Machine learning theory

Convex learning problems

Hamid Beigy

Sharif university of technology

June 8, 2020

Introduction

- Convex learning comprises an important family of learning problems, because most of what we can learn efficiently.
- Linear regression with the squared loss is a convex problem for regression.
- logistic regression is a convex problem for classification.
- Halfspaces with the $0-1$ loss, which is a computationally hard problem to learn in unrealizable case, is non-convex.
- In general, a convex learning problem is a problem.
  1. whose hypothesis class is a convex set and
  2. whose loss function is a convex function for each example.
- Other properties of the loss function that facilitate successful learning are
  1. Lipschitzness
  2. Smoothness
- In this session, we study the learnability of
  1. Convex-Smooth problems
  2. Lipschitz-Bounded problems

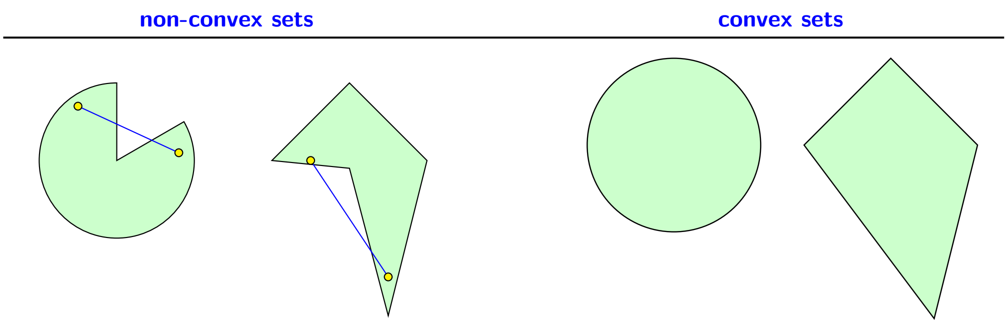
Convexity

Definition (Convex set)

A set C in a vector space is convex if for any two vectors $\mathbf{u}, \mathbf{v} \in C$, the line segment between $\mathbf{u}$ and $\mathbf{v}$ is contained in set C. That is, for any $\alpha \in [0, 1]$, the convex combination $\alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in C$. Given $\alpha \in [0, 1]$, the combination, $\alpha\mathbf{u} + (1-\alpha)\mathbf{v}$ of the points $\mathbf{u}$, $\mathbf{v}$ is called a **convex combination**.

Example (Convex and non-convex sets)

Some examples of convex and non-convex sets in $\mathbb{R}^2$



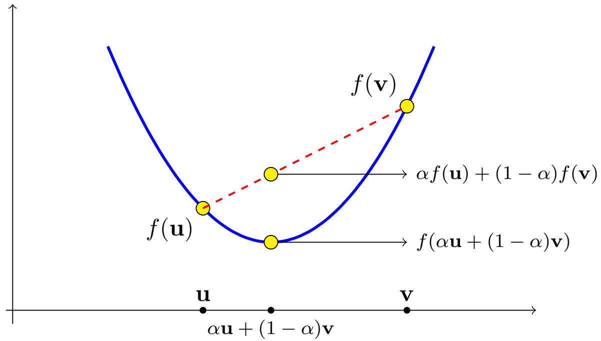
Definition (Convex function)

Let C be a convex set. Function $f : C \mapsto C$ is **convex** if for any two vectors $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0, 1]$,

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

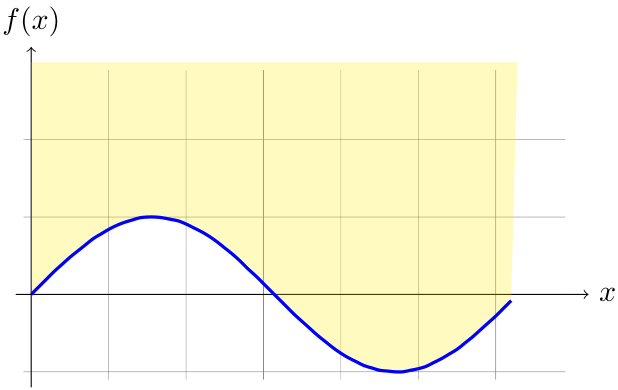
In words, f is convex if for any $\mathbf{u}, \mathbf{v} \in C$, the graph of f between $\mathbf{u}$ and $\mathbf{v}$ lies below the line segment joining $f(\mathbf{u})$ and $f(\mathbf{v})$.

Example (Convex function)



A function f is convex if and only if its epigraph is a convex set.

$$epigraph(f) = \{(\mathbf{x}, \beta) \mid f(\mathbf{x}) \leq \beta\}.$$



1. If f is convex then **every local minimum of f is also a global minimum**.
   - Let $B(\mathbf{u}, r) = \{\mathbf{v} \mid \|\mathbf{v} - \mathbf{u}\| \leq r\}$ be a ball of radius r centered around $\mathbf{u}$.
   - $f(\mathbf{u})$ is a local minimum of f at $\mathbf{u}$ if $\exists r > 0$ such that $\forall \mathbf{v} \in B(\mathbf{u}, r)$, we have $f(\mathbf{v}) \geq f(\mathbf{u})$.
   - It follows that for any $\mathbf{v}$ (not necessarily in B), there is a small enough $\alpha > 0$ such that $\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u}) \in B(\mathbf{u}, r)$ and therefore

$$f(\mathbf{u}) \leq f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u})).$$

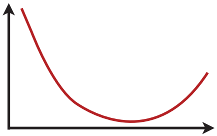
   - If f is convex, we also have that

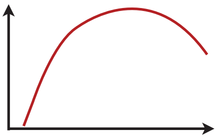
$$f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u})) = f(\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}) \leq (1 - \alpha)f(\mathbf{u}) + \alpha f(\mathbf{v}).$$

   - Combining these two equations and rearranging terms, we conclude that

$$f(\mathbf{u}) \leq f(\mathbf{v}).$$

   - This holds for every $\mathbf{v}$, hence $f(\mathbf{u})$ is also **a global minimum of f**.



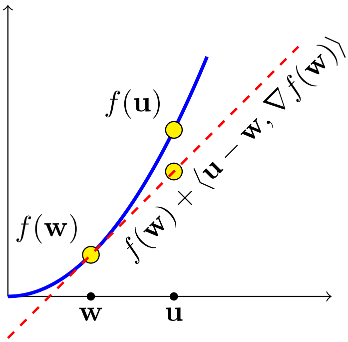


2. If f is convex and differentiable, then

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

where $\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \dots, \frac{\partial f(\mathbf{w})}{\partial w_n} \right)$ is the gradient of f at $\mathbf{w}$.

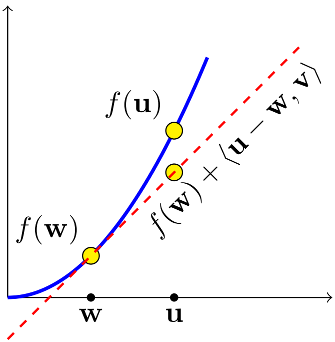
- If f is convex, for every $\mathbf{w}$, we can construct a tangent to f at $\mathbf{w}$ that lies below f everywhere.
- If f is differentiable, this tangent is the linear function $l(\mathbf{u}) = f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle$.



- $\mathbf{v}$ is sub-gradient of f at $\mathbf{w}$ if $\forall \mathbf{u}$, $\quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle$
- The differential set, $\partial f(\mathbf{w})$, is the set of sub-gradients of f at $\mathbf{w}$.
  where $\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_n} \right)$ is the gradient of f at $\mathbf{w}$.

Lemma

Function f is convex iff for every $\mathbf{w}$, $\partial f(\mathbf{w}) \neq 0$.



- f is locally flat around $\mathbf{w}$ ($\mathbf{0}$ is a sub-gradient) iff $\mathbf{w}$ is aglobal minimizer.

Lemma (Convexity of a scaler function)

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a scalar twice differential function, and f', f'' be its first and second derivatives, respectively. Then, the following are equivalent:

1. *f is convex.*
2. *f' is monotonically nondecreasing.*
3. *f'' is nonnegative.*

Example (convexity of scaler functions)

1. The scaler function $f(x) = x^2$ is convex, because $f'(x) = 2x$ and $f''(x) = 2 > 0$.
2. The scaler function $f(x) = \log(1 + e^x)$ is convex, because
   - $f'(x) = \dfrac{e^x}{1 + e^x} = \dfrac{1}{e^{-x} + 1}$ is a monotonically increasing function since the exponent function is a monotonically increasing function.
   - $f''(x) = \dfrac{e^{-x}}{(e^{-x} + 1)^2} = f(x)(1 - f(x))$ is nonnegative.

Lemma (Convexity of composition of a convex scalar function with a linear function)

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ can be written as $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + y)$, for some $\mathbf{x} \in \mathbb{R}^n, y \in \mathbb{R}$ and $g : \mathbb{R} \mapsto \mathbb{R}$. Then convexity of g implies the convexity of f.

Proof (Convexity of composition of a convex scalar function with a linear function).

Let $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n$ and $\alpha \in [0, 1]$. We have

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

where the last inequality follows from the convexity of g. □

Example (Convexity of composition of a convex scalar function with a linear function)

1. Given some $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, let $f(\mathbf{w}) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$. 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 $\mathbf{x} \in \mathbb{R}^n$ and $y \in \{-1, +1\}$, let $f(\mathbf{w}) = \log(1 + \exp(-y \langle \mathbf{w}, \mathbf{x} \rangle))$. 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

Lemma (Convexity of maximum and sum of convex functions)

Let $f_i : \mathbb{R}^n \mapsto \mathbb{R} (1 \leq i \leq r)$ be convex functions. Following functions $g : \mathbb{R}^n \mapsto \mathbb{R}$ are convex.

1. $g(\mathbf{x}) = \max_{i \in \{1,\ldots,r\}} f_i(\mathbf{x})$.
2. $g(\mathbf{x}) = \sum_{i=1}^{r} w_i f_i(\mathbf{x})$, *where* $\forall i, w_i \geq 0$.

Proof (Convexity of maximum and sum of convex functions).

1. The first claim follows by

$$
\begin{aligned}
g(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) &= \max_i f_i(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \max_i \left[\alpha f_i(\mathbf{u}) + (1-\alpha) f_i(\mathbf{v})\right] \\
&= \alpha \max_i f_i(\mathbf{u}) + (1-\alpha) \max_i f_i(\mathbf{v}) = \alpha g(\mathbf{u}) + (1-\alpha) g(\mathbf{v}).
\end{aligned}
$$

2. The second claim follows by

$$
\begin{aligned}
g(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) &= \sum_{i=1}^{r} w_i f_i(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \sum_{i=1}^{r} w_i \left[\alpha f_i(\mathbf{u}) + (1-\alpha) f_i(\mathbf{v})\right] \\
&= \alpha \sum_{i=1}^{r} w_i f_i(\mathbf{u}) + (1-\alpha) \sum_{i=1}^{r} w_i f_i(\mathbf{v}) = \alpha g(\mathbf{u}) + (1-\alpha) g(\mathbf{v}).
\end{aligned}
$$

□

Function $g(x) = |x|$ is convex, because $g(x) = \max\{f_1(x), f_2(x)\}$, where both $f_1(x) = x$ and $f_2(x) = -x$ are **convex**.

Lipschitzness

- Definition of Lipschitzness is w.r.t Euclidean norm $\mathbb{R}^n$, but it can be defined w.r.t any norm.

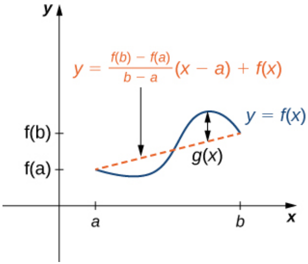
Definition (Lipschitzness)

Function $f : \mathbb{R}^n \mapsto \mathbb{R}^k$ is ρ-**Lipschitz** if for all $\mathbf{w}_1, \mathbf{w}_2 \in C$ we have $\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$.

- A Lipschitz function cannot change too fast. If $f : \mathbb{R} \mapsto \mathbb{R}$ is differentiable, then by the mean value theorem we have $f(w_1) - f(w_2) = f'(u)(w_1 - w_2)$, where u is a point between w_1 and w_2.

Theorem (Mean-Value Theorem)

If $f(x)$ is defined and continuous on the interval $[a, b]$ and differentiable on (a, b), then there is at least one number c in the interval (a, b) (that is $a < c < b$) such that $f'(c) = \dfrac{f(b) - f(a)}{b - a}$.



- If f' is bounded everywhere (in absolute value) by ρ, then f is ρ-Lipschitz.

Example (Lipschitzness)

1. Function $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$, because (using triangle inequality)

$$|x_1| - |x_2| = |x_1 - x_2 + x_2| - |x_2| \leq |x_1 - x_2| + |x_2| - |x_2| = |x_1 - x_2|.$$

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

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

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

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

4. Function $f(x) = x^2$ is ρ-Lipschitz over set $C = \left\{x \mid |x| \leq \frac{\rho}{2}\right\}$. For x_1, x_2, we have

$$\left|x_1^2 - x_2^2\right| = |x_1 - x_2||x_1 + x_2| \leq 2\frac{\rho}{2}|x_1 - x_2| = \rho|x_1 - x_2|.$$

5. Linear function $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined by $f(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle + b$, where $\mathbf{v} \in \mathbb{R}^n$ is $\|\mathbf{v}\|$ $-$Lipschitz. By using Cauchy-Schwartz inequality, we have

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

The following Lemma shows that composition of Lipschitz functions preserves Lipschitzness.

Lemma (Composition of Lipschitz functions)

Let $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$, where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz. The f is $(\rho_1\rho_2)$-Lipschitz. In particular, if g_2 is the linear function, $g_2(\mathbf{x}) = \langle \mathbf{v}, \mathbf{x} \rangle + b$, for some $\mathbf{v} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, then f is $(\rho_1 \|\mathbf{v}\|)$-Lipschitz.

Proof (Composition of Lipschitz functions).

$$\begin{aligned} |f(\mathbf{w}_1) - f(\mathbf{w}_2)| &= |g_1(g_2(\mathbf{w}_1)) - g_1(g_2(\mathbf{w}_2))| \\ &\leq \rho_1 \|g_2(\mathbf{w}_1) - g_2(\mathbf{w}_2)\| \\ &\leq \rho_1\rho_2 \|\mathbf{w}_1 - \mathbf{w}_2\| . \end{aligned}$$

□

Smoothness

- The definition of a smooth function relies on the notion of gradient.
- Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a differentiable function at $\mathbf{w}$ and its gradient as

$$\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_n} \right).$$

- Smoothness of f is defined as

Definition (Smoothness)

A differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is β-smooth if its gradient is β-**Lipschitz**; namely, for all $\mathbf{v}, \mathbf{w}$ we have $\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|$.

- **Show that smoothness implies that or all $\mathbf{v}, \mathbf{w}$ we have**

$$f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|^2 . \tag{1}$$

  while convexity of f implies that

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

- When a function is both convex and smooth, we have both upper and lower bounds on the difference between the function and its first order approximation.
- Setting $\mathbf{v} = \mathbf{w} - \frac{1}{\beta} \nabla f(\mathbf{w})$ in rhs of (1), we obtain

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

- We had

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

- Let $f(\mathbf{v}) \geq 0$ for all $\mathbf{v}$, then smoothness implies that

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

- A function that satisfies this property is also called a self-bounded function.

Example (Smooth functions)

1. Function $f(x) = x^2$ is 2-smooth. This can be shown from $f'(x) = 2x$.
2. Function $f(x) = \log(1 + e^x)$ is $\left(\frac{1}{4}\right)$-smooth. Since $f'(x) = \frac{1}{1+e^{-x}}$, we have

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

   Hence f' is $\left(\frac{1}{4}\right)$-Lipshitz.

Lemma (Composition of smooth scaler function)

Let $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + b)$, where $g : \mathbb{R} \mapsto \mathbb{R}$ is a β-smooth function and $\mathbf{x} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then, f is $\left(\beta \left\| \mathbf{x} \right\|^2\right)$-smooth.

Proof (Composition of smooth scaler function).

1. By using the chain rule we have $\nabla f(\mathbf{w}) = g'(\langle \mathbf{w}, \mathbf{x} \rangle + b)\, \mathbf{x}$.
2. Using smoothness of g and Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
f(\mathbf{v}) &= g(\langle \mathbf{v}, \mathbf{x} \rangle + b) \\
&\leq g(\langle \mathbf{w}, \mathbf{x} \rangle + b) + g'(\langle \mathbf{v}, \mathbf{x} \rangle + b) \langle \mathbf{v} - \mathbf{w}, \mathbf{x} \rangle + \frac{\beta}{2} (\langle \mathbf{v} - \mathbf{w}, \mathbf{x} \rangle)^2 \\
&\leq g(\langle \mathbf{w}, \mathbf{x} \rangle + b) + g'(\langle \mathbf{v}, \mathbf{x} \rangle + b) \langle \mathbf{v} - \mathbf{w}, \mathbf{x} \rangle + \frac{\beta}{2} (\|\mathbf{v} - \mathbf{w}\| \|\mathbf{x}\|)^2 \\
&\leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta \|\mathbf{x}\|^2}{2} \|\mathbf{v} - \mathbf{w}\|^2 .
\end{aligned}$$

□

Example (Smooth functions)

1. For any $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, let $f(\mathbf{w}) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$. Then, f is $\left(2 \left\| \mathbf{x} \right\|^2\right)$-smooth.
2. For any $\mathbf{x} \in \mathbb{R}^n$ and $y \in \{\pm 1\}$, let $f(x) = \log(1 + exp(-y \langle \mathbf{w}, \mathbf{x} \rangle))$. Then, f is $\left(\frac{\|\mathbf{x}\|^2}{4}\right)$-smooth.

Convex learning problems

- Approximately solve

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbb{C}} f(\mathbf{w})$$

where $\mathbb{C}$ is a convex set and f is a convex function.

Example (Convex optimization)

The linear regression problem can be defined as the following convex optimization problem.

$$\operatorname*{argmin}_{\|\mathbf{w}\|\leq 1} \frac{1}{m}\sum_{i=1}^{m}\left[\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i\right]^2$$

- An special case is unconstrained minimization $\mathbb{C} = \mathbb{R}^n$.
- Can reduce one to another
  1. Adding the function $I_{\mathbb{C}}(\mathbf{w})$ to the objective eliminates the constraint.
  2. Adding the constraint $f(\mathbf{w}) \leq f^* + \epsilon$ eliminates the objective.

Definition (Agnostic PAC learnability)

A hypothesis class H is **agnostic PAC learnable** with respect to a set $\mathcal{Z}$ and a loss function $\ell : H \times \mathcal{Z} \mapsto \mathbb{R}_+$, if there exist a function $m_H : (0,1)^2 \mapsto \mathbb{N}$ and a learning algorithm A with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{Z}$, when running the learning algorithm on $m \geq m_H(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns $h \in H$ such that, with probability of at least $(1-\delta)$ (over the choice of the m training examples),

$$\mathbf{R}(h) \leq \min_{h' \in H} \hat{\mathbf{R}}(h) + \epsilon,$$

where $\mathbf{R}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$.

In this definition, we have

1. a hypothesis class H,
2. a set of examples $\mathcal{Z}$, and
3. a loss function $\ell : H \times \mathcal{Z} \mapsto \mathbb{R}_+$

Now, we consider hypothesis classes H that are subsets of the Euclidean space $\mathbb{R}^n$, therefore, denote a hypothesis in H by $\mathbf{w}$.

Definition (Convex learning problems)

A learning problem $(H, \mathcal{Z}, \ell)$ is called convex if

1. the hypothesis class H is a convex set, and
2. for all $z \in \mathcal{Z}$, the loss function, $\ell(., z)$, is a convex function, where, for any z, $\ell(., z)$ denotes the function $f : H \mapsto \mathbb{R}$ defined by $f(\mathbf{w}) = \ell(\mathbf{w}, z)$.

Example (Linear regression with the squared loss)

1. The domain set $\mathcal{X} \subset \mathbb{R}^n$ and the label set $\mathcal{Y} \subset \mathbb{R}$ is the set of real numbers.
2. We need to learn a linear function $h : \mathbb{R}^n \mapsto \mathbb{R}$ that best approximates the relationship between our variables.
3. Let H be the set of homogeneous linear functions $H = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \mid \mathbf{w} \in \mathbb{R}^n\}$.
4. Let the squared loss function $\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^2$ used to measure error.
5. This is a convex learning problem because
   - Each linear function is parameterized by a vector $\mathbf{w} \in \mathbb{R}^n$. Hence, $H = \mathbb{R}^n$.
   - The set of examples is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{n+1}$.
   - The loss function is $\ell(\mathbf{w}, (\mathbf{x}, y)) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$.
   - Clearly, H is a convex set and $\ell(., .)$ is also convex with respect to its first argument.

Lemma (Convex learning problems)

If ℓ is a convex loss function and the class H is convex, then the erm_H problem, of minimizing the empirical loss over H, is a convex optimization problem (that is, a problem of minimizing a convex function over a convex set).

Proof (Convex learning problems).

1. The erm_H problem is defined as

$$erm_H(S) = \underset{\mathbf{w} \in H}{\text{argmin}}\ \hat{\mathbf{R}}(\mathbf{w})$$

2. Since, for a sample $S = \{z_1, \ldots, z_m\}$, for every $\mathbf{w}$, and $\hat{\mathbf{R}}(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{w}, z_i)$, Lemma (Convexity of a scaler function) implies that $\hat{\mathbf{R}}(\mathbf{w})$ is a convex function.
3. Therefore, the erm_H rule is a problem of minimizing a convex function subject to the constraint that the solution should be in a convex set.

□

- We have seen that for many cases implementing the *erm* rule for convex learning problems can be done **efficiently**.
- Is convexity a sufficient condition for the learnability of a problem?
- In VC theory, we saw that halfspaces in n- dimension are learnable (perhaps inefficiently).
- Using discretization trick, if the problem is of n parameters, it is learnable with a sample complexity being a function of n.
- That is, for a constant n, the problem should be learnable.
- Maybe all convex learning problems over $\mathbb{R}^n$, are learnable?
- Answer is negative even when n is low (Show that linear regression is not learnable even if $n = 1$).
- Hence, all convex learning problems over $\mathbb{R}^n$ are not learnable.
- Under some additional restricting conditions that hold in many practical scenarios, convex problems are learnable.
- A possible solution to this problem is to add another constraint on the hypothesis class.
- In addition to the convexity requirement, we require that H will be bounded (i.e. For some predefined scalar B, every hypothesis $\mathbf{w} \in H$ satisfies $\|\mathbf{w}\| \leq B$).
- Boundedness and convexity alone are still not sufficient for ensuring that the problem is learnable (Show that a linear regression with squared loss and $H = \{w \mid |w| \leq 1\} \subset \mathbb{R}$ is not learnability).

Definition (Convex-Lipschitz-bounded learning problems)

A learning problem $(H, \mathcal{Z}, \ell)$ is called convex-Lipschitz-bounded, with parameters ρ, B if the following hold.

1. The hypothesis class H is a convex set, and for all $\mathbf{w} \in H$ we have $\|\mathbf{w}\| \leq B$.
2. For all $z \in \mathcal{Z}$, the loss function, $\ell(., z)$, is a convex and ρ-Lipschitz function.

Example (Linear regression with absolute-value loss)

1. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq \rho\}$ and $\mathcal{Y} \subset \mathbb{R}$.
2. Let $H = \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{w}\| \leq B\}$.
3. Let loss function be $\ell(\mathbf{w}, (\mathbf{x}, y)) = |\langle \mathbf{w}, \mathbf{x} \rangle - y|$.
4. Then, this problem is Convex-Lipschitz-bounded with parameters ρ, B.

Definition (Convex-smooth-bounded learning problems)

A learning problem $(H, \mathcal{Z}, \ell)$ is called convex-smooth-bounded, with parameters β, B if the following hold.

1. The hypothesis class H is a convex set, and for all $\mathbf{w} \in H$ we have $\|\mathbf{w}\| \leq B$.
2. For all $z \in \mathcal{Z}$, the loss function, $\ell(., z)$, is a convex, nonnegative and β-smooth function.

Example (Linear regression with squared loss)

1. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq \beta/2\}$ and $\mathcal{Y} \subset \mathbb{R}$.
2. Let $H = \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{w}\| \leq B\}$.
3. Let loss function be $\ell(\mathbf{w}, (\mathbf{x}, y)) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$.
4. Then, this problem is Convex-smooth-bounded with parameters β, B.

Lemma (Learnability of Convex-Lipschitz/-smooth-bounded learning problems)

The following two families of learning problems are learnable.

1. *Convex-smooth-bounded learning problems.*
2. *Convex-Lipschitz-bounded learning problems.*

That is, the properties of convexity, boundedness, and Lipschitzness or smoothness of the loss function are sufficient for learnability.

Surrogate loss functions

- In many cases, loss function is not convex and, hence, implementing the ERM rule is hard.
- Consider the problem of learning halfspaces with respect to 0-1 loss.

$$\ell^{0-1}(\mathbf{w}, (\mathbf{x}, y)) = \mathbb{I}\left[y \neq \operatorname{sgn}\left(\langle \mathbf{w}, \mathbf{x} \rangle\right)\right] = \mathbb{I}\left[y \langle \mathbf{w}, \mathbf{x} \rangle \leq 0\right].$$

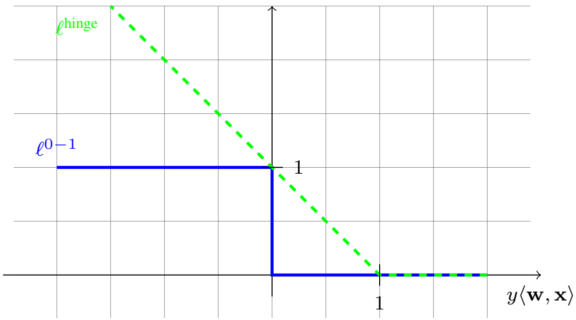
- This loss function is not convex with respect to $\mathbf{w}$.
- When trying to minimize $\hat{\mathbf{R}}(\mathbf{w})$ with respect to this loss function we might encounter local minima.
- We also showed that, solving the ERM problem with respect to the 0-1 loss in the unrealizable case is known to be NP-hard.
- One popular approach is to upper bound the nonconvex loss function by a convex surrogate loss function.
- The requirements from a convex surrogate loss are as follows:
  1. It should be convex.
  2. It should upper bound the original loss.

Hinge-loss

- Hinge-loss function is defined as

$$\ell^{hinge}(\mathbf{w}, (\mathbf{x}, y)) \triangleq \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}.$$

- Hinge-loss has the following two properties
  1. For all $\mathbf{w}$ and all $(\mathbf{x}, y)$, we have $\ell^{0-1}(\mathbf{w}, (\mathbf{x}, y)) \leq \ell^{hinge}(\mathbf{w}, (\mathbf{x}, y))$.
  2. Hinge-loss is a convex function.



- Hence, the hinge loss satisfies the requirements of a convex surrogate loss function for the zero-one loss.

- Suppose we have a learner for hinge-loss that guarantees

$$\mathbf{R}^{hinge}(A(S)) \leq \min_{\mathbf{w} \in H} \mathbf{R}^{hinge}(\mathbf{w}) + \epsilon.$$

- Using the surrogate property,

$$\mathbf{R}^{0-1}(A(S)) \leq \min_{\mathbf{w} \in H} \mathbf{R}^{hinge}(\mathbf{w}) + \epsilon.$$

- We can further rewrite the upper bound as

$$\mathbf{R}^{0-1}(A(S)) \leq \min_{\mathbf{w} \in H} \mathbf{R}^{0-1}(\mathbf{w}) + \left(\min_{\mathbf{w} \in H} \mathbf{R}^{hinge}(\mathbf{w}) - \min_{\mathbf{w} \in H} \mathbf{R}^{0-1}(\mathbf{w}) \right) + \epsilon$$
$$= \epsilon_{approximation} + \epsilon_{optimization} + \epsilon_{estimation}$$

- The optimization error is a result of our inability to minimize the training loss with respect to the original loss.

Assignments

1. Please specify that the following learning problems belong to which category of problems.
   - Support vector regression (SVR)
   - Kernel ridge regression
   - Least absolute shrinkage and selection operator (Lasso)
   - Support vector machine (SVM)
   - Logistic regression
   - AdaBoost

   Prove your claim.
2. Prove Lemma Learnability of Convex-Lipschitz/-smooth-bounded learning problems.

Summary

Summary

- We introduced two families of learning problems:
  1. Convex-Lipschitz-bounded learning problems.
  2. Convex-smooth-bounded learning problems.
- There are some generic learning algorithms such as stochastic gradient descent algorithm for solving these problem. (Please read Chapter 14)
- We also introduced the notion of convex surrogate loss function, which enables us also to utilize the convex machinery for nonconvex problems.

Readings

1. Chapters 12 and 14 of Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning : From theory to algorithms*. Cambridge University Press, 2014.

References

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning : From theory to algorithms*. Cambridge University Press, 2014.

Questions?