Dimensionality Reduction & Subspace Embeddings

Monday, March 2, 2020 1:11 PM

Goal:

Linear regression $\min_x \|Ax-b\|_2$

$\nabla f = 0 \iff A^TAx - A^Tb = 0.$

So $x = (A^TA)^{-1}A^Tb$

Computational complexity:

$O(nd^2 + d^\omega)$.

Q. Can we do it faster?

Iterative methods.

Richardson Iteration:

$$x^{(k+1)} = x^{(k)} - \left(A^TAx^{(k)} - A^Tb\right) = \left(I - A^TA\right)x^{(k)} + A^Tb$$

Thm. $\kappa = \dfrac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)}$ with $A^TA \prec I$

$\left(\quad\right) \|x^{(k)} - x^*\|$

$\lambda_{min}(A^TA)$

Then $\|x^{(k+1)} - x^*\|_2 \leq \left(1 - \frac{1}{\kappa}\right) \|x^{(k)} - x^*\|_2$.

$$A^Tb + (I - A^TA)A^Tb + \left(\quad\right)^2 A^Tb + \dots$$

$$\rightarrow (A^TA)^{-1} = \frac{1}{(I - (I - A^TA))} = \left(I + (I - A^TA) + \dots\right) A^Tb.$$

We will prove a more general theorem using a more general algorithm.
(above could be very slow if κ is large.)

Thm
Suppose we know M s.t.

$$A^TA \precsim M \precsim \kappa \cdot A^TA$$

Let $x^{(k+1)} = x^{(k)} - M^{-1}(A^TAx^{(k)} - A^Tb)$

Then
$$\|x^{(k+1)} - x^*\|_M \leq \left(1 - \frac{1}{\kappa}\right) \|x^{(k)} - x^*\|_M.$$

Note $\|y\|_M^2 = y^T M y$.

Pf.

$$x^{(k+1)} - x^* = x^k - x^* - M^{-1}\left(A^T A x^{(k)} - A^T A x^*\right)$$

$$= \left(I - M^{-1} A^T A\right)\left(x^{(k)} - x^*\right)$$

$$\|x^{(k+1)} - x^*\|_M^2 = \left(x^{(k)} - x^*\right)^T \left(I - A^T A M^{-1}\right) M \left(I - M^{-1} A^T A\right)\left(x^{(k)} - x^*\right)$$

$$= \left(x^{(k)} - x^*\right)^T M^{\frac{1}{2}} \left(I - M^{-\frac{1}{2}} A^T A M^{-\frac{1}{2}}\right)\left(I - M^{-\frac{1}{2}} A^T A M^{-\frac{1}{2}}\right) M^{\frac{1}{2}} \left(x^{(k)} - x^*\right)$$

$$= \left(x^{(k)} - x^*\right)^T M^{\frac{1}{2}} (I - H)^2 M^{\frac{1}{2}} \left(x^{(k)} - x^*\right)$$

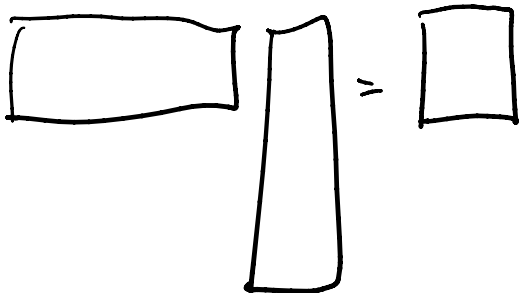
$\frac{1}{k} I \lesssim H \lesssim I \quad \Rightarrow \quad I - H \lesssim \left(1 - \frac{1}{k}\right) \cdot I$

$$\leq \left(1 - \frac{1}{k}\right)^2 \|x^{(k)} - x^*\|_M^2 .$$

What M to choose?

Goal is to approximate $A^T A$.

s.t. $\|Ax\|^2 \approx \|Bx\|^2$, $M = B^T B$..



At. $\|Ax\| \approx \|Bx\|$, [illegible]

There is a perfect M.

$$A = U\Sigma V^T \qquad \forall y \in \{Ax\}$$

$$\|U^T y\|^2 = \|U^T U \Sigma V^T x\|^2 = x^T A^T A x = \|Ax\|^2$$

But finding this U needs SVD. Typically more expensive.

How about random Π ? ΠA

$$M = (\Pi A)^T (\Pi A).$$

Is this any good? what size of Π ?

Low-distortion embedding

Π is a ε-low-dist. emb. for a set S of vectors in $\mathbb{R}^n$

if $\forall y \in S$

$$(1-\varepsilon)\|y\|^2 \le \|\Pi y\|^2 \le (1+\varepsilon)\|y\|^2$$

dimension of Π = # rows.

Version of 11 - 7 Nov.

S for us is the subspace $\{Ax\}$.

We know $\Pi = U$ is perfect ($\varepsilon = 0$).

Oblivious Subspace Embedding.

Random matrix Π is a (d, ε, δ)-OSE for a fixed d-dim subspace S if it preserves $\|\ \|^2$ to within $(1 \pm \varepsilon)$ $\forall y \in S$ with prob. at least $1 - \delta$.

Alternatively. $\forall\, U \in \mathbb{R}^{n \times d}$

$$\Pr\left(\| U^T \Pi^T \Pi U - I_{d \times d} \|_{op} \geq \varepsilon \right) \leq \delta$$

Pf. $S = \{Uz\}$

$$(1-\varepsilon)\|y\|^2 \leq \|\Pi y\|^2 \leq (1+\varepsilon)\|y\|^2$$

$\Leftrightarrow$

$$(1-\varepsilon) U^T U \preceq U^T \Pi^T \Pi U \preceq (1+\varepsilon) U^T U$$

$U^T U = I$

$\Leftrightarrow$ $\quad \| U^T \Pi^T \Pi U - I \|_{op} \leq \varepsilon.$

$\Leftrightarrow \quad \| U'\Pi'\Pi U - I \|_{op} = \varepsilon.$

Thm [Johnson-Lindenstrauss] $\Pi_{ij} \sim N(0, \frac{1}{\sqrt{m}})$ with $m = O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ rows is a $(1, \varepsilon, \delta)$-OSE.

i.e. $\Pr\left(\left|\|\Pi x\|^2 - 1\right| \geq \varepsilon\right) \leq \delta.$

Thm A $(1, \varepsilon, \delta)$-OSE is a $(d, 4\varepsilon, 5^d \delta)$-OSE.

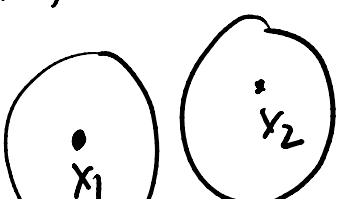
(So it suffices to handle $d = 1$).

Lemma [ε-net]. $\exists\, N \subseteq S^{n-1}$ s.t. $\forall x \in B^n$, $\exists x_i \in N$ s.t. $\|x - x_i\| \leq \varepsilon$ and $|N| \leq \left(1 + \frac{2}{\varepsilon}\right)^n$.

Pf. start with any $x \in S^{n-1}$.

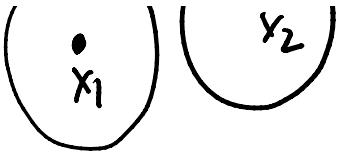
[while $\exists x$ s.t. $\forall x_i \in N \ \|x - x_i\| > \varepsilon$
add x to N

At the end $\forall x \in S^{n-1}$, $\exists x_i \in N$ s.t. $\|x - x_i\| \leq \varepsilon$.



$x_i + \frac{\varepsilon}{2} B_n$ are disjoint

$\subset (1 + \varepsilon) B_n$



$$\bigcup_i x_i + \frac{\varepsilon}{2} B_n \subseteq \left(1+\frac{\varepsilon}{2}\right) B_n$$

$$\therefore |N| \le \frac{\mathrm{Vol}\left(\left(1+\frac{\varepsilon}{2}\right) B_n\right)}{\mathrm{Vol}\left(\frac{\varepsilon}{2} B_n\right)} = \left(\frac{1+\frac{\varepsilon}{2}}{\frac{\varepsilon}{2}}\right)^n = \left(1+\frac{2}{\varepsilon}\right)^n.$$

Lemma 2. $\forall x \in B_n$, $\exists t_1 \dots t_i$ $t_i \le \frac{1}{2^i}$ s.t.

$$x = \sum_i t_i x_i \qquad x_i \in N.$$

Pf: Take N with $\varepsilon = \frac{1}{2}$.

$\forall x$, $\exists x_1$ s.t. $\|x - x_1\| \le \frac{1}{2}$

$\|x\|=1$

$\therefore \exists x_2, t_2, t_2 \le \frac{1}{2}$ s.t. $\|x - x_1 - t x_2\| \le \frac{1}{4}$

(applied to $\frac{1}{2}B^n$) continue to get conclusion.

Pf. (of Thm OSE):

$$x^T(U^T\Pi^T\Pi U - I)x = \sum_{i,j} t_i t_j \, x_i (U^T\Pi^T\Pi U - I) x_j$$

$$\le \sum_{i,j} t_i t_j \max_{x_i, x_j} x_i (U^T\Pi^T\Pi U - I) x_j$$

$$\le 1 \cdot \max x^T (U^T\Pi^T\Pi U - I) x$$

$$\leq 4 \cdot \max_{x \in N} x^T \left| U^T \Pi^T \Pi U - I \right| x$$

$$= 4 \max_{x \in UN} x^T \left| \Pi^T \Pi - I \right| x$$

$$= 4 \max_{x \in UN} \left| \|\Pi x\|^2 - 1 \right|$$

Since Π is an $(1, \varepsilon, \delta)$-OSE

$$\Pr\left(\left| \|\Pi x\|^2 - 1 \right| \geq \varepsilon \right) \leq \delta. \quad \text{for any single } x$$

And for all the $|N| \leq 5^d$ x's in UN,

$$\Pr\left(\forall x \in UN \left| \|\Pi x\|^2 - 1 \right| \geq \varepsilon \right) \leq 5^d \cdot \delta.$$

$\therefore$ Π is a $(d, 4\varepsilon, 5^d \delta)$-OSE.

$\therefore$ using $m = O\left(\frac{1}{\varepsilon^2} \left(d + \log \frac{1}{\delta} \right) \right)$ rows

suffices to get a (d, ε, δ)-OSE.

When $\Pi_{ij} \sim N(0, \frac{1}{m})$ or $\Pi_{ij} = \pm \frac{1}{\sqrt{m}}$.

$\varepsilon = \theta(1)$ suffices for linear regression

---

s.t. [illegible] ΠA takes $O(nd^2)$

But. computing ΠA takes $O(nd^2)$

So no saving on $A^T A$.

How about a sparse random matrix?

$\Pi_{ij} = \pm \frac{1}{\sqrt{s}}$ w.p. $\frac{s}{m}$ and 0 o.w. Π has m rows.

Th. Π as above is a (d, ε, δ)-OSE for

$s = O\left(\frac{1}{\varepsilon^2} \log^2 \frac{d}{\delta}\right)$ and $m = O\left(\frac{d}{\varepsilon^2} \log \frac{d}{\delta}\right)$.

$$U^T \Pi^T \Pi U = \sum_{r=1}^{m} (\Pi U)_r^T (\Pi U)_r$$

We will use 2 thms + 1 Lemma to analyze this sum.

Th 1 (Matrix Chernoff) $M_1, M_2, \ldots \in \mathbb{R}^{n \times n}$, $M_i \succeq 0$, $\mathbb{E} M_i = I$, $M_i \preceq R \cdot I$

$$(1 - O(\varepsilon)) I \preceq \frac{1}{T} \sum M_j \preceq (1 + O(\varepsilon)) I$$

where $T \geq \frac{R}{\varepsilon^2} \log \frac{n}{\delta}$.

Th 2. (Hanson-Wright) σ_i iid $\mathbb{E}\sigma_i = 0$ $|\sigma_i| \le 1$.

Then $\left|\sigma^T A \sigma - \mathbb{E}\sigma^T A \sigma\right| \le C \cdot \left(\|A\|_F \sqrt{\log \frac{1}{\delta}} + \|A\|_\phi \log \frac{1}{\delta}\right)$

w.p. $1-\delta$.

Lemma. $M_r = m\, U^T \Pi_r \Pi_r^T U$

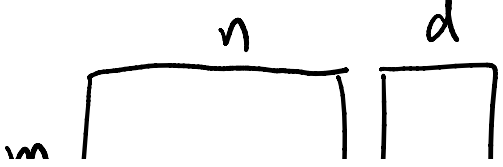
$M_r \succeq 0$ $\quad \mathbb{E} M_r = I$

$M_r \lesssim m \cdot \Pi_r^T U U^T \Pi_r \cdot I$

For $\Delta \ge \frac{m}{n} \log \frac{1}{\delta}, \frac{1}{\varepsilon^2} \log^2 \frac{1}{\delta}$, $m \ge \frac{d}{\varepsilon^2} \log \frac{1}{\delta}$.

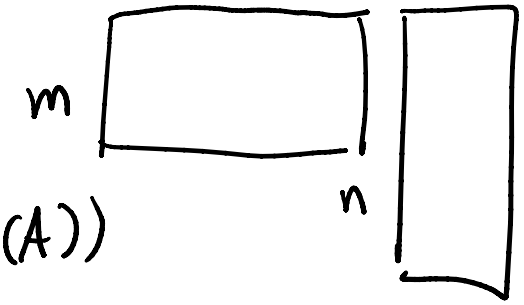
$$\Pi_r^T U U^T \Pi_r \le \frac{\varepsilon^2}{\log \frac{n}{\delta}} \quad \text{w.p. } 1-\delta.$$

Now the running time for ΠA is



~~nnz~~ · n · $\frac{A}{nnz}$

$O(A.nnz(A))$



to get an $\tilde{O}(d) \times d$ matrix.

$\tilde{O}(nnz(A) + d^{\omega})$.
