Machine learning theory

Convex learning problems

Hamid Beigy

Sharif university of technology

May 28, 2022

Table of contents

Introduction

1. 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.
2. In general, a convex learning problem is a problem.
   - whose hypothesis class is a convex set and
   - whose loss function is a convex function for each example.

1. Other properties of the loss function that facilitate successful learning are
   - Lipschitzness
   - Smoothness
2. Are convex smooth learning problems and convex Lipschitz-pounded problems learnable?
3. In this session, we study the learnability of
   - Convex smooth-bounded problems
   - Convex Lipschitz-pounded problems

Mathematical backgrounds

1. For convex learning problems, we need to study the following properties of loss functions.
   - Convexity
   - Lipshitzness
   - Smoothness
   - Strong convexity

Mathematical backgrounds

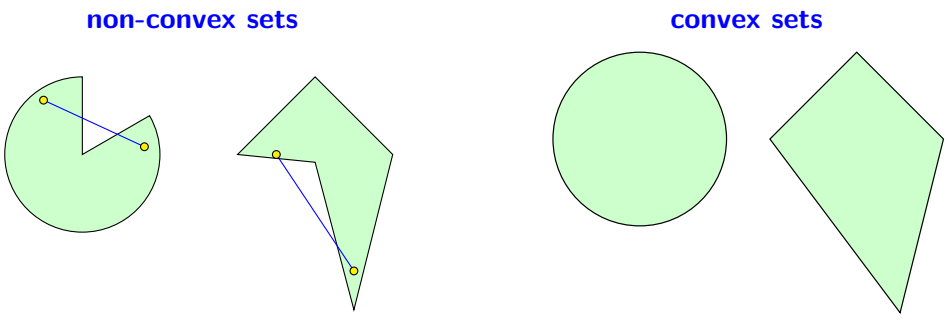
Convexity

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 $\mathbb{R}^2$



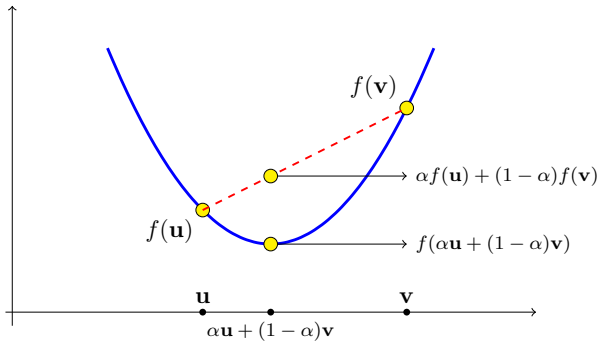
Definition (Convex function)

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 $\alpha \in [0, 1]$,

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

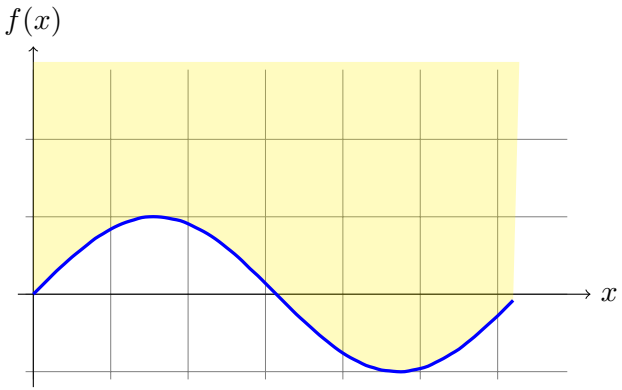
In words, f is convex if for any $\mathbf{u}, \mathbf{v} \in C$, the graph of f between $\mathbf{u}$ and $\mathbf{v}$ lies below the line segment joining $f(\mathbf{u})$ and $f(\mathbf{v})$.

Example (Convex function)



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

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



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

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

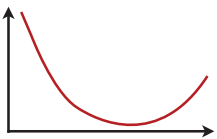
   - If f is convex, we also have that

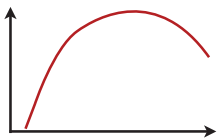
$$f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u})) = f(\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}) \leq (1 - \alpha)f(\mathbf{u}) + \alpha f(\mathbf{v}).$$

   - 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**.



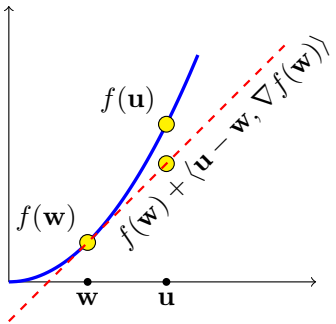


2. If f is convex and differentiable, then

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

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

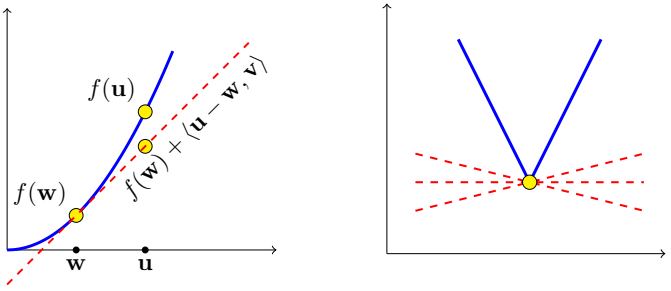
- If f is convex, for every $\mathbf{w}$, we can construct a tangent to f at $\mathbf{w}$ that lies below f everywhere.
- If f is differentiable, this tangent is the linear function $l(\mathbf{u}) = f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle$.



1. $\mathbf{v}$ is sub-gradient of f at $\mathbf{w}$ if $\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle$
2. 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 $\mathbf{w}$, $\partial f(\mathbf{w}) \neq 0$.



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

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

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

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

$$g(\alpha\mathbf{u}+(1-\alpha)\mathbf{v}) = \max_i f_i(\alpha\mathbf{u}+(1-\alpha)\mathbf{v}) \leq \max_i [\alpha f_i(\mathbf{u})+(1-\alpha)f_i(\mathbf{v})]$$
$$\leq \alpha \max_i f_i(\mathbf{u}) + (1-\alpha)\max_i f_i(\mathbf{v}) = \alpha g(\mathbf{u}) + (1-\alpha)g(\mathbf{v}).$$

2. The second claim follows by

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

□

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**.

Mathematical backgrounds

Lipschitzness

1. 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

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

2. A Lipschitz function cannot change too fast.
3. If $f : \mathbb{R} \mapsto \mathbb{R}$ is differentiable, then by the mean value theorem we have

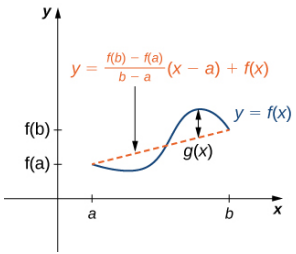
$$f(w_1) - f(w_2) = f'(u)(w_1 - w_2),$$

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

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$



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

Lemma

If f is convex then f is ρ-Lipschitz iff the norm of all sub-gradients of f is at most ρ.

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

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

Example (Lipschitzness)

4. Function $f(x) = x^2$ is ρ-Lipschitz over set $C = \left\{x \;\middle|\; |x| \leq \frac{\rho}{2}\right\}$. Indeed, 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).

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

□

Mathematical backgrounds

Smoothness

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

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

3. 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 $\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|$.

4. **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

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

1. 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.
2. Setting $\mathbf{v} = \mathbf{w} - \frac{1}{\beta}\nabla f(\mathbf{w})$ in rhs of (1), we obtain

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

3. We had

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

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

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

5. 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 $f'(x) = \dfrac{1}{1 + e^{-x}}$, we have

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

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

Convex learning problems

1. Approximately solve

$$\arg\min_{\mathbf{w}\in\mathbb{C}} f(\mathbf{w})$$

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.

$$\arg\min_{\|\mathbf{w}\|\leq 1} \frac{1}{m}\sum_{i=1}^{m}[\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i]^2$$

2. Special cases
   - Feasibility problem: f is a constant function.
   - Unconstrained minimization: $\mathbb{C} = \mathbb{R}^n$.
3. Can reduce one to another
   - Adding the function $I_{\mathbb{C}}(\mathbf{w})$ to the objective eliminates the constraint.
   - 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

$$erm_H(S) = \arg\min_{\mathbf{w} \in H} \hat{\mathbf{R}}(\mathbf{w})$$

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.

□

1. We have seen that for many cases implementing the *erm* rule for convex learning problems can be done **efficiently**.
2. Is convexity a sufficient condition for the learnability of a problem?
3. In VC theory, we saw that halfspaces in n- dimension are learnable (perhaps inefficiently).
4. Using discretization trick, if the problem is of n parameters, it is learnable with a sample complexity being a function of n.
5. That is, for a constant n, the problem should be learnable.
6. Maybe all convex learning problems over $\mathbb{R}^n$, are learnable?
7. Answer is negative even when n is low.

Example (Nonlearnability of linear regression)

1. Let $H = \mathbb{R}$ and $\ell(w, (x, y)) = (wx - y)^2$.
2. This is a convex learning problem.
3. Let A be any deterministic algorithm (successful PAC learner for this problem).
4. So, there exists a function $m(.,.)$, such that for every distribution $\mathcal{D}, \epsilon, \delta$, if A receives a sample of size $m \geq m(\epsilon, \delta)$, it should output $\hat{w} = A(S)$ such that

$$\mathbb{P}\left[\mathbf{R}(\hat{w}) - \hat{\mathbf{R}}(\hat{w}) \leq \epsilon\right] \geq 1 - \delta. \tag{2}$$

5. Choose $\epsilon = 0.01$ and $\delta = 0.5$ and let $m \geq m(\epsilon, \delta)$ and set $\mu = \frac{log(100/99)}{2m}$.
6. We can define two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ that A is likely to fail on at least one of them.
7. $\mathcal{D}_1$ and $\mathcal{D}_2$ are supported on two examples $z_1 = (1, 0)$ and $z_2 = (\mu, -1)$, where

$$\mathcal{D}_1(z_1) = \mu \qquad \mathcal{D}_1(z_2) = 1 - \mu$$
$$\mathcal{D}_2(z_1) = 0 \qquad \mathcal{D}_2(z_2) = 1.$$

Example (Nonlearnability of linear regression (cont.))

8. For $\mathcal{D}_1$ and $\mathcal{D}_2$, the probability that all examples of the training set will be z_2 is at least 99%.
9. This probability under $\mathcal{D}_1$ equals to $(1-\mu)^m \geq e^{-2\mu m} = 0.99$.
10. Let $\hat{w} = A(S)$, where S containing all z_2 examples.
11. We will argue that no matter what value $\hat{w}$ takes, the condition(2) will be violated under $\mathcal{D}_1$ or $\mathcal{D}_2$, and therefore **the problem is not PAC learnable**.
    - Suppose $\hat{w} < -\frac{1}{2\mu}$. We can show that condition (2) is violated under $\mathcal{D}_1$. In particular,
    $$\mathbf{R}_{\mathcal{D}_1}(\hat{w}) \geq \mathcal{D}_1(z_1)\ell(\hat{w}, z_1) = \mu(\hat{w})^2,$$
    whereas
    $$\mathbf{R}_{\mathcal{D}_1}(w) = \mathbf{R}_{\mathcal{D}_1}(0) = \mathcal{D}_1(z_1)\ell(0, z_2) = (1-\mu),$$
    and therefore
    $$\mathbf{R}_{\mathcal{D}_1}(\hat{w}) - \min_w \mathbf{R}_{\mathcal{D}_1}(w) \geq \frac{1}{4\mu} - (1-\mu) \geq \epsilon.$$
    - Therefore, **such algorithm A fails on $\mathcal{D}_1$**.

Example (Nonlearnability of linear regression (cont.))

12. We will argue that no matter what value $\hat{w}$ takes, the condition(2) will be violated under $\mathcal{D}_1$ or $\mathcal{D}_2$, and therefore **the problem is not PAC learnable**.
    - Suppose $\hat{w} \geq -\frac{1}{2\mu}$. We can show that condition (2) is violated under $\mathcal{D}_2$. In particular,
    $$\mathbf{R}_{\mathcal{D}_2}(\hat{w}) \geq \ell(\hat{w}, z_2) = (\hat{w}\mu + 1)^2 \geq \frac{1}{4},$$
    whereas
    $$\mathbf{R}_{\mathcal{D}_2}(w, \mathcal{D}_2) = 0,$$
    and therefore
    $$\mathbf{R}_{\mathcal{D}_2}(\hat{w}) - \min_w \mathbf{R}_{\mathcal{D}_2}(w) \geq \epsilon.$$
    - Therefore, **such algorithm A fails on $\mathcal{D}_2$**.
13. **In summary, we have shown that for every A there exists a distribution on which A fails, which implies that the problem is not PAC learnable.**

1. Hence, all convex learning problems over $\mathbb{R}^n$ are not learnable.
2. Under some additional restricting conditions that hold in many practical scenarios, convex problems are learnable.
3. A possible solution to this problem is to add another constraint on the hypothesis class.
4. 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 $\|\mathbf{w}\| \leq B$).
5. Boundedness and convexity alone are still not sufficient for ensuring that the problem is learnable .
6. **Homework:** Show that a linear regression with squared loss and $H = \{w \mid |w| \leq 1\} \subset \mathbb{R}$ is not learnabile.

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 $z \in \mathcal{Z}$, loss function, $\ell(., z)$, is a convex, nonnegative and β-smooth function.

Example (Linear regression with squared loss)

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

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

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

1. **convexity**,
2. **boundedness**, *and*
3. **Lipschitzness** *or* **smoothness**

of the loss function are **sufficient** *for learnability.*

Regularization and stability

1. Regularized loss minimization (RLM) is a learning rule in which we jointly minimize the empirical risk and a regularization function.
2. A regularization function is a mapping $R : \mathbb{R} \mapsto \mathbb{R}$, and the regularized loss minimization rule outputs a hypothesis

$$\arg\min_{\mathbf{w}} \hat{\mathbf{R}}(\mathbf{w}) + R(\mathbf{w})$$

3. Regularization function measures the complexity of hypotheses.
4. RLM is similar to MDL and SRM.
5. There are many possible regularization functions one can use, reflecting some prior belief about the problem.
6. We consider on one of the most simple regularization functions $R(\mathbf{w}) = \|\mathbf{w}\|^2$.
7. The learning rule becomes

$$A(S) = \arg\min_{\mathbf{w}} \hat{\mathbf{R}}(\mathbf{w}) + \lambda \|\mathbf{w}\|^2$$

8. This type of regularization function is often called **Tikhonov regularization**.

Example (Ridge regression)

1. Applying RLM with Tikhonov regularization to linear regression with squared loss, we obtain

$$\arg\min_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i\right)^2 + \lambda \left\|\mathbf{w}\right\|_2^2$$

2. Setting gradient of objective to zero, we obtain

$$(2\lambda m\mathbf{I} + \mathbf{A}) = \mathbf{b}$$

where $\mathbf{A}$ and $\mathbf{b}$ are

$$\mathbf{A} = \left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^\top\right) \qquad \mathbf{b} = \sum_{i=1}^{m} y_i\mathbf{x}_i$$

3. $\mathbf{A}$ is positive semidefinite, and $2\lambda m\mathbf{I} + \mathbf{A}$ has all its eigenvalues bounded below by $2\lambda m$.
4. Hence, this matrix is invertible, we have $\mathbf{w} = (2\lambda m\mathbf{I} + \mathbf{A})^{-1}\mathbf{b}$.

Regularization and stability

Stability of learning algorithms

1. A learning algorithm is stable if a **small change of input does not change output much**.
2. Given training set S and an additional example $\mathbf{z}'$, let $S^{(i)}$ be training set obtained by replacing ith example of S with $\mathbf{z}'$: $S^{(i)} = (\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}', \mathbf{z}_{i+1}, \ldots, \mathbf{z}_m)$.
3. A small change of input means feeding A with $S^{(i)}$ instead of S.
4. We measure effect of this change on output of A by comparing $\ell(A(S), z_i)$ with $\ell(A(S^{(i)}), z_i)$.
5. A good learning algorithm will have $\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \geq 0$.
6. If $\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)$ is is very large, we suspect that the learning algorithm might overfit.
7. The reason is learning algorithm **drastically changes its prediction** on $\mathbf{z}_i$ if it observes $\mathbf{z}_i$ in the training set.

Theorem

Let $\mathcal{D}$ be a distribution. Let $S = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$ be a sequence of iid examples sampled according to $\mathcal{D}$ and let $\mathbf{z}'$ be another iid example sampled from $\mathcal{D}$. Let $U(m)$ be the uniform distribution over $[m]$. Then, for any learning algorithm A

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\mathbf{R}(A(S)) - \hat{\mathbf{R}}(A(S))\right] = \mathop{\mathbb{E}}_{(S,\mathbf{z}'')\sim\mathcal{D}_1^{m+1}, i\sim U(m)}\left[\ell(A(S^{(i)}), \mathbf{z}_i)) - \ell(A(S), \mathbf{z}_i))\right].$$

Proof.

1. Since S and $\mathbf{z}'$ are both drawn iid from $\mathcal{D}$, for every i, we have

$$\mathop{\mathbb{E}}_{S}[\mathbf{R}(A(S))] = \mathop{\mathbb{E}}_{S,z'}\left[\ell(A(S), \mathbf{z}')\right] = \mathop{\mathbb{E}}_{S,z'}\left[\ell(A(S^{(i)}), \mathbf{z}')\right]$$

2. On the other hand, we can write

$$\mathop{\mathbb{E}}_{S}\left[\hat{\mathbf{R}}(A(S))\right] = \mathop{\mathbb{E}}_{S,i}[\ell(A(S), \mathbf{z}_i)]$$

3. By combining these two equations, we conclude our proof.

□

1. When $\mathbb{E}_{S,z'}\left[\ell(A(S), \mathbf{z}')\right] = \mathbb{E}_{S,z'}\left[\ell(A(S^{(i)}), \mathbf{z}')\right]$ is small, we say that A is stable.
2. This means that **changing a single example in the training set does not lead to a significant change**.

Definition (On-average-replace-one-stable)

Let $\epsilon : \mathbb{N} \mapsto \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm A is **on-average-replace-one-stable** with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$

$$\mathop{\mathbb{E}}_{(S,\mathbf{z}'')\sim\mathcal{D}_1^{m+1},i\sim U(m)}\left[\ell(A(S^{(i)}), \mathbf{z}_i)) - \ell(A(S), \mathbf{z}_i))\right] \leq \epsilon(m).$$

3. The preceding theorem states that a learning algorithm does not overfit **if and only if it is on-average-replace-one-stable**.
4. Of course, a learning algorithm that does not overfit is not necessarily a good learning algorithm.
5. A useful algorithm should find a hypothesis that **fits the training set and does not overfit**.

The main property of Tikhonov regularization is that it makes objective of RLM **strongly convex**.

Definition (Strongly convex functions)

A function f is λ-strongly convex if for all $\mathbf{u}, \mathbf{w}$ and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\left\|\mathbf{w}-\mathbf{u}\right\|^2.$$

Lemma

The following statements hold for strongly convex functions.

1. *The function $f(\mathbf{w}) = \left\|\mathbf{w}\right\|^2$ is 2λ-strongly convex.*
2. *If function f is λ-strongly convex and function g is convex, then $f+g$ is λ-strongly convex.*
3. *If function f is λ-strongly convex and $\mathbf{u}$ is a minimizer of f, then for any $\mathbf{w}$*

$$f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}\left\|w-u\right\|^2.$$

1. We now turn to prove that **RLM** (algorithm A) is stable.
2. The goal is to bound $|\hat{\mathbf{R}}_S(A(S^{(i)})) - \hat{\mathbf{R}}_S(A(S))|$ for Tikhonov regularization.
3. Define $f_S(\mathbf{w}) = \hat{\mathbf{R}}_S(\mathbf{w}) + \lambda \|\mathbf{w}\|^2$ and $A(S) = \arg\min_{\mathbf{w}} f_S(\mathbf{w})$.
4. Since $f_S(\mathbf{w})$ is 2λ-strongly convex, then for any $\mathbf{v}$, we have

$$f_S(\mathbf{v}) - f_S(A(S)) \geq \lambda \|\mathbf{v} - A(S)\|^2 \tag{3}$$

5. Also for any $\mathbf{v}, \mathbf{u}$ and i, we have

$$\begin{aligned} f_S(\mathbf{v}) - f_S(\mathbf{u}) &= \hat{\mathbf{R}}_S(\mathbf{v}) + \lambda \|\mathbf{v}\|^2 - (\hat{\mathbf{R}}_S(\mathbf{u}) + \lambda \|\mathbf{u}\|^2) \\ &= \hat{\mathbf{R}}_{S^{(i)}}(\mathbf{v}) + \lambda \|\mathbf{v}\|^2 - (\hat{\mathbf{R}}_{S^{(i)}}(u) + \lambda \|\mathbf{v}\|^2) \\ &+ \frac{\ell(\mathbf{v}, \mathbf{z}_i) - \ell(\mathbf{u}, \mathbf{z}_i)}{m} + \frac{\ell(\mathbf{u}, \mathbf{z}') - \ell(\mathbf{v}, \mathbf{z}')}{m} \end{aligned}$$

1. Let $\mathbf{v} = A(S^{(i)})$ and $\mathbf{u} = A(S)$, thus we obtain

$$f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}), \mathbf{z}_i) - \ell(A(S), \mathbf{z}_i)}{m} + \frac{\ell(A(S), \mathbf{z}') - \ell(A(S^{(i)}), \mathbf{z}')}{m}.$$

2. Combining this with inequality (3), we obtain

$$\lambda \left\| A(S^{(i)}) - A(S) \right\|^2 \leq \frac{\ell(A(S^{(i)}), \mathbf{z}_i) - \ell(A(S), \mathbf{z}_i)}{m} + \frac{\ell(A(S), \mathbf{z}') - \ell(A(S^{(i)}), \mathbf{z}')}{m}. \tag{4}$$

Theorem (Stability for Lipschitz loss function)

Assume that the loss function is convex and ρ-Lipschitz. Then, RLM rule with the regularizer $\lambda \|\mathbf{w}\|^2$ is **on-average-replace-one-stable** *with rate $\frac{2\rho^2}{\lambda m}$.*

Proof (Stability for Lipschitz loss function).

1. Since the loss function is convex and ρ-Lipschitz. Then

$$\begin{aligned} \ell(A(S^{(i)}), \mathbf{z}_i) - \ell(A(S), \mathbf{z}_i) &\leq \rho \left\| A(S^{(i)}) - A(S) \right\| \\ \ell(A(S), \mathbf{z}') - \ell(A(S^{(i)}), \mathbf{z}') &\leq \rho \left\| A(S^{(i)}) - A(S) \right\|. \end{aligned} \tag{5}$$

2. Plugging these inequalities into (4), we obtain

$$\lambda \left\| A(S^{(i)}) - A(S) \right\|^2 \leq 2\rho \frac{\left\| A(S^{(i)}) - A(S) \right\|}{m} \quad \Leftrightarrow \left\| A(S^{(i)}) - A(S) \right\| \leq \frac{2\rho}{\lambda m}.$$

3. Plugging these inequalities into (5), we obtain

$$\ell(A(S^{(i)}), \mathbf{z}_i) - \ell(A(S), \mathbf{z}_i) \leq \frac{2\rho^2}{\lambda m} \qquad \text{which holds for all } S, \mathbf{z}', i.$$

□

Theorem

Assume that the loss function is β-smooth and nonnegative. Then, RLM rule with the regularizer $\lambda \|\mathbf{w}\|^2$ where $\lambda \geq \frac{2\beta}{m}$ satisfies

$$\begin{aligned}\mathbb{E}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\right] &\leq \frac{48\beta}{\lambda m}\mathbb{E}\left[\hat{\mathbf{R}}_S(A(S))\right] \\ &\leq \frac{48\beta}{\lambda m}C\end{aligned}$$

where $\hat{\mathbf{R}}_S(A(S)) \leq C$.

1. We have shown that the stability term decreases, when λ increases.
2. Also the empirical risk increases with λ.
3. Hence, there is a tradeoff between fitting and overfitting.
4. We will choose a value of λ to derive a new bound for the true risk.

Theorem (Oracle inequality)

Assume that the convex and ρ-Lipschitz. Then, the RLM rule with the regularization function $\lambda \|\mathbf{w}\|^2$ satisfies

$$\forall \mathbf{w}^* : \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} [\mathbf{R}(A(S))] \leq \mathbf{R}(\mathbf{w}^*) + \lambda \|\mathbf{w}^*\|^2 + \frac{2\rho^2}{\lambda m}$$

5. This bound is often called an **oracle inequality**.
6. If we think of $\mathbf{w}^*$ as a hypothesis with low risk, the bound tells us how many examples are needed so that $A(S)$ will be almost as good as $\mathbf{w}^*$, had we known the norm of $\mathbf{w}^*$.
7. We do not know $\|\mathbf{w}^*\|^2$ and tune λ using validation set.

Proof (Oracle inequality).

1. We can rewrite the expected risk of a learning algorithm as

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathbf{R}(A(S))] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\hat{\mathbf{R}}(A(S))\right] + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\mathbf{R}(A(S)) - \hat{\mathbf{R}}(A(S))\right] \tag{6}$$

2. Fix some arbitrary vector $\mathbf{w}^*$. We have

$$\hat{\mathbf{R}}(A(S)) \leq \hat{\mathbf{R}}(A(S)) + \lambda\|\mathbf{w}\|^2 \leq \hat{\mathbf{R}}(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2$$

3. Taking expectation and using $\mathbb{E}_S\left[\hat{\mathbf{R}}(\mathbf{w}^*)\right] = \mathbf{R}(\mathbf{w}^*)$, yields

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\hat{\mathbf{R}}(A(S))\right] \leq \mathbf{R}(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2.$$

4. Using equation (6), we obtain

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathbf{R}(A(S))] \leq \mathbf{R}(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2 + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\mathbf{R}(A(S)) - \hat{\mathbf{R}}(A(S))\right].$$

5. Applying Theorem (Stability for Lipschitz loss function) finishes the proof.

□

Theorem

Let $(H, \mathcal{Z}, \ell)$ be a convex-Lipschitz-bounded learning problem with parameters ρ, B. For any training set size m, let $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$. Then,the RLM rule with the regularization function $\lambda \|\mathbf{w}^*\|^2$ satisfies

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathbf{R}(A(S))] \leq \min_{\mathbf{w}\in H} \mathbf{R}(\mathbf{w}) + \rho B\sqrt{\frac{8}{m}}.$$

In particular for every $\epsilon > 0$, if $m > \frac{8\rho^2 B^2}{\epsilon^2}$ then for every distribution $\mathcal{D}$, we have

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathbf{R}(A(S))] \leq \min_{\mathbf{w}\in H} \mathbf{R}(\mathbf{w}) + \epsilon.$$

Proof.

The proof follows directly by setting $\mathbf{w}^*$ to $\arg\min_{\mathbf{w}\in H} \mathbf{R}(\mathbf{w})$, inserting λ in Theorem (Stability for Lipschitz loss function) and using $\|\mathbf{w}\| \leq B$. □

Regularization and stability

Learnability of convex learning problems

Convex-Lipschitz-bound problems are PAC learnable, as the following Theorem shows.

Theorem

If an algorithm A guarantees that for $m \geq m_H(\epsilon)$ and every distribution $\mathcal{D}$ the following inequality holds:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\mathbf{R}(A(S))] \leq \min_{h\in H} \mathbf{R}(h) + \epsilon,$$

then the problem is PAC learnable by A .

Proof.

1. Let $\delta \in (0,1)$ and $m \geq m_H(\epsilon, \delta)$.
2. Define $X = \mathbf{R}(A(S)) - \min_{h\in H} \mathbf{R}(h)$.
3. Then $X \geq 0$ and by our assumption $\mathbb{E}[X] \geq \epsilon\delta$.
4. Using Markov's inequality, we obtain

$$\mathbb{P}\left[\mathbf{R}(A(S)) \geq \min_{h\in H} \mathbf{R}(h) + \epsilon\right] = \mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X \geq \epsilon]}{\epsilon} \leq \frac{\epsilon\delta}{\epsilon} = \delta.$$

□

Theorem

Assume that the convex, β-smooth, and nonnegative. Then, the RLM rule with the regularization function $\lambda \|\mathbf{w}\|^2$ for $\lambda \geq \frac{2\beta}{m}$, satisfies the following for all $\mathbf{w}^*$

$$\begin{aligned}\mathop{\mathbb{E}}_{S}[\mathbf{R}(A(S))] &\leq \left(1 + \frac{48\beta}{\lambda m}\right) \mathop{\mathbb{E}}_{S}\left[\hat{\mathbf{R}}(A(S))\right] \\ &\leq \left(1 + \frac{48\beta}{\lambda m}\right)\left(\mathbf{R}(\mathbf{w}^*) + \lambda \|\mathbf{w}^*\|^2\right)\end{aligned}$$

Theorem

Let $(H, \mathcal{Z}, \ell)$ be a convex-smooth-bounded learning problem with parameters β, B. Assume in addition that $\ell(0, z) \leq 1$ for all $z \in \mathcal{Z}$. For any $\epsilon \in (0, 1)$, let $m \geq \frac{150\beta B^2}{\epsilon^2}$ and set $\lambda = \frac{\epsilon}{3B^2}$. Then, for every distribution $\mathcal{D}$

$$\mathop{\mathbb{E}}_{S}[\mathbf{R}(A(S))] \leq \min_{\mathbf{w} \in H} \mathbf{R}(\mathbf{w}) + \epsilon.$$

Surrogate loss functions

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

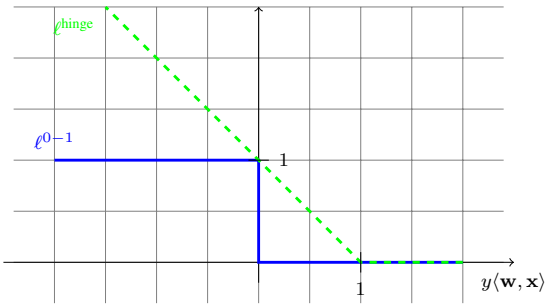
$$\ell^{0-1}(\mathbf{w}, (\mathbf{x}, y)) = \mathbb{I}\left[y \neq \operatorname{sgn}\left(\langle \mathbf{w}, \mathbf{x} \rangle\right)\right] = \mathbb{I}\left[y \langle \mathbf{w}, \mathbf{x} \rangle \leq 0\right].$$

3. This loss function is not convex with respect to $\mathbf{w}$.
4. When trying to minimize $\hat{\mathbf{R}}(\mathbf{w})$ with respect to this loss function we might encounter local minima.
5. 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.
6. One popular approach is to upper bound the nonconvex loss function by a convex surrogate loss function.
7. The requirements from a convex surrogate loss are as follows:
   - It should be convex.
   - It should upper bound the original loss.

1. Hinge-loss function is defined as

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

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



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

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

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

2. Using the surrogate property,

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

3. We can further rewrite the upper bound as

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

4. 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

1. We introduced two families of learning problems:
   - Convex-Lipschitz-bounded learning problems.
   - Convex-smooth-bounded learning problems.
2. There are some generic learning algorithms such as stochastic gradient descent algorithm for solving these problem. (Please read Chapter 14)
3. We have shown that **stable algorithms do not overfit**.
4. We have shown convex-Lipschitz-bounded learning problems and convex-Lipschitz-bounded learning problems are **PAC learnable**.
5. 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, 13, and 14 of Shai Shalev-Shwartz and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

References


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

Questions?