|u\|_2\|v\|_2}{\|u+v\|_2^2} \\
&\leq (1+\epsilon/2)(\|u+v\|_2^2) + \|u+v\|_2^2 \frac{(10/\sqrt{r})\gamma\|u\|_2^2 + (10\sqrt{\gamma/r})\|u\|_2^2}{\|u+v\|_2^2}.
\end{aligned}$$

If $\gamma/r \leq c\epsilon^2$ for a small enough constant c, we then obtain that

$$\alpha + \|Sv\|_2^2 + 2\langle Su, Sv\rangle \leq (1+\epsilon)\|u+v\|_2^2.$$

Similarly, we can show that if $\gamma/r \leq c\epsilon^2$ for a small enough constant c, we have

$$\alpha + \|Sv\|_2^2 + 2\langle Su, Sv\rangle \geq (1-\epsilon)\|u+v\|_2^2.$$

Thus, a countsketch matrix S with $r \geq \Omega(\gamma/\epsilon^2)$ rows suffices to output a $1 \pm \epsilon$ approximation for $\|u+v\|_2^2$. Such a sketch can then be maintained in $O(\gamma/\epsilon^2 \cdot \log(n))$ space.