Convex learning problems

- Let X be a feature space.
- Let Y be a space of labels.
- Let H be a space of predictors.
- Let $\ell : H \times X \times Y \to \mathbb{R}$ be a loss function:

$$\ell(h, x, y) = \text{loss of } h \text{ on example } (x, y)$$

- We saw some examples

1) Zero-one loss

$$\ell(h, x, y) = \mathbb{1}[h(x) \neq y]$$

2) Square loss

$$\ell(w, x, y) = (w^T x - y)^2$$

3) Logistic loss

$$\ell(w, x, y) = \log\left(1 + e^{-y w^T x}\right)$$

- If $S = ((x_1, y_1), \ldots, (x_m, y_m))$ is a labeled example, we define <u>empirical loss</u>

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, x_i, y_i)$$

- If D is a distribution over $X \times Y$, we define <u>generalization error</u>

$$L_D(h) = \mathop{\mathbb{E}}_{(x,y) \sim D} \left[\ell(h, x, y) \right]$$

- $L_S(h)$, $L_D(h)$ are generalizations of $\widehat{err_S}(h)$ and $err_D(h)$

- Often instead h, we use w. Especially, when $w \in \mathbb{R}^d$ is a "weight vector".

- The notions of PAC, Agnostic PAC, non-uniform learnability can be generalized to arbitrary loss functions.

- ERM algorithm

$$\hat{h} = \underset{h \in H}{\operatorname{argmin}} \; L_S(h)$$

- Our goal will be to find minimizer of $L_D(h)$.

- As usual, we have only access to i.i.d. sample S from D

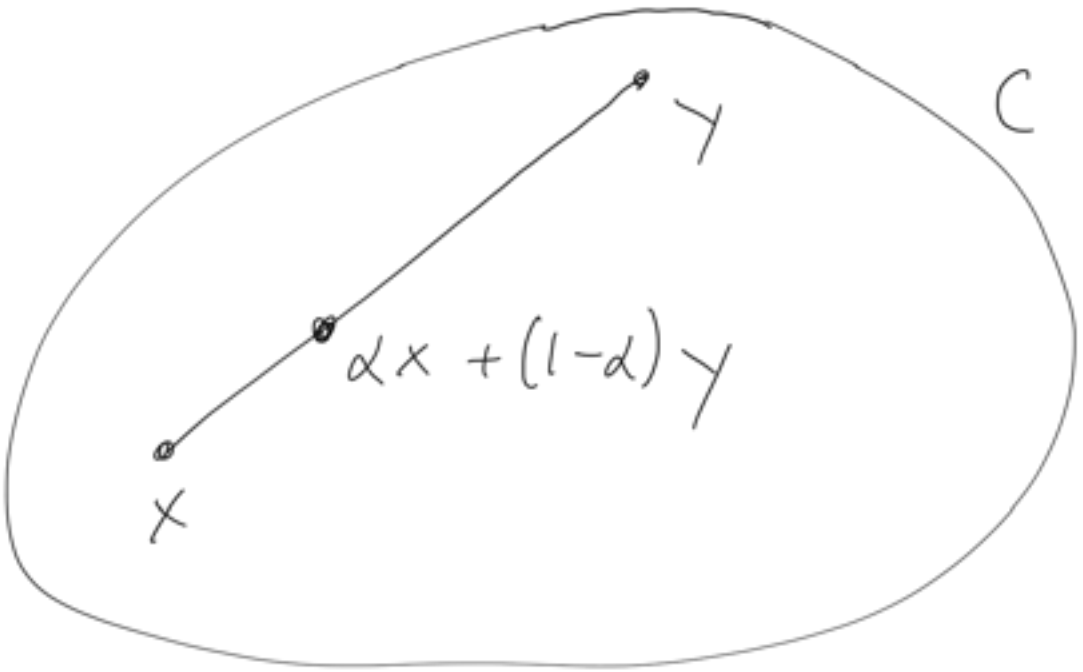
- we will look at convex loss functions
- Convex loss functions have a lot of useful properties
  1) ERM is computationally easy
  2) It is possible to find $h \in H$ with "low" generalization error.

Convex analysis

Definition: A set $C \subset \mathbb{R}^d$ is

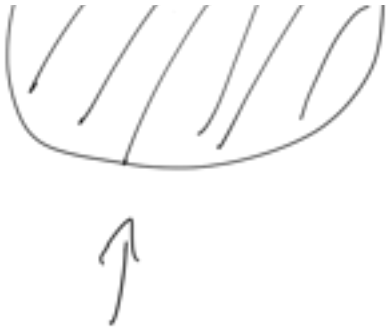
called convex if for every $x, y \in C$ and every $\alpha \in [0,1]$

$$\underbrace{\alpha x + (1-\alpha) y}_{\text{Convex combination of } x, y} \in C$$



Examples:

convex



not convex

- Ball of radius r centered at $c \in \mathbb{R}^d$ is convex:

$$B_{r,c} = \{ x \in \mathbb{R}^d : \|x - c\| \leq r \}$$

- Half-space is convex:

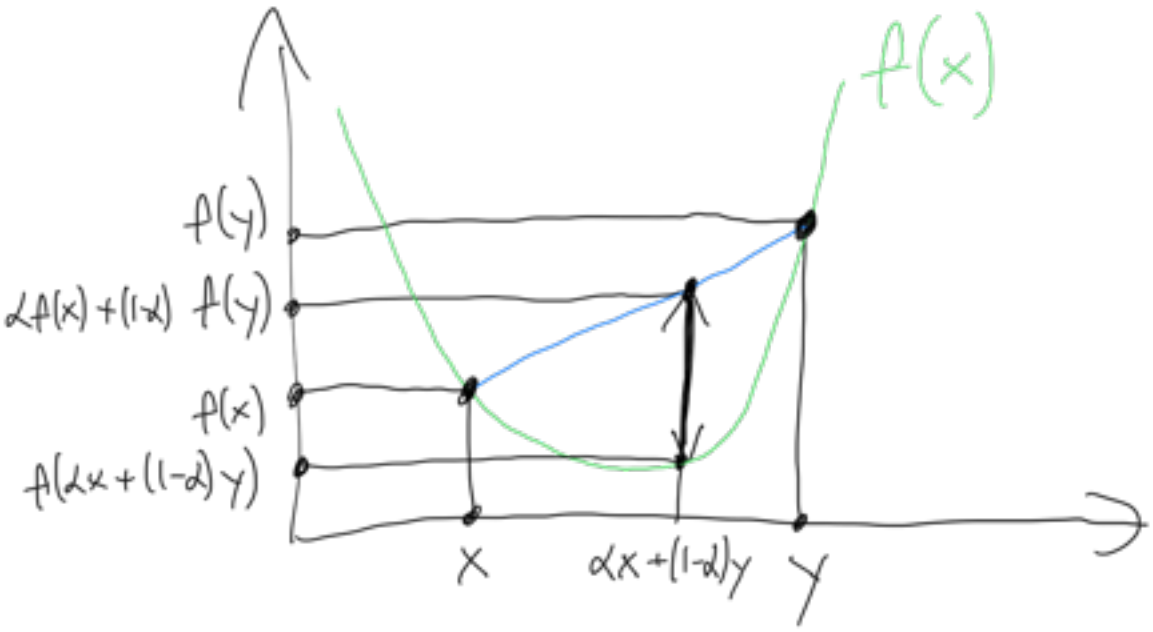
$$H_{w,b} = \{ x \in \mathbb{R}^d : w^T x \geq b \}$$

- $\mathbb{R}^d$ is convex

Definition: Let $C \subseteq \mathbb{R}^d$ be

a convex set. A function $f: C \to \mathbb{R}$ is called <u>convex</u> if for every $x, y \in C$ and every $\alpha \in [0,1]$,

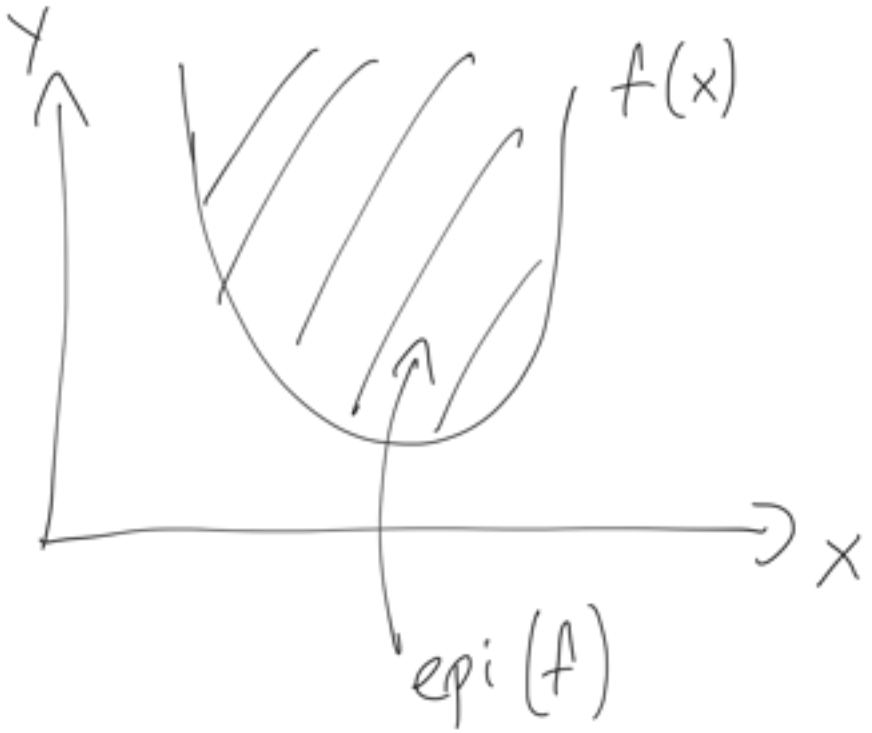
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y).$$



<u>Definition</u>: Let $C \subseteq \mathbb{R}^d$ and $f: C \to \mathbb{R}$. <u>Epigraph</u> of f

is the subset of $\mathbb{R}^{d+1}$

$$epi(f) = \{(x, y) \in C \times \mathbb{R} : y \geq f(x)\}$$



Claim: Let $C \subseteq \mathbb{R}^d$ and $f: C \to \mathbb{R}$. The function f is convex if and only if $epi(f)$ is convex.

Proof:

$\Rightarrow$:

- Suppose $f: C \to \mathbb{R}^d$ is convex.

- By definition, C is also convex
- Let $(x, y), (x', y') \in \text{epi}(f)$
- Thus $y \geq f(x)$ and $y' \geq f(x')$
- Let $\alpha \in [0, 1]$
- $\alpha y + (1-\alpha) y' \geq \alpha f(x) + (1-\alpha) f(x')$

adding the two inequalities

$\geq f(\alpha x + (1-\alpha) x)$

by convexity of f

$\Leftarrow$:

- Suppose $\text{epi}(f)$ is convex.

- Suppose $x, x' \in C$ and $\alpha \in [0,1]$
- $(x, f(x)),\ (x', f(x')) \in epi(f)$
- Since $epi(f)$ is convex,

$$(\alpha x + (1-\alpha)x',\ \alpha f(x) + (1-\alpha) f(x')) \in epi(f)$$

- Thus, by definition of $epi(f)$

$$\alpha f(x) + (1-\alpha) f(x') \geq f(\alpha x + (1-\alpha)x').$$

Claim: Let $C \subseteq \mathbb{R}^d$ be a convex set and $f: C \to \mathbb{R}$ be a convex function.

If $x \in C$ is a local minimum

of f the x is also a global minimum of f.

Proof:

- By contradiction, suppose x is not a global minimum.
- There exists $x' \in C$ such that $f(x') < f(x)$.
- Since x is bcal minimum, there exists a ball $B_{x,r}$ such that for all $x'' \in B_{x,r}$

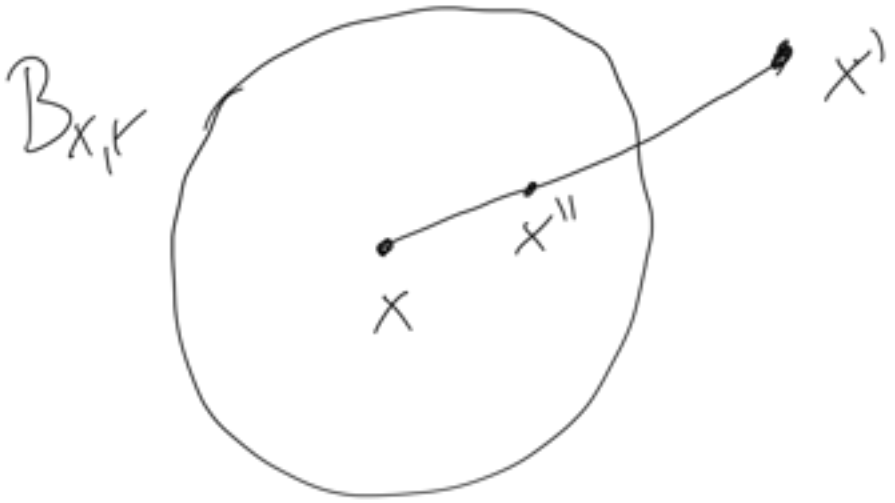
$$f(x'') \geq f(x).$$

- There exist $\alpha \in (0,1)$

such that

$$\alpha x + (1-\alpha)x' \in B_{x,r}$$

- Let $x'' = \alpha x + (1-\alpha)x'$



- Since f is convex

$$f(x'') = f(\alpha x + (1-\alpha) f(x'))$$

$$\leq \alpha f(x) + (1-\alpha) f(x')$$

$$< \alpha f(x) + (1-\alpha) f(x)$$

$$< f'(x)$$

since $\alpha \in (0,1)$ and $f(x') < f(x)$.

- Thus $f(x'') < f(x)$ which contradicts local optimality of X.

∎

Claim: Let $C \subseteq \mathbb{R}^d$ be a convex set. Let $f: C \to \mathbb{R}$ be a convex differentiable function. For any $x \in \text{int}(C)$ and any $y \in C$

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

Proof:

- For any $\alpha \in [0, 1)$

$$(1-\alpha)x + \alpha y \in \text{int}(C)$$

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

- Let $g: [0, 1] \to \mathbb{R}$

$$g(\alpha) = f((1-\alpha)x + \alpha y) - (1-\alpha)f(x) - \alpha f(y)$$

- We have

$$g(0) = 0$$

$$g(\alpha) \leq 0$$

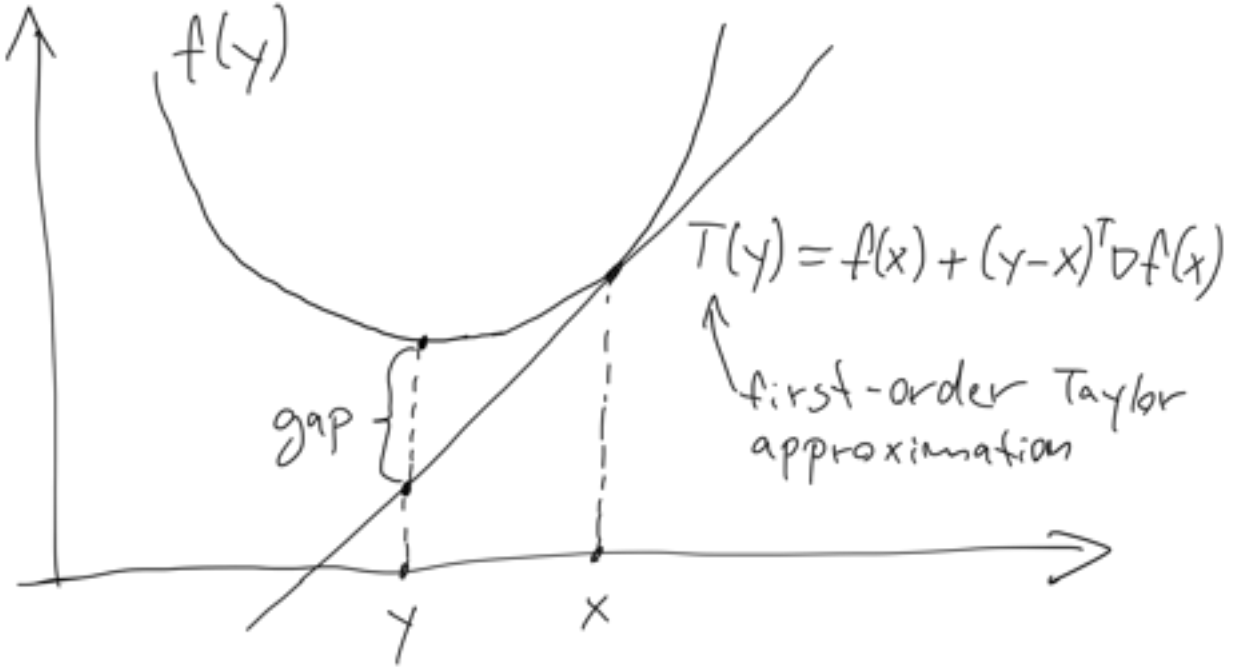
- Therefore $g'(0) \leq 0$

- $g'(\alpha) = (y-x)^T \nabla f((1-\alpha)x + \alpha y)$

$$+ f'(x) - f(y)$$

- Thus for $\alpha = 0$

$$(y-x)^T \nabla f(x) + f(x) - f(y) \leq 0.$$



Lemma:

Let $I \subseteq \mathbb{R}$ be an interval.

Let $f: I \to \mathbb{R}$ be a function.

The following are equivalent

1) f is convex

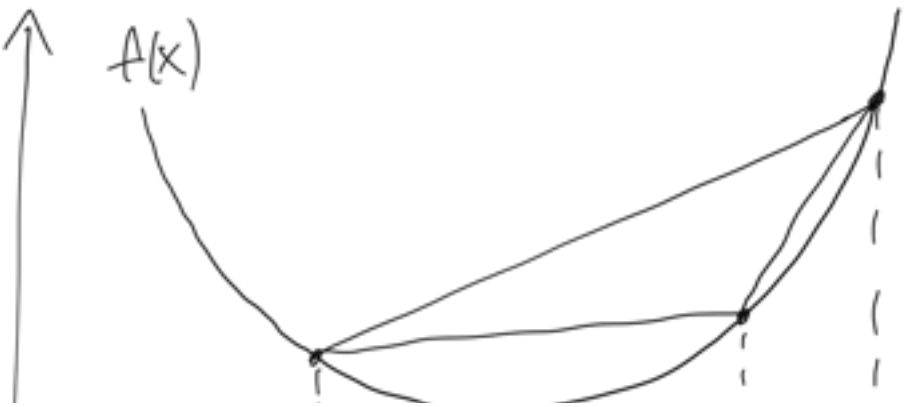
2) For any $x_1, x_2, x_3 \in I$, such that $x_1 < x_2 < x_3$

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_1)}{x_3 - x_1}$$

3) For any $x_1, x_2, x_3 \in I$, such that $x_1 < x_2 < x_3$

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2}$$

Proof:



- Convexity of f is equivalent to:

  For any $x_1, x_2, x_3 \in I$ such that $x_1 < x_2 < x_3$,

  $$f(x_2) \leq \alpha f(x_1) + \beta f(x_3) \quad (*)$$

  where $x_2 = \alpha x_1 + \beta x_3$

  and $\alpha = \dfrac{x_3 - x_2}{x_3 - x_1}$, $\beta = \dfrac{x_2 - x_1}{x_3 - x_1}$

- The condition $(*)$ is equivalent to

$$f(x_2) \leq \frac{x_3 - x_2}{x_3 - x_1} f(x_1) + \frac{x_2 - x_1}{x_3 - x_1} f(x_2)$$

- That is equivalent to

$$(x_3 - x_1) f(x_2) \leq (x_3 - x_2) f(x_1) + (x_2 - x_1) f(x_3)$$

$(x_3 - x_1) f(x_2) \leq (x_3 - x_2) f(x_1) + (x_2 - x_1) f(x_3)$

- That is equivalent to

a)

$$(x_3 - x_1) f(x_2) - (x_3 - x_1) f(x_1) \leq (x_2 - x_1) f(x_3) - (x_2 - x_1) f(x_1)$$

and also

b)

$$(x_3 - x_2) f(x_2) - (x_3 - x_2) f(x_1) \leq (x_2 - x_1) f(x_3) - (x_2 - x_1) f(x_2)$$

- a) $\Leftrightarrow$ 2) and b) $\Leftrightarrow$ 3)

Lemma:

Let $I \subseteq \mathbb{R}$ be an open interval.

Let $f: I \to \mathbb{R}$ be a twice-differentiable function.

Then, the following are equivalent:

1) f is convex
2) f' is non-decreasing
3) f'' is non-negative

Proof:

1) $\Rightarrow$ 2)

- Let $x_1 < x_2 < x_3 < x_4$

- Then

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2} \leq \frac{f(x_4) - f(x_3)}{x_4 - x_3}$$

- Ignore the middle fraction

- Let $x \to x^+$

and let $x_2 \to x_1^+$

and let $x_3 \to x_4^-$

- $f'(x_1) \leq f'(x_4)$

2) $\Rightarrow$ 1)

- Let $x_1 < x_2 < x_3$
- By mean value theorem there exists u, v such that

$$x_1 < u < x_2 < v < x_3$$

and

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(u)$$

and

$$\frac{f(x_3) - f(x_2)}{x_3 - x_2} = f'(v)$$

- Since f' is non-decreasing,
  $f'(u) \leq f'(v)$

- Thus

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_1}$$

- Thus f is convex.

3) $\Rightarrow$ 2) :

- If $x < y$ there exists
  u such that $x < u < y$

$$f''(u) = \frac{f'(y) - f'(x)}{y - x}$$

- Since $f''(u) \geq 0$ and $y - x > 0$,
  we must have $f'(y) - f'(x) \geq 0$.

3) $\Rightarrow$ 2) :

- $f''(x) = \lim_{y \to x} \frac{f'(y) - f'(x)}{y - x}$

- since f' is non-decreasing,

$$\frac{f'(y) - f'(x)}{y - x} \geq 0$$

for any x, y, $x \neq y$.

- Thus

$$f''(x) = \lim_{y \to x} \frac{f'(y) - f'(x)}{y - x} \geq 0.$$

∎

Examples:

1) $f(x) = x^2$ is convex

since $f''(x) = 2 > 0$

2) $f(x) = \ln(1 + e^x)$ is convex

since

$$\frac{e^x}{\quad} \quad 1$$

$$f'(x) = \frac{e^{-x}}{1+e^{-x}} = \frac{1}{e^{x}+1}$$

Lemma:

Let $C \subseteq \mathbb{R}^d$ be a convex set.

Let $f_1, f_2, \ldots, f_n : C \to \mathbb{R}$ be convex functions.

Let $w_1, w_2, \ldots, w_n$ be non-negative real numbers

Then, the following functions are also convex:

1) $g(x) = \max_{i=1,\ldots,n} f_i(x)$

2) $h(x) = \sum_{i=1}^{n} w_i f_i(x)$

2) $h(x) = \sum_{i=1}^{2} w_i \cdot h_i(x)$

Proof: Left as an exercise.

Lemma:

Let $g: \mathbb{R} \to \mathbb{R}$ be a convex function.

Let $c \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Then the function $f: \mathbb{R}^d \to \mathbb{R}$,

$$f(x) = g(c^T x + b)$$

is convex.

Proof:

$$f(\alpha x + (1-\alpha) y) = g(\alpha c^T x + (1-\alpha) c^T y + b)$$

$$= g(\alpha(c^T x + b) + (1-\alpha)(c^T y + b))$$

$$= g(c^T x + b) + (1-\alpha) g(c^T y + b)$$

$$= \alpha f(x) + (1-\alpha) f(y)$$

Examples:

1) $f(w) = (w^T x_i - y_i)^2$

is convex

2) $f(w) = \ln\left(1 + e^{-y_i w^T x_i}\right)$

is convex

3) $f(w) = \sum_{i=1}^{m} (w^T x_i - y_i)^2$

is convex

m ($y_i w^T x_i$)

4) $f(w) = \sum_{i=1} \ln\left(1 + e^{-y_i w^T x_i}\right)$

is convex

Lipschitz property

Definition: Let $C \subseteq \mathbb{R}^d$.

Let $f: \mathbb{R}^d \to \mathbb{R}^k$ be any function.

Let $\rho \geq 0$ be a real number.

The function f is called

ρ-Lipschitz over C if

for any $x, y \in C$

$$\|f(x) - f(y)\| \leq \rho \|x - y\|.$$

Note: Lipschitz function is

Continuous. The reverse is

Also

false.

Note: The definition depends on choice of the two norms

$$\|f(x) - f(y)\| \leq \rho \|x - y\|$$

a norm on $\mathbb{R}^k$ — a norm on $\mathbb{R}^d$

We use Euclidean norm for both.

Other choices are possible...

Lemma:

Let $I \subseteq \mathbb{R}$ be a non-trivial interval.

Let $f: \mathbb{R} \to \mathbb{R}$ be a differentiable function. Then, the following are

equivalent

1) f is ρ-Lischitz on I

2) $\forall x \in I \quad |f'(x)| \leq \rho$

Proof:

2) $\Rightarrow$ 1):

- Let $x, y \in I$

- $f(x) - f(y) = f'(v)(x-y)$
  for some v between x, y.

- $|f(x) - f(y)| = |f'(v)| \cdot |x-y|$

- Since f is ρ-Lipchitz on I
  $|f(x)-f(y)| \leq \rho |x-y|$

- Therefore $|f'(v)| \leq \rho$

1) $\Rightarrow$ 2)

- Let $x \in I$

- $$|f'(x)| = \left| \lim_{y \to x} \frac{f(y) - f(x)}{y - x} \right|$$

$$= \lim_{y \to x} \left| \frac{f(y) - f(x)}{y - x} \right|$$

$$= \lim_{y \to x} \frac{|f(y) - f(x)|}{|y - x|}$$

$$\leq \rho$$

∎

Examples:

1) $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$

2) $f(x) = \ln(1 + e^x)$ is 1-Lipchitz

over $\mathbb{R}$ since

$$|f'(x)| = \left| \frac{1}{e^{-x}+1} \right| \leq 1$$

3) $f(x) = x^2$ is <u>not</u> ρ-Lipchitz over $\mathbb{R}$ for any ρ, since $|f'(x)| = |2x|$ is unbounded.

4) $f(x) = x^2$ is ρ-Lipchitz over $[-\rho/2, \rho/2]$ for any $\rho \geq 0$.

5) $f(x) = c^T x + b$ is $\|c\|$-Lipchitz since

$$|f(x) - f(y)| = |c^T(x-y)|$$

$$\leq \|c\| \cdot \|x - y\|$$

Cauchy-Schwartz inequality

Lemma:

Let $f: \mathbb{R}^a \to \mathbb{R}^b$ be ρ_1-Lipschitz over $\mathbb{R}^a$.

Let $g: \mathbb{R}^b \to \mathbb{R}^c$ be ρ_2-Lipschitz. over $\mathbb{R}^b$.

Then $h: \mathbb{R}^a \to \mathbb{R}^c$, $h(x) = g(f(x))$ is $\rho_1\rho_2$-Lischitz ever $\mathbb{R}^a$.

Proof:

$$\|h(x) - h(y)\| = \|g(f(x)) - g(f(y))\|$$

$$\leq \rho_2 \|f(x) - f(y)\|$$

$$\leq \rho_2 \rho_1 \|x - y\|$$

Smooth functions

Definition:
Let $f: \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Let $\beta \geq 0$.
The function f is called β-smooth if $\nabla f: \mathbb{R}^d \to \mathbb{R}^d$ is β-Lipschitz over $\mathbb{R}^d$.

That is, for all $x, y \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

Lemma:
Let $f: \mathbb{R}^d \to \mathbb{R}$ be β-smooth.

Then for any $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{\beta}{2}\|x-y\|^2$$

Proof for $d=1$:

(For $d>1$, the concept Hessian is needed.)

- Since f is β-smooth, f' is β-Lipschitz. Therefore $|f''(v)| \leq \beta$ for any $v \in \mathbb{R}$
- Let $x, y \in \mathbb{R}$.
- By Taylor's theorem, there exist v between x, y such that

$$f(y) = f(x) + f'(x)\cdot(x-y) + \underbrace{f''(v)}_{\leq \beta} \cdot \frac{(x-y)^2}{2}$$

$\leq \beta$

Examples:

1) $f(x) = x^2$ is 2-smooth since $f''(x) = 2$.

2) $f(x) = \ln(1+e^x)$ is $\frac{1}{4}$-smooth, since

$$f'(x) = \frac{1}{1+e^{-x}}, \quad f''(x) = \frac{-e^{-x}}{(1+e^{-x})^2}$$

$$|f''(x)| = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{(1+e^x)(1+e^{-x})}$$

1

$$= \frac{}{(1+A)\left(1+\frac{1}{A}\right)}$$

$$\leq \frac{1}{4}$$

since

$$(1+A)\left(1+\frac{1}{A}\right) \geq 4$$

$$\Updownarrow$$

$$A + \frac{1}{A} \geq 2$$

$$\Updownarrow$$

$$\frac{A+\frac{1}{A}}{2} \geq 1$$

$$\Updownarrow$$

$$\frac{A+\frac{1}{A}}{2} \geq \sqrt{A \cdot \frac{1}{A}}$$

Lemma:

Let $g: \mathbb{R} \to \mathbb{R}$ be β-smooth.

Let $c \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Then, the function $f: \mathbb{R}^d \to \mathbb{R}$,

$$f(x) = g(c^T x + b)$$

is $\beta \|c\|^2$- smooth.

Proof:

- $\nabla f(x) = g'(c^T x + b)\, c$

- For any $x, y \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\| = \|g'(c^T x + b)c - g'(c^T y + b)c\|$$

$$= \|c\| \cdot |g'(c^T x + b) - g'(c^T y + b)|$$

$$\leq \|c\| \cdot \beta |(c^T x + b) - (c^T y + b)|$$

$$= \|c\| \cdot \beta \, |c^T(x-y)|$$

$$\leq \|c\| \cdot \beta \cdot \|c\| \cdot \|x-y\|$$

$$= \left(\beta \|c\|^2\right) \cdot \|x-y\|$$

∎

Note: Suppose $f(w) = \ell(w, x, y)$ is convex / Lipschitz / smooth.

Then

$$g(w) = L_D(w) = \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(w, x, y)\right]$$

$$h(w) = L_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, x_i, y_i)$$

are convex / Lipschitz / smooth

Non-learnability of least squares

- Let the dimension be 1
- Let $\ell : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be

$$\ell(w, x, y) = (wx - y)^2$$

- Agnostically PAC learnability for ℓ:

($\exists$ algorithm A) ($\forall \epsilon, \delta \in (0,1)$, $\exists$ sample size m, $\forall$ dist. over $X \times Y$ D)

With probability at least $1-\delta$,

$$L_D(A(S)) \leq \min_{w \in \mathbb{R}} L_D(w) + \epsilon$$

where $S = ((x_1, y_1) \ldots (x_m, y_m))$ is an i.i.d. sample from D.

Theorem:

ℓ is NOT agnostically PAC learnable.

Proof:

- By contradiction assume ℓ is agnostically PAC learnable
- There exist a learning algorithm $A : (\mathbb{R} \times \mathbb{R})^* \to \mathbb{R}$
- Choose $\varepsilon = \frac{1}{100}$, $\delta = \frac{1}{2}$.
- Let $m := m(\varepsilon, \delta)$ be the required sample size such that $\forall D$

$$L_D(A(S)) \leq \min_w L_D(w) + \varepsilon$$

$w \in \mathbb{R}^d$

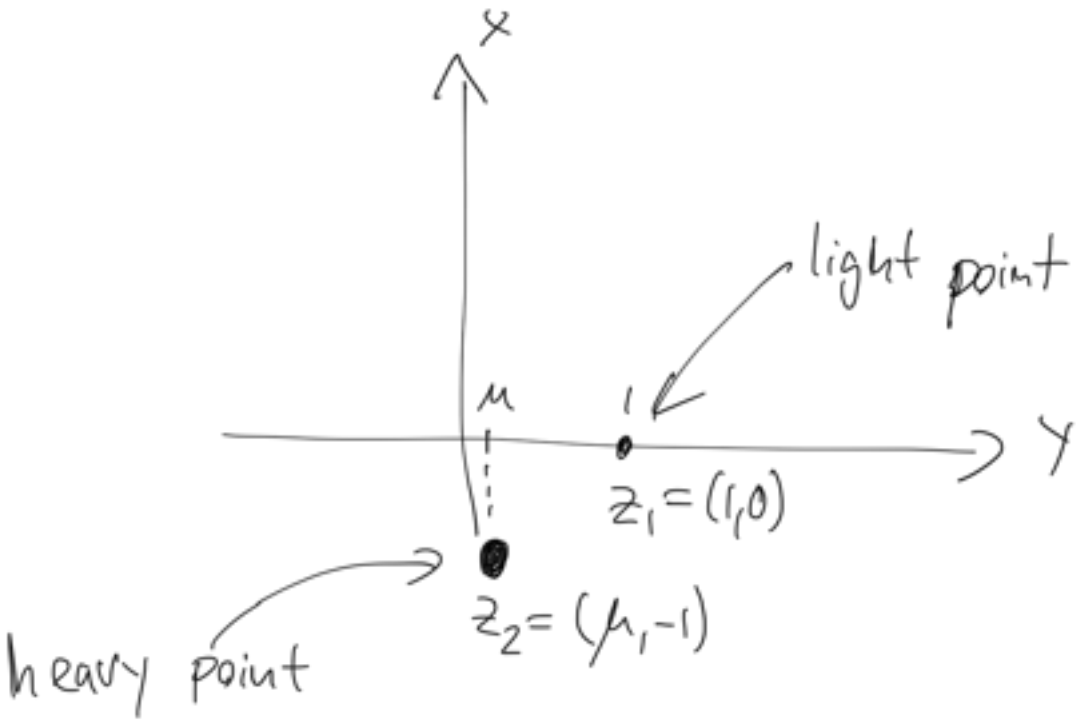
- Let $\mu = 1 - \sqrt[m]{\frac{99}{100}}$
- Note $\mu \leq \frac{1}{100}$ for $m \geq 1$.
- Let $z_1 = (1, 0) \in \mathbb{R} \times \mathbb{R}$

  $z_2 = (\mu, -1) \in \mathbb{R} \times \mathbb{R}$
- We define two distributions over $\mathbb{R} \times \mathbb{R}$, D_1 and D_2

$D_1(z_1) = \mu$

$D_1(z_2) = 1 - \mu$

$D_2(z_2) = 1$



- $L_{D_1}(w) = \mu (w-0)^2 + (1-\mu)(w/\mu + 1)^2$

- $L_{D_2}(w) = (w/\mu + 1)^2$

- μ is very small. It is very likely that an i.i.d. sample from D_1 (or D_2) consists entirely of m copies of z_2.

- Suppose

$$\hat{w} = A(\underbrace{(z_2, z_2, \ldots, z_2)}_{m \text{ times}})$$

Case 1: $\hat{w} \geq -\frac{1}{2\mu}$

- $L_{D_2}(\hat{w}) = (\hat{w}\mu + 1)^2$

$$\geq \left(-\frac{1}{2\mu} \cdot \mu + 1\right)^2$$

$$= \left(-\frac{1}{2} + 1\right)$$

$$= \left(\frac{1}{2}\right)^2$$

$$= \frac{1}{4}$$

- $\min_{w \in \mathbb{R}} L_{D_2}(w) = L_{D_2}\left(-\frac{1}{\mu}\right) = 0$

- $L_{D_2}(\hat{w}) \not\leq \min L_D(w) + \varepsilon$

$\geq 1/4$ $\quad = 0$ $\quad = \frac{1}{100}$

Case 2: $\hat{w} \leq -\frac{1}{2\mu}$

- With probability

$$(1-\mu)^m = \left(\sqrt[m]{\frac{99}{100}}\right)^m = \frac{99}{100}$$

an i.i.d. sample S from D_1 consists of m copies of z_2 (i.e. no z_1 appears in S)

- $L_{D_1}(\hat{w}) \geq \mu(\hat{w} - 0)^2$

$$= \mu(\hat{w})^2$$

$$\geq \mu \frac{1}{4\mu^2}$$

$$= \frac{1}{4\mu}$$

$$\geq 400$$

- $\min_{w \in \mathbb{R}^d} L_{D_1}(w) \leq L_{D_1}(0) = 1 - \mu \leq 1$

- $L_{D_1}(\hat{w}) \not\leq \min_{w \in \mathbb{R}} L_{D_1}(w) + \varepsilon$

  ≥ 400 $\quad \leq 1 \quad$ $= \frac{1}{100}$

Surrogate losses

- Zero-one loss for halfspaces

$$\ell_{01}(w, x, y) = \mathbb{1}[\text{sign}(w^T x) \neq y]$$

- ERM for ℓ_{01} is NP-hard

- People use so called surrogate losses instead

- A surrogate loss ℓ has to

  1) Upper bound ℓ_{01}

  2) Be convex in w

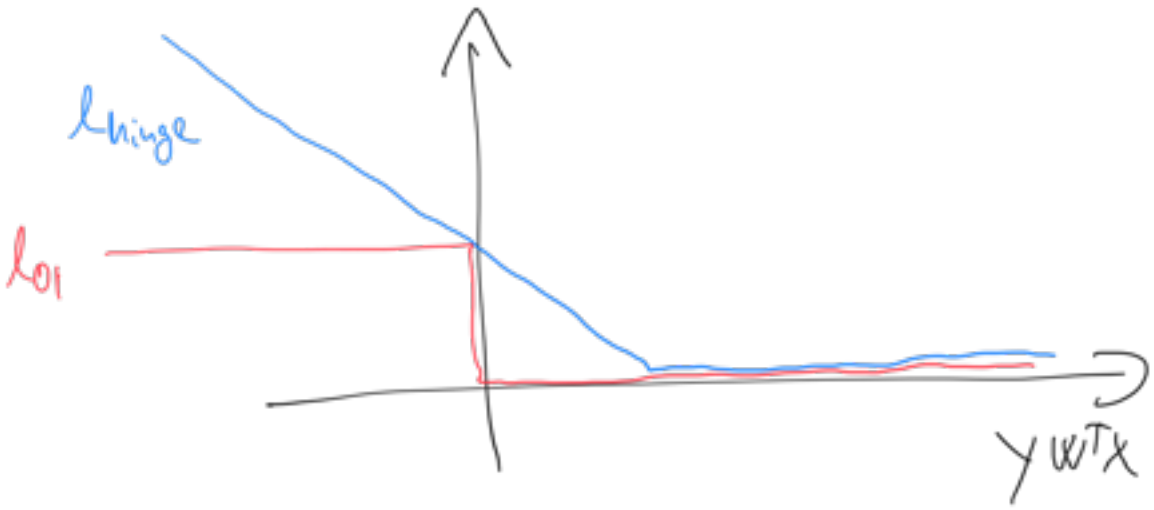
- Formally

1) $\ell_{01}(w, x, y) \leq \ell(w, x, y)$

2) $f(w) = \ell(w, x, y)$ is convex for any x, y

- Popular choice is hinge loss

$$\ell_{hinge}(w, x, y) = \max(0, 1 - y w^T x)$$



- ERM for surrogate losses is efficient.

- However we don't find