Foundations for Machine Learning

L. Y. Stefanus

TU Dresden, June-July 2018

Reference

- Shai Shalev-Shwartz and Shai Ben-David. UNDERSTANDING MACHINE LEARNING: From Theory to Algorithms. Cambridge University Press, 2014.

Convex Learning Problems

Convex Learning Problems

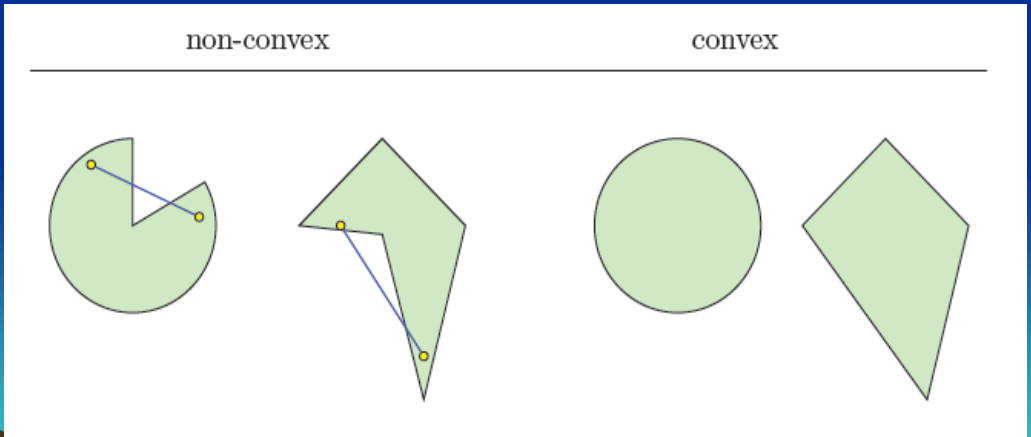
- Convex learning problems consist of an important family of learning problems, which can be implemented efficiently.

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

The combination $(1 - \alpha)\mathbf{u} + \alpha\mathbf{v}$ is called a convex combination.



- Given two points **u** and **v**, the expression

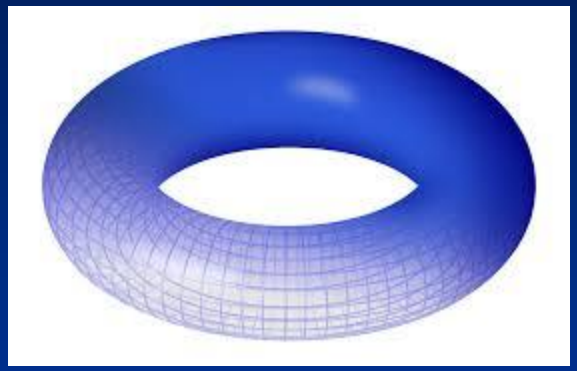
  $L(\alpha) = (1 - \alpha)\mathbf{u} + \alpha\mathbf{v}$ with $0 \leq \alpha \leq 1$

  describes the line segment with end points **u** and **v**.

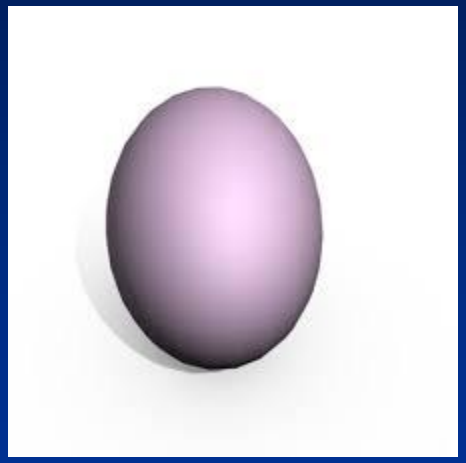
L(0) = **u**

L(1) = **v**

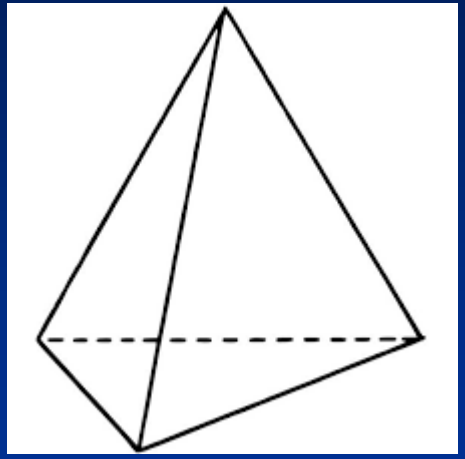
Which solid set is convex?



torus



ellipsoid

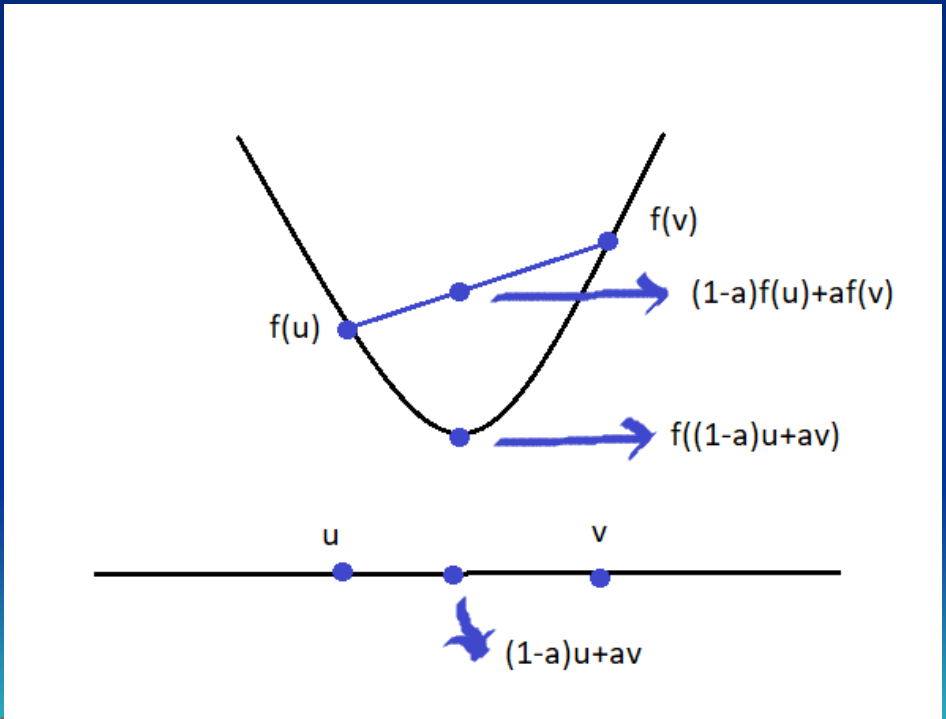


tetrahedron

Definition (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 $a \in [0,1]$,

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

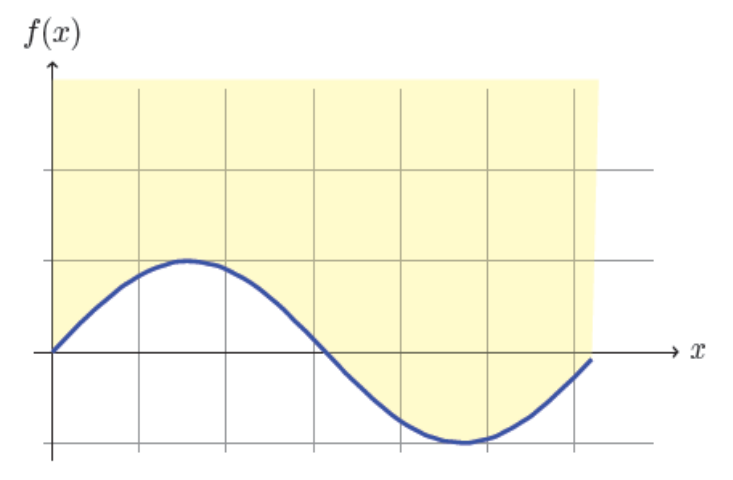


In words, f is convex if for any **u**, **v**, the graph of f between **u** and **v** lies below the line segment joining f(**u**) and f(**v**).

- The epigraph of a function f is the set

$$\text{epigraph}(f) = \{(x,b): f(x) \leq b\}$$

- A function f is convex if and only if its epigraph is a convex set.
- An illustration of a non-convex function $f: R \rightarrow R$, along with its epigraph, is as follows.



Global Minimum of a Convex Function

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.

- If f is a scalar differentiable function, there is an easy way to check if 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: The scalar function $f(x) = x^2$ is convex. We have that $f'(x) = 2x$ and $f''(x) = 2 > 0$.

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 : R $\rightarrow$ 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.

Notasi: (inner product)

$$< \mathbf{v}, \mathbf{w} > = \sum_{i=1}^{d} v_i w_i$$

Examples

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

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

- 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,

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

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

The function $f(x) = x^2$ is 2-smooth. This follows directly from the fact that $f'(x) = 2x$.

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 $f(\mathbf{w}) = \ell(\mathbf{w}, z)$).

- In the definition above, H can be an arbitrary set.

Indeed, we consider hypothesis classes H that are subsets of the Euclidean space R^d. That is, every hypothesis is some real-valued vector.
We, therefore, denote a hypothesis in H by w.

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).*

This means that such problems can be solved efficiently using generic optimization algorithms.

Learnability of Convex 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.

Learnability of Convex Learning Problems

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.

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 study further this claim later, by introducing algorithms that learn these problems successfully.