Machine learning theory

Convex learning problems

Hamid Beigy

Sharif university of technology

June 8, 2020
1. Introduction

2. Convexity

3. Lipschitzness

4. Smoothness

5. Convex learning problems

6. Surrogate loss functions

7. Assignments

8. Summary
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 \( u, v \in C \), the line segment between \( u \) and \( v \) is contained in set \( C \). That is, for any \( \alpha \in [0, 1] \), the convex combination \( \alpha u + (1 - \alpha)v \in C \).

Given \( \alpha \in [0, 1] \), the combination, \( \alpha u + (1 - \alpha)v \) of the points \( u, v \) is called a **convex combination**.

**Example (Convex and non-convex sets)**
Some examples of convex and non-convex sets in \( \mathbb{R}^2 \)

<table>
<thead>
<tr>
<th>non-convex sets</th>
<th>convex sets</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Non-Convex Set 1" /></td>
<td><img src="image2" alt="Convex Set 1" /></td>
</tr>
<tr>
<td><img src="image3" alt="Non-Convex Set 2" /></td>
<td><img src="image4" alt="Convex Set 2" /></td>
</tr>
<tr>
<td><img src="image5" alt="Non-Convex Set 3" /></td>
<td><img src="image6" alt="Convex Set 3" /></td>
</tr>
</tbody>
</table>
**Definition (Convex function)**

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

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

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

**Example (Convex function)**

![Diagram showing the convex function property](image)
A function $f$ is convex if and only if its epigraph is a convex set.

$$\text{epigraph}(f) = \{(x, \beta) \mid f(x) \leq \beta\}.$$
1. If $f$ is convex then every local minimum of $f$ is also a global minimum.
   - Let $B(u, r) = \{v \mid \|v - u\| \leq r\}$ be a ball of radius $r$ centered around $u$.
   - $f(u)$ is a local minimum of $f$ at $u$ if $\exists r > 0$ such that $\forall v \in B(u, r)$, we have $f(v) \geq f(u)$.
   - It follows that for any $v$ (not necessarily in $B$), there is a small enough $\alpha > 0$ such that $u + \alpha(v - u) \in B(u, r)$ and therefore
     \[ f(u) \leq f(u + \alpha(v - u)). \]
   - If $f$ is convex, we also have that
     \[ f(u + \alpha(v - u)) = f(\alpha u + (1 - \alpha)v) \leq (1 - \alpha)f(u) + \alpha f(v). \]
   - Combining these two equations and rearranging terms, we conclude that
     \[ f(u) \leq f(v). \]
   - This holds for every $v$, hence $f(u)$ is also a global minimum of $f$. 

![Graphs of convex and concave functions](https://example.com/graphs.png)
2. If \( f \) is convex and differentiable, then

\[
\forall u, \quad f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle
\]

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

- If \( f \) is convex, for every \( w \), we can construct a tangent to \( f \) at \( w \) that lies below \( f \) everywhere.
- If \( f \) is differentiable, this tangent is the linear function \( l(u) = f(w) + \langle \nabla f(w), u - w \rangle \).
v is sub-gradient of f at w if \( \forall u, \quad f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle \)

- The differential set, \( \partial f(w) \), is the set of sub-gradients of f at w.
  - where \( \nabla f(w) = \left( \frac{\partial f(w)}{\partial w_1}, \ldots, \frac{\partial f(w)}{\partial w_n} \right) \) is the gradient of f at w.

**Lemma**

*Function f is convex iff for every w, \( \partial f(w) \neq 0 \).*

- f is locally flat around w (0 is a sub-gradient) iff 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
   - $f'(x) = \frac{e^x}{1 + e^x} = \frac{1}{e^{-x} + 1}$ is a monotonically increasing function since the exponent function is a monotonically increasing function.
   - $f''(x) = \frac{e^{-x}}{(e^{-x} + 1)^2} = f(x)(1 - f(x))$ 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 \( f(w) = g(\langle w, x \rangle + y) \), for some \( x \in \mathbb{R}^n \), \( y \in \mathbb{R} \) 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 \( w_1, w_2 \in \mathbb{R}^n \) and \( \alpha \in [0, 1] \). We have

\[
\begin{align*}
f(\alpha w_1 + (1 - \alpha) w_2) &= g(\langle \alpha w_1 + (1 - \alpha) w_2, x \rangle + y) \\
&= g(\alpha \langle w_1, x \rangle + (1 - \alpha) \langle w_2, x \rangle + y) \\
&= g(\alpha (\langle w_1, x \rangle + y) + (1 - \alpha)(\langle w_2, x \rangle + y)) \\
&\leq \alpha g(\langle w_1, x \rangle + y) + (1 - \alpha)g(\langle w_2, x \rangle + y).
\end{align*}
\]

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 \( x \in \mathbb{R}^n \) and \( y \in \mathbb{R} \), let \( f(w) = (\langle w, x \rangle - y)^2 \). 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 \( x \in \mathbb{R}^n \) and \( y \in \{-1, +1\} \), let \( f(w) = \log(1 + \exp(-y \langle w, x \rangle)) \). 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. $g(x) = \max_{i \in \{1, \ldots, r\}} f_i(x)$.
2. $g(x) = \sum_{i=1}^r w_i f_i(x)$, where $\forall i, w_i \geq 0$.

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

1. The first claim follows by

$$g(\alpha u + (1 - \alpha)v) = \max_i f_i(\alpha u + (1 - \alpha)v) \leq \max [\alpha f_i(u) + (1 - \alpha)f_i(v)]$$

$$= \alpha \max_i f_i(u) + (1 - \alpha) \max_i f_i(v) = \alpha g(u) + (1 - \alpha)g(v).$$

2. The second claim follows by

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

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

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.
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 $\rho$-Lipschitz if for all $w_1, w_2 \in C$ we have $\|f(w_1) - f(w_2)\| \leq \rho \|w_1 - w_2\|$.

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

If $f'$ is bounded everywhere (in absolute value) by $\rho$, then $f$ is $\rho$-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 \( \rho \)-Lipschitz over \( \mathbb{R} \) for any \( \rho \). 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 \( \rho \)-Lipschitz over set \( C = \{ x \mid |x| \leq \frac{\rho}{2} \} \). For \( x_1, x_2 \), we have
   \[
   |x_1^2 - x_2^2| = |x_1 - x_2||x_1 + x_2| \leq 2 \frac{\rho}{2} |x_1 - x_2| = \rho|x_1 - x_2|.
   \]

5. Linear function \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) defined by \( f(w) = \langle v, w \rangle + b \), where \( v \in \mathbb{R}^n \) is \( ||v|| \)-Lipschitz. By using Cauchy-Schwartz inequality, we have
   \[
   |f(w_1) - f(w_2)| = |\langle v, w_1 - w_2 \rangle| \leq ||v|| ||w_1 - w_2||.
   \]
The following Lemma shows that composition of Lipschitz functions preserves Lipschitzness.

**Lemma (Composition of Lipschitz functions)**

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

**Proof (Composition of Lipschitz functions).**

\[
|f(w_1) - f(w_2)| = |g_1(g_2(w_1)) - g_1(g_2(w_2))| \\
\leq \rho_1 \|g_2(w_1) - g_2(w_2)\| \\
\leq \rho_1 \rho_2 \|w_1 - w_2\|.
\]
Smoothness
The definition of a smooth function relies on the notion of gradient. Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) be a differentiable function at \( w \) and its gradient as

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

Smoothness of \( f \) is defined as

**Definition (Smoothness)**

A differentiable function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is \( \beta \)-smooth if its gradient is \( \beta \)-Lipschitz; namely, for all \( v, w \) we have

\[
\|\nabla f(v) - \nabla f(w)\| \leq \beta \|v - w\|.
\]

Show that smoothness implies that for all \( v, w \) we have

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

while convexity of \( f \) implies that

\[
f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle.
\]

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

\[
\frac{1}{2\beta} \|\nabla f(w)\|^2 \leq f(w) - f(v).
\]
We had
\[
\frac{1}{2\beta} \| \nabla f(w) \|^2 \leq f(w) - f(v).
\]
Let \( f(v) \geq 0 \) for all \( v \), then smoothness implies that
\[
\| \nabla f(w) \|^2 \leq 2\beta f(w).
\]
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 \( f(w) = g(\langle w, x \rangle + b) \), where \( g : \mathbb{R} \mapsto \mathbb{R} \) is a \( \beta \)-smooth function and \( x \in \mathbb{R}^n \) and \( b \in \mathbb{R} \). Then, \( f \) is \( (\beta \|x\|^2) \)-smooth.

**Proof (Composition of smooth scaler function).**

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

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

**Example (Smooth functions)**

1. For any \( x \in \mathbb{R}^n \) and \( y \in \mathbb{R} \), let \( f(w) = (\langle w, x \rangle - y)^2 \). Then, \( f \) is \( (2 \|x\|^2) \)-smooth.

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

\[ \arg \min_{w \in C} f(w) \]

where \( C \) 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.

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

An special case is **unconstrained minimization** \( C = \mathbb{R}^n \).

Can reduce one to another

1. Adding the function \( l_C(w) \) to the objective eliminates the constraint.
2. Adding the constraint \( f(w) \leq f^* + \epsilon \) 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 \( D \) over \( \mathcal{Z} \), when running the learning algorithm on \( m \geq m_H(\epsilon, \delta) \) i.i.d. examples generated by \( D \), the algorithm returns \( h \in H \) such that, with probability of at least \( (1 - \delta) \) (over the choice of the \( m \) training examples),

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

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

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 \( w \).
Definition (Convex learning problems)

A learning problem \((H, Z, \ell)\) is called convex if

1. the hypothesis class \(H\) is a convex set, and
2. for all \(z \in Z\), 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 \(f(w) = \ell(w, z)\).

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 \(H = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^n\}\).
4. Let the squared loss function \(\ell(h, (x, y)) = (h(x) - y)^2\) used to measure error.
5. This is a convex learning problem because
   - Each linear function is parameterized by a vector \(w \in \mathbb{R}^n\). Hence, \(H = \mathbb{R}^n\).
   - The set of examples is \(Z = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{n+1}\).
   - The loss function is \(\ell(w, (x, y)) = \langle w, x \rangle - y^2\).
   - Clearly, \(H\) is a convex set and \(\ell(., .)\) is also convex with respect to its first argument.
Lemma (Convex learning problems)

If \( \ell \) is a convex loss function and the class \( H \) is convex, then the \( \text{erm}_H \) 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 \( \text{erm}_H \) problem is defined as

\[
\text{erm}_H(S) = \arg\min_{w \in H} \hat{R}(w)
\]

2. Since, for a sample \( S = \{z_1, \ldots, z_m\} \), for every \( w \), and \( \hat{R}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, z_i) \), Lemma (Convexity of a scalar function) implies that \( \hat{R}(w) \) is a convex function.

3. Therefore, the \( \text{erm}_H \) 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 $w \in H$ satisfies $\|w\| \leq B$).

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 \((H, Z, \ell)\) is called convex-Lipschitz-bounded, with parameters \(\rho, B\) if the following hold.

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

### Example (Linear regression with absolute-value loss)

1. Let \(X = \{x \in \mathbb{R}^n \mid \|x\| \leq \rho\}\) and \(Y \subset \mathbb{R}\).
2. Let \(H = \{w \in \mathbb{R}^n \mid \|w\| \leq B\}\).
3. Let loss function be \(\ell(w, (x, y)) = |\langle w, x \rangle - y|\).
4. Then, this problem is Convex-Lipschitz-bounded with parameters \(\rho, B\).
Definition (Convex-smooth-bounded learning problems)
A learning problem \((H, \mathcal{Z}, \ell)\) is called convex-smooth-bounded, with parameters \(\beta, B\) if the following hold.

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

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

Lemma (Learnability of Convex-Lipschitz/-smooth-bounded learning problems)
The following two families of learning problems are learnable.


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

- In many cases, the 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.

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

- This loss function is not convex with respect to \( w \).
- When trying to minimize \( \hat{R}(w) \) 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

\[ \ell_{\text{hinge}}(w, (x, y)) \triangleq \max\{0, 1 - y \langle w, x \rangle \}. \]

- Hinge-loss has the following two properties
  1. For all \( w \) and all \((x, y)\), we have \( \ell^{0-1}(w, (x, y)) \leq \ell_{\text{hinge}}(w, (x, y)) \).
  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
\[ R_{\text{hinge}}(A(S)) \leq \min_{w \in H} R_{\text{hinge}}(w) + \epsilon. \]

Using the surrogate property,
\[ R^{0-1}(A(S)) \leq \min_{w \in H} R_{\text{hinge}}(w) + \epsilon. \]

We can further rewrite the upper bound as
\[
R^{0-1}(A(S)) \leq \min_{w \in H} R^{0-1}(w) + \left( \min_{w \in H} R_{\text{hinge}}(w) - \min_{w \in H} R^{0-1}(w) \right) + \epsilon \\
= \epsilon_{\text{approximation}} + \epsilon_{\text{optimization}} + \epsilon_{\text{estimation}}
\]

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

Summary
We introduced two families of learning problems:

There are some generic learning algorithms such as stochastic gradient descent algorithm for solving these problems. (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.
Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning : From theory to algorithms*. Cambridge University Press, 2014.
Questions?