# APPM 5360, Spring 2023 - Written Homework 5

Eappen Nelluvelil; Collaborators: Jack, Bisman, Logan, Tyler

March 3, 2023

1. Derive

$$\nabla \ell\left(\mathbf{w}\right) = -\sum_{i=1}^{n} \sigma\left(-y_i \mathbf{w}^{\mathbf{T}} \mathbf{x_i}\right) y_i \mathbf{x_i}$$

using calculus.

**Hint**: First derive $\sigma'\left(a\right) = \sigma\left(a\right)\left(1 - \sigma\left(a\right)\right)$.

Note that

$$\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{w_i}} &= \sum_{k=1}^{n} \frac{\partial}{\partial \mathbf{w_i}} \left[\log\left(1 + e^{-y_k \sum_{j=1}^{n} \mathbf{w_j} \mathbf{x_{k,j}}}\right)\right] \\
&= \sum_{k=1}^{n} \frac{e^{-y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}}}{1 + e^{-y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}}} \left(-y_k \mathbf{x_{k,i}}\right) \\
&= -\sum_{k=1}^{n} \sigma\left(-y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}\right) y_k \mathbf{x_{k,i}},
\end{aligned}$$

which implies that $\nabla \ell\left(\mathbf{w}\right) = -\sum_{i=1}^{n} \sigma\left(-y_i \mathbf{w}^{\mathbf{T}} \mathbf{x_i}\right) y_i \mathbf{x_i}$, as desired.

2. (a) Derive $\nabla^2 \ell\left(\mathbf{w}\right) = \sum_{i=1}^{n} \mu_i \left(1 - \mu_i\right) \mathbf{x_i} \mathbf{x_i^T} = X^T S X$, where $S = \text{diag}\left(\mu_i \left(1 - \mu_i\right)\right)$.

We first derive that $\sigma'\left(x\right) = \sigma\left(x\right)\left(1 - \sigma\left(x\right)\right)$. First, note that $\sigma\left(x\right) = e^x \left(1 + e^x\right)^{-1}$, so

$$\begin{aligned}
\sigma'\left(x\right) &= e^x \left(1 + e^x\right)^{-1} - e^x \left(1 + e^x\right)^{-2} \\
&= \frac{e^x}{1 + e^x} - \frac{\left(e^x\right)^2}{\left(1 + e^x\right)^2} \\
&= \frac{e^x}{1 + e^x}\left(1 - \frac{e^x}{1 + e^x}\right) \\
&= \sigma\left(x\right)\left(1 - \sigma\left(x\right)\right).
\end{aligned}$$

Then, we have that

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mathbf{w_j} \partial \mathbf{w_i}} &= -\sum_{k=1}^{n} \frac{\partial}{\partial \mathbf{w_j}} \left[y_k x_{k,i} \sigma\left(-y_k \sum_{\ell=1}^{n} w_\ell x_{k,\ell}\right)\right] \\
&= -\sum_{k=1}^{n} \left(y_k x_{k,i}\right) \sigma\left(-y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}\right)\left(1 - \sigma\left(-y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}\right)\right)\left(-y_k x_{k,j}\right) \\
&= \sum_{k=1}^{n} \sigma\left(-y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}\right)\left(1 - \sigma\left(-y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}\right)\right) x_{k,i} x_{k,j} \\
&= \sum_{k=1}^{n} \sigma\left(y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}\right)\left(1 - \sigma\left(y_k \mathbf{w}^{\mathbf{T}} \mathbf{x_k}\right)\right) x_{k,i} x_{k,j},
\end{aligned}$$

where we used the property of the sigmoid function that $\sigma(a) = 1 - \sigma(-a)$ to obtain the last line. Using the definition of matrix multiplication, this implies that $\nabla^2 \ell(\mathbf{w}) = \sum_{i=1}^{n} \mu_i (1 - \mu_i) \mathbf{x_i} \mathbf{x_i^T}$, which can be written as $X^T S X$, where $X$ and $S$ are defined per the problem description.

(b) Assuming all data is real-valued (no $\pm\infty$), then $0 < \mu_i < 1$. Is $\ell(\mathbf{w})$ convex? strictly convex? strongly convex? Is $\nabla \ell(\mathbf{w})$ Lipschitz continuous? If so, what constant? Prove/disprove your answers (you may assume that $X$ is full-rank and $n \times p$ with $n \geq p$).

**Hint**: the Lipschitz constant can be found as $L = \sup_{\mathbf{w}} \left|\left|\nabla^2 \ell(\mathbf{w})\right|\right|$, where $||\cdot||$ is the spectral norm.

i. Is $\ell(\mathbf{w})$ convex?
Yes. Let $x \in \mathbb{R}^p$ be given, with $x \neq 0$. Since $X$ is full rank, $Xx = 0$ if and only if $x = 0$, so we have that $Xx \neq 0$. We then compute $x^T X^T S X x = (Xx)^T S (Xx)$. Since $S$ is a diagonal matrix with positive diagonal entries, we have that $(Xx)^T S (Xx) > 0$. Furthermore, $X^T S X$ is symmetric by construction, so we have that $X^T S X$ is symmetric, positive-definite, which implies that $\ell(\mathbf{w})$ is convex.

ii. Is $\ell(\mathbf{w})$ strictly convex?
Because $X^T S X$ is symmetric, positive-definite, we have that it is strictly convex.

iii. Is $\ell(\mathbf{w})$ strongly convex?
No, $\ell(\mathbf{w})$ is not strongly convex. For $\ell(\mathbf{w})$ to be strongly convex, we require that $\nabla^2 \ell(\mathbf{w}) - mI$ be symmetric, positive definite for some $m > 0$. However, note that we cannot find such an $m$ for $X^T S X$. The diagonal matrix $S$ depends on $\mathbf{w}$, and we can make $\mu_i (1 - \mu_i)$ arbitrarily close to 0 for each $i = 1, 2, \ldots, n$ by choosing $\mathbf{w}$ in such a way that $\sigma\left(y_i \mathbf{w^T x_i}\right)$ gets arbitrarily close to 1, i.e., $\mu_i$ gets arbitrarily close to 1 and $(1 - \mu_i)$ gets arbitrarily cose to 0, or vice versa. Thus, we can make $S$ get arbitrarily close to the zero matrix, which prevents us from obtaining a uniform bound $m$ that works for all $\mathbf{w} \in \mathbb{R}^{\mathbf{p}}$.

iv. Is $\nabla \ell(\mathbf{w})$ Lipschitz continuous?
Yes, $\ell(\mathbf{w})$ Lipschitz continuous. We assume our matrix $X$ is fixed and full rank. Using the hint, we have that

$$\begin{aligned}
L &= \sup_{\mathbf{w}} \left|\left|\nabla^2 \ell(\mathbf{w})\right|\right| \\
&= \sup_{\mathbf{w}} \left|\left|X^T S X\right|\right| \\
&\leq ||X||^2 \sup_{\mathbf{w}} ||S|| \quad \text{since } X \text{ is not dependent on } \mathbf{w} \text{ and } \left|\left|X^T\right|\right| = ||X|| \\
&\leq ||X||^2 \max_i \mu_i (1 - \mu_i) \\
&\leq \frac{1}{4} ||X||^2,
\end{aligned}$$

since $\mu(1 - \mu)$ is maximized at $\mu = \frac{1}{2}$.
We can achieve the supremum by taking $\mathbf{w} = \mathbf{0}$ as $\mu_i = \sigma(0) = \frac{1}{2}$, and in this case $S = \frac{1}{4} I$, and we have that $\left|\left|X^T S X\right|\right| = \frac{1}{4} ||X||^2$. Since $X$ is full rank, $||X|| > 0$, so $\nabla \ell(\mathbf{w})$ is $L$-Lipschitz continuous with $L = \frac{1}{4} ||X||^2$.

3. Problem 4.11 (a), (b), (d).

Formulate the following problems as LPs. Explain in detail the relation between the optimal solution of each problem and the solution of its equivalent LP.

(a) Minimize $||Ax - b||_\infty$ ($\ell_\infty$-norm approximation)
We can convert this to an LP as follows:

$$\begin{aligned}
&\min \quad t \\
\text{such that} \quad &a_i^T x - b_i \leq t, \quad i = 1, 2, \ldots, m \\
&-a_i^T x + b_i \leq t, \quad i = 1, 2, \ldots, m,
\end{aligned}$$

where $t \in \mathbb{R}$ and $x \in \mathbb{R}^n$. This is an application of the epigraph trick from Boyd and Vandenberghe's textbook.

The optimal solution to the original problem and its equivalent LP formulation will not be the same as the solution to the original problem is in $\mathbb{R}^n$, whereas the optimal solution to the equivalent LP formulation is in $\mathbb{R}^{n+1}$. However, the $x$ that satisfies the LP formulation will be the same $x$ that satisfies the original problem. Furthermore, the optimal objective values for both problems will be the same, e.g., if the LP formulation had a smaller objective value than the original problem, then it would be the case that we could use the optimal $x$ from the LP formulation to make $||Ax - b||_\infty$ smaller, and vice versa.

(b) Minimize $||Ax - b||_1$ ($\ell_1$-norm approximation)

Note that for any vector $x$, $||x||_1 = \sum_{i=1}^{n} |x_i|$. Using this, we can convert this to be an LP as follows:

$$\begin{aligned} \text{min} \quad & \mathbf{1^T t} \\ \text{such that} \quad & a_i^T x - b_i \leq t_i, \quad i = 1, 2, \ldots, m \\ & -a_i^T x + b_i \leq t_i, \quad i = 1, 2, \ldots, m, \end{aligned}$$

where $t \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$.

Note that if $t$ solves the LP formulation, the entries of $t$ will necessarily be non-negative, and furthermore, $||t|| = ||Ax - b||_1$ at the optimal $x$ that minimizes the $\ell^1$-norm.

(c) Minimize $||x||_1$ subject to $||Ax - b||_\infty \leq 1$.

We first write $x = x^+ - x^-$, where $x^+, x^- \geq 0$, and we can recast the original problem as an LP as follows:

$$\begin{aligned} \text{min} \quad & \mathbf{1^T x^+} + \mathbf{1^T x^-} \\ \text{such that} \quad & a_i^T x^+ - a_i^T x^- - b_i \leq 1 \quad i = 1, 2, \ldots, m \\ & -a_i^T x^+ + a_i^T x^- + b_i \leq 1, \quad i = 1, 2, \ldots, m \end{aligned}$$

Again, we use the epigraph trick covered in the textbook to rewrite the constraint. Also note that $||x||_1 \leq ||x^+||_1 + ||x^-||$ by the triangle inequality. However, since we are looking to minimize $||x||_1$, we can find a way to pick $x^+$ and $x^-$ such that $x = x^+ - x^-$ and $||x^+||_1 + ||x^-||_1 = ||x||_1$. For example, in the one-dimensional case, we can write any number as the difference of two non-negative numbers, e.g., $5 = 7 - 2$, and clearly $|5| \leq |7| + |-2|$, but we can rewrite $5 = 5 - 0$, in which case, we get equality of norms, as desired. Thus the optimal objective values of both problems will be the same, and so will the optimal solutions since we wrote $x$ in the form given above.

4. Brainstorm 3 possible project ideas (a title for each one), and write a few sentences with more detail on at least of one of these ideas. These ideas are not binding.

(a) LP relaxations of integer programming

Integer programming belongs to the subclass of combinatorial optimization problems, and is NP-hard. However, we can perform an LP relaxation on an integer programming problem and solve the relaxed problem in polynomial time. Although the relaxed problem will, in general, not have the same optimal solution as the original problem, we can solve the relaxed problem in polynomial time, e.g., via the simplex method or an interior point method, and gain information about the optimal solution to the original problem. Furthermore, we can refine our relaxed problem solutions using branch-and-bound techniques and cutting plane techniques. One idea would be to further explore these techniques and basically give a lecture on how these techniques are used in practice.

(b) Numerically comparing different deep learning optimization models

Deep learning problems are typically non-convex problems, and most researchers use heuristics and first-order methods that are basically "fancier" versions of gradient descent to train deep learning models.

One idea would be to explore why researchers use these methods to solve such problems and train models, despite these methods having been proven to not converge under certain conditions that can be encountered in practice.

(c) Bayesian optimization

If we are maximizing an unknown or expensive-to-evaluate function over a low-dimensional search space, we can use Bayesian optimization techniques to build a probabilistic model that helps us determine where to evaluate $f$ in order to build an understanding of it. Bayesian optimization techniques are used in training expensive-to-evaluate machine learning models that depend in complicated ways on hyerparameters, for example. One idea would be to explore this field and present a lecture on some these techniques.

5. List at least 2 potential partners (and their emails) for the project, and **at least one of these potential partners should be someone you did not know from before the class**. Make sure to get the partners' permission!

   Some potential partners for this project are Jack, Logan, and Tyler.

6. Read/skim (but do not solve) the problems from chapter 4. In particular, 4.12, 4.15, 4.23, 4.24, 4.26, 4.28, 4.40, 4.44/2.37, 4.45, 4.57, 4.59.