Convex Functions
Showing a function is convex
Gradient Methods

- No tutorial this week.
- A4 out tomorrow, due Oct 15th.
- return course eval on monday.

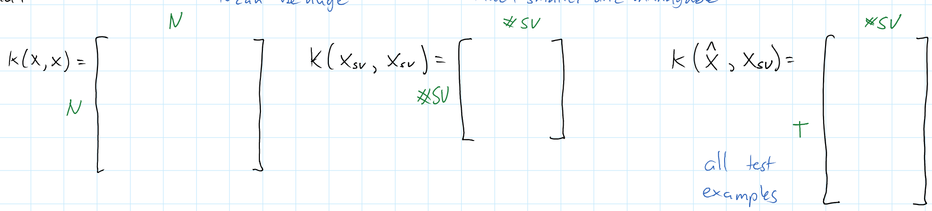
- why do support vectors make methods more efficient?

Usual

N can be huge

Much smaller and managable

much better than N



Convex Functions (zero-order definition)

- solve in poly time

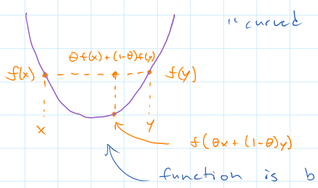
Function "f" is <u>convex</u> if dom(f) is convex, and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

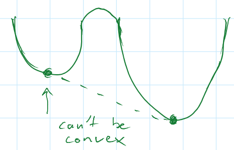
"below the chord", "epigraph is convex set"

"curved upwards"

Examples $f(x) =$
- $\exp(ax)$
- $x \log(x)$, $x > 0$
- $ax^2 + bx$, $a \geq 0$
- $\max_i x_i$
- $\log(\sum \exp(x_i))$
- $-\log(x)$



- Implies local minima are global minima.
(and global minima are convex set)



Strictly convex: "strict inequality"

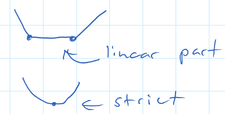
$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$

- implies at most one global minimum

"concave": $(-f)$ is convex

"log convex": $(\log(f(x)))$ is convex

linear is not strict.



negative log : turns concave

$p(y|x)$: log concave

$\log p(y|x)$: concave

NLL : $-\log p(y|x)$: convex

Convex Functions (first order definition)

- If 'f' is differentiable, equivalent to (everywhere)

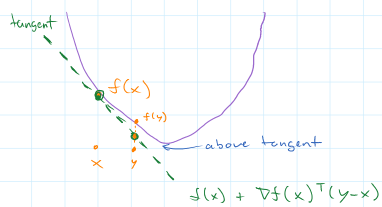
$$f(y) \geq f(x) + \nabla f(x)^T (y-x), \quad \forall y,x \in \text{dom}(f)$$

" f is above its tangent "

- can see that any stationary point $(\nabla f(x^*) = 0)$ is a global optimum,

$$f(y) \geq f(x^*) + 0, \quad \forall y$$

all other points larger than stationary pt, x^*



Convex Functions (second order definition)

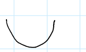
- If 'f' is twice differentiable, equivalent to

$$\nabla^2 f(x) \succeq 0, \quad \forall x \quad (f''(x) \geq 0, \text{ for all } x \in \mathbb{R})$$

" f is curved upward everywhere"

second derivative is positive,
Hessian positive semi definite

- usually the best way to show a simple function is convex

Ex1 $f(x) = x^2$  convex

$f'(x) = 2x$ 

$f''(x) = 2$ ≥ 0, convex

Ex2 $f(x) = \frac{1}{2} x^T A x + b^T x + c$, $A \succcurlyeq 0$

$f'(x) = Ax + b$

$f''(x) = A \succcurlyeq 0$, convex

Ex3 $f(x) = \frac{1}{2} \| Ax - b \|^2$

$\nabla f(x) = A^T A x - A^T b$

$\nabla^2 f(x) = A^T A$ "matrix norm"

⊛ $= \sum_{i=1}^{N} a_i a_i^T$ rank 1 matrix.

$x^T A x \succcurlyeq 0, \quad \forall x$ positive semi definite definition

$\Rightarrow x^T \left(\sum_{i=1}^{N} a_i a_i^T \right) x$

$\Rightarrow \sum_{i=1}^{N} \underbrace{(x^T a_i)}_{\text{scalar}} \underbrace{(a_i^T x)}_{\text{scalar}} \succcurlyeq 0$

"bunch of numbers, squared"

Operations that preserve convexity

let f_1, f_2 be convex functions

1. Non-negative weighted sum: $w_1 f_1(x) + w_2 f_2(x)$ $\quad w_1, w_2 \geq 0$
2. Composition with Affine function: $f_1(Ax + b)$
3. Pointwise Maximum: $\max\{f_1(x), f_2(x)\}$

⊛ other composition rules exist

Show that SVMs are convex

$$\underbrace{\tfrac{1}{2} \|w\|^2}_{f_1} + \underbrace{c \sum_{i=1}^{N} \max\{0, 1 - y_i \bar{w}^T \bar{x}_i\}}_{f_2}$$

① $f_1 + f_2$

$f_1 = w^T w = w^T I w$ ✓ good. similar to: $w^T A w$, $A = I$, $I \succcurlyeq 0$

$\nabla^2(w^T w) = 2I$ eigen are $\Rightarrow 1$

$f_2 = \max\{0, 1 - y_i \bar{w}^T \bar{x}_i\}$ ignore $c \sum_{i=1}^{N}$, as per ① again

③ $f_1 = 0$ ✓ good

$f_2 = 1 - y_i \bar{w}^T \bar{x}_i$ ✓ good (linear)

Gradient Method

- minimize convex 'f'

  ie MLE, MAP, SVM

- Generate a sequence

  $x^1, x^2, x^3 \ldots$

  such that $x^t \longrightarrow x^*$ as $t \to \infty$

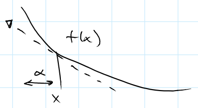
$\operatorname{argmin}_x f(x) = x^*$ (optimizer)

- Gradient Method

$$x^{t+1} = x^t - \alpha^t \nabla f(x^t)$$

step size (small)

gradient descent
steepest descent



∇ is negative

Stochastic Gradient Method

$$\underset{x}{\text{argmin}} \; \frac{1}{N} \sum_{i=1}^{N} f_i(x) \quad = f(x)$$

applies to all models we've looked at (except knn)

$$x^{t+1} = x^t - \alpha^t \underbrace{\nabla f_i(x^t)}$$

replacement ok

picked randomly in $i \in \{1 : N\}$

"zone of confusion": if you're far away, all points probably go in the same direction.

$$E[\nabla f_i(x^t)] = \sum_{i=1}^{N} p(i) \cdot \nabla f_i(x^t)$$

$p(i) = 1/N$ pick any with equal likelyhood

$$= \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x^t)$$

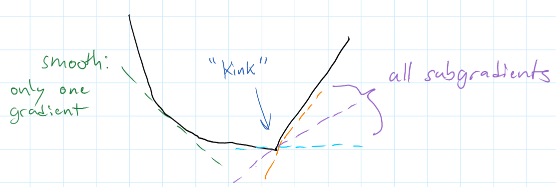
"on average, implements the gradient method"

notes: step size should go to zero, or you can oscillate

Subgradient

- a "subgradient" of 'f' at 'x' is a vector 'd' such that

$$f(y) \geq f(x) + d^T (y - x)$$



- Set of subgradients is the "sub-differential"

$$\partial f(x)$$

Ex: $|x|$

$$\partial |x| = \begin{cases} \{1\} & x > 0 \\ \{-1\} & x < 0 \\ [-1, 1] & x = 0 \end{cases}$$

zero is in here

$\{1\}$: only one gradient, as it is differentiable

$[-1$ to $1]$: sub-differential

x^* is Global optimum if $0 \in \partial f(x^*)$