AM 221: Advanced Optimization Spring 2018

Kojin Oshiba *Problem Set 7 (updated) — Due Wednesday, Mar. 21st at 23:59*

Instructions:

- All your solutions should be prepared in LaTeX and the PDF and .tex should be submitted to Canvas. Please submit all your files as ONE archive of filetype zip, tgz, or tar.gz.
- Name the file [your-first-name]_[your-last-name].[filetype]. For example, I would call my submission rasmus_kyng.zip.
- INCLUDE your name in the submisson pdf and any files with code.
- If the TFs cannot easily deduce your identity from your files alone, they may decide not to grade your submission.
- For each question, a well-written and correct answer will be selected a sample solution for the entire class to enjoy. If you prefer that we do not use your solutions, please indicate this clearly on the first page of your assignment.

1. Entropy maximization. In this problem, we will consider the *entropy maximization problem.* Let us consider a probability distribution $\mathbf{x} \in \mathbb{R}^n$ over a finite set of size n. We have $\mathbf{x} \geq 0$ and $\sum_{i=1}^n x_i = 1$. The entropy of $\mathbf{x}$ is defined by:

$$H(\mathbf{x}) = \sum_{i=1}^n x_i \log \frac{1}{x_i}$$

We are interested in maximizing entropy, or equivalently, solving the following problem:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} & \sum_{i=1}^n x_i \log x_i \\ \text{s.t} & \sum_{i=1}^n x_i = 1 \\ & \mathbf{x} \geq 0 \end{aligned}$$

a. Prove Jensen's inequality: let $f : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex function, let $\mathbf{x}_1, \dots \mathbf{x}_m$ by m vectors in $\mathbb{R}^n$, and let $\lambda_1, \dots, \lambda_m$ be such that $\sum_{i=1}^m \lambda_i = 1$ and $\lambda_i \geq 0$, $1 \leq i \leq m$, then:

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i)$$

and prove that the inequality is an equality if and only if $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_m$.

b. Using Jensen's inequality, what is the optimal solution to the entropy maximization problem above? Specify both the distribution $\mathbf{x}$ of maximum entropy and the value of its entropy.

c. We now add the constraint $A\mathbf{x} \leq b$ to the entropy maximization problem, where $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. The problem now becomes:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \ & \sum_{i=1}^{n} x_i \log x_i \\ \text{s.t} \ & \sum_{i=1}^{n} x_i = 1 \\ & \mathbf{x} \geq 0 \\ & A\mathbf{x} \leq \mathbf{b} \end{aligned}$$

Show that the dual of this problem can be written in the following form:

$$\begin{aligned} \max_{\nu\in\mathbb{R}^m} \ & -\mathbf{b}^\intercal \nu - \log\left(\sum_{i=1}^{n} e^{-\mathbf{a}_i^\intercal \nu}\right) \\ \text{s.t} \ & \nu \geq 0 \end{aligned}$$

where $\mathbf{a}_i$ is the ith column of A. Assuming that strong duality holds for this problem, re-derive the result of part b. by considering a pair of primal/dual optimal solutions.

a. We show this by induction. When $m = 2$, because f is convex,

$$f(\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2) \leq \lambda_1 f(\mathbf{x}_1) + \lambda_2 f(\mathbf{x}_2)$$

Suppose $f\left(\sum_{i=1}^m \lambda_i\mathbf{x}_i\right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i)$ for $m = k$. Then, for $m = k+1$,

$$\begin{aligned} f\left(\sum_{i=1}^{m} \lambda_i\mathbf{x}_i\right) &= f(\lambda_{k+1}\mathbf{x}_{k+1} + \sum_{i=1}^{k} \lambda_i\mathbf{x}_i) \\ &= f(\lambda_{k+1}\mathbf{x}_{k+1} + (1-\lambda_{k+1})\sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}}\mathbf{x}_i \\ &\leq \lambda_{k+1} f(\mathbf{x}_{k+1}) + (1-\lambda_{k+1}) f(\sum_{i=1}^{k} \frac{1}{1-\lambda_{k+1}}\mathbf{x}_i) \end{aligned}$$

Since $\sum_{i=1}^{k+1} \lambda_i = 1, 1-\lambda_{k+1} = \sum_{i=1}^{k} \lambda_i$. Hence, $\sum_{i=1}^{k} \frac{1}{1-\lambda_{k+1}} = 1$.
So, $f(\sum_{i=1}^{k} \frac{1}{1-\lambda_{k+1}}\mathbf{x}_i) \leq \frac{1}{1-\lambda_{k+1}} \sum_{i=1}^{k} \lambda_i f(\mathbf{x}_i)$. Therefore,

$$\begin{aligned} f(\sum_{i=1}^{m} \lambda_i\mathbf{x}_i) &\leq \lambda_{k+1} f(\mathbf{x}_{k+1}) + \sum_{i=1}^{k} \lambda_i f(\mathbf{x}_i) \\ &= \sum_{i=1}^{k+1} \lambda_i f(\mathbf{x}_i) \end{aligned}$$

When $\mathbf{x}_1 = ... = \mathbf{x}_m$,

$$f\left(\sum_{i=1}^{m} \lambda_i \mathbf{x}_i\right) = f(\mathbf{x}_i), \sum_{i=1}^{m} \lambda_i f(\mathbf{x}_i) = f(\mathbf{x}_i)$$

The converse is true by induction analogous to the above proof: using the fact that in the base case, $f(\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2) = \lambda_1 f(\mathbf{x}_1) + \lambda_2 f(\mathbf{x}_2)$ implies $\mathbf{x}_1 = \mathbf{x}_2$ because f is strictly convex.

b. First,

$$min \sum_{i=1}^{n} x_i log x_i = min\ n \sum_{i=1}^{n} \frac{1}{n} x_i log x_i$$

Since $f(x) = xlogx$ is convex for $0 \leq x \leq 1$,

$$\sum_{i=1}^{n} \frac{1}{n} x_i log x_i \geq f(\sum_{i=1}^{n} \frac{1}{n} x_i) = (\sum_{i=1}^{n} \frac{1}{n} x_i) log(\sum_{i=1}^{n} \frac{1}{n} x_i)$$

When $x_1 = ... = x_n = x$,

$$-n(\sum_{i=1}^{n} \frac{1}{n} x_i) log(\sum_{i=1}^{n} \frac{1}{n} x_i) = -nxlogx = logn$$

c. The langrangian is

$$L(\mathbf{x}, \lambda, \nu) = \sum_{i=1}^{n} x_i log x_i + \nu^{\intercal}(A\mathbf{x} - \mathbf{b}) + \lambda(1^{\intercal}\mathbf{x} - 1)$$

$$= -b^{\intercal} - \lambda + \sum_{i=1}^{n} x_i(log x_i + a_i^{\intercal}\nu + \lambda)$$

$$\frac{\partial}{\partial x_i} L(\mathbf{x}, \lambda, \nu) = \frac{\partial}{\partial x_i} x_i(log x_i + a_i^{\intercal}\nu + \lambda)$$

To minimize $L(x, \lambda, \nu)$, set this to 0, which results in

$$x_i^* = exp(-a_i^{\intercal}\nu - \lambda - 1)$$

Hence,

$$inf_{\mathbf{x}} L(\mathbf{x}, \nu, \lambda) = L(\mathbf{x}^*, \nu, \lambda) = -b^{\intercal}\nu - \lambda - \sum_{i=1}^{n} exp(-a_i^{\intercal}\nu - \lambda - 1)$$

Now, let's find the optimal λ.

$$\frac{\partial}{\partial \lambda} L(\mathbf{x}^*, \nu, \lambda) = -1 + \sum_{i=1}^{n} exp(-a_i^{\intercal}\nu - \lambda - 1)$$

Setting this to 0 and solving for λ results in

$$\lambda * = log(\sum_{i=1}^{n} exp(-a_i^\intercal \nu - 1)$$

Hence,

$$\begin{aligned} L(\mathbf{x}^*, \nu, \lambda *) &= -b^\intercal \nu - \lambda * + \sum_{i=1}^{n} x_i(log x_i + a_i^\intercal \nu + \lambda) \\ &= -b^\intercal \nu - \lambda * + 1 \\ &= -b^\intercal \nu + log(\sum_{i=1}^{n} x_i exp(-a_i^\intercal \nu)) \end{aligned}$$

2. Minimum volume ellipsoid. An ellipsoid in $\mathbb{R}^d$ is the image of the unit ball by a linear invertible map, i.e a set $\mathcal{E}$ defined by:

$$\mathcal{E} = \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^d,\ \|\mathbf{x}\|_2 \leq 1\}$$

for some invertible linear map $A : \mathbb{R}^d \mapsto \mathbb{R}^d$. In this case, we define the volume of the ellipsoid to be $|\det A|$. An equivalent parametrization of the ellipsoid is:

$$\mathcal{E} = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{y}^\intercal W \mathbf{y} \leq 1\}$$

with $W = (A^{-1})^\intercal A^{-1}$. Note that W is symmetric positive definite and that under this parametrization, the volume of the ellipsoid is $(\det W)^{-1/2}$.

Let us denote by $\mathbf{S}_d^{++}$ the set of symmetric positive definite matrices of size $d \times d$. Given n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$, the *minimum volume ellipsoid* problem consists in finding the ellipsoid of minimum volume containing all points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, that is:

$$\begin{aligned} \min_{W \in \mathbf{S}_d^{++}} \quad & (\det W)^{-1/2} \\ \text{s.t. } & \mathbf{x}_i^\intercal W \mathbf{x}_i \leq 1,\ 1 \leq i \leq n \end{aligned}$$

a. Show that $\mathbf{S}_d^{++}$ is convex.

b. Let us define $d : \mathbf{S}_d^{++} \to \mathbb{R}$ by $d(W) = (\det W)^{-1/2}$. Is d convex over $\mathbf{S}_d^{++}$?

Using the fact that log is increasing over $\mathbb{R}^+ \setminus \{0\}$, we consider the following problem which is equivalent to the minimum volume ellipsoid problem:

$$\begin{aligned} \min_{W \in \mathbf{S}_d^{++}} \quad & \log\det(W^{-1}) \\ \text{s.t. } & \mathbf{x}_i^\intercal W \mathbf{x}_i \leq 1,\ 1 \leq i \leq n \end{aligned} \tag{1}$$

c. Show that the function f defined by $f(W) = \log\det(W^{-1})$ is convex and differentiable over S_d^{++} and that $\nabla f(W) = -W^{-1}$.

d. Show that the dual of problem (1) is:

$$\begin{aligned}&\max_{\lambda\in\mathbb{R}^n}\ \log\det\left(\sum_{i=1}^n \lambda_i \mathbf{x}_i\mathbf{x}_i^\intercal\right)-\sum_{i=1}^n\lambda_i+d\\ &\text{s.t. } \lambda\geq 0,\ \sum_{i=1}^n \lambda_i\mathbf{x}_i\mathbf{x}_i^\intercal\in\mathbf{S}_d^{++}\end{aligned}$$

e. Show that the dual can be further simplified to:

$$\begin{aligned}&\max_{\lambda\in\mathbb{R}^n}\ \log\det\left(\sum_{i=1}^n \lambda_i \mathbf{x}_i\mathbf{x}_i^\intercal\right)+d\log d\\ &\text{s.t. } \lambda\geq 0,\ \sum_{i=1}^n \lambda_i\mathbf{x}_i\mathbf{x}_i^\intercal\in\mathbf{S}_d^{++}, \sum_{i=1}^n\lambda_i=1\end{aligned}\tag{2}$$

a. For $A_1, A_2 \in S_d^{++}$,

$$(\lambda A_1+(1-\lambda)A_2)^\intercal=\lambda A_1^\intercal+(1-\lambda)A_2^\intercal=\lambda A_1+(1-\lambda)A_2$$

Moreover,

$$\mathbf{x}^\intercal(\lambda A_1+(1-\lambda)A_2)\mathbf{x}=\lambda\mathbf{x}^\intercal A_1\mathbf{x}+(1-\lambda)\mathbf{x}^\intercal A_2\mathbf{x}>0$$

Hence,

$$\lambda A_1+(1-\lambda)A_2\in S_d^{++}$$

Thus, S_d^{++} is convex.

b.

$$(detW)^{-\frac{1}{2}}=exp(-\frac{1}{2}logdet(W))$$

Since$logdet$ is concave from (c), $-logdet$ is convex. Moreover, we have that exp is convex and monotonically increasing. Since $f\circ g$ is convex if f is convex and g is convex and monotonically increasing, we have that $(detW)^{-\frac{1}{2}}$ is convex.

c. To show this, it suffices to show that $logdet(W)$ is concave. To show this, let $g(t)=f(W+tX)$ where $W,X\in S_d^{++}$. Since S_d^{++} is convex, $W+tX\in S_d^{++}$ as well. Now,

$$\begin{aligned}g(t)&=logdet(W+tX)\\&=logdet(W^{1/2}(I+tW^{-1/2}XW^{1/2})W^{1/2}\\&=\sum_{i=1}^n log(1+t\lambda_i)+logdetW\end{aligned}$$

where λ_i's are eigen values of $W^{-1/2}XW^{1/2})W^{1/2}$. Hence,

$$g''(t)=-\sum_{i=1}^n\frac{\lambda_i^2}{(1+t\lambda_i)^2}\leq 0$$

Hence, we have that f is concave.

d.
$$L(W,\lambda) = logdet(W^{-1}) + \sum_{i=1}^{n} \lambda_i(\mathbf{x}_i^\intercal W \mathbf{x}_i - 1)$$

Hence,
$$\frac{\partial L(W,\lambda)}{\partial W} = -W^{-1} + \sum_{i=1}^{n} \lambda_i \mathbf{x}_i^\intercal \mathbf{x}_i$$

Setting this to 0 reuslts in
$$W^{*-1} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i^\intercal \mathbf{x}_i$$

So,
$$L(W^*,\lambda) = logdet(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i^\intercal \mathbf{x}_i) + \sum_{i=1}^{n} \lambda_i \mathbf{x}_i^\intercal W \mathbf{x}_i - \sum_{i=1}^{n} \lambda_i$$

Now,
$$\begin{aligned} tr(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i^\intercal W \mathbf{x}_i) &= \sum_{i=1}^{n} \lambda_i tr(\mathbf{x}_i^\intercal W \mathbf{x}_i) \\ &= \sum_{i=1}^{n} \lambda_i tr(W \mathbf{x}_i \mathbf{x}_i^\intercal) \\ &= tr(W \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\intercal) \\ &= tr(I) \\ &= d \end{aligned}$$

Hence,
$$L(W^*,\lambda) = logdet(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i^\intercal \mathbf{x}_i) + d - \sum_{i=1}^{n} \lambda_i$$

which is the objective we would like to have. Finally the constraint $\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\intercal \in \mathbf{S}_d^{++}$ comes from the fact that we want W^{-1} to be positive definite.

e. Let $\sum_{i=1}^{n} \lambda'_i = 1, \lambda' \geq 0$. This will correspond to the λ_i's in e, as we will show. Let $S = \sum_{i=1}^{n} \lambda_i =$ be the sum of λ_i's in d. Let $\lambda_i = \lambda'_i S$. Then, we have that

$$logdet(\sum_{i=1}^{n} \lambda \mathbf{x}_i \mathbf{x}_i^\intercal) - \sum_{i=1}^{n} \lambda_i + d = logdet(S \sum_{i=1}^{n} \lambda'_i \mathbf{x}_i \mathbf{x}_i^\intercal) - S + d$$

Maximizing this with respect to S by taking the derivative and setting this to 0, we have $S = d$. Hence, the objective now becomes

$$logdet(\sum_{i=1}^{n} \lambda'_i \mathbf{x}_i \mathbf{x}_i^\intercal) - dlogd$$

with the constraint that $\sum_{i=1}^{n} \lambda'_i = 1$ We can rewrite λ'_i as λ_i to get the notations as specified in the problem.

3. Support vector machines. In this problem, we will use a dataset on forged banknotes. We will use support vector machines to construct a classifier that tries to predict if notes are forged. The dataset is available at `http://rasmuskyng.com/am221_spring18/psets/hw7/banknotes.data`. In each line, the first four columns contain measurements from a banknote (real numbers) and the last column is a binary (0 or 1) variable indicating if the banknote was forged. Denoting by $\mathbf{x}^i \in \mathbb{R}^4$ the measurements from banknote i, the goal is to construct a classifier which takes $\mathbf{x}^i$ as input and predicts the last column $y^i \in \{0, 1\}$.

First, convert the labels to $\hat{y}^i \in \{-1, 1\}$, i.e. $\hat{y}^i = 2y^i - 1$. As seen in class, finding a separating hyperplane now amounts to finding $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $\hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1$, for $1 \leq i \leq n$, where $\mathbf{x}_1, \dots \mathbf{x}_n$ are the (modified) data points.

As seen in class, the optimization problem for support vector machines now takes the following form:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \ & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t } & \hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1, \ 1 \leq i \leq n \end{aligned}$$

In cases where the dataset is not linearly separable, it is not possible to find $\mathbf{w}$ satisfying the constraints of the above problem. In particular, we might have $\hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b) < 1$ for some i. If this is the case, there exists $\xi_i \geq 0$ such that $\hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b) + \xi_i \geq 1$. The number ξ_i quantifies the "misclassification" of data point i. Since we want to discourage these misclassifications, we incorporate them into the objective function and consider the following optimizing problem instead:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d,\, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \ & \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{n} \xi_i \\ \text{s.t } & \hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b) + \xi_i \geq 1, \ 1 \leq i \leq n \\ & \xi_i \geq 0, \ 1 \leq i \leq n \end{aligned} \tag{3}$$

where λ is a parameter that we can choose depending on how much we want to penalize misclassified data points.

a. Reuse your implementation of the perceptron algorithm from HW2, Question 5d and run it on the banknote dataset. Which behavior do you observe? Can you explain why?

b. Use a convex solver to solve the convex program (3). Note that the objective function is quadratic, so you can use a function specific to quadratic problems. In CVXOPT, this is the cvxopt.solvers.qp function. Solve the problem for different values of λ and plot the classification accuracy (fraction of the data points that were correctly classified) as a function of λ. How do you explain the shape of this plot?

c. Show that the program (3) is equivalent to the following problem:

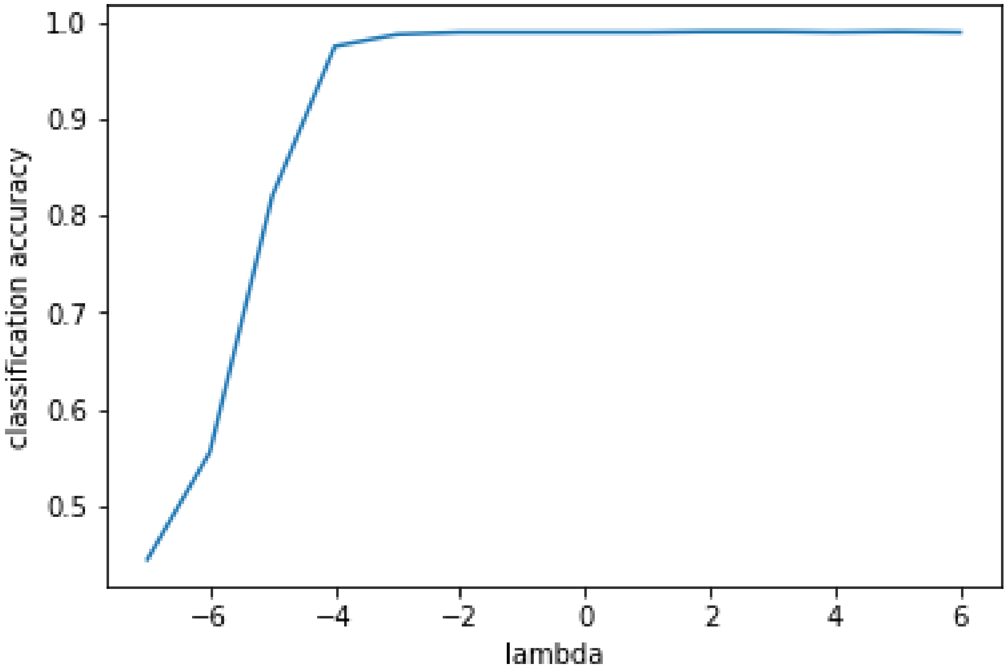
$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{n} \max(0, 1 - \hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b)) \tag{4}$$

d. **[Optional, for bonus credits]** The advantage of problem (4) is that it is unconstrained. So we can use subgradient descent to solve it. Run the subgradient descent algorithm to solve Problem (4) for the best value of λ found in part b.

e. **[Optional, for bonus credits]** Note that (4) also has a "separable" objective function as seen in Stochastic Gradient Descent (section 7). Implement the Stochastic Gradient descent algorithm and use it to solve (4). Compare the number of iterations required to reach the same accuracy with gradient descent (part d.) and stochastic gradient descent (part e.).

a. The implementation is at 'hw7.py'. The perceptron doesn't hult, since the data is not linearly separable.

b. I observe that the larger lambda is the higher the accuracy is. This makes sense because as lambda increases, we are penalizing the size of the slack variables more, meaning that we are being more strict to the points misclassified.



c.

$$
\begin{aligned}
&\xi_i \geq 0, \hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b) + \xi_i \geq 1 \\
\Leftrightarrow &\xi_i \geq 0, \xi_i \geq 1 - \hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b) \\
\Leftrightarrow &\xi_i \geq max(0, 1 - \hat{y}^i(\mathbf{w}^\intercal \mathbf{x}_i + b))
\end{aligned}
$$

Hence, we can replace ξ_i in the minimization problem with its lower bound.