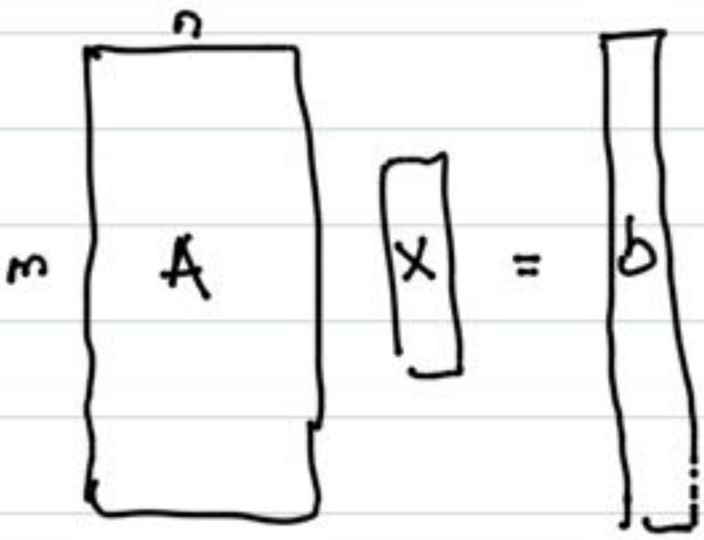
# Kaczmarz, following Strohmer, Vershynin 2006

Many modern applications of randomization are aimed at mitigating the scaling challenges that come up when doing data analysis on large datasets. Linear system solving is a basic primitive for data analysis

Ex: Tomographic imaging (CT, for example) works by sending radiation through an unknown object from a variety of vantage points and measuring the attenuation factor. An inverse problem is then solved to determine the make-up of the object. This process can be modeled as the solution of a linear system: $m$ constraints correspond to $m$ different measurements, and the $n$ unknowns describe the objects

We will consider a randomized algorithm for solving large ($m, n$ both large), overdetermined ($m > n$), consistent (there is a solution), full-rank ($A$ has rk $n$) linear systems




We model this as finding
$x^* = \text{argmin} \|Ax - b\|_2$
where
$x^*$ is the unique soln

We can solve using classical direct algorithms in time $O(mn^2)$ (say QR method), but m & n are large, so this is too expensive.

Thus we look for iterative algorithms, which do a small amount of work at each iteration to turn an estimated solution $x_k$ into a better estimate $x_{k+1}$. The idea is that we can get reasonably accurate solutions in much less time than you get the 'exact' solution using direct methods.
The goal is to design an iterative algorithm so that

$$\|x^* - x_k\|_2 \to 0 \quad \text{as } k \to \infty, \text{ fast}$$

The classical iterative alg for solving linear system is the Conjugate-Gradient method (CG).
Each iteration costs on the order of a mat-vec product, $O(mn)$ at most (less for sparse A)

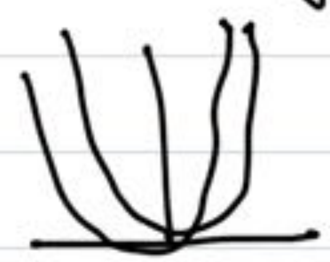
And CG satisfies a bound of the form

$$\|x^* - x_{k+1}\|_{A^TA} \leq \rho^k \|x^* - x_k\|_{A^TA}$$

where $\rho \leq \frac{K(A)-1}{K(A)+1}$. This is called exponential or linear convergence, and $\rho$ is the convergence factor.

Here $K(A)$, the condition number of A, measures the difficulty of solving this linear system

directions w/ very big ratio of curvatures
$\Rightarrow$ hard to solve & $K(A) \gg 1$



all directions have same curvatures
$\Rightarrow$ easy to solve & $K(A) = 1$

$\|x\|_{A^TA} = \|Ax\|_2^2$, so in particular

$$\|x^* - x\|_{A^TA} = \|A(x^* - x)\|_2 = \|b - Ax\|_2$$

and CG guarantees that the <u>residual error</u> decreases exponentially, and the rate depends on $K(A)$, a measure of the difficulty of the problem

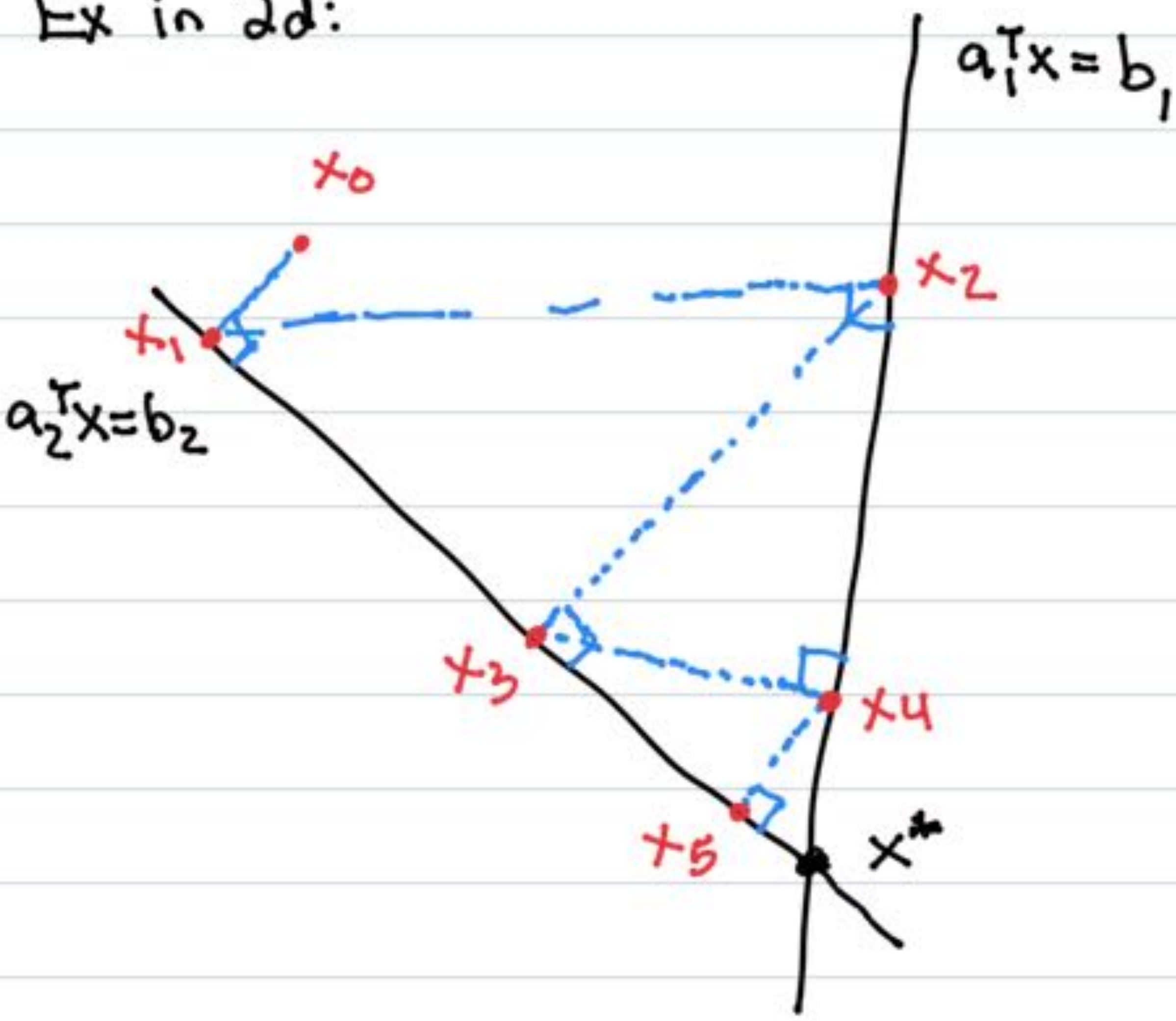
Our randomized algorithm can outperform CG in terms of convergence rate and does not require access to all of A at each iteration. In fact, at each iteration it needs access only one row of A, so it can be applied in the streaming setting.

## Kaczmarz algorithm

pick an $x_0$
for $k=0,\dots,T-1$
  pick $(a_i, b_i)$ one of the constraints, unif at random
  $$x_{k+1} = x_k - \left(\frac{a_i^T x_k - b_i}{\|a_i\|_2^2}\right) a_i$$
end
Output: $x_{k+1}$

The intuition is that we're picking a constraint at each time and modifying the current estimate so that it exactly satisfies this constraint. These constraint sets are convex, so we're projecting each estimate onto those sets. Eventually this procedure is guaranteed to converge onto their intersection. We will show this rate of convergence is linear, like CG

Ex in 2d:




To describe convergence ratio, we define some quantities:

Frobenius norm: $\|A\|_F^2 = \sum a_{ij}^2$

minimum singular value: $\sigma_n(A) = \min_x \frac{\|Ax\|_2}{\|x\|_2}$

largest row norm / $(2,\infty)$ operator norm:

$$\|A\|_{2\to\infty} = \max_{\|x\|_2=1} \|Ax\|_\infty = \max_i \|a_i\|_2^2$$

Thm

After $T$ iterations of the ~~CG~~ Kacz algorithm,

$$\mathbb{E}\|x_T - x^*\|_2^2 \le \left(1 - \frac{\sigma_n(A)^2}{m\|A\|_{2\to\infty}^2}\right)^T \|x_T - x_0\|_2^2$$

Prf

The idea is to show that one iteration decreases the approximation error by a factor of at least $\left(1 - \frac{\sigma_n^2}{m\|A\|^2_{2\to\infty}}\right)$ in Expectation.

To do so, note that

$$\|x_{k+1} - x^*\|_2^2 = \|x_{k+1} - x_k + x_k - x^*\|_2^2$$
$$= \|x_{k+1} - x_k\|_2^2 + \|x_k - x^*\|_2^2 + 2\langle x_{k+1} - x_k, x_k - x^*\rangle$$

Now recall that $x_{k+1} = x_k - \frac{(a_i^T x_k - b_i) a_i}{\|a_i\|_2^2}$

$$= x_k - \frac{a_i^T(x_k - x^*) a_i}{\|a_i\|_2^2}$$
$$:= x_k - \alpha a_i$$

so $\|x_{k+1} - x^*\|_2^2 = \alpha^2 \|a_i\|_2^2$

and $2\langle x_{k+1} - x_k, x_k - x^*\rangle = 2\langle -\alpha a_i, x_k - x^*\rangle$

$$= -2\alpha\langle a_i, x_k - x^*\rangle$$
$$= -2\alpha^2 \|a_i\|_2^2$$

So,

$$\|x_{k+1} - x^*\|_2^2 = \|x_k - x^*\|_2^2 - \alpha^2 \|a_i\|_2^2$$

$$= \|x_k - x^*\|_2^2 - \left[\frac{a_i^T(x_k - x^*)}{\|a_i\|_2^2}\right]^2 \|a_i\|_2^2$$

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

This is the key equation showing that the error decreases between iterations by an amount determined by our selection of constraints. Now we want to see what our specific sampling scheme guarantees

$$\mathbb{E}\left[\|x_{k+1} - x^*\|_2^2 \mid x_k\right] = \|x_k - x^*\|_2^2 - \sum_{i=1}^{m} \frac{1}{m\|a_i\|_2^2}(a_i^T(x_k - x^*))^2$$

$$\leq \|x_k - x^*\|_2^2 - \frac{1}{m\|A\|_{2\to\infty}^2} \sum_{i=1}^{m} (a_i^T(x_k - x^*))^2$$

$$= \|x_k - x^*\|_2^2 - \frac{1}{m\|A\|_{2\to\infty}^2} \|A(x_k - x^*)\|_2^2$$

$$\leq \|x_k - x^*\|_2^2 - \frac{\sigma_n(A)}{m\|A\|_{2\to\infty}^2} \|x_k - x^*\|_2^2$$

$$= \left(1 - \frac{\sigma_n(A)}{m\|A\|_{2\to\infty}^2}\right) \|x_k - x^*\|_2^2$$

Note that we considered the expectation over simply the selection of the $(k+1)$st constraint set. In particular, $x_k$ is still a random variable. What we have shown is

$$\mathbb{E}\left[f(x_{k+1}) \mid x_k\right] \leq \rho f(x_k)$$

where $f(x) = \|x - x^*\|_2^2$ and $\rho = \left(1 - \frac{\sigma_n(A)^2}{m\|A\|_{2\to\infty}}\right)$

By the law of total expectation,

$$\mathbb{E} f(x_{k+1}) = \mathbb{E}\left[\mathbb{E}\left[f(x_{k+1}) \mid x_k\right]\right] \leq \rho \mathbb{E} f(x_k)$$

Now we can use induction to conclude that

$$\mathbb{E} f(x_T) \leq \rho^T f(x_0)$$

∎

Note that the key point of this result is to establish the fact

$$\|x_i - x^*\|_2^2 \leq \|x_{i-1} - x^*\|_2^2 - \frac{|a_r^T(x_{i-1} - x^*)|^2}{\|a_r\|_2^2}$$

$r$ is index of the selected constraint

relating the error of the next iterate to that of the previous iterate, using knowledge of the problem itself, regression.

Randomness only comes in in the choice of the constraint (r) to project onto, and the analysis from that point involves conditional expectation and two facts:

$$\|Ax\|_2 \geq \sigma_{min}(A)\|x\|_2$$

$$\frac{1}{\|a_r\|_2^2} \geq \frac{1}{\|A\|_{2\to\infty}^2} \quad \text{for all } r \in [m]$$

One can get different (better, or worse) convergence rates by using nonuniform sampling distributions over the rows.