Approximate solutions to systems of linear equations:
Approximation, error, norms, and gradient descent.

What is an approximate solution to a problem?

An answer that is "close" to correct.

Measure close by a distance, $\|x-\tilde{x}\|$
for some norm $\|\cdot\|$.

We first think of the Euclidean norm - the standard notion of length

$$\|v\|_2 = \left(\sum_i v(i)^2\right)^{1/2} = \sqrt{v^T v}$$

Other common norms are $\|v\|_1 = \sum_i |v(i)|$ 1-norm

and $\|v\|_\infty = \max_i |v(i)|$ ∞-norm or max-norm

for $1<p<\infty$, $\|v\|_p = \left(\sum_i |v(i)|^p\right)^{1/p}$, called a p-norm.

<u>Def</u> $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is a <u>norm</u> if

a. $\|v\| \geq 0$ for all v.

b. $\|v\| = 0$ iff $v = \bar{0}$

c. For $c \in \mathbb{R}$, $\|c \cdot v\| = |c| \cdot \|v\|$

d. $\|v+w\| \leq \|v\| + \|w\|$ for all v, w
(triangle inequality)

Let's check that $\|\cdot\|_1$ and $\|\cdot\|_2$ satisfy property d.

$$\|v+w\|_1 = \sum_i |v(i)+w(i)| \le \sum_i |v(i)| + |w(i)| = \|v\|_1 + \|w\|_1$$

To show $\|v+w\|_2 \le \|v\|_2 + \|w\|_2$, will show

$$\|v+w\|_2^2 \le (\|v\|_2 + \|w\|_2)^2$$

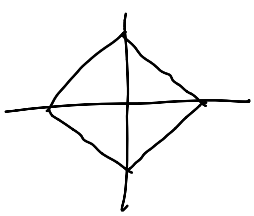
$$\iff (v+w)^T(v+w) \le \|v\|_2^2 + 2\|v\|_2\|w\|_2 + \|w\|_2^2$$

$$\iff v^Tv + 2v^Tw + w^Tw \le v^Tv + 2\|v\|_2\|w\|_2 + w^Tw$$

$$\iff v^Tw \le \|v\|_2\|w\|_2 \leftarrow \text{The Cauchy-Schwartz inequality.}$$

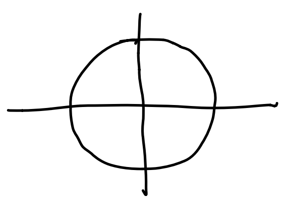
So, $\|\cdot\|_2$ is a norm is equivalent to Cauchy-Schwartz.

Understand norms by examing v s.t. $\|v\| \le 1$



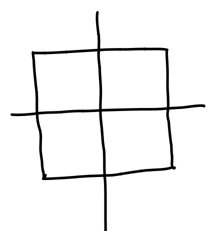
$\|v\|_1 \le 1$

Generalized Octahedron



$\|v\|_2 \le 1$

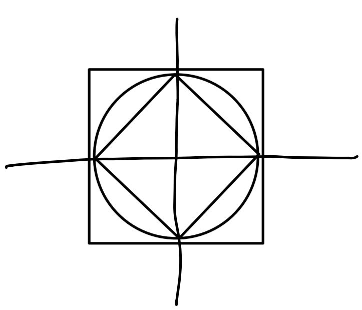
Ball



$\|v\|_\infty \le 1$

Cube

Basic relations.



$$\Rightarrow \|v\|_\infty \le \|v\|_2 \le \|v\|_1$$

proof $\|v\|_\infty^2 = \max_i v(i)^2 \le \sum_i v(i)^2 = \|v\|_2^2$

$$\|v\|_2^2 = \sum_i v(i)^2 \le \sum_i v(i)^2 + \sum_{i \ne j} |v(i)||v(j)| = \|v\|_1^2$$

Thm $\|v\|_1 \le \sqrt{n}\,\|v\|_2$, $\|v\|_2 \le \sqrt{n}\,\|v\|_\infty$

proof Let $w(i) = |v(i)|$ for all i. So, $\|w\|_1 = \|v\|_1 = 1^T w$

and $\|w\|_2 = \|v\|_2$.

$$\|w\|_1 = 1^T w \le \|1_n\|_2 \|w\|_2 \text{ by Cauchy Schwartz}$$

$$= \sqrt{n}\,\|w\|_2$$

$$\|v\|_2^2 = \sum_{i=1}^n v(i)^2 \le \sum_{i=1}^n \|v\|_\infty^2 = n\|v\|_\infty$$

For matrices. The Frobenius norm, $\|M\|_F$, treats a matrix like a vector: $\|M\|_F = \left(\sum_{i,j} M(i,j)^2\right)^{1/2}$

We often use operator norms, like

$$\|M\|_2 = \max_{x \ne 0} \frac{\|Mx\|_2}{\|x\|_2}$$

Measures how much M can increase length of a vector.

More later...

Consider problem of computing $f(x)$ $\quad x \in \mathbb{R}^m, \; f(x) \in \mathbb{R}^n$

Our code might compute an approximate solution, $\tilde{f}(x)$.

The absolute forward error is $\|f(x) - \tilde{f}(x)\|$

The relative forward error is $\dfrac{\|f(x) - \tilde{f}(x)\|}{\|f(x)\|}$

Scale error by norm of solution

The <u>absolute</u> <u>backward</u> error is

$$\min \{ \|\tilde{X} - X\| \text{ s.t. } f(\tilde{X}) = \tilde{f}(X) \}$$

The closest problem $\tilde{X}$ whose answer is $\tilde{f}(X)$.

Relative backward:

$$\min \left\{ \frac{\|\tilde{X} - X\|}{\|X\|} \text{ s.t. } f(\tilde{X}) = \tilde{f}(X) \right\}$$

<u>Example</u> Fix invertible matrix A, and let $f_A(b) = \{y : Ay = b\}$

That is, $f_A(b) = A^{-1}y$. b is playing the role of X

If our alg returns $\tilde{y}$,

forward error is $\|y - \tilde{y}\|$

backward error is $\|b - A^{-1}\tilde{y}\|$

because if $\tilde{b} = A^{-1}\tilde{y}$, $f_A(\tilde{b}) = \tilde{y}$

Advantage of backward error is that we can compute it.

To compute forward error we would need to know y.

Fast approximate solutions to $Ax=b$ by gradient descent. Assume A square (n-by-n), invertible.

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ $\quad f(x)=0$ iff $Ax=b$,

so try to minimize f.

$$f(x) = \frac{1}{2}(Ax-b)^T(Ax-b) = \frac{1}{2}x^TA^TAx - b^TAx + \frac{1}{2}b^Tb$$

$$\frac{1}{2}x^TMx - c^Tx + \frac{1}{2}b^Tb, \quad \text{for } M = A^TA \text{ and } c = A^Tb$$

note: M is symmetric

Lem $\nabla f = Mx - c$

proof $\nabla c^Tx = c$ because $\frac{\partial}{\partial x(j)} \sum_i c(i)\,x(i) = c(j)$

$\nabla x^TMx = Mx$ because $x^TMx = \sum_{i,j=1}^{n} M(i,j)\,x(i)\,x(j)$

And $$\frac{\partial}{\partial x(k)} \sum_{i,j=1}^{n} M(i,j)\,x(i)\,x(j) =$$

$$\frac{\partial}{\partial x(k)}\Big(M(k,k)\,x(k)^2 + 2\sum_{i\neq k} M(k,i)\,x(i)\,x(k)\Big)$$

$$= 2M(k,k)\,x(k) + 2\sum_{i\neq k} M(k,i)\,x(i)$$

$$= 2\sum_{i=1}^{n} M(k,i)\,x(i) = k\text{th component of } Mx$$

when $\nabla f = 0$, $\quad Mx = c \iff A^T A x = A^T b$

$\iff Ax = b$,

because A^T is invertible.

If $\nabla f \neq 0$, move in direction of ∇f

That is, move to $\hat{x} = x - \alpha (\nabla f)(x)$ for some $\alpha \in \mathbb{R}$

Will choose the α that minimizes $f(\hat{x})$

For general f this is called a line search.

For this problem, we can compute it directly.

Let $g = (\nabla f)(x)$

$$f(\hat{x}) = \frac{1}{2}(x - \alpha g)^T M (x - \alpha g) - c^T(x - \alpha g) + \frac{1}{2} b^T b$$

$$= \frac{1}{2} x^T M x - \alpha g^T M x + \frac{1}{2}\alpha^2 g^T M g - c^T x + \alpha c^T g + \frac{1}{2} b^T b$$

Is quadratic in α, so can minimize by taking deriv in α and setting it to zero.

Deriv in α is:

$$- g^T M x + c^T g + \alpha g^T M g = \alpha g^T M g - g^T (Mx - c)$$

$$= \alpha g^T M g - g^T g$$

So, set $\alpha = \dfrac{g^T g}{g^T M g}$

And, improvement is $f(x) - f(x - \alpha g)$

$$= \alpha g^T M x + \alpha g^T c - \tfrac{1}{2}\alpha^2 g^T M g$$

$$= \alpha g^T g - \tfrac{1}{2}\alpha^2 g^T M g$$

$$= \frac{1}{2}\frac{(g^T g)^2}{g^T M g} = \frac{1}{2}\frac{(g^T g)^2}{(Ag)^T (Ag)}$$, so is positive.

To get a nice expression,

<u>claim</u> $f(x) = \frac{1}{2} g^T M^{-1} g$

<u>proof</u> $g = A^T A x - A^T b \quad M = A^T A \quad M^{-1} = A^{-1}(A^T)^{-1}$

so, $g^T M^{-1} g = \left((A^T)^{-1} g\right)^T \left((A^T)^{-1} g\right)$

$$= (Ax - b)^T (Ax - b) = 2f(x).$$

So, write $$\frac{f(x) - f(\hat{x})}{f(x)} = \frac{(g^T g)^2}{(g^T M g)(g^T M^{-1} g)}$$

Or, $$f(\hat{x}) = f(x)\left(1 - \frac{(g^T g)^2}{(g^T M g)(g^T M^{-1} g)}\right)$$

$$\frac{g^T g}{g^T M g} = \frac{g^T g}{(Ag)^T (Ag)} \geq \frac{1}{\|A\|_2^2} \quad \text{and} \quad \frac{g^T g}{g^T M^{-1} g} \geq \frac{1}{\|(A^T)^{-1}\|_2^2}$$

So, $f(\hat{x}) \leq f(x)\left(1 - \frac{1}{\|A\|_2^2 \, \|(A^T)^{-1}\|_2^2}\right)$

Next lecture we will show $\|A^T\|_2^2 = \|A\|_2^2$,

and define $\kappa(A) = \|A\|_2 \, \|A^{-1}\|_2$

to be the condition number of A.

If start with x_0 and let x_t be result of t iterations,

$$f(x_t) \leq f(x_0)\left(1 - \frac{1}{\kappa(A)^2}\right)^t$$

$$\leq f(x_0) \exp\left(\frac{-t}{\kappa(A)^2}\right), \quad \text{as } 1-z \leq \exp(-z)$$

This algorithm is fast if $\kappa(A)$ is small.

\# operations per iteration $\approx$ # nonzeros in A.

Standard alternative is Gaussian elimination, which takes time $\sim n^3$ (or $n^{2.37\ldots}$)

Note: The Conjugate Gradient is an improvement of this algorithm that makes improvement like $\left(1 - \frac{1}{\kappa(A)}\right)^t$, which is much better.

Bounds for these are usually stated in terms of $\kappa(M) = \kappa(A)^2$.