# Handout 1: Elements of convex learning problems

Lecturer & author: Georgios P. Karagiannis      georgios.karagiannis@durham.ac.uk

**Aim.** To introduce elements of convexity, Lipschitzbness, and smoothmess that can be used for the analysis of stochastic gradient related learning algorithms.

**Reading list & references:**

- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.
- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

## 1. MOTIVATIONS

*Note* 1. Introducing convexity and smoothness in the learning problems makes easier the (theoretical) analysis of the problem and its solution.

*Note* 2. Most of the ML problems discussed in the course (eg, Artificial neural networks, Gaussian process regression) are usually non-convex.

*Note* 3. Non convex problems can Treatments of convex learning methods to handle non-convex problems can be done via surrogates –to be discussed.

## 2. CONVEX LEARNING PROBLEMS

**Definition 4.** Convex learning problem is a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ that the hypothesis class $\mathcal{H}$ is a convex set, and the loss function $\ell$ is a convex function for each example $z \in \mathcal{Z}$.

**Example 5.** Multiple linear regression $< w, x > \to y$ with $y \in \mathbb{R}$ , hypothesis class $\mathcal{H} = \left\{ w \in \mathbb{R}^d \right\}$ and loss $\ell\left(w, (x, y)\right) = (< w, x > -y)^2$ with

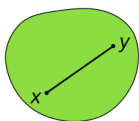$$w^* = \arg \min_{\forall w} \mathrm{E}\left(< w, x > -y\right)^2$$

or

$$w^{**} = \arg \min_{\forall w} \frac{1}{m} \sum_{i=1}^{m} (< w, x_i > -y)^2$$

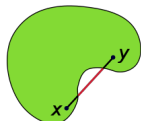is a convex learning problem for reasons that will be discussed below.

## 3. CONVEXITY

**Definition 6.** A set $C$ is convex if for any $u, v \in C$, the line segment between $u$ and $v$ is contained in $C$. Namely,

- for any $u, v \in C$ and for any $\alpha \in [0, 1]$ we have that $\alpha u + (1 - \alpha) v \in C$.
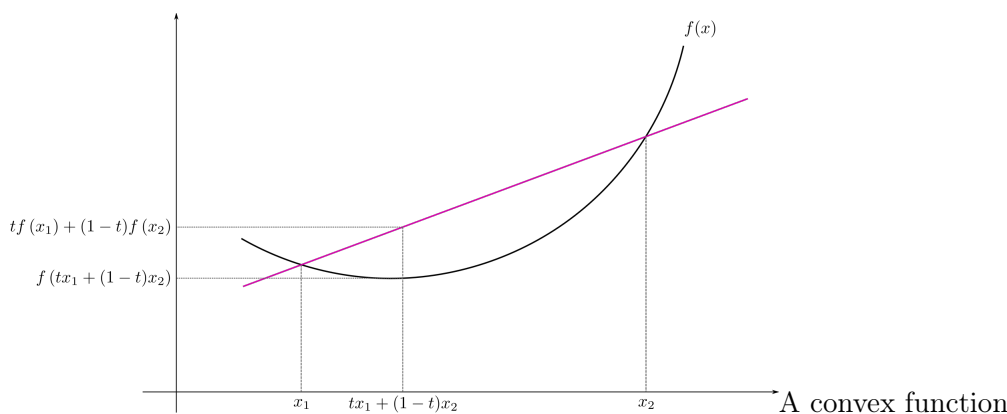
A convex set          A non-convex set

**Definition 7.** Let $C$ be a convex set. A function $f : C \to R$ is convex function if for any $u, v \in C$ and for any $\alpha \in [0, 1]$

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



A convex function

**Example 8.** The function $f : \mathbb{R} \to \mathbb{R}_+$ with $f(x) = x^2$ is convex function. For any $u, v \in C$ and for any $\alpha \in [0, 1]$ it is

$$\left(\alpha u + (1 - \alpha) v\right)^2 \leq \alpha^2 \left(u\right)^2 + (1 - \alpha)^2 \left(v\right)^2 + 2\alpha u (1 - \alpha) v \leq \alpha \left(u\right)^2 + (1 - \alpha) \left(v\right)^2$$

**Proposition 9.** *Every local minimum of a convex function is the global minimum.*

**Proposition 10.** *Let $f : C \to \mathbb{R}$ be convex function. The tangent of $f$ at $w \in C$ is below $f$, namely*

$$\forall u \in C \quad f\left(u\right) \geq f\left(w\right) + < \nabla f\left(w\right), u - w >$$

**Proposition 11.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ such that $f\left(w\right) = g\left(< w, x > + y\right)$ for some $x \in \mathbb{R}^d$, $y \in \mathbb{R}$. If $g$ is convex function then $f$ is convex function.*

*Proof.* See Exercise 1 in the Exercise sheet.  □

**Example 12.** Consider the regression problem $x \mapsto y$ with $x \in \mathbb{R}^d$, $y \in \mathbb{R}$ and predictor $h\left(x\right) = < w, x >$. The risk $R\left(w\right) = \left(< w, x > + y\right)^2$ is convex because $g\left(a\right) = \left(a\right)^2$ and Proposition 11.

**Example 13.** Let $f_j : \mathbb{R}^d \to \mathbb{R}$ convex functions for $j = 1, ..., r$. Then:

(1) $g\left(x\right) = \max_{\forall j} \left(f_j\left(x\right)\right)$ is a convex function
(2) $g\left(x\right) = \sum_{j=1}^{r} w_j f_j\left(x\right)$ is a convex function where $w_j > 0$

**Solution.**

(1) For any $u, v \in \mathbb{R}^d$ and for any $\alpha \in [0, 1]$

$$g\left(\alpha u + (1-\alpha) v\right) = \max_{\forall j} \left(f_j\left(\alpha u + (1-\alpha) v\right)\right)$$

$$\leq \max_{\forall j}\left(\alpha f_j\left(u\right) + (1-\alpha) f_j\left(v\right)\right) \qquad (f_j \text{ is convex})$$

$$\leq \alpha \max_{\forall j}\left(f_j\left(u\right)\right) + (1-\alpha) \max_{\forall j}\left(f_j\left(v\right)\right) \qquad (\max(\cdot) \text{ is convex})$$

$$\leq \alpha g\left(u\right) + (1-\alpha) g\left(v\right)$$

(2) For any $u, v \in \mathbb{R}^d$ and for any $\alpha \in [0, 1]$

$$g\left(\alpha u + (1-\alpha) v\right) = \sum_{j=1}^{r} w_j f_j\left(\alpha u + (1-\alpha) v\right)$$

$$\leq \alpha \sum_{j=1}^{r} w_j f_j\left(u\right) + (1-\alpha) \sum_{j=1}^{r} w_j f_j\left(v\right) \qquad (f_j \text{ is convex})$$

$$\leq \alpha g\left(u\right) + (1-\alpha) g\left(v\right)$$

**Example 14.** $g\left(x\right) = |x|$ is convex according to Example 13, as $g\left(x\right) = |x| = \max\left(-x, x\right)$.

## 4. LIPSCHITZBNESS

**Definition 15.** Let $C \in \mathbb{R}^d$. Function $f : \mathbb{R}^d \to \mathbb{R}^k$ is $\rho$-Lipschitz over $C$ if for every $w_1, w_2 \in C$ we have that

$$(4.1) \qquad \|f\left(w_1\right) - f\left(w_2\right)\| \leq \rho \|w_1 - w_2\|. \qquad \text{Lipschitz condition}$$

*Conclusion* 16. That means: a Lipschitz function $f\left(x\right)$ cannot change too drastically wrt $x$.

**Example 17.** Consider the function $f : \mathbb{R} \to \mathbb{R}_+$ with $f\left(x\right) = x^2$.

(1) $f$ is not a $\rho$-Lipschitz in $\mathbb{R}$.
(2) $f$ is a $\rho$-Lipschitz in $C = \{x \in \mathbb{R} : |x| < \rho/2\}$.

$$|f\left(x_2\right) - f\left(x_1\right)| = \left|x_2^2 - x_1^2\right| = |\left(x_2 + x_1\right)\left(x_2 - x_1\right)| \leq 2\rho/2\left(x_2 - x_1\right) = \rho|x_2 - x_1|$$

**Solution.**

(1) For $x_1 = 0$ and $x_2 = 1 + \rho$, it is

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

(2) It is

$$|f\left(x_2\right) - f\left(x_1\right)| = \left|x_2^2 - x_1^2\right| = |\left(x_2 + x_1\right)\left(x_2 - x_1\right)| \leq 2\rho/2\left(x_2 - x_1\right) = \rho|x_2 - x_1|$$

**Theorem 18.** *Let functions $g_1$ be $\rho_1$-Lipschitz and $g_2$ be $\rho_2$-Lipschitz. Then $f$ with $f\left(x\right) = g_1\left(g_2\left(x\right)\right)$ is $\rho_1\rho_2$-Lipschitz.*

**Solution.** See Exercise 2 from the exercise

**Example 19.** Let functions $g$ be $\rho$-Lipschitz and $g_2$ has $g_2(x) = <v,x> +b$. Then $f$ with $f(x) = g(<v,x> +b)$ is $\rho$-Lipschitz.

$$|f(w_1) - f(w_2)| = |g_1(<v,x> +b) - g_1(<v,x> +b)| \le \rho_1 |<v,w_1> +b- <v,w_2> -b|$$

$$\le \rho_1 \left| v^\top w_1 - v^\top w_2 \right| = \rho_1 |v| |w_1 - w_2|$$

*Note* 20. So, given Examples 17 and 19, in the linear regression setting using loss $\ell(w,z) = (w^\top x - z)^2$, the loss unction is -Lipschitz for a given $z$.

## 5. SMOOTHNESS

**Definition 21.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $\beta$-smooth if its gradient is $\beta$-Lipschitz; namely for all $v, w \in \mathbb{R}^d$

$$\text{(5.1)} \qquad \|\nabla f(w_1) - \nabla f(w_2)\| \le \beta \|w_1 - w_2\|.$$

**Theorem 22.** *Function $f : \mathbb{R}^d \to \mathbb{R}$ is $\beta$-smooth iff*

$$\text{(5.2)} \qquad f(v) \le f(w) + <\nabla f(w), v - w> + \frac{\beta}{2} \Uparrow v - w \Uparrow^2$$

*Remark* 23. If $f : \mathbb{R}^d \to \mathbb{R}$ is $\beta$-smooth then (5.2) holds, and if it is convex as well then

$$f(v) > f(w) + <\nabla f(w), v - w>$$

holds. Hence if both conditions imply upper and lower bounds

$$f(v) - f(w) \in \left( <\nabla f(w), v - w>, <\nabla f(w), v - w> + \frac{\beta}{2} \|v - w\|^2 \right)$$

*Remark* 24. If $f : \mathbb{R}^d \to \mathbb{R}$ is $\beta$-smooth then for $v, w \in \mathbb{R}^d$ such that $v = w - \frac{1}{\beta} \nabla f(w)$ then by (5.2), it is

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

If additionally $f(w) > 0$ for all $x \in \mathbb{R}^d$ then

$$\|\nabla f(w)\|^2 \le 2\beta f(w)$$

which provides assumptions to bound the gradient.

**Theorem 25.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ with $f(w) = g(<w,x> +y)$ $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be a $\beta$-smooth function. Then $f$ is a $\left( \beta \|x\|^2 \right)$-smooth.*

**Example 26.** Let $f(w) = (<w,x> +y)^2$ for $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Then $f$ is $\left( 2\|x\|^2 \right)$-smooth.

**Solution.** It is $f(w) = g(<w,x> +y)$ for $g(a) = a^2$. $g$ is 2-smooth since

$$\left\| g'(w_1) - g'(w_2) \right\| = \|2w_1 - 2w_2\| \le 2\|w_1 - w_2\|.$$

Hence from (25), $f$ is $\left( 2\|x\|^2 \right)$-smooth.

*Remark* 27. The loss function of a learning problem may be non-convex which implies that the risk function is non-convex. A proper treatment would be to upper bound the non-convex loss function by a convex surrogate loss function.

**Example 28.** consider the problem of learning $w \in \mathcal{H}$ from hypothesis set $\mathcal{H} \subset \mathbb{R}^d$ with respect to the $0 - 1$ loss

$$\ell(w, (x, y)) = 1_{(y\langle w, x\rangle \leq 0)}$$

with $y \in \mathbb{R}$, \nd $x \in \mathbb{R}^d$. Here $\ell(\cdot)$ is non-convex. A convex surrogate loss function can be

$$\tilde{\ell}(w, (x, y)) = \max(0, 1 - y\langle w, x\rangle)$$

which is convex (Example 14) wrt $w$. Note that $\max(\cdot)$ is convex as $\max(1, \alpha u + (1 - \alpha) v) \leq \alpha \max(1, u) + (1 - \alpha) \max(1, v)$. Then we compute

$$\tilde{w}_* = \arg\min_{\forall x} \left(\tilde{R}_g(w)\right) = \arg\min_{\forall x} \left(\mathrm{E}_{(x,y)\sim g}\left(\max(0, 1 - y\langle w, x\rangle)\right)\right)$$

instead of

$$w_* = \arg\min_{\forall x} \left(R_g(w)\right) = \arg\min_{\forall x} \left(\mathrm{E}_{(x,y)\sim g}\left(1_{(y\langle w, x\rangle \leq 0)}\right)\right)$$

Of course by using the surrogate loss instead of the actuall one, we introduce some approximation error in the produced output $\tilde{w}_* \neq w_*$.

*Remark* 29. (Intuitions...) Using a convex surrogate loss function instead the convex one, fascilitates computations but introduces extra error to the solution. If $R_g(\cdot)$ is the risk under the non-convex loss, $\tilde{R}_g(\cdot)$ is the risk under the convex surrogate loss, and $\tilde{w}_{\mathrm{alg}}$ is the output of the learning algorithm under $\tilde{R}_g(\cdot)$ then we have the upper bound

$$R_g(\tilde{w}_{\mathrm{alg}}) \leq \underbrace{\min_{w\in\mathcal{H}}\left(R_g(w)\right)}_{\mathrm{I}} + \underbrace{\left(\min_{w\in\mathcal{H}}\left(\tilde{R}_g(w)\right) - \min_{w\in\mathcal{H}}\left(R_g(w)\right)\right)}_{\mathrm{II}} + \underbrace{\epsilon}_{\mathrm{III}}$$

where term I is the approximation error measuring how well the hypothesis class performs on the generating model, term II is the optimization error due to the use of surrogate loss instead of the actual non-convex one, and term III is the estimation error due to the use of a training set and not the whole generation model.