12

Convex Learning Problems

In this chapter we introduce *convex learning problems*. Convex learning comprises an important family of learning problems, mainly because most of what we can learn efficiently falls into it. We have already encountered linear regression with the squared loss and logistic regression, which are convex problems, and indeed they can be learned efficiently. We have also seen nonconvex problems, such as halfspaces with the 0-1 loss, which is known to be computationally hard to learn in the unrealizable case.

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. We begin the chapter with some required definitions of convexity. Besides convexity, we will define Lipschitzness and smoothness, which are additional properties of the loss function that facilitate successful learning. We next turn to defining convex learning problems and demonstrate the necessity for further constraints such as Boundedness and Lipschitzness or Smoothness. We define these more restricted families of learning problems and claim that Convex-Smooth/Lipschitz-Bounded problems are learnable. These claims will be proven in the next two chapters, in which we will present two learning paradigms that successfully learn all problems that are either convex-Lipschitz-bounded or convex-smooth-bounded.

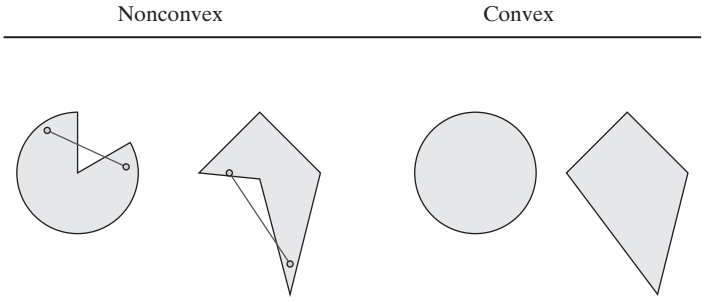
Finally, in Section 12.3, we show how one can handle some nonconvex problems by minimizing "surrogate" loss functions that are convex (instead of the original nonconvex loss function). Surrogate convex loss functions give rise to efficient solutions but might increase the risk of the learned predictor.

12.1 CONVEXITY, LIPSCHITZNESS, AND SMOOTHNESS

12.1.1 Convexity

Definition 12.1 (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 C. That is, for any $\alpha \in [0, 1]$ we have that $\alpha\mathbf{u} + (1 - \alpha)\mathbf{v} \in C$.

Examples of convex and nonconvex sets in $\mathbb{R}^2$ are given in the following. For the nonconvex sets, we depict two points in the set such that the line between the two points is not contained in the set.

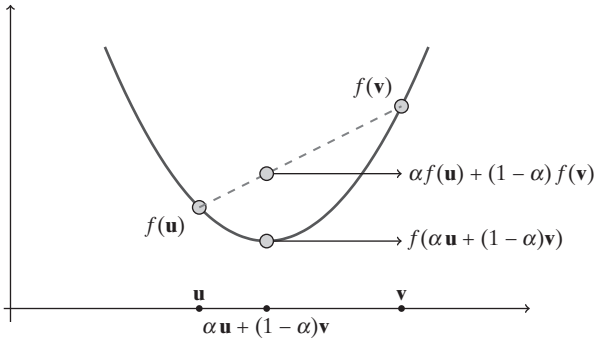


Given $\alpha \in [0,1]$, the combination, $\alpha\mathbf{u} + (1-\alpha)\mathbf{v}$ of the points $\mathbf{u}, \mathbf{v}$ is called a *convex combination*.

Definition 12.2 (Convex Function). Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if for every $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0,1]$,

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

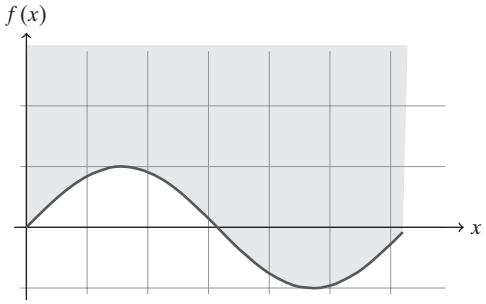
In words, f is convex if for any $\mathbf{u}, \mathbf{v}$, the graph of f between $\mathbf{u}$ and $\mathbf{v}$ lies below the line segment joining $f(\mathbf{u})$ and $f(\mathbf{v})$. An illustration of a convex function, $f : \mathbb{R} \to \mathbb{R}$, is depicted in the following.



The *epigraph* of a function f is the set

$$\text{epigraph(f)} = \{(\mathbf{x}, \beta) : f(\mathbf{x}) \leq \beta\}. \tag{12.1}$$

It is easy to verify that a function f is convex if and only if its epigraph is a convex set. An illustration of a nonconvex function $f : \mathbb{R} \to \mathbb{R}$, along with its epigraph, is given in the following.



An important property of convex functions is that every local minimum of the function is also a global minimum. Formally, let $B(\mathbf{u}, r) = \{\mathbf{v} : \|\mathbf{v} - \mathbf{u}\| \leq r\}$ be a ball of radius r centered around $\mathbf{u}$. We say that $f(\mathbf{u})$ is a local minimum of f at $\mathbf{u}$ if there exists some $r > 0$ such that for all $\mathbf{v} \in B(\mathbf{u}, r)$ 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

$$f(\mathbf{u}) \leq f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u})). \tag{12.2}$$

If f is convex, we also have that

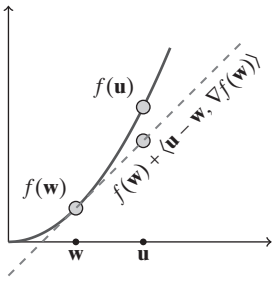
$$f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u})) = f(\alpha\mathbf{v} + (1 - \alpha)\mathbf{u}) \leq (1 - \alpha) f(\mathbf{u}) + \alpha f(\mathbf{v}). \tag{12.3}$$

Combining these two equations and rearranging terms, we conclude that $f(\mathbf{u}) \leq f(\mathbf{v})$. Since this holds for every $\mathbf{v}$, it follows that $f(\mathbf{u})$ is also a global minimum of f.

Another important property of convex functions is that 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 $l(\mathbf{u}) = f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle$, where $\nabla f(\mathbf{w})$ is the gradient of f at $\mathbf{w}$, namely, the vector of partial derivatives of f, $\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right)$. That is, for convex differentiable functions,

$$\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle. \tag{12.4}$$

In Chapter 14 we will generalize this inequality to nondifferentiable functions. An illustration of Equation (12.4) is given in the following.



If f is a scalar differentiable function, there is an easy way to check whether it is convex.

Lemma 12.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be a scalar twice differential function, and let 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 12.1.

- The scalar function $f(x) = x^2$ is convex. To see this, note that $f'(x) = 2x$ and $f''(x) = 2 > 0$.
- The scalar function $f(x) = \log(1 + \exp(x))$ is convex. To see this, observe that $f'(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{\exp(-x)+1}$. This is a monotonically increasing function since the exponent function is a monotonically increasing function.

The following claim shows that the composition of a convex scalar function with a linear function yields a convex vector-valued function.

Claim 12.4. *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ can be written as $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + y)$, for some $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Then, convexity of g implies the convexity of f.*

Proof. Let $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ 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 12.2.

- Given some $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as $f(\mathbf{w}) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$. Then, f is a composition of the function $g(a) = a^2$ onto a linear function, and hence f is a convex function.
- Given some $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{\pm 1\}$, let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as $f(\mathbf{w}) = \log(1 + \exp(-y\langle \mathbf{w}, \mathbf{x} \rangle))$. Then, f is a composition of the function $g(a) = \log(1 + \exp(a))$ onto a linear function, and hence f is a convex function.

Finally, the following lemma shows that the maximum of convex functions is convex and that a weighted sum of convex functions, with nonnegative weights, is also convex.

Claim 12.5. *For $i = 1, \ldots, r$, let $f_i : \mathbb{R}^d \to \mathbb{R}$ be a convex function. The following functions from $\mathbb{R}^d$ to $\mathbb{R}$ are also convex.*

- $g(x) = \max_{i \in [r]} f_i(x)$
- $g(x) = \sum_{i=1}^{r} w_i f_i(x)$, *where for all i, $w_i \geq 0$.*

Proof. The first claim follows by

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

For the second claim

$$\begin{aligned} g(\alpha u + (1-\alpha)v) &= \sum_i w_i f_i(\alpha u + (1-\alpha)v) \\ &\leq \sum_i w_i [\alpha f_i(u) + (1-\alpha) f_i(v)] \\ &= \alpha \sum_i w_i f_i(u) + (1-\alpha) \sum_i w_i f_i(v) \\ &= \alpha g(u) + (1-\alpha) g(v). \end{aligned}$$

□

Example 12.3. The function $g(x) = |x|$ is convex. To see this, note that $g(x) = \max\{x, -x\}$ and that both the function $f_1(x) = x$ and $f_2(x) = -x$ are convex.

12.1.2 Lipschitzness

The definition of Lipschitzness that follows is with respect to the Euclidean norm over $\mathbb{R}^d$. However, it is possible to define Lipschitzness with respect to any norm.

Definition 12.6 (Lipschitzness). Let $C \subset \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if for every $\mathbf{w}_1, \mathbf{w}_2 \in C$ we have that $\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$.

Intuitively, a Lipschitz function cannot change too fast. Note that if $f : \mathbb{R} \to \mathbb{R}$ 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 some point between w_1 and w_2. It follows that if the derivative of f is everywhere bounded (in absolute value) by ρ, then the function is ρ-Lipschitz.

Example 12.4.

- The function $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$. This follows from the triangle inequality: For every x_1, x_2,

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

  Since this holds for both x_1, x_2 and x_2, x_1, we obtain that $||x_1| - |x_2|| \leq |x_1 - x_2|$.
- The function $f(x) = \log(1 + \exp(x))$ is 1-Lipschitz over $\mathbb{R}$. To see this, observe that

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

- The function $f(x) = x^2$ is not ρ-Lipschitz over $\mathbb{R}$ for any ρ. To see this, take $x_1 = 0$ and $x_2 = 1 + \rho$, then

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

  However, this function is ρ-Lipschitz over the set $C = \{x : |x| \leq \rho/2\}$. Indeed, for any $x_1, x_2 \in C$ we have

$$|x_1^2 - x_2^2| = |x_1 + x_2|\,|x_1 - x_2| \leq 2(\rho/2)\,|x_1 - x_2| = \rho|x_1 - x_2|.$$

- The linear function $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle + b$ where $\mathbf{v} \in \mathbb{R}^d$ is $\|\mathbf{v}\|$-Lipschitz. Indeed, using Cauchy-Schwartz inequality,

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

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

Claim 12.7. *Let $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$, where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz. Then, 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 $\mathbf{v} \in \mathbb{R}^d, b \in \mathbb{R}$, then f is $(\rho_1 \|\mathbf{v}\|)$-Lipschitz.*

Proof.

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

□

12.1.3 Smoothness

The definition of a smooth function relies on the notion of *gradient*. Recall that the gradient of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{w}$, denoted $\nabla f(\mathbf{w})$, is the vector of partial derivatives of f, namely, $\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right)$.

Definition 12.8 (Smoothness). A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz; namely, for all $\mathbf{v}, \mathbf{w}$ we have $\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|$.

It is possible to show that smoothness implies that for all $\mathbf{v}, \mathbf{w}$ we have

$$f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|^2. \tag{12.5}$$

Recall that convexity of f implies that $f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle$. Therefore, 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 the right-hand side of Equation (12.5) and rearranging terms, we obtain

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

If we further assume that $f(\mathbf{v}) \geq 0$ for all $\mathbf{v}$ we conclude that smoothness implies the following:

$$\|\nabla f(\mathbf{w})\|^2 \leq 2\beta f(\mathbf{w}). \tag{12.6}$$

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

Example 12.5.

- The function $f(x) = x^2$ is 2-smooth. This follows directly from the fact that $f'(x) = 2x$. Note that for this particular function Equation (12.5) and Equation (12.6) hold with equality.
- The function $f(x) = \log(1 + \exp(x))$ is $(1/4)$-smooth. Indeed, since $f'(x) = \frac{1}{1+\exp(-x)}$ we have that

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

  Hence, f' is $(1/4)$-Lipschitz. Since this function is nonnegative, Equation (12.6) holds as well.

The following claim shows that a composition of a smooth scalar function over a linear function preserves smoothness.

Claim 12.9. *Let $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x}\rangle + b)$, where $g : \mathbb{R} \to \mathbb{R}$ is a β-smooth function, $\mathbf{x} \in \mathbb{R}^d$, and $b \in \mathbb{R}$. Then, f is $(\beta \|\mathbf{x}\|^2)$-smooth.*

Proof. By the chain rule we have that $\nabla f(\mathbf{w}) = g'(\langle \mathbf{w}, \mathbf{x}\rangle + b)\mathbf{x}$, where g' is the derivative of g. Using the smoothness of g and the Cauchy-Schwartz inequality we therefore obtain

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

□

Example 12.6.

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

12.2 CONVEX LEARNING PROBLEMS

Recall that in our general definition of learning (Definition 3.4 in Chapter 3), we have a hypothesis class $\mathcal{H}$, a set of examples Z, and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$. So far in the book we have mainly thought of Z as being the product of an instance

space and a target space, $Z = \mathcal{X} \times \mathcal{Y}$, and $\mathcal{H}$ being a set of functions from $\mathcal{X}$ to $\mathcal{Y}$. However, $\mathcal{H}$ can be an arbitrary set. Indeed, throughout this chapter, we consider hypothesis classes $\mathcal{H}$ that are subsets of the Euclidean space $\mathbb{R}^d$. That is, every hypothesis is some real-valued vector. We shall, therefore, denote a hypothesis in $\mathcal{H}$ by $\mathbf{w}$. Now we can finally define convex learning problems:

Definition 12.10 (Convex Learning Problem). A learning problem, $(\mathcal{H}, Z, \ell)$, is called convex if the hypothesis class $\mathcal{H}$ is a convex set and for all $z \in Z$, the loss function, $\ell(\cdot, z)$, is a convex function (where, for any z, $\ell(\cdot, z)$ denotes the function $f : \mathcal{H} \to \mathbb{R}$ defined by $f(\mathbf{w}) = \ell(\mathbf{w}, z)$).

Example 12.7 (Linear Regression with the Squared Loss). Recall that linear regression is a tool for modeling the relationship between some "explanatory" variables and some real valued outcome (see Chapter 9). The domain set $\mathcal{X}$ is a subset of $\mathbb{R}^d$, for some d, and the label set $\mathcal{Y}$ is the set of real numbers. We would like to learn a linear function $h : \mathbb{R}^d \to \mathbb{R}$ that best approximates the relationship between our variables. In Chapter 9 we defined the hypothesis class as the set of homogenous linear functions, $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathbb{R}^d\}$, and used the squared loss function, $\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^2$. However, we can equivalently model the learning problem as a convex learning problem as follows. Each linear function is parameterized by a vector $\mathbf{w} \in \mathbb{R}^d$. Hence, we can define $\mathcal{H}$ to be the set of all such parameters, namely, $\mathcal{H} = \mathbb{R}^d$. The set of examples is $Z = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$, and the loss function is $\ell(\mathbf{w}, (\mathbf{x}, y)) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$. Clearly, the set $\mathcal{H}$ is a convex set. The loss function is also convex with respect to its first argument (see Example 12.2).

Lemma 12.11. *If ℓ is a convex loss function and the class $\mathcal{H}$ is convex, then the* $\mathrm{ERM}_{\mathcal{H}}$ *problem, of minimizing the empirical loss over $\mathcal{H}$, is a convex optimization problem (that is, a problem of minimizing a convex function over a convex set).*

Proof. Recall that the $\mathrm{ERM}_{\mathcal{H}}$ problem is defined by

$$\mathrm{ERM}_{\mathcal{H}}(S) = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{H}} L_S(\mathbf{w}).$$

Since, for a sample $S = z_1, \ldots, z_m$, for every $\mathbf{w}$, $L_S(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{w}, z_i)$, Claim 12.5 implies that $L_S(\mathbf{w})$ is a convex function. Therefore, the ERM rule is a problem of minimizing a convex function subject to the constraint that the solution should be in a convex set. □

Under mild conditions, such problems can be solved efficiently using generic optimization algorithms. In particular, in Chapter 14 we will present a very simple algorithm for minimizing convex functions.

12.2.1 Learnability of Convex Learning Problems

We have argued that for many cases, implementing the ERM rule for convex learning problems can be done efficiently. But is convexity a sufficient condition for the learnability of a problem?

To make the quesion more specific: In VC theory, we saw that halfspaces in d-dimension are learnable (perhaps inefficiently). We also argued in Chapter 9 using

the "discretization trick" that if the problem is of d parameters, it is learnable with a sample complexity being a function of d. That is, for a constant d, the problem should be learnable. So, maybe all convex learning problems over $\mathbb{R}^d$, are learnable?

Example 12.8 later shows that the answer is negative, even when d is low. Not all convex learning problems over $\mathbb{R}^d$ are learnable. There is no contradiction to VC theory since VC theory only deals with binary classification while here we consider a wide family of problems. There is also no contradiction to the "discretization trick" as there we assumed that the loss function is bounded and also assumed that a representation of each parameter using a finite number of bits suffices. As we will show later, under some additional restricting conditions that hold in many practical scenarios, convex problems are learnable.

Example 12.8 (Nonlearnability of Linear Regression Even If $d = 1$). Let $\mathcal{H} = \mathbb{R}$, and the loss be the squared loss: $\ell(w, (x, y)) = (wx - y)^2$ (we're referring to the homogenous case). Let A be any deterministic algorithm.[1] Assume, by way of contradiction, that A is a successful PAC learner for this problem. That is, there exists a function $m(\cdot, \cdot)$, such that for every distribution $\mathcal{D}$ and for every ϵ, δ if A receives a training set of size $m \geq m(\epsilon, \delta)$, it should output, with probability of at least $1 - \delta$, a hypothesis $\hat{w} = A(S)$, such that $L_{\mathcal{D}}(\hat{w}) - \min_w L_{\mathcal{D}}(w) \leq \epsilon$.

Choose $\epsilon = 1/100, \delta = 1/2$, let $m \geq m(\epsilon, \delta)$, and set $\mu = \frac{\log(100/99)}{2m}$. We will define two distributions, and will show that A is likely to fail on at least one of them. The first distribution, $\mathcal{D}_1$, is supported on two examples, $z_1 = (1, 0)$ and $z_2 = (\mu, -1)$, where the probability mass of the first example is μ while the probability mass of the second example is $1 - \mu$. The second distribution, $\mathcal{D}_2$, is supported entirely on z_2.

Observe that for both distributions, the probability that all examples of the training set will be of the second type is at least 99%. This is trivially true for $\mathcal{D}_2$, whereas for $\mathcal{D}_1$, the probability of this event is

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

Since we assume that A is a deterministic algorithm, upon receiving a training set of m examples, each of which is $(\mu, -1)$, the algorithm will output some $\hat{w}$. Now, if $\hat{w} < -1/(2\mu)$, we will set the distribution to be $\mathcal{D}_1$. Hence,

$$L_{\mathcal{D}_1}(\hat{w}) \geq \mu(\hat{w})^2 \geq 1/(4\mu).$$

However,

$$\min_w L_{\mathcal{D}_1}(w) \leq L_{\mathcal{D}_1}(0) = (1 - \mu).$$

It follows that

$$L_{\mathcal{D}_1}(\hat{w}) - \min_w L_{\mathcal{D}_1}(w) \geq \frac{1}{4\mu} - (1 - \mu) > \epsilon.$$

Therefore, such algorithm A fails on $\mathcal{D}_1$. On the other hand, if $\hat{w} \geq -1/(2\mu)$ then we'll set the distribution to be $\mathcal{D}_2$. Then we have that $L_{\mathcal{D}_2}(\hat{w}) \geq 1/4$ while $\min_w L_{\mathcal{D}_2}(w) = 0$, so A fails on $\mathcal{D}_2$. 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] Namely, given S the output of A is determined. This requirement is for the sake of simplicity. A slightly more involved argument will show that nondeterministic algorithms will also fail to learn the problem.

A possible solution to this problem is to add another constraint on the hypothesis class. In addition to the convexity requirement, we require that $\mathcal{H}$ will be *bounded*; namely, we assume that for some predefined scalar B, every hypothesis $\mathbf{w} \in \mathcal{H}$ satisfies $\|\mathbf{w}\| \leq B$.

Boundedness and convexity alone are still not sufficient for ensuring that the problem is learnable, as the following example demonstrates.

Example 12.9. As in Example 12.8, consider a regression problem with the squared loss. However, this time let $\mathcal{H} = \{w : |w| \leq 1\} \subset \mathbb{R}$ be a bounded hypothesis class. It is easy to verify that $\mathcal{H}$ is convex. The argument will be the same as in Example 12.8, except that now the two distributions, $\mathcal{D}_1, \mathcal{D}_2$ will be supported on $z_1 = (1/\mu, 0)$ and $z_2 = (1, -1)$. If the algorithm A returns $\hat{w} < -1/2$ upon receiving m examples of the second type, then we will set the distribution to be $\mathcal{D}_1$ and have that

$$L_{\mathcal{D}_1}(\hat{w}) - \min_w L_{\mathcal{D}_1}(w) \geq \mu(\hat{w}/\mu)^2 - L_{\mathcal{D}_1}(0) \geq 1/(4\mu) - (1-\mu) > \epsilon.$$

Similarly, if $\hat{w} \geq -1/2$ we will set the distribution to be $\mathcal{D}_2$ and have that

$$L_{\mathcal{D}_2}(\hat{w}) - \min_w L_{\mathcal{D}_2}(w) \geq (-1/2+1)^2 - 0 > \epsilon.$$

This example shows that we need additional assumptions on the learning problem, and this time the solution is in Lipschitzness or smoothness of the loss function. This motivates a definition of two families of learning problems, convex-Lipschitz-bounded and convex-smooth-bounded, which are defined later.

12.2.2 Convex-Lipschitz/Smooth-Bounded Learning Problems

Definition 12.12 (Convex-Lipschitz-Bounded Learning Problem). A learning problem, $(\mathcal{H}, Z, \ell)$, is called Convex-Lipschitz-Bounded, with parameters ρ, B if the following holds:

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

Example 12.10. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \rho\}$ and $\mathcal{Y} = \mathbb{R}$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$ and let the loss function be $\ell(\mathbf{w}, (\mathbf{x}, y)) = |\langle \mathbf{w}, \mathbf{x} \rangle - y|$. This corresponds to a regression problem with the absolute-value loss, where we assume that the instances are in a ball of radius ρ and we restrict the hypotheses to be homogenous linear functions defined by a vector $\mathbf{w}$ whose norm is bounded by B. Then, the resulting problem is Convex-Lipschitz-Bounded with parameters ρ, B.

Definition 12.13 (Convex-Smooth-Bounded Learning Problem). A learning problem, $(\mathcal{H}, Z, \ell)$, is called Convex-Smooth-Bounded, with parameters β, B if the following holds:

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

Note that we also required that the loss function is nonnegative. This is needed to ensure that the loss function is self-bounded, as described in the previous section.

Example 12.11. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \beta/2\}$ and $\mathcal{Y} = \mathbb{R}$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$ and let the loss function be $\ell(\mathbf{w}, (\mathbf{x}, y)) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$. This corresponds to a regression problem with the squared loss, where we assume that the instances are in a ball of radius $\beta/2$ and we restrict the hypotheses to be homogenous linear functions defined by a vector $\mathbf{w}$ whose norm is bounded by B. Then, the resulting problem is Convex-Smooth-Bounded with parameters β, B.

We claim that these 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. We will prove this claim in the next chapters by introducing algorithms that learn these problems successfully.

12.3 SURROGATE LOSS FUNCTIONS

As mentioned, and as we will see in the next chapters, convex problems can be learned effficiently. However, in many cases, the natural loss function is not convex and, in particular, implementing the ERM rule is hard.

As an example, consider the problem of learning the hypothesis class of halfspaces with respect to the 0−1 loss. That is,

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

This loss function is not convex with respect to $\mathbf{w}$ and indeed, when trying to minimize the empirical risk with respect to this loss function we might encounter local minima (see Exercise 12.1). Furthermore, as discussed in Chapter 8, solving the ERM problem with respect to the 0−1 loss in the unrealizable case is known to be NP-hard.

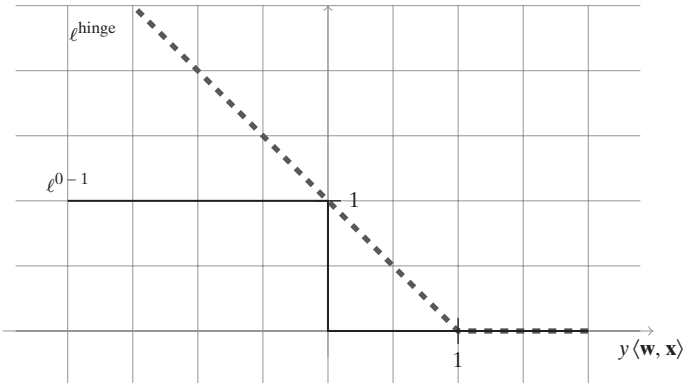
To circumvent the hardness result, one popular approach is to upper bound the nonconvex loss function by a convex surrogate loss function. As its name indicates, the requirements from a convex surrogate loss are as follows:

1. It should be convex.
2. It should upper bound the original loss.

For example, in the context of learning halfspaces, we can define the so-called hinge loss as a convex surrogate for the 0−1 loss, as follows:

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

Clearly, for all $\mathbf{w}$ and all $(\mathbf{x}, y)$, $\ell^{0-1}(\mathbf{w}, (\mathbf{x}, y)) \leq \ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y))$. In addition, the convexity of the hinge loss follows directly from Claim 12.5. Hence, the hinge loss satisfies the requirements of a convex surrogate loss function for the zero-one loss. An illustration of the functions ℓ^{0-1} and ℓ^{hinge} is given in the following.



Once we have defined the surrogate convex loss, we can learn the problem with respect to it. The generalization requirement from a hinge loss learner will have the form

$$L_{\mathcal{D}}^{\text{hinge}}(A(S)) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon,$$

where $L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell^{\text{hinge}}(\mathbf{w},(\mathbf{x},y))]$. Using the surrogate property, we can lower bound the left-hand side by $L_{\mathcal{D}}^{0-1}(A(S))$, which yields

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

We can further rewrite the upper bound as follows:

$$L_{\mathcal{D}}^{0-1}(A(S)) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w}) + \left(\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})\right) + \epsilon.$$

That is, the $0-1$ error of the learned predictor is upper bounded by three terms:

- *Approximation error*: This is the term $\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})$, which measures how well the hypothesis class performs on the distribution. We already elaborated on this error term in Chapter 5.
- *Estimation error*: This is the error that results from the fact that we only receive a training set and do not observe the distribution $\mathcal{D}$. We already elaborated on this error term in Chapter 5.
- *Optimization error*: This is the term $\left(\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})\right)$ that measures the difference between the approximation error with respect to the surrogate loss and the approximation error with respect to the original loss. The optimization error is a result of our inability to minimize the training loss with respect to the original loss. The size of this error depends on the specific distribution of the data and on the specific surrogate loss we are using.

12.4 SUMMARY

We introduced two families of learning problems: convex-Lipschitz-bounded and convex-smooth-bounded. In the next two chapters we will describe two generic

learning algorithms for these families. We also introduced the notion of convex surrogate loss function, which enables us also to utilize the convex machinery for nonconvex problems.

12.5 BIBLIOGRAPHIC REMARKS

There are several excellent books on convex analysis and optimization (Boyd & Vandenberghe 2004, Borwein & Lewis 2006, Bertsekas 1999, Hiriart-Urruty & Lemaréchal 1993). Regarding learning problems, the family of convex-Lipschitz-bounded problems was first studied by Zinkevich (2003) in the context of online learning and by Shalev-Shwartz, Shamir, Sridharan, and Srebro ((2009)) in the context of PAC learning.

12.6 EXERCISES

12.1 Construct an example showing that the $0-1$ loss function may suffer from local minima; namely, construct a training sample $S \in (X \times \{\pm 1\})^m$ (say, for $X = \mathbb{R}^2$), for which there exist a vector $\mathbf{w}$ and some $\epsilon > 0$ such that
   1. For any $\mathbf{w}'$ such that $\|w - w'\| \leq \epsilon$ we have $L_S(\mathbf{w}) \leq L_S(\mathbf{w}')$ (where the loss here is the $0-1$ loss). This means that $\mathbf{w}$ is a local minimum of L_S.
   2. There exists some $\mathbf{w}^*$ such that $L_S(\mathbf{w}^*) < L_S(\mathbf{w})$. This means that $\mathbf{w}$ is not a global minimum of L_S.

12.2 Consider the learning problem of logistic regression: Let $\mathcal{H} = \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq B\}$, for some scalar $B > 0$, let $\mathcal{Y} = \{\pm 1\}$, and let the loss function ℓ be defined as $\ell(\mathbf{w}, (\mathbf{x}, y)) = \log(1 + \exp(-y\langle \mathbf{w}, \mathbf{x}\rangle))$. Show that the resulting learning problem is both convex-Lipschitz-bounded and convex-smooth-bounded. Specify the parameters of Lipschitzness and smoothness.

12.3 Consider the problem of learning halfspaces with the hinge loss. We limit our domain to the Euclidean ball with radius R. That is, $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq R\}$. The label set is $\mathcal{Y} = \{\pm 1\}$ and the loss function ℓ is defined by $\ell(\mathbf{w}, (\mathbf{x}, y)) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}$. We already know that the loss function is convex. Show that it is R-Lipschitz.

12.4 **(*) Convex-Lipschitz-Boundedness Is Not Sufficient for Computational Efficiency:** In the next chapter we show that from the statistical perspective, all convex-Lipschitz-bounded problems are learnable (in the agnostic PAC model). However, our main motivation to learn such problems resulted from the computational perspective – convex optimization is often efficiently solvable. Yet the goal of this exercise is to show that convexity alone is not sufficient for efficiency. We show that even for the case $d = 1$, there is a convex-Lipschitz-bounded problem which cannot be learned by any computable learner.

   Let the hypothesis class be $\mathcal{H} = [0, 1]$ and let the example domain, Z, be the set of all Turing machines. Define the loss function as follows. For every Turing machine $T \in Z$, let $\ell(0, T) = 1$ if T halts on the input 0 and $\ell(0, T) = 0$ if T doesn't halt on the input 0. Similarly, let $\ell(1, T) = 0$ if T halts on the input 0 and $\ell(1, T) = 1$ if T doesn't halt on the input 0. Finally, for $h \in (0, 1)$, let $\ell(h, T) = h\ell(0, T) + (1 - h)\ell(1, T)$.
   1. Show that the resulting learning problem is convex-Lipschitz-bounded.
   2. Show that no computable algorithm can learn the problem.