Convex Learning Problems

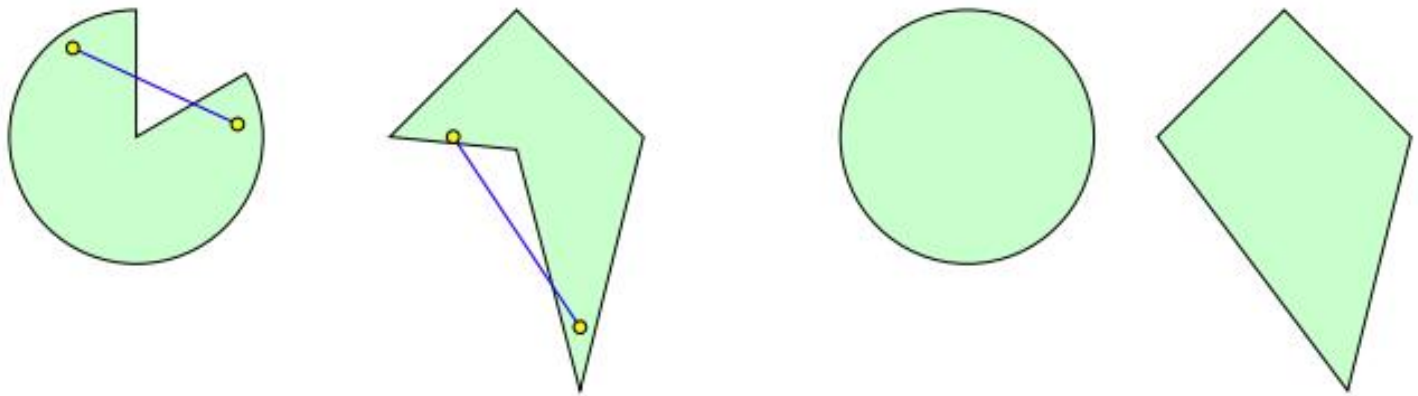
Convex Learning Problems

- A broad class of learning problems which have efficient learning algorithms

- Sample complexity
- Computational complexity

Convex Sets

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 $\alpha\mathbf{u} + (1 - \alpha)\mathbf{v} \in C$.

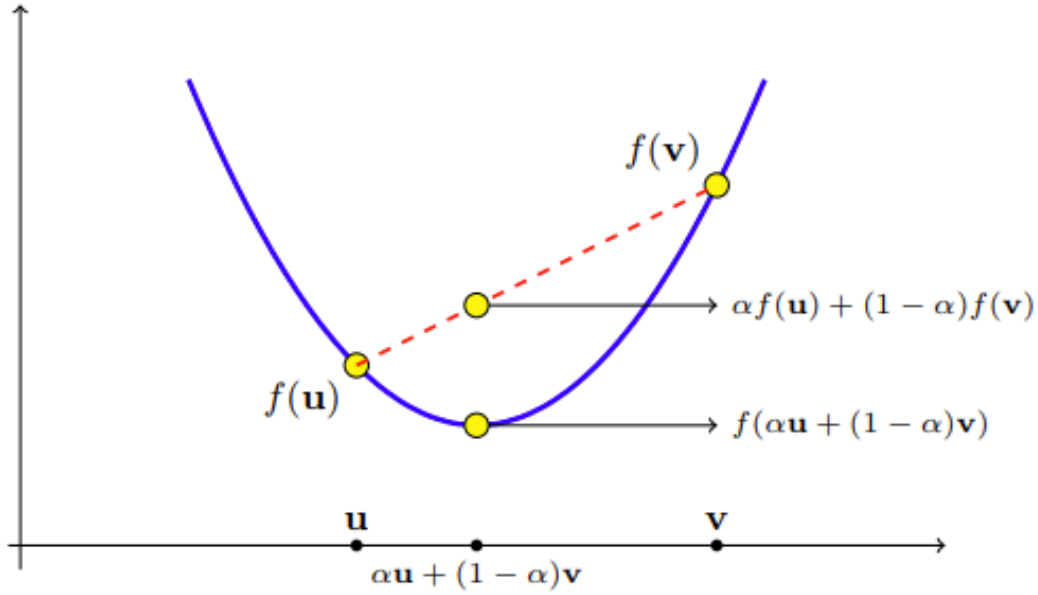


Given $\alpha \in [0, 1]$, the combination, $\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}$ of the points $\mathbf{u}, \mathbf{v}$ is called a *convex combination.*

Convex Functions

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

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

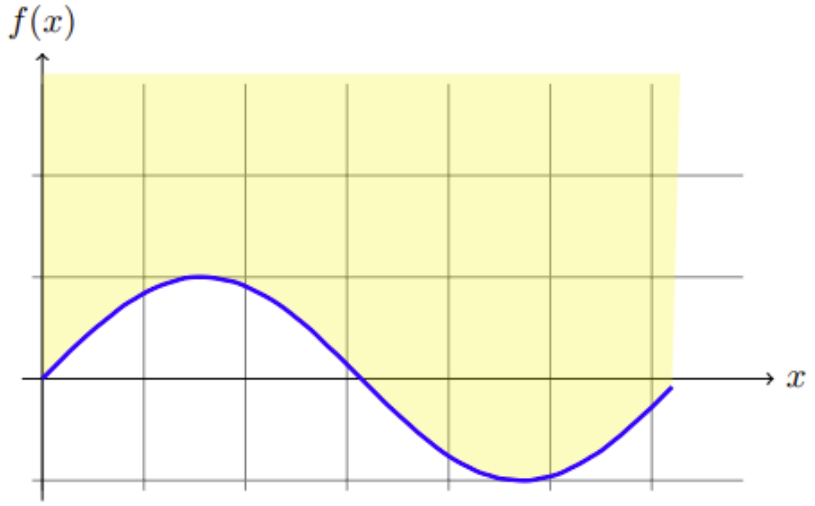


Relation between Convex Function and Set

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.



Local and Global Minima

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

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

Local and Global Minima

If f is convex, we also have that

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

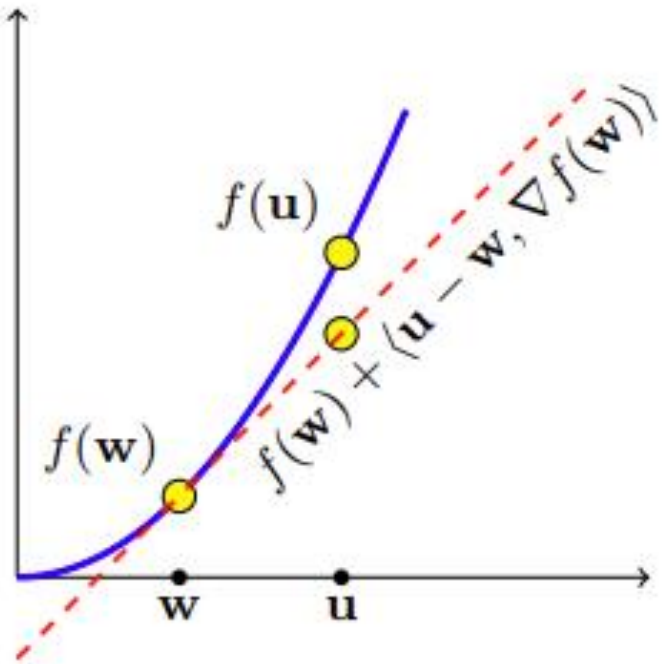
Combining these two equations and rearranging terms, we conclude that $f(\mathbf{u}) \leq f(\mathbf{v})$. Since this holds for every $\mathbf{v}$, it follows that $f(\mathbf{u})$ is also a global minimum of f.

Derivatives of Convex Functions

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,

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

Derivatives of Convex Functions



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

Derivative of Convex Functions

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*

Composition of Convex Functions

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.

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, $w_i \geq 0$.

Lipschitzness

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.

Composition of Lipschitz Functions

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

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

□

Smoothness

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

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

Self Bounded Functions

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.

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

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.

ERM on Convex Problems

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).*

Convex 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

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.

Convex-Smooth-Bounded Learning Problem

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.

Surrogate Loss Function

- In many cases, the natural loss function is not convex and, in particular, implementing the ERM rule is 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:
  - It should be convex
  - It should upper bound the original loss.