HOMEWORK SET 3

CSCI 5525 Advanced Machine Learning (Spring 2021)

Due 11:59 pm, Mar 31 2021

Instruction Typesetting your homework in LaTeX is optional but encouraged, and you need to submit it as a single PDF file in Canvas. For programming, include all your codes and running results in a single Jupyter notebook file and submit it alongside the main PDF (since Jupyter notebook also allows text editing, feel free to put your textual answers inside the Jupyter notebook sometimes). No late submission will be accepted.

For each problem, your should acknowledge your collaborators if any. For problems containing multiple subproblems, there are often close logic connections between the subproblems. So always remember to build on previous ones, rather than work from scratch.

Reminder about notations We will use small letters (e.g., u) for scalars, small boldface letters (e.g., $\boldsymbol{a}$) for vectors, and capital boldface letters (e.g., $\boldsymbol{A}$) for matrices. For a matrix $\boldsymbol{A}$, $\boldsymbol{a}^i$ (supscripting) means its i-th row as a *row vector*, and $\boldsymbol{a}_j$ (subscripting) means the j-the column as a column vector, and a_{ij} means its (i, j)-th element. $\mathbb{R}$ is the set of real numbers. $\mathbb{R}^n$ is the space of n-dimensional real vectors, and similarly $\mathbb{R}^{m \times n}$ is the space of $m \times n$ real matrices. The dotted equal sign $\doteq$ means defining.

Problem 1 (Convex analysis & KKT conditions; 7/12) Recall that a set S is said to be convex if every line segment connecting two distinct points lies in the set, i.e., $\{t\boldsymbol{x} + (1-t)\,\boldsymbol{y} : t \in [0,1]\} \subset S$ for all $\boldsymbol{x}, \boldsymbol{y} \in S$.

Let X be a convex set. A function $f : X \mapsto \mathbb{R}$ is said to be convex if every chord connecting any two distinct points on the graph of f lies above (i.e., not below) the graph, i.e.,

$$f\left(t\boldsymbol{x} + (1-t)\,\boldsymbol{y}\right) \le t f\left(\boldsymbol{x}\right) + (1-t)\, f\left(\boldsymbol{y}\right) \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in X \text{ and } \forall\, t \in [0,1]. \tag{1}$$

To verify a function is convex, it is important to check 1) the domain X is a convex set, and 2) f satisfies the condition in Eq. (1).

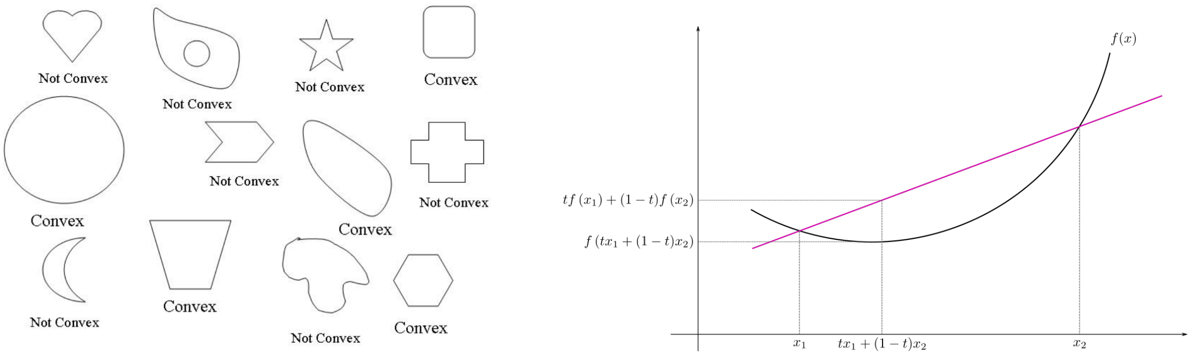

Figure 1: Illustration of convex sets (left; image credit: http://www2.econ.iastate.edu/classes/econ500/hallam/documents/Convex_Opt_000.pdf) and convex functions (right; image credit: wikipedia).

(a) Any local minimizer of a convex function is also a global minimizer. Prove this. (Hint: for any $\boldsymbol{x}, \boldsymbol{y}$, the line segment $\{t\boldsymbol{x} + (1-t)\,\boldsymbol{y} : t \in [0,1]\}$ intersects arbitrarily small balls around $\boldsymbol{x}$. Review the notes on linear predictions if you forget the difference between *minimizer* and *minimum*.) (1/12)

(b) In the definition of convex functions, when every chord connecting any two distinct points lies *strictly* above the the graph except for the two end points, the function is called *strictly convex*, i.e.,

$$f\left(t\boldsymbol{x}+(1-t)\,\boldsymbol{y}\right) < tf\left(\boldsymbol{x}\right)+(1-t)\,f\left(\boldsymbol{y}\right) \quad \forall\, \boldsymbol{x},\boldsymbol{y}\in X \text{ and } \forall\, t\in(0,1). \tag{2}$$

Prove that a strictly convex function has a unique global minimizer. (0.5/12)

Let X be a convex set. A first-order differentiable function $f : X \mapsto \mathbb{R}$ is convex if and only if

$$f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\geq\left\langle\nabla f\left(\boldsymbol{x}\right),\boldsymbol{y}-\boldsymbol{x}\right\rangle \quad \forall\, \boldsymbol{x},\boldsymbol{y}\in X, \tag{3}$$

is strictly convex if and only if

$$f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)>\left\langle\nabla f\left(\boldsymbol{x}\right),\boldsymbol{y}-\boldsymbol{x}\right\rangle \quad \forall \text{ distinct } \boldsymbol{x},\boldsymbol{y}\in X. \tag{4}$$

A second-order differentiable function $g : X \mapsto \mathbb{R}$ is convex if and only if

$$\nabla^2 g\left(\boldsymbol{x}\right)\succeq \mathbf{0} \quad \forall\, \boldsymbol{x}\in X, \tag{5}$$

is strictly convex if (the converse not true)

$$\nabla^2 g\left(\boldsymbol{x}\right)\succ \mathbf{0} \quad \forall\, \boldsymbol{x}\in X. \tag{6}$$

Is $f\left(\boldsymbol{w}\right)=\frac{1}{2}\left\|\boldsymbol{w}\right\|_2^2$ strictly convex? Why or why not? Is the Hinge loss $\phi\left(z\right)=\max\left(0,1-z\right)$ strictly convex? Why or why not? (0.5/12)

(c) Show that if $f\left(\boldsymbol{x}\right) : X \mapsto \mathbb{R}$ is a convex function, the sublevel set $\{\boldsymbol{x}\in X : f\left(\boldsymbol{x}\right)\leq 0\}$ is a convex set. (0.5/12) Is $\{\boldsymbol{x}\in X : f\left(\boldsymbol{x}\right)=0\}$ a convex set? Why or why not? What about when $f\left(\boldsymbol{x}\right)$ is an affine function, i.e., $f\left(\boldsymbol{x}\right)=\left\langle\boldsymbol{w},\boldsymbol{x}\right\rangle+b$ for certain $\boldsymbol{w}$ and b? (0.5/12)

(d) Recall operations that preserve convexity for sets: for any two convex sets S_1, S_2,

- the set product $S_1\times S_2=\{(\boldsymbol{x}_1,\boldsymbol{x}_2) : \boldsymbol{x}_1\in S_1, \boldsymbol{x}_2\in S_2\}$ is convex;
- the set summation $S_1+S_2=\{\boldsymbol{x}_1+\boldsymbol{x}_2 : \boldsymbol{x}_1\in S_1, \boldsymbol{x}_2\in S_2\}$ is convex;
- the set projection $\{(x_1,\dots,x_k) : (x_1,\dots,x_k,x_{k+1},\dots,x_n)\in S_1\}$ is convex for any $k\in[n]$.

In addition,

- for a family of sets $\{S_i\}_{i=1}^K$, the set intersection $\bigcap_{i=1}^K S_i$ is convex.

Moreover, all valid vector and matrix norms, particularly vector ℓ_p norms for all $p\geq 1$, and the matrix Frobenius norm are all convex functions. This is because valid norms should by definition satisfy the triangular inequality: i.e., $\left\|\boldsymbol{x}+\boldsymbol{y}\right\|_\diamond\leq\left\|\boldsymbol{x}\right\|_\diamond+\left\|\boldsymbol{y}\right\|_\diamond$ for any norm $\left\|\cdot\right\|_\diamond$ under consideration. Operations that preserve convexity of functions include:

- positive combinations: $\sum_{i=1}^K \alpha_i f_i\left(\boldsymbol{x}\right)$ is convex over $\bigcap_{i=1}^K S_i$ for all $\alpha_i\geq 0$ if $f_i\left(\boldsymbol{x}\right)$ is convex over S_i for all $i\in[K]$.
- pointwise maximization: $\max_{i\in[K]} f_i\left(\boldsymbol{x}\right)$ is convex over $\bigcap_{i=1}^K S_i$ if $f_i\left(\boldsymbol{x}\right)$ is convex over S_i for all $i\in[K]$.
- composition with affine functions: $f\left(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\right)$ is convex over X (which is the domain of the affine function $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$) if f is convex over the range of $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$, i.e., $\{\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b} : \boldsymbol{x}\in X\}$.

- composition: $h \circ g$ is convex over X if both $g : X \mapsto \mathbb{R}$ and $h : \mathbb{R} \mapsto \mathbb{R}$ are twice differentiable and either of the following holds:
  - h is convex and non-decreasing and g is convex,
  - h is convex and non-increasing and $-g$ is convex.
- partial minimization: Let $f(\boldsymbol{x}, \boldsymbol{y})$ be a convex function over $X \times Y$, where both X and Y are convex sets and hence $X \times Y$ is also a convex set. The partial minimization $\inf_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y})$ [1] is convex over X.

(i) Prove that the Lasso objective

$$f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1 \tag{7}$$

is convex. You're supposed to use facts and properties covered above only and try to work from first principles. Directly stating that say, $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ is convex does not work. (1/12)

(ii) For any convex subset $S \subset \mathbb{R}^n$, the distance function to S induced by the ℓ_p norm is defined as

$$d_{\ell_p}(\boldsymbol{x}; S) \doteq \inf_{\boldsymbol{y} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|_p . \tag{8}$$

Show that $d_{\ell_p}(\boldsymbol{x}; S)$ is convex. Similar to (i), you're supposed to prove this from first principles. (1/12)

(e) Consider a convex optimization problem:

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) \quad \text{s.t.} \; g_i(\boldsymbol{x}) \le 0, \forall\, i \in \mathcal{I} \quad \text{and} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0}, \tag{9}$$

where f and g_i's are convex functions and $\mathcal{I}$ is the index set for the inequality constraints. Define the Lagrangian function as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \doteq f(\boldsymbol{x}) + \sum_{i \in \mathcal{I}} \pi_i g_i(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \rangle , \tag{10}$$

where $\boldsymbol{x}$ is called the primal variable, and $\boldsymbol{\pi} \ge \boldsymbol{0}$ and $\boldsymbol{\lambda}$ are called the dual variables. The KKT optimality condition states:

> Suppose f and $g_i \;\forall\, i$ are continuously differentiable, and the constraint set is *strictly feasible* (also called the *Slater's condition*), i.e., there exists an $\boldsymbol{x}_0$ so that $g_i(\boldsymbol{x}_0) < 0$ for all $i \in \mathcal{I}$ and $\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{b} = \boldsymbol{0}$. Then, $\boldsymbol{x}_*$ is a global minimizer if and only if there exist dual variables $\boldsymbol{\lambda}_*$ and π_* so that the following hold simultaneously:
>
> - **stationarity**: $\partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}_*, \boldsymbol{\pi}_*, \boldsymbol{\lambda}_*) = \boldsymbol{0}$
> - **feasibility**: primal feasibility $g_i(\boldsymbol{x}_*) \le 0 \;\forall\, i \in \mathcal{I}$ and $\boldsymbol{A}\boldsymbol{x}_* + \boldsymbol{b} = \boldsymbol{0}$, and dual feasibility: $\boldsymbol{\pi}_* \ge \boldsymbol{0}$
> - **complementary slackness**: $(\pi_*)_i g_i(\boldsymbol{x}_*) = 0$ for all $i \in \mathcal{I}$, or equivalently $\sum_{i \in \mathcal{I}} (\pi_*)_i g_i(\boldsymbol{x}_*) = 0$

[1] Recall that $\inf$ can be roughly treated as $\min$, but for $\min$ the minimum value must be achieved by a point inside the domain, whereas for $\inf$ we take a limit point that can be outside. An example is minimizing $f(x) = x^2$ over $(-\infty, 0)$. Here, $\min f(x)$ does not make sense, but $\inf f(x) = 0$. So, in general, it is safer to use $\inf$, instead of $\min$, when we mean to perform minimization.

Make use of the KKT condition to solve the following projection problems.

(i) Projection onto a hyperplane. Prove that ℓ_2 projection of a point $\boldsymbol{y}$ onto an affine hyperplane $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$, where $\boldsymbol{A}$ has full row rank, is $\left(\boldsymbol{I} - \boldsymbol{A}^{\intercal}\left(\boldsymbol{A}\boldsymbol{A}^{\intercal}\right)^{-1}\boldsymbol{A}\right)\boldsymbol{y} + \boldsymbol{A}^{\intercal}\left(\boldsymbol{A}\boldsymbol{A}^{\intercal}\right)^{-1}\boldsymbol{b}$, i.e., the global minimizer to

$$\min_{\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \mathrm{s.\,t.}\ \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{11}$$

is $\boldsymbol{x}_* = \left(\boldsymbol{I} - \boldsymbol{A}^{\intercal}\left(\boldsymbol{A}\boldsymbol{A}^{\intercal}\right)^{-1}\boldsymbol{A}\right)\boldsymbol{y} + \boldsymbol{A}^{\intercal}\left(\boldsymbol{A}\boldsymbol{A}^{\intercal}\right)^{-1}\boldsymbol{b}$. (1/12)

(ii) Projection onto an ℓ_2 ball. Prove that the global minimizer $\boldsymbol{x}_*$ to

$$\min_{\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \mathrm{s.\,t.}\ \ \|\boldsymbol{x}\|_2^2 \le 1 \tag{12}$$

is

$$\boldsymbol{x}_* = \begin{cases} \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} & \|\boldsymbol{y}\|_2 \ge 1 \\ \boldsymbol{y} & \|\boldsymbol{y}\|_2 < 1 \end{cases}. \tag{13}$$

You can draw a picture on $\mathbb{R}^2$ to see if this solution makes sense. (1/12)

Problem 2 (Linear SVMs; 5/12) Assume a training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ with $y_i \in \{1, -1\}$ for all i. We consider the hard-margin SVM formulation

$$\min_{\boldsymbol{w}, b} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2 \quad \mathrm{s.\,t.}\ \ y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right) \ge 1 \ \forall\, i \tag{14}$$

and the *generalized* soft-margin SVM formulation

$$\min_{\boldsymbol{w}, b, \xi_i' s} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^N \xi_i^p \quad \mathrm{s.\,t.}\ \ y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right) \ge 1 - \xi_i, \xi_i \ge 0 \ \forall\, i, \tag{15}$$

where p can take any value from $[1, \infty)$.

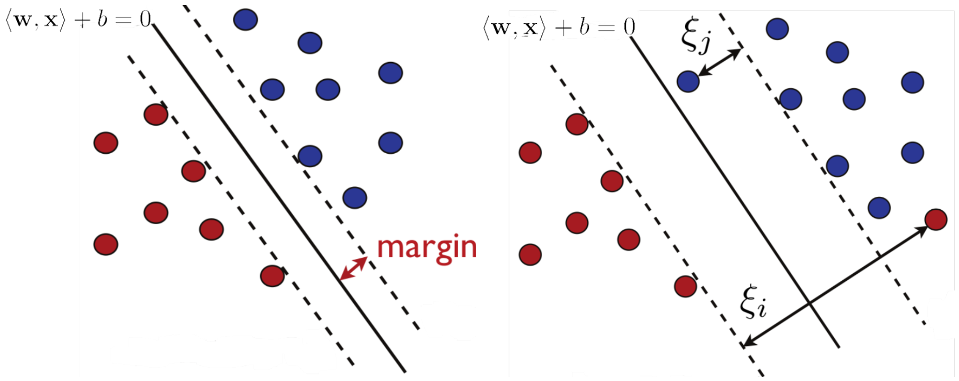
$\langle \mathrm{w}, \mathrm{x}\rangle + b = 0$
margin
$\langle \mathrm{w}, \mathrm{x}\rangle + b = 0$
ξ_j
ξ_i

Figure 2: Illustration of hard-margin (left) and soft-margin (right) SVMs. Figures adapted from Chap 5 of [Moh18].

(a) Suppose the training set is linearly separable. Recall that in our lecture on hard-margin SVM, we made a handwavy argument that the two marginal hyperplanes must have the same distance to the separating (also called maximum-margin) hyperplane $\{\boldsymbol{x} : \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0\}$. Prove it using the KKT condition for optimality, and also write down the set representations of the two hyperplanes. (1/12)

(b) Prove that both the hard-margin and soft-margin SVM problems with $p > 1$ have unique minimizers. (Hint: think of strict convexity) (1/12)

(c) Apply the KKT condition to the soft-margin SVM with $p = 1$, and conclude that any support vector/point is either on the marginal hyperplane, or an outlier with a strictly positive slackness variable, i.e., $\xi > 0$. (1/12)

(d) What happens when we apply the hard-margin SVM to a training set that is not linearly separable? (1/12)

(e) What happens when we apply the soft-margin SVM with $p = 1$ to a training set that is linearly separable? Particularly, do we expect to obtain the same hyperplane as that obtained if hard-margin SVM is applied? (Hint: draw a picture when thinking about this, and consider how varying C from a small to a large value changes the two types of support vectors in (c).) (1/12)

References

[Moh18] Mehryar Mohri, *Foundations of machine learning*, 2 ed., The MIT Press, Cambridge, Massachusetts, 2018.