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

Handout 1: 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.

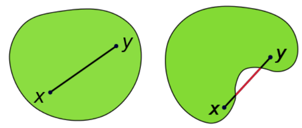
2. Convexity

Definition 3. 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 $\alpha u + (1-\alpha) v \in C$.

Definition 4. 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 $\alpha \in [0, 1]$

$$f(\alpha u + (1-\alpha) v) \le \alpha f(u) + (1-\alpha) f(v)$$



(A) Convex set (B) non-convex set

FIGURE 2.1. Convex set / non-convex set

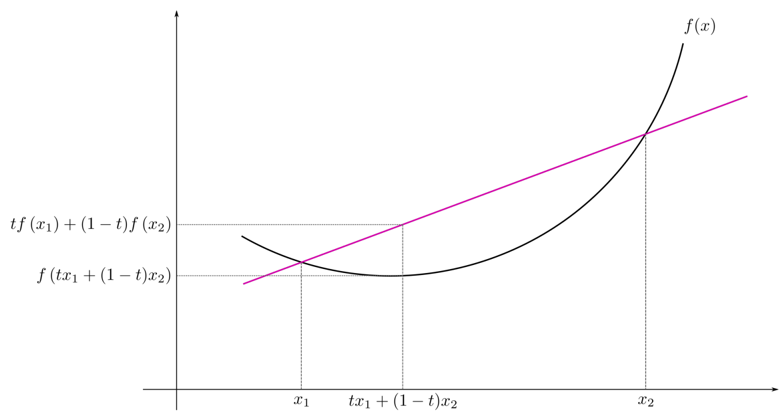


FIGURE 2.2. A convex function

Example 5. 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

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

Proposition 6. *Every local minimum of a convex function is the global minimum.*

Proposition 7. *Let $f : C \to \mathbb{R}$ be convex function. The tangent of f at $w \in C$ is below f, namely*

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

Proposition 8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(w) = g(\langle w, x \rangle + y)$ 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 9. 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 8.

Proposition 10. *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 11. $g(x) = |x|$ is convex according to Example 10, as $g(x) = |x| = \max(-x, x)$.

3. Strong convexity

Note 12. Strong convexity is a central concept in regularization, e.g. LASSO, as it makes a convex loss function strongly convex by adding a shrinkage term. The re

Definition 13. (Strongly convex functions) A function f is λ-strongly convex function is for all w, u, and $\alpha \in (0,1)$ we have

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

Proposition 14.

(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

Example 15. If $f$is λ-strongly convex and u is a minimizer of f then for any w

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

Hint:: Use the definition, and set $\alpha \to 0$.

Solution. From the definition I have

$$
\begin{aligned}
f(aw + (1-\alpha) u) \leq af(w) + (1-\alpha) f(u) - \frac{\lambda}{2} \alpha (1-\alpha) \|w-u\|^2 \Leftrightarrow \\
\frac{f(aw + (1-\alpha) u) - f(u)}{\alpha} \leq f(w) - f(u) - \frac{\lambda}{2} \alpha (1-\alpha) \|w-u\|^2 \overset{\alpha \to 0}{\Leftrightarrow} \\
\frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) \Big|_{\alpha=0} \leq f(w) - f(u) - \frac{\lambda}{2} \|w-u\|^2
\end{aligned}
$$

u is the minimizer of f, then 0 is the minimizer of $g(\alpha) = f(aw + (1-\alpha)u)$, hence $g'(0) = 0$.

4. Lipschitzness

Definition 16. 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

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

Conclusion 17. That means: a Lipschitz function $f(x)$ cannot change too drastically wrt x.

Example 18. 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 19. *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 2 from the exercise sheet

Example 20. 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 21. So, given Examples 18 and 20, 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 22. 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 23. *Function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth iff*

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

Remark 24. If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth then (5.2) holds, and if it is convex as well then

$$f(v) \ge f(w) + \langle \nabla f(w), v - w \rangle$$

holds. Hence together these conditions imply upper and lower bounds

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

Example 25. If $f : \mathbb{R}^d \to \mathbb{R}$ is β-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 $f(x) > 0$ 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 26. *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 3 from the Exercise sheet □

Example 27. 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 (26), f is $\left(2 \|x\|^2\right)$-smooth.

Example 28. 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 27.

6. Convex learning problems

Definition 29. 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 30. 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 Example 10.

Definition 31. 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 $\|w\| \le B$, and the loss function $\ell(\cdot, z)$ is convex and ρ-Lipschitz function for all $z \in \mathcal{Z}$.

Example 32. 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 10, and 18(2).

Definition 33. 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 34. 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 10, and 28.

7. Non-convex learning problems (surrogate treatment)

Remark 35. 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 36. 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

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

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

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

which is convex (Example 10) wrt w. Note that:

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

Then we can 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 37. (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

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

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.