Geometric view of $Ax=b$. Let $a_1^T \ldots a_n^T$ be rows of A:

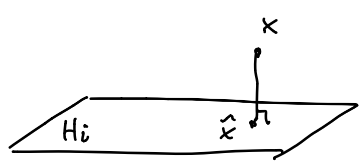
$$\begin{pmatrix} — a_1^T — \\ \vdots \\ — a_n^T — \end{pmatrix} \qquad \text{So } Ax=b \iff a_i^T x = b_i \text{ for all } i.$$

$a_i^T x = b_i$ is the equation defining a hyperplane in $\mathbb{R}^n$.

some people call it an affine subspace.

Kaczmarz: if there is an i s.t. $a_i^T x \neq b_i$,

move x to the closest $\hat{x}$ s.t. $a_i^T \hat{x} = b_i$.



$H_i = \{ y : a_i^T y = b_i \}$

Obtain by $\hat{x} = x - \gamma a_i$, for some γ

Solve for γ by $a_i^T \hat{x} = b_i \Rightarrow a_i^T (x - \gamma a_i) = b_i$

$$\Rightarrow \gamma a_i^T a_i = a_i^T x - b_i \Rightarrow \gamma = \frac{a_i^T x - b_i}{\|a_i\|^2}$$

<u>Claim</u> (not necessary, but useful)

$x - \gamma a_i$ is closest point to x on H_i.

<u>proof</u> 1: $\operatorname{dist}(x, x - \gamma a_i) = \|\gamma a_i\|_2 = \dfrac{|a_i^T x - b_i|}{\|a_i\|_2^2} \cdot \|a_i\|_2$

$$= \frac{|a_i^T x - b_i|}{\|a_i\|_2} \qquad (1)$$

whereas if $x - \delta \in H_i$,

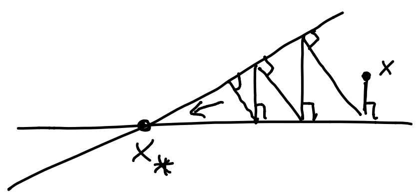
$a_i^T (x - \delta) = b_i \Rightarrow a_i^T x - b_i = a_i^T \delta$

$$\Rightarrow \quad |a_i^T x - b_i| \le \|a_i\|_2 \cdot \|\delta\|_2$$

$$\Rightarrow \quad \frac{|a_i^T x - b_i|}{\|a_i\|_2} \le \|\delta\|_2$$

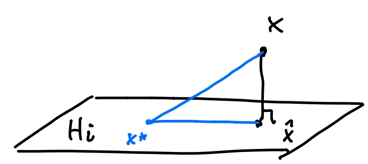
proof 2: $a_i / \|a_i\|$ is unit normal to H_i

Kaczmarz. Do this repeatedly.



Let x_* be solution to $Ax_* = b$.

How much does $\|x - x_*\|_2$ decrease?



Claim $\|x - x_*\|_2^2 = \|x_* - \hat{x}\|_2^2 + \|x - \hat{x}\|_2^2$

by Pythagorean theorem because $x_* - \hat{x} \perp x - \hat{x}$

Here is an algebraic proof: $x - \hat{x} = \gamma a_i$,

whereas $\gamma a_i^T (x_* - \hat{x}) = \gamma (b_i - b_i) = 0$

So, (i) above implies

$$\|\hat{x} - x_*\|_2^2 = \|x - x_*\|_2^2 - \frac{(a_i^T x - b_i)^2}{\|a_i\|_2^2}$$

Leads to question: how choose i.

Originally, people tried $1, 2, \dots, n$.

Strohmer & Vershynin '06

suggested a non-uniform random choice

pick i with probability proportional to $\|a_i\|_2^2$

Leads to a very clean analysis.

Recall $\|A\|_F^2 = \sum_i \|a_i\|_2^2$.

So, set prob of i to $p_i = \frac{\|a_i\|_2^2}{\|A\|_F^2}$

Then $\mathbb{E}\left[\|x - x_*\|_2^2 - \|\hat{x} - x_*\|_2^2\right]$

$$= \sum_i p_i \frac{(a_i^T x - b)^2}{\|a_i\|_2^2} = \sum_i \frac{\|a_i\|_2^2}{\|A\|_F^2} \frac{(a_i^T x - a_i^T x_*)^2}{\|a_i\|_2^2}$$

$$= \frac{1}{\|A\|_F^2} \sum_i (a_i^T x - a_i^T x_*)^2 = \frac{1}{\|A\|_F^2} \|A(x - x_*)\|_2^2$$

$$\geq \frac{\sigma_n(A)}{\|A\|_F^2} \|x - x_*\|_2^2 = \frac{1}{\|A\|_F^2 \|A^{-1}\|_2^2} \|x - x_*\|_2^2$$

Implies $\mathbb{E}\left[\|\hat{x}-x_*\|_2^2\right] \le \left(1-\frac{1}{\|A\|_F^2\|A^{-1}\|_2^2}\right)\|x-x_*\|_2^2$

If start at $x_0=0$, and x_t is tth iterate,

$$\mathbb{E}\left[\|x_t-x_*\|_2^2\right] \le \left(1-\frac{1}{\|A\|_F^2\|A^{-1}\|_2^2}\right)^t \|x_*\|_2^2$$

Time: when pick i, need to compute $a_i^T x$

and subtract αa_i from x,

which takes time $\theta(n_i)$,

where $n_i = \#$ non-zero entries in a_i.

Let's compare this to GD and its improvement CG

each step requires mult by A, which takes time $\sum_i n_i$

$$\text{and (GD): } \|A(x-x_*)\|_2^2 \le \left(1-\frac{1}{\|A\|_2^2\cdot\|A^{-1}\|_2^2}\right)^t \|x_*\|_2^2$$

$$\text{CG: } \lesssim 2\left(1-\frac{2}{\|A\|_2\|A^{-1}\|_2}\right)^{2t} \|x_*\|_2^2$$

So, CG needs fewer iterations, but each takes longer.

Sometimes Kaczmarz is a win.

(Note: most analyses of CG (conjugate gradients) state bounds in terms of $\kappa(A^TA)$, whereas we are using $\kappa(A)$)

Understanding perturbed matrices.

Consider $B = A + \varepsilon R$ where R has independent $N(0,1)$ entries   Normal / Gaussian

Recall x is $N(0,\sigma^2)$ - normal, mean 0, variance σ^2

has density $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$

If x is $N(0,1)$, εx is $N(0,\varepsilon^2)$.

And if $x(1), \dots, x(n)$ are independent $N(0,1)$,

then $t^T x = \sum_i t(i)\, x(i)$ is $N(0, \|t\|_2^2)$

Is unlikely $\kappa(B)$ is big   we prove with $n^{3/2}$

$$\Pr\left[\kappa(A+\varepsilon R) \geq \left(\|A\|_2 + \varepsilon\sqrt{n}\right)/\lambda \right] \lesssim \sqrt{n}\,\lambda/\varepsilon$$

$\|A+\varepsilon R\|_2$ is approximately $\|A\|_2$, plus $\varepsilon\sqrt{n}$.

$\sigma_n(A+\varepsilon R)$ is unlikely to be small.

For $B = \begin{pmatrix} | & & | \\ b_1 & \cdots & b_n \\ | & & | \end{pmatrix}$ define

$$\text{height}(i) = \text{dist}\left(b_i, \text{span}(b_j : j \neq i)\right)$$

$= t^T b_i$ where t is the unit normal to $\text{span}(b_j : j \neq i)$

i.e. $t^T b_j = 0$ for $j \neq i$, $\|t\| = 1$

Geometry Lem 1 $\exists i$ s.t. $\text{height}(i) \le \sqrt{n}\,\sigma_n(B)$

proof let v be s.t. $\|v\|_2 = 1$ and $\|Bv\|_2 = \sigma_n$

Let i be s.t. $|v(i)| = \|v\|_\infty$, so $|v(i)| \ge 1/\sqrt{n}$.

Now $\left\| \sum_j v(j)\, b_j \right\| = \sigma_n$

$$\left\| \sum_j \frac{v(j)}{v(i)}\, b_j \right\| = \frac{\sigma_n}{|v(i)|} \le \sqrt{n}\,\sigma_n$$

$$\|$$

$$\left\| b_i - \sum_{j \ne i} \frac{v(j)}{v(i)}\, b_j \right\| \text{ and } \sum_{j \ne i} \frac{v(j)}{v(i)}\, b_j \in \operatorname{span}(b_j : j \ne i)$$

Probability Lem 2 for each i,

$$\Pr[\text{height}(i) \le \lambda] \le \lambda/\varepsilon$$

proof of lem Sample b_j for $j \ne i$. Having fixed these,

Let t be the unit normal to $\operatorname{span}(b_j : j \ne i)$

so $\text{height}(i) = t^T b_i$

$b_i = a_i + \varepsilon \tau_i$ so $\Pr[\text{height}(i) \le \lambda] = \Pr[|t^T b_i| \le \lambda]$

$$= \Pr\left[\varepsilon t^T \tau_i \in \left[-t^T a_i - \lambda,\ -t^T a_i + \lambda\right]\right]$$

$$= \frac{1}{\sqrt{2\pi}\,\varepsilon} \int_{-t^T a_i - \lambda}^{-t^T a_i + \lambda} e^{-x^2/2\varepsilon^2}\, dx \le \frac{2\lambda}{\sqrt{2\pi}\,\varepsilon} \le \lambda/\varepsilon$$

Thm $\Pr[\sigma_n(B) \le \lambda] \le n^{3/2}\lambda/\varepsilon$

proof lem 1 $\Rightarrow$

$$\Pr[\sigma_n(B) \le \lambda] \le \Pr[\exists i \text{ s.t. } \text{height}(i) \le \sqrt{n}\,\lambda]$$

$$\le \sum_i \Pr[\text{height}(i) \le \sqrt{n}\,\lambda]$$

$$\le \sum_i \sqrt{n}\,\lambda/\varepsilon \quad , \quad \text{by lem } 2$$

$$\le n^{3/2}\lambda/\varepsilon$$

Note: Improved to $2.35\sqrt{n}\,\lambda/\varepsilon$ in SST'06
and a tight $\sqrt{n}\,\lambda/\varepsilon$ is BKMS'19

Plot it?