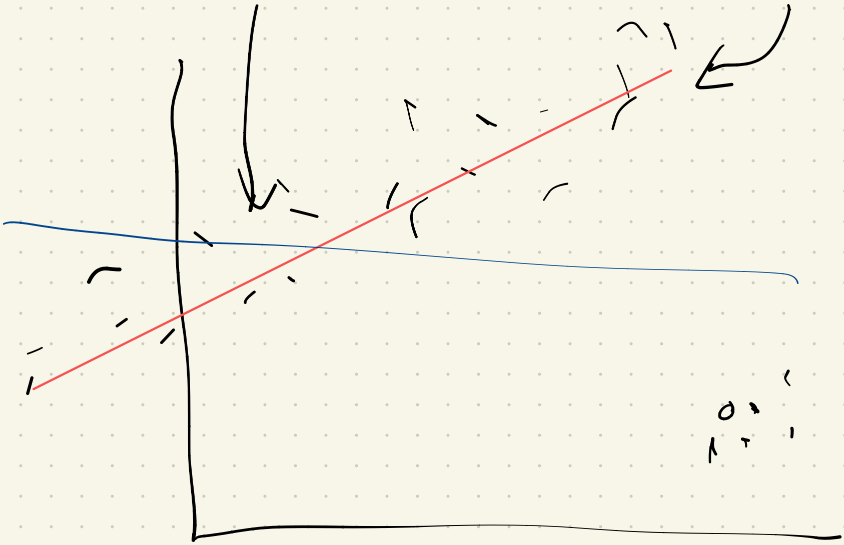
Amath 515: Lecture 2

Jan 6th

* Calc prelims
* Necessary cond for opt
* Start on convex sets.

Last time: two examples of regression
Linear

① Gaussian ② Cauchy



linear

In both models, $y_i = \boxed{x_0 + x_1 a_i} + \boxed{\varepsilon_i}$

① ②

Density fcn for choices of ε_i influenced our objective fcns:

For 1 we wrote

$$f(x) = \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - x_0 - x_1 a_i)^2$$

For 2 we wrote

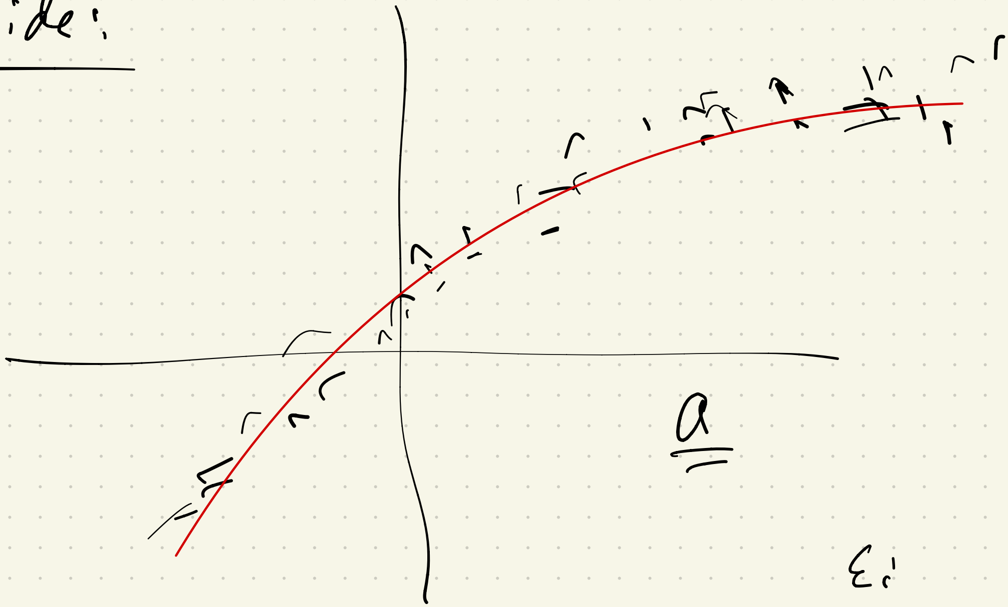
$$f(x) = \sum_{i=1}^{m} \ln\left(1 + (y_i - x_0 - x_1 a_i)^2\right)$$

Defined $A = \begin{bmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_m \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$

Then $f(x)$ in 1 is equal to

$$\frac{1}{2\sigma^2} \|Ax - y\|^2, \quad x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$$

Aside:



$$y_i = \underline{x_0 + x_1 a_i + x_2 a_i^2} + \underline{noise}$$

$$A = \begin{bmatrix} 1 & a_1 & a_1^2 \\ \vdots & \vdots & \vdots \\ 1 & a_m & a_m^2 \end{bmatrix}, \quad x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$$

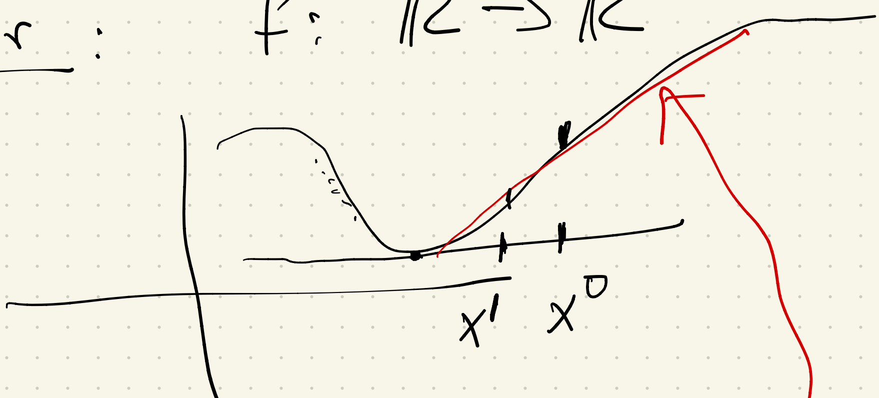
$$\frac{1}{2\sigma^2} \| Ax - y \|^2$$

* More generally, can also do splines (try to show you)

Today:

* Review of some calculus & preliminaries
* Necessary conditions for optimality.

Primer: $f: \mathbb{R} \to \mathbb{R}$



Can write $f(x) \approx \underline{\underline{f(x^0) + f'(x^0)(x-x^0)}}$

+ higher order terms

To make f smaller, try

$\underline{\underline{x}} - x^0 \approx -\underline{\underline{\alpha}}\, f'(x^0)$

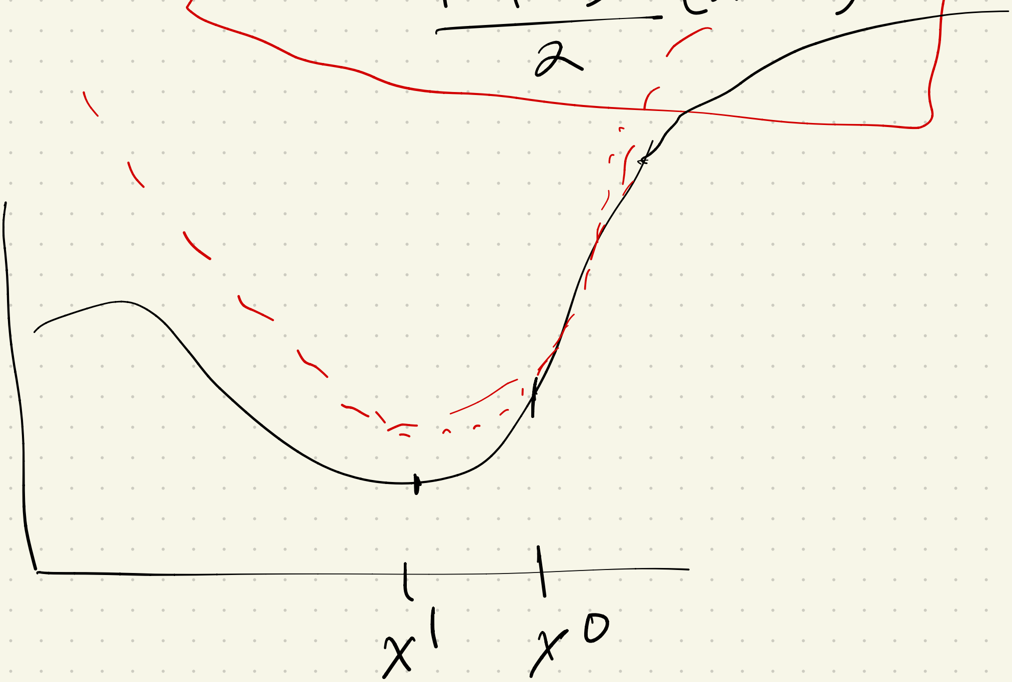
(Later)

$$f(x) \approx f(x^0) - \alpha \left(f'(x)\right)^2 + \text{higher order}$$

$$x^1 = x_0 - \alpha f'(x^0) \quad \text{steepest descent}$$

Can also write

$$f(x) \approx f(x^0) + f'(x^0)(x-x^0) + \frac{f''(x^0)}{2}(x-x^0)^2$$



Can minimize quadratic wrt x

Take deriv and set it equal to 0

$$f'(x^0) + f''(x^0)(x - x^0) = 0$$

Solve for $\underline{\underline{x}}$

$$\boxed{x^1 = x^0 - \frac{f'(x^0)}{f''(x^0)}}$$

Newton's method.

<u>Def:</u> $r(t)$ is called $o(t)$ (little o of t) if

$$\lim_{t \to 0} \frac{r(t)}{t} = 0$$

<u>ex:</u> t^2 is $o(t)$, $\frac{t^2}{t} = t$

$t^{1.1}$ is $o(t)$ $\frac{t^{1.1}}{t} = t^{0.1}$

Def: A function $f: \mathbb{R}^n \to \mathbb{R}$ is <u>differentiable</u> at a point x if there exists a vector $v \in \mathbb{R}^n$ with

$$f(x+h) = f(x) + v^T h + o(\|h\|)$$

for all $h \in \mathbb{R}^n$

In this case, we call v the <u>gradient</u> of f at x, written $\underline{\nabla f(x)}$, think as a column vector.

Ex: $f(x) = x_1 + x_2 x_3 + \sin(x_4)$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_4} \end{bmatrix} = \begin{bmatrix} 1 \\ x_3 \\ x_2 \\ \cos(x_4) \end{bmatrix}$$

$$v^T h = \langle v, h \rangle = \sum_{i=1}^{n} v_i h_i$$

Euclidean inner product

$$\langle h, h \rangle = h^T h = \sum h_i^2 = \underline{\underline{\|h\|_2^2}}$$

<u>Def:</u> A function $F: \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at x if $\exists$ a matrix $\underline{\underline{J}}$ satisfying $^{m \times n}$

$$F(x+h) = F(x) + Jh + o(\|h\|^{\cancel{2}})$$

for any $h \in \mathbb{R}^n$

In this case J is called the <u>Jacobian</u>, written $\nabla F(x)$

$$f(x+h) \approx f(x) + \underline{\nabla f(x)}^{T} h + \dots$$

$$F(x+h) = F(x) + \underline{\nabla F(x)}\, h + \dots$$

$$\nabla F(x) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix}$$

Def: $f : \mathbb{R}^n \to \mathbb{R}$ is twice diff'ble at x

when ∇f is itself differentiable at x, so

$$\nabla f(x+h) = \nabla f(x) + \boxed{\nabla^2 f}\, h + o(\|h\|)$$

Jacobian of gradient is called the Hessian, $\nabla^2 f(x)$

Recall $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$

So $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$

If $x \to \nabla^2 f(x)$ is continuous, we have equality of mixed partials (order doesn't matter).

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & & \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1^2} \end{bmatrix}$$

when $x \to \nabla^2 f(x)$ cont, $\nabla^2 f(x)$ is <u>symmetric</u>, $\nabla^2 f^T = \nabla^2 f$

A^T has its columns as rows of A

$$A = \begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix}, \quad A^T = \begin{bmatrix} — a_1^T — \\ — a_2^T — \end{bmatrix}$$

def A symmetric matrix A is called <u>positive semidefinite</u> when $x^T A x \geq 0$ for any x, equivalently all <u>eigenvalues</u> of A are ≥ 0.

and is called <u>positive definite</u> when $x^T A x > 0$ for $x \neq 0$, or equivalently all eigenvalues are <u>positive</u>.

When f is twice diff'ble at x, we have

$$f(x+h) = \underline{f(x) + \nabla f(x)^T h} \\ \underline{+ \frac{1}{2} h^T \nabla^2 f(x) h} \\ + o(\|h\|^2)$$

Necessary conditions for optimality

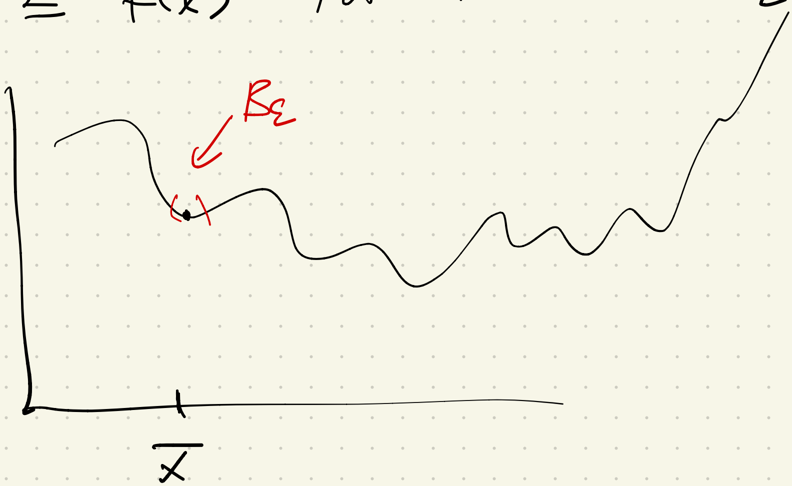
Consider the problem $\underset{x}{\min}\, f(x)$, $f: \mathbb{R}^n \to \mathbb{R}$

f differentiable everywhere.

Def: $\bar{x}$ is a local minimum for f

if $\exists$ a small ball B_ε, with

$f(\bar{x}) \leq f(x)$ for all $x \in B_\varepsilon$.



Thm: If $\bar{x}$ is a local min for f, then $\nabla f(\bar{x}) = 0$.

Pf: $f(\bar{x}+tv) = f(\bar{x}) + t\nabla f(\bar{x})^T v + o(t)$

Consider $v = -\nabla f(\bar{x})$

$$f(\bar{x} - t\nabla f(\bar{x})) = f(\bar{x}) - t\|\nabla f(\bar{x})\|^2 + o(t)$$

$$0 \leq \frac{f(\bar{x} - t\nabla f(\bar{x})) - f(\bar{x})}{t}$$ for small enough t

$$= -\frac{\cancel{t}\|\nabla f(\bar{x})\|^2}{\cancel{t}} + \frac{o(t)}{t}$$

Take limit $t \to 0$

$$0 \leq -\|\nabla f(\bar{x})\|^2 = 0$$

So $\nabla f(\bar{x}) = 0$ ✓

Thm 2.8: Second order $\boxed{\text{necessary}}$ + sufficient conditions.

$\hookrightarrow$ book with index

If f is twice differentiable, then

(a) $\bar{x}$ a local min for f $\Rightarrow$

$\nabla f(\bar{x}) = 0$, and $\nabla^2 f(\bar{x})$ is positive semidefinite.

(done)

(B) $\nabla f(\bar{x}) = 0$, $\nabla^2 f(\bar{x}) \succeq 0$ $\Rightarrow$ $\bar{x}$ local min.

Pf of (a):

By thm 2.7, $\nabla f(\bar{x}) = 0$

$$f(\bar{x} + tv) = f(\bar{x}) + 0 + \frac{1}{2} t^2 v^T \nabla^2 f(\bar{x}) v + o(t^2)$$

$0 = \nabla f(\bar{x})^T v$

for all v

Again $0 \leq \dfrac{f(\bar{x}+tv) - f(\bar{x})}{t^2}$ for small t

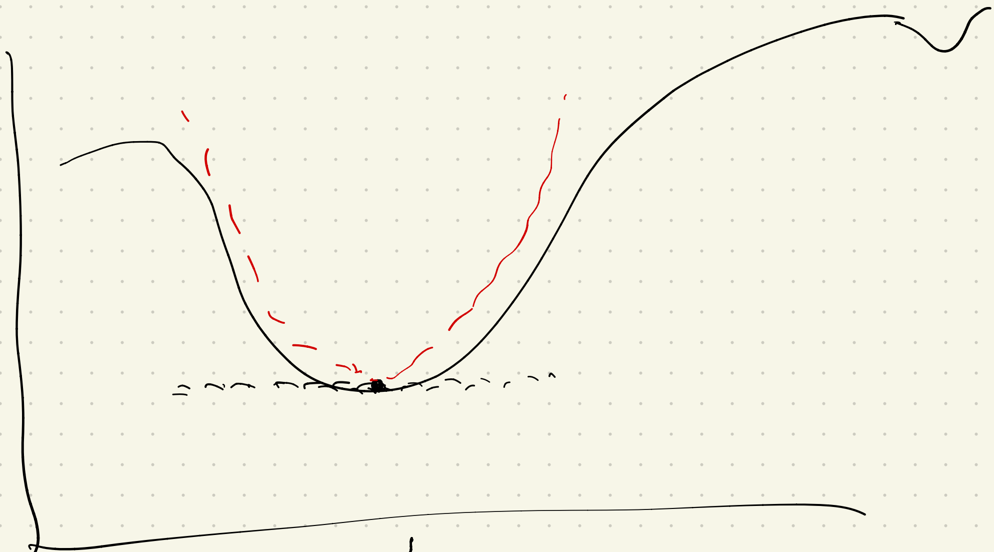
$$= \frac{1}{2} \cancel{t^2} v^T \nabla^2 f(\bar{x}) v + \frac{o(t^2)}{t^2}$$

$$0 \leq \frac{1}{2} v^T \nabla^2 f(\bar{x}) v + \frac{o(t^2)}{t^2}$$

take $t \to 0$, $0 \leq v^T \nabla^2 f(\bar{x}) v$

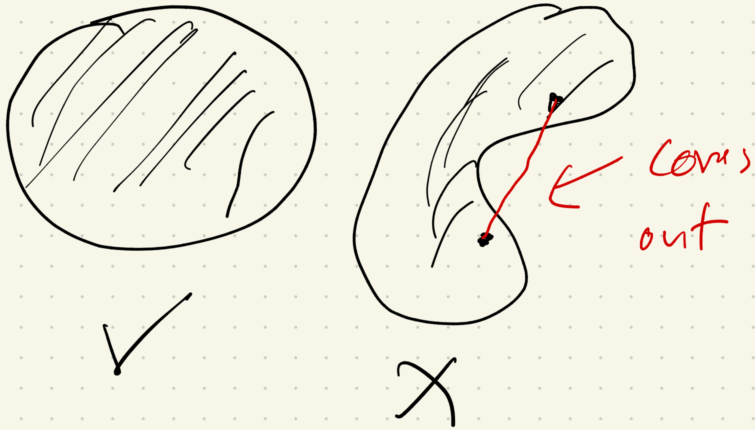
for all v

This is def of pos semidef.



Convex sets & functions.

Sets:



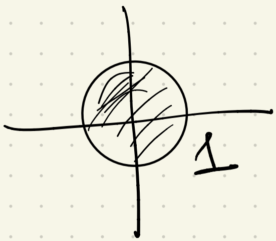
A set C is <u>convex</u> if it contains all line segments between any two points in C.

<u>Algebraically</u>, If $x, y \in C$ then the segment $\underbrace{\lambda x + (1-\lambda) y, \ \lambda \in (0,1)}_{\text{a convex combination}}$, is contained in C.

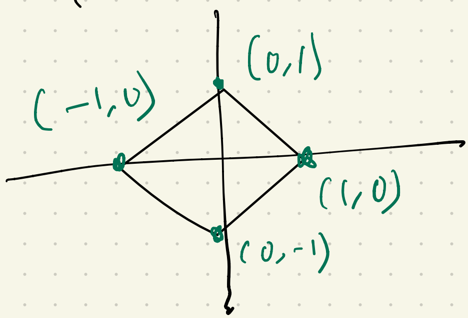
Planes, lines, half-spaces are convex.

Norm balls are convex.

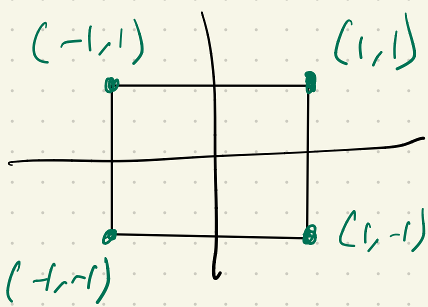
① $\|x\|_2 = \sqrt{x^T x}$, $\mathbb{B}_2 = \{x : \|x\|_2 \le 1\}$



② $\|x\|_1 = \sum_{i=1}^{n} |x_i|$, $\mathbb{B}_1 = \{x : \|x\|_1 \le 1\}$



③ $\|x\|_\infty = \max_i |x_i|$, $\mathbb{B}_\infty = \{x : \|x\|_\infty \le 1\}$



Ex: An intersection $C_1 \cap C_2$ is convex when C_1, C_2 convex.