Dimensionality Reduction & Subspace Embeddings

Monday, March 2, 2020 1:11 PM

Goal.

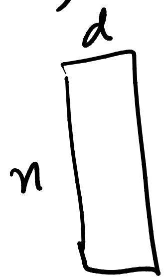
Linear regression $\min_x \|Ax-b\|_2$

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

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

Computational complexity:

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



Q. Can we do it faster?

Iterative methods.

Richardson Iteration:

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

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

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

$$A^Tb + (I - A^TA)A^Tb + (\qquad)^2 A^Tb + \dots$$

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

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

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

Thm Suppose we know M s.t.

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

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

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

Note $\|y\|_M^2 = y^TMy$.

<u>Pf.</u> $$x^{(k+1)} - x^* = x^k - x^* - M^{-1}(A^TAx^{(k)} - A^TAx^*)$$

$$= (I - M^{-1}A^TA)(x^{(k)} - x^*)$$

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

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

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

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

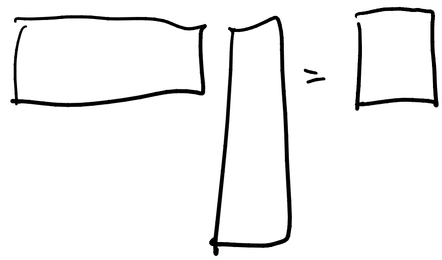
$$\frac{1}{\kappa} I \preceq H \preceq I \quad \Rightarrow \quad I - H \preceq \left(1-\frac{1}{\kappa}\right) \cdot I$$

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

What M to choose?

Goal is to approximate A^TA.

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



There is a perfect M.

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

$\|U^Ty\|^2 = \|U^TU\Sigma V^Tx\|^2 = x^TA^TAx = \|Ax\|^2$

But finding this U needs SVD. Typically more expensive.

How about random Π? $\quad \Pi A$

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

Is this any good? What size of Π?

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 $\forall y \in S$

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

dimension of Π = # rows.

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. $S = \{Uz\}$

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

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

$\iff$

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

$U^T U = I$

$\iff$
$$\| U^T \Pi^T \Pi U - I \|_{op} \le \varepsilon.$$

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. $\Pr\left(\left| \|\Pi x\|^2 - 1 \right| \ge \varepsilon \right) \le \delta.$

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

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

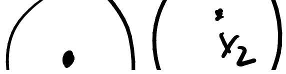
Lemma [ε-net]. $\exists N \subseteq S^{n-1}$ s.t. $\forall x \in B^n$, $\exists x_1 \in N$ s.t. $\|x - x_1\| \le \varepsilon$ and $|N| \le (1 + \frac{2}{\varepsilon})^n$.

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

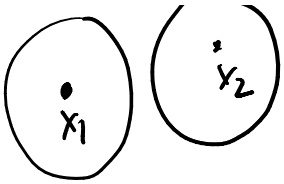
[while $\exists x$ s.t. $\forall x_i \in N$ $\|x - x_i\| > \varepsilon$

  add x to N

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



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



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

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

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

<u>Lemma 2</u>. $\forall x \in B_n$, $\exists t_1 \dots t_i$ $\quad t_i \leq \frac{1}{2^i}$ s.t.

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

<u>Pf.</u> Take N with $\varepsilon = \frac{1}{2}$.

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

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

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

Pf. (of Thm OSE):

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

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

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

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

$$x \in UN$$

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

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

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

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.

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.

Thm. Π as above is a (d, ε, δ) - OSE for

Thm. Π as above is a (d, ε, δ)-OSE for

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

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

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

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

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

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

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

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

w.p. $1 - \delta$.

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

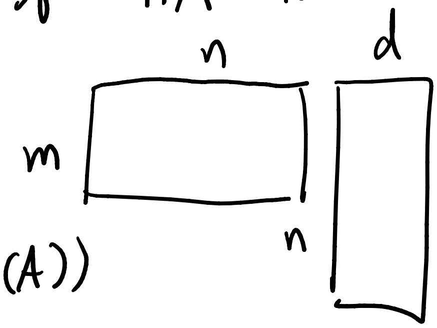
$M_r \succcurlyeq 0$ $\quad \mathbb{E} M_r = I$

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

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

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

Now the running time for ΠA is



$\cancel{m}\cdot n \cdot \frac{s}{\cancel{m}}$ $O(s\,\mathrm{nnz}(A))$

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

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

Pf of Lemma. Π_r has $\frac{sn}{m}$ nonzeros in expectation and at most $\frac{2sn}{m}$ nonzeros

(CHERNOFF: $\Pr\left(\sum X_i > \mathbb{E}(\sum X_i)(1+\varepsilon)\right) \leq e^{-c\varepsilon^2 \mathbb{E}(X_i)}$).
$|X_i| \leq 1$

with prob $1-\delta$ assuming $\mathbb{E}(X) = \frac{sn}{m} > C\cdot\log\frac{1}{\delta}$.

Let $\sigma = \Pi_r\big|_I$ $\quad I = \{i : (\Pi_r)_i \neq 0\}$. $|I| \leq 2\frac{sn}{m}$.

$_I$

$\sigma_i = \pm \frac{1}{\sqrt{s}}$ $\qquad P = (UU^T)_{I\times I}$

$$\Pi_r^T U U^T \Pi_r = \sigma^T P \sigma \qquad P \succcurlyeq 0.$$

By the H-W inequality

$$|\sigma^T P\sigma - \mathbb{E}(\sigma^T P \sigma)| \le \frac{C}{s}\left(\|P\|_F \sqrt{\log\frac{1}{\delta}} + \|P\| \log\frac{1}{\delta}\right)$$

$\|P\|_{op} \le \|UU^T\|_{op} \le 1$.

$\|P\|_F \le \sqrt{\operatorname{tr} P}$ $\qquad$ P is a $2\frac{sn}{m}$ diagonal block of UU^T

$\operatorname{tr}(UU^T) = d$

$\mathbb{E}(\operatorname{tr}(P)) = \frac{\frac{2sn}{m}}{n} \cdot d = \frac{2sd}{m}$.

W.p. $1-\delta$ $\operatorname{tr} P \le 4\frac{sd}{m}$ (Chernoff again)

Also $\mathbb{E}(\sigma^T P \sigma) = \frac{1}{s}\operatorname{tr}(P)$

So $\Pi_r^T U U^T \Pi_r \le \frac{C}{s}\left(\operatorname{tr}(P) + \sqrt{\operatorname{tr} P}\sqrt{\log\frac{1}{\delta}} + \log\frac{1}{\delta}\right)$

$$\le C\cdot\left(\frac{d}{m} + \sqrt{\frac{d}{ms}\log\frac{1}{\delta}} + \log\frac{1}{\delta}\right).$$

$m \geq C\cdot\frac{d}{\varepsilon^2}\log\frac{d}{\delta}$, $\quad s \geq \frac{C}{\varepsilon^2}\log^2\left(\frac{d}{\delta}\right)$ $\qquad \le \frac{\varepsilon^2}{\log\frac{d}{\delta}}$.

---

matrix chernoff

To prove the theorem, we can now apply <u>matrix-chernoff</u>.

Another proof (classical) of JL $(1, \varepsilon, \delta)$-OSE.

<u>Lemma.</u> $\Pi_{ij} \sim N(0, \frac{1}{m})$. Then for any fixed $x \in \mathbb{R}^n$

$$\Pr\left(\left| \|\Pi x\|^2 - \|x\|^2 \right| \geq \varepsilon \|x\|^2 \right) \leq 2 e^{-m \cdot \frac{(\varepsilon^2 - \varepsilon^3)}{4}}.$$

<u>Pf.</u> $\|\Pi x\|^2 = \sum_{r=1}^{m} (\Pi_r^T x)^2$ $\qquad Y_r \sim \Pi_r^T x \sim N(0, \sigma^2)$

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 \cdot \frac{1}{m} = \frac{\|x\|^2}{m}.$$

$$Y = \sum_{r=1}^{m} Y_r^2$$

$\mathbb{E}(Y) = \|x\|^2$ $\qquad$ Y has a Chi-squared distribution.

$$\Pr(Y > t\,\mathbb{E}(Y)) = \Pr\left(e^{\alpha Y} > e^{\alpha t \mathbb{E}(Y)}\right) \leq \frac{\mathbb{E}(e^{\alpha Y})}{e^{\alpha t \mathbb{E}(Y)}}.$$

$$\mathbb{E}(e^{\alpha Y}) = \prod_{r=1}^{m} \mathbb{E}(e^{\alpha Y_r})$$

$$\mathbb{E}(e^{\alpha Y_r}) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\alpha y^2} \cdot e^{-\frac{y^2}{2\sigma^2}} \, dy$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}\left(\frac{1}{\sigma^2} - 2\alpha\right)} \, dy$$

$(2\alpha\sigma^2 < 1)$ $\qquad = \frac{1}{\ldots} \sqrt{(1-2\alpha\sigma^2)} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2 / (1-2\alpha\sigma^2)}} \, dy$

$(2\alpha\sigma^2 < 1)$ $$= \frac{1}{\sqrt{1-2\alpha\sigma^2}} \cdot \frac{\sqrt{(1-2\alpha\sigma^2)}}{\sqrt{2\pi}\,\sigma} \int_{-\infty} e^{-\frac{y^2}{2\sigma^2}/1-2\alpha\sigma^2} \cdot dy.$$

$$= \frac{1}{\sqrt{1-2\alpha\sigma^2}}.$$

$$\therefore \Pr(Y > t\,\mathbb{E}(Y)) \le \frac{e^{-\alpha t}}{(1-2\alpha\sigma^2)^{\frac{m}{2}}} = \frac{e^{-\alpha(1+\varepsilon)}}{(1-2\frac{\alpha}{m})^m}$$

$$= \left(\frac{e^{-2\alpha(1+\varepsilon)}}{(1-2\alpha)}\right)^{\frac{m}{2}}$$

we can choose α.

$$-2(1+\varepsilon)\frac{1}{(1-2\alpha)} + \frac{2}{(1-2\alpha)^2} = 0$$

$$(1-2\alpha) = \frac{1}{1+\varepsilon} \implies \alpha = \frac{\varepsilon}{2(1+\varepsilon)}$$

$$= \left(e^{-\varepsilon}(1+\varepsilon)\right)^m$$

$$\le e^{-\frac{(\varepsilon^2-\varepsilon^3)}{4}\cdot m}.$$

Other direction is the same, except

$$\Pr(Y < t\,E(Y)) = \Pr(e^{-\alpha Y} > e^{-\alpha t E(Y)}) \text{ for } \alpha \ge 0.$$

Subspace embedding : Now

Another approach to subspace embedding : row sampling.