COMS E6998-9: Algorithms for Massive Data (Spring'19) Feb 7, 2019

Lecture 6: Dimension Reduction

Instructor: *Alex Andoni* Scribes: *Minghao Chen, Bicheng Gao*

1 Linearity

Definition 1. *$S_R : \mathbb{R}^n \to \mathbb{R}^k$ is linear sketch $\iff$ S_R is linear function, where R is random seed we choose.*

We could think a linear function is a special sketch.

Example 2. *Assume that we have $1, 2, \cdots, l$ routers, and each of them has its own data steaming. Then we could compute their frequency vectors f_i separately.Now, if we'd like to get information about the entire data traffic. We don't have to compute the sum of f_i and then sketch it. We could compute their sketch separately and sum of them to get the same result. In formula, Assume that $f_1, f_2 \cdots f_l$ is a sequence of frequency vectors, $S(f_1 + f_2 + \cdots + f_l) = \sum_{i=1}^{l} S(f_i)$.*

Example 3. *For G.S.M(General Turnstile Streaming), we have a sequence updates $(i, \delta_i), \delta_i \in \mathbb{R}, i \in [n], f' = f + \delta_i e_i$.Then $S(f') = S(f + \delta_i e_i) = S(f) + S(\delta_i e_i)$.*

We want to estimate the information contain in f using sketch S. Without the linearity, every time we see a new i, we have to update f and get f'. With the linear sketch , we don't have to update the f', we just update the sketch by adding $S(\delta_i e_i)$ to the old one. Therefore, we just need to store the sketch of f.

Example 4. *For a linear sketch, we have $f_1, f_2 \in \mathbb{R}^n$, $S(f_1 - f_2) = S(f_1) - S(f_2)$. To be specific, for l_2 norm, we use sketch T.o.W. to estimate it. It is a linear sketch and $E_{T.o.W.}(S(f_1 - f_2)) \approx \|f_1 - f_2\|_2^2$.*

We could consider sketch as approximately functional compression. $S(f_1), S(f_2)$ are used to estimate the information containing in f_1, f_2.
According to lecture 3, we are able to get a $(1 + \epsilon)$-approximation using $O(1/\epsilon^2)$ counters. In words, $E_{T.o.W.+}(S(f_1) - S(f_2)) \in (1 \pm \epsilon)||f_1 - f_2||_2^2$ with Probability 90%.

Observation 5. *Now, what if we expect the Probability to be $1 - \delta$, where δ is relatively small.*
Can we use Median Trick here?

Yes, but Media Trick **is not** dimension reduction, since it takes the median instead of l_2 norm.

2 Johnson-Lindenstrauss '84

Theorem 6. *$\forall \epsilon > 0, \forall k \in \mathbb{N}, \exists$linear sketch $\phi : \mathbb{R}^n \to \mathbb{R}^k, s.t. \forall x, y \in \mathbb{R}^n, Pr[||\phi(x) - \phi(y)||_2 \in (1 \pm \epsilon)||x - y||_2] \geq 1 - e^{-\frac{\epsilon^2 k}{9}}$.This is equivalent to $1 - \delta$ probability, when $k = O(\frac{\log \frac{1}{\delta}}{\epsilon^2})$*

Original theorem(old version of JL): ϕ is random linear subspace.

Proof. Because ϕ is linear sketch, we have $||\phi(x)-\phi(y)||_2 = ||\phi(x-y)||_2$. Our goal is to prove $||\phi(x)-\phi(y)||_2 = ||\phi(x-y)||_2 \approx ||x-y||_2$. If we are able to show $||\phi(x)||_2 \approx ||x||_2$. Then, we can easily get the former one.
Now we take $\phi = \frac{1}{\sqrt{k}}Gx$, where G_{ij} is Gaussian random variable.[1][2]
First, consider $k=1$, $\phi(x) = \sum_{i=1}^{n} G_{1j}x_j$

Fact 7. *Stability of Gaussian r.v.:* $\sum_{i=1}^{n} g_i x_i \sim g||x||^2$, *where* $g_1, g_2, \cdots, g_n \sim$ *standard G.r.v.*

Proof.

1. $(g_1, g_2, \cdots, g_n)$ is a distribution, and it is spherical symmetric.
2. $p.d.f.(g_1, g_2, \cdots, g_n) = (\frac{1}{2\pi})^{\frac{n}{2}} e^{-\frac{g_1^2}{2}} e^{-\frac{g_2^2}{2}} \cdots e^{-\frac{g_n^2}{2}} = (\frac{1}{2\pi})^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n} g_i^2}{2}} = (\frac{1}{2\pi})^{\frac{n}{2}} e^{-\frac{||g||_2^2}{2}}$

Therefore, $\sum_{i=1}^{n} G_{1j}x_j = g_1 \circ ||x||$, where $g_1 \sim N(0,1)$ □

Now, consider $k>1$:

$$\phi(x) = \frac{1}{\sqrt{k}}Gx \sim \frac{1}{\sqrt{k}}(g_1 \circ ||x||_2, g_2 \circ ||x||_2, \cdots, g_k \circ ||x||_2) \sim \frac{||x||_2}{k}(g_1, g_2, \cdots, g_k)$$

$$||\phi(x)||_2^2 = ||x||_2^2 \cdot \frac{1}{k} \cdot (g_1^2 + g_2^2 + \cdots + g_k^2)$$

Therefore, We only have to prove that $\frac{1}{k} \cdot (g_1^2 + g_2^2 + \cdots + g_k^2) \approx 1$
Notice that $\frac{1}{k} \cdot (g_1^2 + g_2^2 + \cdots + g_k^2) \sim \chi_k^2$ with k freedom degree.

Fact 8. $Pr[\chi_k^2 \notin (1 \pm \epsilon)] \leq 2 \cdot e^{\frac{k}{4}(\epsilon^2 - \epsilon^2)}$

As long as we choose $\epsilon < \frac{1}{3}$, we get J.L. Theorem. □

Corollary 9. *Fix* $N \in \mathbb{N}$, *consider* $x_1, x_2, ..., x_n \in R^n$. *Then* $\exists \phi : R^n \to R^k$, $k = O(\frac{\lg N}{\epsilon^2})$, *s.t*, $\forall i, j \in [N], ||\phi(x_i) - \phi(x_j)|| \in (1 \pm \epsilon)||x_i - x_j||_2$.

(In fact, a random ϕ (from JL theorem) works with probability greater than $1 - \frac{1}{N}$.)

Proof. Set $\delta = \frac{1}{N^3}$ in JL theorem $\Rightarrow Pr[\forall x_i, x_j, ||\phi(x_i) - \phi(x_j)|| \in (1 \pm \epsilon)||x_i - x_j||] >= 1 - \delta = 1 - \frac{1}{N^3}$. By the union bound over all pairs of $i, j \in [N]$, we can get $Pr[\forall i, j \in [N]||\phi(x_i) - \phi(x_j)|| \in (1 \pm \epsilon)||x_i - x_j||_2] >= 1 - \frac{N^2}{N^3} = 1 - \frac{1}{N}$. □

Fact 10. *Can not be the case that* $\exists \phi$ *that works for all sets of* N *points (unless* $k >= n$*).*
Why?

$$\begin{aligned} &\phi : R^n \to R^k, k < n \\ \Rightarrow &\exists x, y \in R^n, x \neq y, s.t, \phi(x) = \phi(y) \\ \Rightarrow &\phi \textit{ does not preserve dist } ||x-y||_2 \textit{ up to any approximation} \end{aligned} \tag{1}$$

Observation 11. *Time to compute* $\phi(x)$ *is* $O(nk)$. *Since in JL,* $\phi(x) = \frac{1}{\sqrt{k}} \cdot G \cdot x$

[1] $p.d.f. = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
[2] side notes: We could write T.o.W. as the same form, $\phi(x) = \frac{1}{\sqrt{k}}RX$, where $R_{ij} \in \{\pm 1\}$

Observation 12. *Are there dimension reduction in other norms, where $k = O(\frac{\lg \frac{1}{\delta}}{\epsilon^2})$ is the target dimension, for $(1 \pm \epsilon)$ with $1 - \delta$ probability?*

The general answer is **No**, but we there could be some sketch instead dimension reduction could do this.

Theorem 13. *$\exists$ linear sketch $S : R^n \to R^k$ and estimator E s.t, $Pr[E(S(x)) \in (1 \pm \epsilon) \cdot ||x||_1] >= 90\%$ and $k = O(\frac{1}{\epsilon^2})$.*

Proof. From observation 14 to Fact 17.

Observation 14. *In JL, we have $\phi(x) = \frac{1}{\sqrt{k}} \cdot G \cdot x$. And we can say $\phi_1(x) = \frac{1}{\sqrt{k}}(G_1 \cdot x)$ where $G = \begin{bmatrix} G_1 \\ \cdots \\ G_k \end{bmatrix}$. Then we know $\phi_1(x) \sim \frac{1}{\sqrt{k}} \cdot g_1 \cdot ||x||_2$, where $g_1 \in N(0,1)$.*

Fact 15. *$\sum_{i=1}^{n} c_i x_i = c \cdot ||x||_1$, where $c, c_1, c_2, \cdots, c_n \sim$ random variable in Cauchy distribution*[3]

Definition 16.

$$\begin{aligned} \phi(x) &= \frac{1}{k} \cdot C \cdot x \\ &\sim \frac{1}{k} \cdot (c_1 \cdot ||x||_1, c_2 \cdot ||x||_1, \cdots, c_k \cdot ||x||_1) = \frac{||x||_1}{k} \cdot (c_1, c_2, ..., c_k). \end{aligned} \tag{2}$$

where $c_i \sim$ Cauchy random variable

Set estimator

$$\begin{aligned} E[\phi(x)] &= k \cdot \mathbf{median}[\phi_i(x)] \\ &= \mathbf{median}||x||_1 \cdot |c_j| \\ &= ||x||_1 \cdot \mathbf{median}_{j=1..k}|c_j| \end{aligned} \tag{3}$$

We want the median part $\in (1 \pm \epsilon)$ with probability $>= 90\%$. And we know the fact that

Fact 17. *$Pr[\textbf{\textit{median}}_{j=1..k}|c_j| \in (1 + \epsilon) >= 90\%]$ as long as $k = \Omega(\frac{1}{\epsilon^2})$.*

$\square$

[3]the p.d.f of standard Cauchy distribution is $p(x) = \frac{1}{\pi(x^2+1)}$