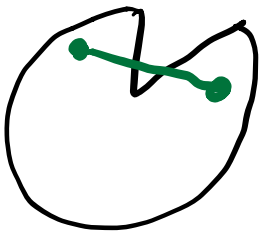
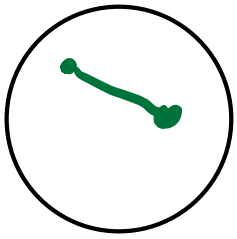
Convexity

Convex Set: A set C in a vector space is convex if for any two vectors $\vec{u}, \vec{v}$ in C, the line segment between $\vec{u}$ and $\vec{v}$ is contained in C.
That is, $\forall \alpha \in [0,1]$ we have $\alpha\vec{u} + (1-\alpha)\vec{v} \in C$.

Non-Convex



Convex



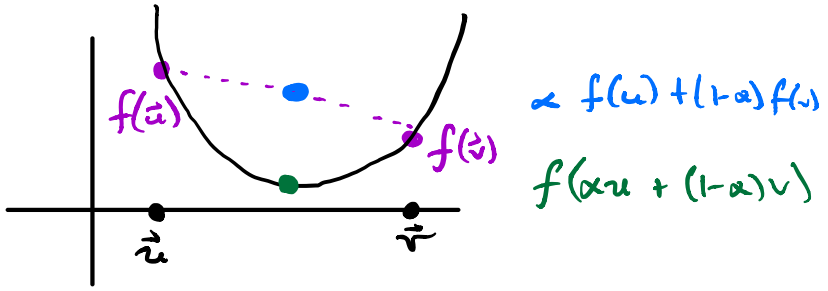
Given $\alpha \in [0,1]$, $\alpha\vec{u} + (1-\alpha)\vec{v}$ is called a **convex combination**

Convex Function: Let C be a convex set. A function $f: C \to \mathbb{R}$ is convex if for every $\vec{u}, \vec{v} \in C$ and $\alpha \in [0,1]$

$$f(\alpha\vec{u} + (1-\alpha)\vec{v}) \leq \alpha f(\vec{u}) + (1-\alpha) f(\vec{v})$$

i.e. f is convex if for any $\vec{u}, \vec{v}$ the graph of f between $\vec{u}, \vec{v}$ lies below the line segment joining $f(\vec{u})$ and $f(\vec{v})$.

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



The epigraph of a function is the set

$$\text{epigraph}(f) = \{(\vec{x}, \beta) \mid f(\vec{x}) \leq \beta\}$$

Note: f is convex if and only if its epigraph is a convex set

An important property of convex functions is that every local minimum of the function is also a global minimum.

Formally, let

$$B(\vec{u}, r) = \{\vec{v} \mid \|\vec{v} - \vec{u}\| < r\}$$

be the ball of radius r centered at $\vec{u}$.
We say $f(\vec{u})$ is a local minimum of f at $\vec{u}$ if $\exists r > 0$ such that $\forall \vec{v} \in B(\vec{u}, r)$ we have $f(\vec{v}) \geq f(\vec{u})$.

It follows that $\forall \vec{v}$ (not necessarily in B), there is a small enough $\alpha > 0$ s.t.

$$\vec{u} + \alpha(\vec{v} - \vec{u}) \in B(\vec{u}, r)$$

and therefore

$$f(\vec{u}) \leq f(\vec{u} + \alpha(\vec{v} - \vec{u})). \tag{1}$$

If f convex, also have

$$f(\vec{u} + \alpha(\vec{v} - \vec{u})) = f(\alpha\vec{v} + (1-\alpha)\vec{u}) \leq (1-\alpha) f(\vec{u}) + \alpha f(\vec{v}) \tag{2}$$

Combing (1), (2):

$$f(\vec{u}) \geq f(\vec{v})$$

Since this holds for every $\vec{v}$, it follows $f(\vec{w})$ is ALSO a global minimum of f.

Another important property of convex functions is that for every $\vec{w}$ we can construct a tangent to to f at $\vec{w}$ that lies below f everywhere.

If f is differentiable, this tangent is the linear function

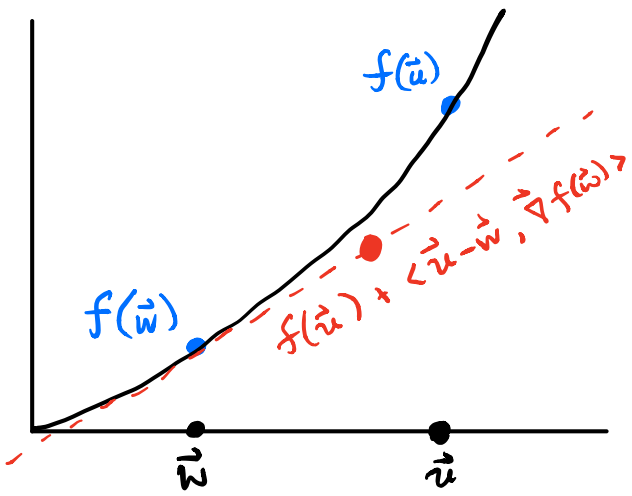
$$L(\vec{u}) = f(\vec{w}) + \langle \vec{\nabla} f(\vec{w}), \vec{u} - \vec{w} \rangle$$

where

$$\vec{\nabla} f(\vec{w}) = \left(\frac{\partial f(\vec{w})}{\partial w_1} \quad \cdots \quad \frac{\partial f(\vec{w})}{\partial w_d} \right)$$

That is, for convex differentiable functions,

$$\forall \vec{u}, \quad f(\vec{u}) \geq f(\vec{w}) + \langle \vec{\nabla} f(\vec{w}), \vec{u} - \vec{w} \rangle$$



If f is scalar differentiable function, there is an EASY way to check if it is convex

Lemma: Let $f: \mathbb{R} \to \mathbb{R}$ be a scalar twice differentiable function. Then the following are equivalent:

1. f is convex
2. f' is monotonically non-decreasing
3. f'' is non-negative

Claim: Assume $f: \mathbb{R}^d \to \mathbb{R}$ can be written as

$$f(\vec{w}) = g(\langle \vec{w}, \vec{x} \rangle + y),$$

for some $\vec{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$, and $g: \mathbb{R} \to \mathbb{R}$.
Then, convexity of $g \Rightarrow$ convexity of f.

Proof: Let $\vec{w}_1, \vec{w}_2 \in \mathbb{R}^d$ and $\alpha \in [0, 1]$.

$$\begin{aligned} f(\alpha \vec{w}_1 + (1-\alpha)\vec{w}_2) &= g(\langle \alpha \vec{w}_1 + (1-\alpha)\vec{w}_2, \vec{x} \rangle + y) \\ &= g(\alpha \langle \vec{w}_1, \vec{x} \rangle + (1-\alpha)\langle \vec{w}_2, \vec{x} \rangle + y) \\ &= g(\alpha(\langle \vec{w}_1, \vec{x} \rangle + y) + (1-\alpha)(\langle \vec{w}_2, \vec{x} \rangle + y)) \\ &\leq \alpha g(\langle \vec{w}_1, \vec{x} \rangle + y) + (1-\alpha) g(\langle \vec{w}_2, \vec{x} \rangle + y) \end{aligned}$$

Examples

① Given $\vec{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, let $f: \mathbb{R}^d \to \mathbb{R}$ be defined by.

$\vec{w} \mapsto \langle \vec{w}, \vec{x} \rangle - y^2$

Then, f is a composition of the function $g(a) = a^2$ onto a linear function, and hence f is a convex function.

② Given some $\vec{x} \in \mathbb{R}^d$ and $y \in \{-1, 1\}$, let

$$f: \mathbb{R}^d \to \mathbb{R}$$

be defined by

$$\vec{w} \longmapsto \log\left(1 + e^{-y\langle \vec{w}, \vec{x} \rangle}\right)$$

Then, f is a composition of the function $g(a) = \log(1 + e^a)$ onto a linear function, and hence f is a convex function.

Claim: For $i = 1, \ldots, r$, let $f_i : \mathbb{R}^d \to \mathbb{R}$ be a convex function. The following functions from $\mathbb{R}^d$ to $\mathbb{R}$ are also convex.

① $g(x) = \max_{i \in [r]} f_i(x)$

② $\sum_{i=1}^{r} w_i f_i(x)$, where $w_i \geq 0 \ \forall \ i$

Proof

① $g(\alpha \vec{u} + (1-\alpha)\vec{v}) = \max_i f_i(\alpha \vec{u} + (1-\alpha)\vec{v})$

$\leq \max_i [\alpha f_i(\vec{u}) + (1-\alpha) f_i(\vec{v})]$

$\leq \alpha \max_i f_i(\vec{u}) + (1-\alpha) f_i(\vec{v})$

$= \alpha g(\vec{u}) + (1-\alpha) g(\vec{v})$

② $g(\alpha\vec{u} + (1-\alpha)\vec{v}) = \sum_i w_i f_i(\alpha\vec{u} + (1-\alpha)\vec{v})$

$\leq \sum_i w_i [\alpha f_i(\vec{u}) + (1-\alpha) f_i(\vec{v})]$

$= \alpha \sum_i w_i f_i(\vec{u}) + (1-\alpha) \sum_i w_i f_i(\vec{v})$

$= \alpha g(\vec{u}) + (1-\alpha) g(\vec{v})$

Lipschitzness

Definition: Let $C \subset \mathbb{R}^d$. A function $f: \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if $\forall \vec{w}_1, \vec{w}_2 \in C$ we have that

$$\|f(\vec{w}_1) - f(\vec{w}_2)\| \leq \rho \|\vec{w}_1 - \vec{w}_2\|$$

Note: If $f: \mathbb{R} \to \mathbb{R}$ differentiable, then by mean value theorem we have

$$f(w_1) - f(w_2) = f'(u)(w_1 - w_2),$$

where $u \in (w_1, w_2)$.

It follows that if the derivative of f is everywhere bounded by ρ, then the function is ρ-Lipschitz.

"Intuitively," a Lipschitz function cannot change to fast.

Examples

① $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$. To see this, observe that $\forall\, x_1, x_2 \in \mathbb{R}$

$$\begin{aligned} |x_1| - |x_2| &= |x_1 - x_2 + x_2| - |x_2| \\ &\leq |x_1 - x_2| + |x_2| - |x_2| \\ &= |x_1 - x_2| \end{aligned}$$

Thus

$$||x_1| - |x_2|| \leq |x_1 - x_2|$$

② $f(x) = \log(1 + e^x)$ is 1-Lipschitz over $\mathbb{R}$.

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

③ $f(x) = x^2$ is **NOT** ρ-Lipschitz over $\mathbb{R}$ for any ρ. Take $x_1 = 0$, $x_2 = 1 + \rho$, then

$$f(x_2) - f(x_1) = (1+\rho)^2 > \rho(1+\rho) = \rho\,|x_2 - x_1|$$

④ $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(\vec{w}) = \langle \vec{v}, \vec{w} \rangle + b$ where $\vec{v} \in \mathbb{R}^d$ is $\|\vec{v}\|$-Lipschitz.

Cauchy-Schwarz

$$|f(\vec{w}_1) - f(\vec{w}_2)| = |\langle \vec{v}, \vec{w}_1 - \vec{w}_2 \rangle| \leq \|\vec{v}\|\,\|\vec{w}_1 - \vec{w}_2\|$$

Claim: Let $f(x) = g_1(g_2(x))$, where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz.
Then, f is $(\rho_2 \cdot \rho_2)$-Lipschitz.
In particular, if g_2 is the linear function,

$$g_2(\vec{x}) = \langle \vec{v}, \vec{x} \rangle + b$$

for some $\vec{v} \in \mathbb{R}^d$, $b \in \mathbb{R}$, then f is $\rho_1 \|\vec{v}\|$-Lipschitz

Proof

$$|f(\vec{w}_1) - f(\vec{w}_2)| = |g_1(g_2(\vec{w}_1)) - g_1(g_2(\vec{w}_2))|$$
$$\leq \rho_1 \|g_2(\vec{w}_1) - g_2(\vec{w}_2)\|$$
$$\leq \rho_1 \rho_2 \|\vec{w}_1 - \vec{w}_2\|$$

Smoothness

A differentiable function $f: \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz; namely $\forall \vec{v}, \vec{w}$ we have

$$\|\vec{\nabla} f(\vec{v}) - \vec{\nabla} f(\vec{w})\| \leq \beta \|\vec{v} - \vec{w}\|$$

Smoothness implies that for all $\vec{v}, \vec{w}$ we have

$$f(\vec{v}) \leq f(\vec{w}) + \langle \vec{\nabla} f(\vec{w}), \vec{v} - \vec{w} \rangle + \frac{\beta}{2} \|\vec{v} - \vec{w}\|^2$$

Recall that convexity of f implies that

$$f(\vec{v}) \geq f(\vec{w}) + \langle \vec{\nabla} f(\vec{w}), \vec{v} - \vec{w} \rangle$$

This means that when a function is both convex AND smooth, we have upper & lower bounds on the difference between the function and its first order approximation.

Setting $\vec{v} = \vec{w} - \frac{1}{\beta} \vec{\nabla} f(\vec{w})$, we get

$$\frac{1}{2\beta} \| \vec{\nabla} f(\vec{w}) \|^2 \leq f(\vec{w}) - f(\vec{v})$$

Assuming $f(\vec{v}) \geq 0 \ \forall \ \vec{v}$, we conclude that smoothness implies

$$\| \vec{\nabla} f(\vec{w}) \|^2 \leq 2\beta f(\vec{w})$$

Functions which satisfy this property are considered self-bounded

Examples

① $f(x) = x^2$ is 2-smooth.

② The function

$$f(x) = \log(1 + e^x)$$

is $1/4$ smooth.

Observe

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

$$|f''(x)| = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{(1+e^{-x})(1+e^{x})} \leq \frac{1}{4}$$

Hence f' is $\frac{1}{4}$-Lipschitz.

Claim: Let $f(\vec{w}) = g(\langle \vec{w}, \vec{x} \rangle + b)$, where $g: \mathbb{R} \to \mathbb{R}$ is a β-smooth function, $\vec{x} \in \mathbb{R}^d$, and $b \in \mathbb{R}$. Then, f is $(\beta \|\vec{x}\|^2)$-smooth.

Proof: By the chain rule we have that

$$\vec{\nabla} f(\vec{w}) = g'(\langle \vec{w}, \vec{x} \rangle + b)\, \vec{x},$$

where g' is the derivative of g.

Using the smoothness of g and Cauchy-Schwarz,

$$f(\vec{v}) = g(\langle \vec{v}, \vec{x} \rangle + b)$$

$$\leq g(\langle \vec{w}, \vec{x} \rangle + b) + g'(\langle \vec{w}, \vec{x} \rangle + b)\langle \vec{v} - \vec{w}, \vec{x} \rangle + \frac{\beta}{2}(\langle \vec{v} - \vec{w}, \vec{x} \rangle)^2$$

$$\leq g(\langle \vec{w}, \vec{x} \rangle + b) + g'(\langle \vec{w}, \vec{x} \rangle + b)\langle \vec{v} - \vec{w}, \vec{x} \rangle + \frac{\beta}{2}(\|\vec{v} - \vec{w}\|\|\vec{x}\|)^2$$

$$= f(\vec{w}) + \langle \vec{\nabla} f(\vec{w}), \vec{v} - \vec{w} \rangle + \frac{\beta \|\vec{x}\|^2}{2}\|\vec{v} - \vec{w}\|^2$$

Convex Learning Problems

A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called convex if the hypothesis class $\mathcal{H}$ is a convex set, and $\forall z \in \mathcal{Z}$, the loss function, $\ell(\cdot, z)$, is a convex function (where, for any z, $\ell(\cdot, z)$ denotes the function $f: \mathcal{H} \to \mathbb{R}$ defined by $f(\vec{w}) = \ell(\vec{w}, z)$).

Example - Linear Regression with Squared Loss

Domain Set $\mathcal{X} \subset \mathbb{R}^d$

Label Set $\mathcal{Y} = \mathbb{R}$

We would like to learn a linear function $h: \mathbb{R}^d \to \mathbb{R}$ that best approximates the relationship between our variables.

Can model as a convex learning problem.

Each linear function is parameterized by a vector $\vec{w} \in \mathbb{R}^d$.

Hence, define $\mathcal{H} := \mathbb{R}^d$.

The set of examples is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R} \cong \mathbb{R}^{d+1}$, and the loss function is $\ell(\vec{w}, (\vec{x}, y)) = (\langle \vec{w}, \vec{x} \rangle - y)^2$.

Clearly, $\mathcal{H}$ is convex, and $\ell(\cdot, z)$ is a convex loss function.

Lemma: If ℓ is a convex loss function and the class $\mathcal{H}$ is convex, then the $ERM_{\mathcal{H}}$ problem, of minimizing the empirical loss over $\mathcal{H}$, is a convex optimization problem.

Proof: Recall, the $ERM_{\mathcal{H}}$ problem is defined by

$$ERM_{\mathcal{H}}(S) = \underset{\vec{w} \in \mathcal{H}}{\operatorname{argmin}} \; L_S(\vec{w})$$

Since, for a sample $S = z_1, \ldots, z_m$, for every $\vec{w}$

$$L_S(\vec{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\vec{w}, z_i)$$

is a convex function (from claim above).

Therefore, the ERM rule is a problem of minimizing a convex function subject to the constraint that the solution should be in a convex set.

Note: Not all convex learning problems over $\mathbb{R}^d$ are learnable!!!

Example - Nonlearnability of Linear Regression ($\forall d$)

Let $\mathcal{H} = \mathbb{R}$, and the loss be the squared loss

$$\ell(w, (x, y)) = (wx - y)^2 \quad \} \text{ the homogeneous case}$$

Let $A :=$ ANY deterministic algorithm

Assume, towards a contradiction, that A is a successful PAC learner for this problem.
That is, $\exists\ m(\cdot,\cdot)$ such that $\forall$ distributions D and for every ε, δ if A receives a training set of size $m \geq m(\varepsilon, \delta)$, it should output, w/ probability of at least $1-\delta$, a hypothesis $\hat{w} = A(S)$, such that

$$L_D(\hat{w}) - \min_w L_D(w) \leq \varepsilon.$$

Choose $\varepsilon = 1/100$, $\delta = 1/2$, let $m \geq m(\varepsilon, \delta)$ and set $\mu = \frac{\log(100/99)}{2m}$.

Define two distributions.

D_1 is supported on 2 examples, $z_1 = (1, 0)$; $z_2 = (\mu, -1)$ where the probability mass of z_1 is μ and of z_2 is $1-\mu$.

D_2 is supported entirely on Z_2.

Observe, that for both distributions,

$$\Pr\left(\begin{Bmatrix}\text{all examples of training set will}\\ \text{be of the second type}\end{Bmatrix}\right) \geq 0.99$$

Trivially true for D_2, whereas for D_1 the probability of the event is

$$(1-\mu)^m \geq e^{-2\mu m} = 0.99$$

Since A was assumed deterministic, upon receiving a training set of m examples, each of which is $(\mu, -1)$, the algorithm will output some $\hat{w}$.

Now, if $\hat{w} < \frac{-1}{2\mu}$, we set distribution to D_1. Hence,

$$L_{D_1}(\hat{w}) \geq \mu(\hat{w})^2 \geq \frac{1}{4\mu}$$

However,

$$\min_w L_{D_1}(w) \leq L_{D_1}(0) = 1-\mu$$

It follows that

$$L_{D_2}(\hat{w}) - \min_w L_{D_2}(w) \geq \frac{1}{4\mu} - (1-\mu) > \varepsilon \Bigg\} \text{A fails on } D_1$$

On the other hand, if $\hat{w} \geq -1/2\mu$ set distribution to P_2. Then

$$L_{P_2}(\hat{w}) \geq \frac{1}{4}$$

and

$$\min_w L_{P_2}(w) = 0$$

$\Rightarrow$ A fails on P_2.

In summary, we have shown that $\forall$ A $\exists$ D on which A fails $\Rightarrow$ problem NOT PAC learnable.

A possible solution to this is to add another constraint to the hypothesis class. In ADDITION to the convexity requirement, we require that $\mathcal{H}$ will be <u>bounded</u>

Namely, we assume every hypothesis $\vec{w} \in \mathcal{H}$ satisfies $\|\vec{w}\| \leq B$ for some predefined scalar B.

However, boundedness and convexity ALONE are NOT enough for ensuring a problem is learnable.

Example

Consider a regression problem with the squared loss. Let $\mathcal{H} = \{w : |w| \leq 1\} \subset \mathbb{R}$ be a bounded hypothesis class. (Note $\mathcal{H}$ is convex!)

The argument is the same as the last example, except now the two distributions D_1, D_2 will be supported on $z_1 = (1/\mu, 0)$ and $z_2 = (1, -1)$.

If A returns $\hat{w} < -1/2$ upon receiving m examples of the second type, then we will set the distribution to D_1 and have that

$$L_{D_1}(\hat{w}) - \min_w L_{D_1}(w) \geq \mu(\hat{w}/\mu)^2 - L_{D_1}(0) \geq \frac{1}{4\mu} - (1-\mu) > \varepsilon$$

Similarly, if $\hat{w} \geq -\frac{1}{2}$ we set distribution to be D_2 and have

$$L_{D_2}(\hat{w}) - \min_w L_{D_2}(w) \geq (-\tfrac{1}{2} + 1)^2 - 0 > \varepsilon$$

This shows we need ADDITIONAL assumptions on the learning problem.

This motivates a definition of two families of learning problems.

Convex-Lipschitz / Smooth-Bounded Learning Problems

A learning problem, (H, Z, l), is called Convex-Lipschitz-Bounded, with parameters ρ, B if the following holds:

- The hypothesis class $\mathcal{H}$ is a convex set and $\forall \vec{w} \in \mathcal{H}$ we have $\|\vec{w}\| \leq B$
- $\forall z \in Z$, $l(\cdot, z)$ is a convex and ρ-Lipschitz function

Example

Let $X = \{\vec{x} \in \mathbb{R}^d \mid \|\vec{x}\| \leq \rho\}$ and $Y = \mathbb{R}$.

Let $\mathcal{H} = \{\vec{w} \in \mathbb{R}^d \mid \|\vec{w}\| \leq B\}$ and $l(\vec{w}, (\vec{x}, y)) \triangleq |\langle \vec{w}, \vec{x} \rangle - y|$.

Clear to see the resulting problem is Convex-Lipschitz Bounded w/ parameters ρ, B.

A learning problem (H, Z, l) is called Convex-Smooth-Bounded with parameters β, B if the following holds:

- The hypothesis class $\mathcal{H}$ is a convex set and $\forall \vec{w} \in \mathcal{H}$ we have $\|\vec{w}\| \leq B$
- $\forall z \in Z$, $l(\cdot, z)$ is a convex, nonnegative and β-smooth function

ensures self-boundedness as stated previously!

These two families of learning problems are learnable.