Machine Learning and Neural Networks (MATH3431) Epiphany term, 2024

Lecture notes 2: Elements of convex learning problems

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

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

Reading list & references:

- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
  - Ch. 12 Convex Learning Problems

Further reading

- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.

1. Motivations

Note 1. We introduce concepts of convexity and smoothness that facilitate the analysis and understanding of the learning problems and their solutions that we will discuss (eg stochastic gradient descent, SVM) later on. Also learning problems with such characteristics can be learned more efficiently.

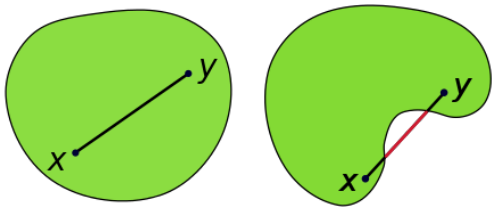
Note 2. Some of the ML problems discussed in the course (eg, Artificial neural networks, Gaussian process regression) are non-convex. To overcome this problem, we will introduce the concept of surrogate loss function that allows a non-convex problem to be handled with the tools introduced int he convex setting.

2. Convexity

Note 3. Convexity is a central concept in learning, e.g. least squares, as it often considers a Euclidean distance as a Risk function to be minimized.

Definition 4. A set C is convex if for any $u, v \in C$ and for any $\alpha \in [0,1]$ we have that $\alpha u + (1-\alpha)\, v \in C$.

Note 5. Namely, a set C is convex if for any $u, v \in C$, the line segment between u and v is contained in C.



(A) (B)

FIGURE 2.1. (2.1a) is a Convex set ; (2.1b) is a non-convex set

Example 6. For instance $\mathbb{R}^d$ for $d \geq 1$ is a 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 $\alpha \in [0, 1]$

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

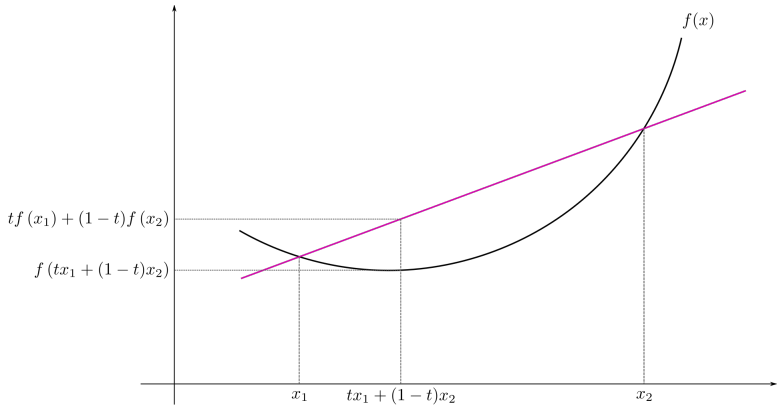


FIGURE 2.2. A convex function

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

$$\left(\alpha u + (1-\alpha)\, v\right)^2 - \alpha\left(u\right)^2 + (1-\alpha)\left(v\right)^2 = -\alpha\left(1-\alpha\right)\left(u-v\right)^2 \leq 0$$

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*

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

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

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

Example 12. Consider the regression problem with regressor $x \in \mathbb{R}^d$, and response $y \in \mathbb{R}$ and predictor rule $h(x) = \langle w, x \rangle$. The loss function $\ell(w,(x,y)) = (\langle w, x \rangle - y)^2$ is convex because $g(a) = (a)^2$ is convex and Proposition 11.

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

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

Solution.

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

$$\begin{aligned} g(\alpha u + (1-\alpha)v) &= \max_{\forall j}(f_j(\alpha u + (1-\alpha)v)) \\ &\leq \max_{\forall j}(\alpha f_j(u) + (1-\alpha) f_j(v)) && (f_j \text{ is convex}) \\ &\leq \alpha \max_{\forall j}(f_j(u)) + (1-\alpha)\max_{\forall j}(f_j(v)) && (\max(\cdot) \text{ is convex}) \\ &\leq \alpha g(u) + (1-\alpha) g(v) \end{aligned}$$

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

$$\begin{aligned} g(\alpha u + (1-\alpha)v) &= \sum_{j=1}^{r} w_j f_j(\alpha u + (1-\alpha)v) \\ &\leq \alpha \sum_{j=1}^{r} w_j f_j(u) + (1-\alpha)\sum_{j=1}^{r} w_j f_j(v) && (f_j \text{ is convex}) \\ &\leq \alpha g(u) + (1-\alpha) g(v) \end{aligned}$$

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

3. Strong convexity

Note 15. Strong convexity is a central concept in regularization, e.g. Ridge, as it makes a convex loss function strongly convex by adding a shrinkage term.

Definition 16. (Strongly convex functions) A function f is λ-strongly convex function is for all w, u, and $\alpha \in (0,1)$ we have

$$f(\alpha w + (1-\alpha)u) \leq a f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2 \tag{3.1}$$

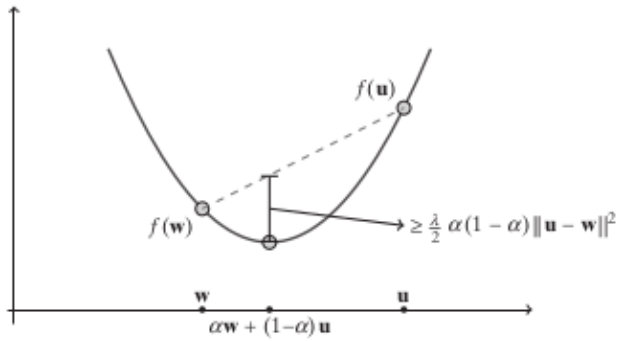


FIGURE 3.1. Strongly convex function

Proposition 17.

(1) *The function* $f(w) = \lambda \|w\|^2$ *is* 2λ-strongly convex
(2) *If* f *is* λ-strongly convex and g is convex then $f + g$ is λ-strongly convex

Proof. Both can be checked from the definition by substitution. □

Lemma 18. *If* f *is* λ-strongly convex and $w^* = \arg\min_w f(w)$ is a minimizer of f then for any w

$$f(w) - f(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

Proof. Exercise **??** in the Exercise sheet. □

4. Lipschitzness

Definition 19. Let $C \in \mathbb{R}^d$. Function $f : \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if for every $w_1, w_2 \in C$ we have that

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

Conclusion 20. That means: a Lipschitz function $f(x)$ cannot change too drastically wrt x.

Example 21. Consider the function $f : \mathbb{R} \to \mathbb{R}_+$ with $f(x) = x^2$.

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

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

Solution.

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

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

(2) It is

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

Theorem 22. *Let functions* g_1 *be* ρ_1*-Lipschitz and* g_2 *be* ρ_2*-Lipschitz. Then* f *with* $f(x) = g_1(g_2(x))$ *is* $\rho_1\rho_2$*-Lipschitz.*

Solution. See Exercise **??** from the exercise sheet

Example 23. Let functions g be ρ-Lipschitz. Then f with $f(x) = g(< v, x > +b)$ is $(\rho|v|)$-Lipschitz.

Solution. It is

$$\begin{aligned}|f(w_1) - f(w_2)| =& |g(< v, w_1 > +b) - g(< v, w_2 > +b)| \leq \rho |< v, w_1 > +b - < v, w_2 > -b| \\ \leq& \rho \left| v^\top w_1 - v^\top w_2 \right| \leq \rho |v| |w_1 - w_2|\end{aligned}$$

Note 24. So, given Examples 21 and 23, in the linear regression setting using loss $\ell(w, z = (x, y)) = \left(w^\top x - y\right)^2$, the loss function is -Lipschitz for a given $z = (x, y)$ and and bounded $\|w\| < \rho$.

5. Smoothness

Definition 25. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz; namely for all $v, w \in \mathbb{R}^d$

$$\|\nabla f(w_1) - \nabla f(w_2)\| \leq \beta \|w_1 - w_2\|. \tag{5.1}$$

Theorem 26. *Function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth iff*

$$f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2 \tag{5.2}$$

Theorem 27. *Let $f : \mathbb{R}^d \to \mathbb{R}$ with $f(w) = g(\langle w, x \rangle + y)$ $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be a β-smooth function. Then f is a $\left(\beta \|x\|^2\right)$-smooth.*

Proof. See Exercise **??** from the Exercise sheet □

Example 28. 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

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

Hence from Theorem 27, f is $\left(2\|x\|^2\right)$-smooth.

Example 29. Consider the regression problem with predictor rule $h(x) = \langle w, x \rangle$, loss function $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$, feature $x \in \mathbb{R}^d$, and target $y \in \mathbb{R}$. Then $\ell(w, \cdot)$ is $\left(2\|x\|^2\right)$-smooth.

Solution. Follows from Example 28.

6. Convex learning problems

Definition 30. 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 ℓ is a convex function for each example $z \in \mathcal{Z}$.

Example 31. Consider the regression problem with predictor rule $h(x) = \langle w, x \rangle$, loss function $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$, feature $x \in \mathbb{R}^d$, and target $y \in \mathbb{R}$. This imposes a convex learning problem due to Examples 6 and 13.

Definition 32. Convex-Lipschitz-Bounded Learning Problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with parameters ρ, and B, is called the learning problem whose the hypothesis class $\mathcal{H}$ is a convex set, for all $w \in \mathcal{H}$ it is $\|w\| \leq B$, and the loss function $\ell(\cdot, z)$ is convex and ρ-Lipschitz function for all $z \in \mathcal{Z}$.

Example 33. Consider the regression problem with predictor rule $h(x) = \langle w, x \rangle$, loss function $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$, feature $x \in \mathbb{R}^d$, and target $y \in \mathbb{R}$. This imposes a Convex-Lipschitz-Bounded Learning Problem if $\mathcal{H} = \left\{ w \in \mathbb{R}^d : \|w\| \leq B \right\}$ due to Examples 13, and 21(2).

Definition 34. Convex-Smooth-Bounded Learning Problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with parameters β, and B, is called the learning problem whose the hypothesis class $\mathcal{H}$ is a convex set, for all $w \in \mathcal{H}$ it is $\|w\| \leq B$, and the loss function $\ell(\cdot, z)$ is convex, nonnegative, and β-smooth function for all $z \in \mathcal{Z}$.

Example 35. Consider the regression problem with predictor rule $h(x) = \langle w, x \rangle$, loss function $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$, feature $x \in \mathbb{R}^d$, and target $y \in \mathbb{R}$. This imposes a Convex-Smooth-Bounded Learning Problem if $\mathcal{H} = \left\{ w \in \mathbb{R}^d : \|w\| \leq B \right\}$ due to Examples 13, and 29.

Proposition 36. *If ℓ is a convex loss function and the class $\mathcal{H}$ is convex, then the $ERM_{\mathcal{H}}$ problem, of minimizing the empirical risk $\hat{R}_{\mathcal{S}}(w)$ over $\mathcal{H}$, is a convex optimization problem (that is, a problem of minimizing a convex function over a convex set).*

Proof. The $\text{ERM}_{\mathcal{H}}$ problem is

$$w^* = \arg\min_{w \in \mathcal{H}} \left\{ \hat{R}_{\mathcal{S}}(w) \right\}$$

given a sample $\mathcal{S} = \{z_1, ..., z_m\}$ for $\hat{R}_{\mathcal{S}}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, z_i)$. $\hat{R}_{\mathcal{S}}(w)$ is a convex function from Proposition (13). Hence ERM rule is a problem of minimizing a convex function subject to the constraint that the solution should be in a convex set. □

Example 37. Multiple linear regression with predictor rule $h(x) = \langle w, x \rangle$, loss function $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$, feature $x \in \mathbb{R}^d$, and target $y \in \mathbb{R}$ where

$$w^* = \arg\min_{w} \text{E} \left(\langle w, x \rangle - y \right)^2$$

or

$$w^{**} = \arg\min_{w} \frac{1}{m} \sum_{i=1}^{m} \left(\langle w, x_i \rangle - y_i \right)^2$$

is a convex learning problem –from Proposition 36.

Note 38. Problems like that in Proposition 36 can be efficiently solved with algorithms such as Stochastic Gradients Descent to be introduced later.

7. Non-convex learning problems (surrogate treatment)

Remark 39. A learning problem may involve non-convex loss function $\ell(w, z)$ which implies a non-convex risk function $R_g(w)$. However, our learning algorithm will be analyzed in the convex setting. A suitable treatment to overcome this difficulty would be to upper bound the non-convex loss function $\ell(w, z)$ by a convex surrogate loss function $\tilde{\ell}(w, z)$ for all w, and use $\tilde{\ell}(w, z)$ instead of $\ell(w, z)$.

Example 40. Consider the binary classification problem with inputs $x \in \mathcal{X}$, outputs $y \in \{-1, +1\}$; we need to learn $w \in \mathcal{H}$ from hypothesis class $\mathcal{H} \subset \mathbb{R}^d$ with respect to the loss

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

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

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

which is convex (Example 13) wrt w. Note that:

- $\tilde{\ell}\left(w, (x, y)\right)$ is convex wrt w ; because $\max(\cdot)$ is convex
- $\ell\left(w, (x, y)\right) \leq \tilde{\ell}\left(w, (x, y)\right)$ for all $w \in \mathcal{H}$

Then we can compute

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

instead of

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

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

Remark 41. (Intuitions...) Using a convex surrogate loss function instead the convex one, facilitates 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 $\tilde{w}_{\text{alg}}$ is the output of the learning algorithm under $\tilde{R}_g(\cdot)$ then we have the upper bound

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

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.

8.

Proposition 42. *If ℓ is a convex loss function and the class $\mathcal{H}$ is convex, then the Ridge $ERM_{\mathcal{H}}$ problem, with learning rule*

$$\mathfrak{A}\left(\mathcal{S}\right) = \arg\min_{w\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}}\left(w\right) + \lambda\left\|w\right\|_2^2\right)$$

is a 2λ-strongly convex learning problem.

Proposition 43. *(*ERM *with Ridge regularization) If ℓ is a convex loss function, the class $\mathcal{H}$ is convex, and $J\left(\cdot;\lambda\right) = \lambda\left\|\cdot\right\|_2^2$ with $\lambda > 0$ then $\hat{R}_{\mathcal{S}}\left(w\right) + \lambda\left\|w\right\|_2^2$ is a 2λ-strongly convex function, and the $ERM_{\mathcal{H}}$ problem*

$$w^* = \arg\min_{w\in\mathcal{H}}\left\{\hat{R}_{\mathcal{S}}\left(w\right) + \lambda\left\|w\right\|_2^2\right\}$$

is a strongly convex optimization problem (i.e. the learning rule is the minimizer of a strongly convex function over a convex set).

Proof. $\hat{R}_{\mathcal{S}}(\cdot)$ is a convex function from Proposition 36, $\lambda \|\cdot\|_2^2$ is 2λ-strongly convex, hence $\hat{R}_{\mathcal{S}}(w) + \lambda \|w\|_2^2$ is a 2λ-strongly convex function. Hence the above $\mathrm{ERM}_{\mathcal{H}}$ problem is a strongly convex optimization problem. □