Honours Math for Machine Learning HW3

Saina Koukpari - McGill University

Problem 1.1

2.1) By the definition of convexity, the statement holds for $k = 2$. Assume $\exists k$ such that $\theta_1 + \cdots + \theta_k = 1$ for $\theta_i \geq 0$ and $\theta_1 x_1 + \cdots \theta_k x_k \in C$. Using the case $k = 2$ to prove for $k+1$ we have that $\exists \Theta_1, \Theta_2 \geq 0$ such that $\Theta_1 + \Theta_2 = 1$ and taking $\mathbf{X}_2 = \theta_1 x_1 + \cdots \theta_k x_k$,

$$\Theta_1 \mathbf{X}_1 + \Theta_2 \left(\theta_1 x_1 + \cdots \theta_k x_k\right) \in C$$

and $\mathbf{X}_1,\ \mathbf{X}_2 = \theta_1 x_2 + \cdots \theta_k x_k \in C$. Then we have $\Theta_1, \Theta_2, \theta_1, \ldots, \Theta_2\theta_k$ which are $k+1$ thetas such that $\Theta_1 + \Theta_2\theta_1 + \cdots + \Theta_2\theta_k = 1$ and all thetas are greater or equal to 0 $\implies$ in $\Theta_1, \Theta_2, \theta_1, \ldots, \Theta_2\theta_k \in C$. Thus the case for $k+1$ holds and our assumption for k is true. ■

2.5) Let $C = \{x \in \mathbf{R}^n | a^T x = b_1\}$ and $D = \{x \in \mathbf{R}^n | a^T x = b_2\}$. The distance between the two hyperplanes C and D is given by

$$\begin{aligned}
\mathbf{dist}(C, D) &= \inf\{\|x_1 - x_2\|_2 \mid x_1 \in C, x_2 \in D\} \\
&= \inf\{\|b_1/a^T - b_2/a^T\|_2\} \\
&= \left\|\frac{b_1 - b_2}{a^T}\right\|_2 \\
&= \frac{|b_1 - b_2|}{\|a^T\|_2}
\end{aligned}$$

2.12) A set C is convex if the line segment between any two points in C also lies in C. This is true for half spaces(Example 2.2.1 in Boyd), i.e. the set of half spaces is a convex set.

a) For $\alpha \in A, \beta \in B$ in half spaces A and B, a slab $\{x \in \mathbb{R}^n | \alpha \leq a^T, x \leq \beta\} = \{x \in \mathbb{R}^n | \alpha \leq a^T x \leq \beta\} = A \bigcap B \neq \varnothing$. Thus it is a convex set since the line segment between any points in the set lie in a half space.

b) For $\alpha_i \in A_i, \beta_i \in B_i$ in half spaces A_i and B_i, $i = 1, ..., n$, a rectangle is the set of finite (n) intersections of half spaces, so it is a convex set.

c) A wedge is a convex set since it is the intersection of two half spaces .

d) Given $\|x - x_0\| \leq \|x - y\|$, by properties of the norm, we have that $\|x - x_0\|^2 \leq \|x - y\|^2$. Then,

$$\begin{aligned}
&\|x - x_0\| \leq \|x - y\| \\
\implies &\|x - x_0\|^2 \leq \|x - y\|^2 \\
\implies &(x - x_0)^T (x - x_0) \leq (x - y)^T (x - y) \\
\implies &x^T x - 2x_0^T x + x_0^T x_0 \leq x^T x - 2y^T x + y^T y \\
\implies &2y^T x - 2x_0^T x \leq y^T y - x_0^T x_0 \\
\implies &2x(y - x_0)^T \leq y^T y - x_0^T x_0 \\
\implies &2(y - x_0)^T x \leq y^T y - x_0^T x_0
\end{aligned}$$

So for fixed y and x_0 the set $\{x \mid \|x - x_0\| \le \|x - y\| \forall y \in S\}$ is the intersection of half spaces, so it is a convex set.

Problem 2.1

i) Consider the piece-wise linear function(non-convex),

$$f(x) = \begin{cases} x & x \le 1 \\ 1 & x \ge 1 \end{cases}$$

Then the minimizer is found where the derivative of the function is equal to zero which in this case occurs when

$$f'(x) = \begin{cases} 1 & x \le 1 \\ 0 & x \ge 1 \end{cases} = 0$$

Then the minimum is either at 0 or 1, hence the function has more than one minimizer.

ii) Consider the function $f(x) = e^x$. This is a convex function on $\mathbb{R}$ by Boyd Examples 3.1.5. Taking the derivative to find the minimzer, we have $f'(x) = e^x = 0$ but $a^x = 0$ if and only if $a = 0$ and $e \ne 0$. Hence we have found a convex function on $\mathbb{R}$ which has no minimum value.

Problem 2.2

i) Since $\ell(h_w(x), y)$ is convex, we know that $h_w(x)$ is convex for all x. Then $h_w(X)$ is convex since we have a matrix composed of x where $h_w(x)$ is convex for all x. Thus by composition of affine mappings, $\ell(h_w(X), y)$ is convex.

ii) By properties of convexity, convexity is preserved under non negative scaling and addition for sums. Thus since $\ell(h_w(X_i), y_i)$ is convex as a function of x for every y (by part (i)), the empirical loss is also convex as a function of x.

Problem 3.1

To reduce the optimality gap by a factor of 10, we want our k-th gap to be less than our initial gap by a factor of 10, i.e. $f(x^k) - p \le \frac{1}{10}(f(x^0) - p)$. Then given our condition number 3, $c = 1 - m/M = 1 - 1/3 = 2/3$, we define the gradient descent rate as $f(x^k) - p \le c^k(f(x^0) - p)$. Thus the minimal amount of iterations required is given by k where $(2/3)^k \le 1/10 \implies k = 6$.

For condition number 100 we have, $c = 1 - 1/100 = 99/100$ and the number of iterations k is given by $(99/100)^k \le 1/10 \implies k = 230$.

Problem 4.1

1) Since the quadratic loss is convex, the gradient descent converges to a minimizer w^* where the minimizer satisfies $\nabla_w \hat{L}(w) = 0$. Then

$$\begin{aligned}
\nabla_w \hat{L}(w) &= \frac{1}{m} \sum_{i=1}^{m} \partial_h \ell(h_w(x_i), y_i) \nabla_w h_w(x_i) \\
&= \frac{1}{m} \sum_{i=1}^{m} (h_i - y_i) \nabla_w(w) \\
&= \frac{1}{m} \sum_{i=1}^{m} (w - y_i) \\
&= \frac{1}{m}(m \cdot w) - \frac{1}{m} \sum_{i=1}^{m} y_i \\
&= 0 \\
\implies w &= \frac{1}{m} \sum_{i=1}^{m} y_i = \overline{y}
\end{aligned}$$

2) $w\hat{L}'(w) = w \cdot \frac{1}{m}\sum_{i=1}^{m}(w - y_i) = w\left(\frac{1}{m}\sum_{i=1}^{m} w - \sum_{i=1}^{m} y_i\right) = w\left(\frac{1}{m}\sum_{i=1}^{m} w - \overline{y}\right)$ (by part (1)) $= w(w - \overline{y})$

3) The gradient descent is given by $w_{k+1} = w_k - h\nabla\hat{L}(w_k)$. Then for $h = 1$,

$$w_1 = w_0 - 1 \cdot \nabla\hat{L}(w_0) = w_0 - (w_0 - \overline{y}) = \overline{y} \quad (\text{ converges in one step })$$

4) For $h = 1/2$, $w_{k+1} = w_k - 1/2\nabla\hat{L}(w_k) = w_k - 1/2(w_k - \overline{y}) = 1/2(w_k + \overline{y})$. Then the rate satisfies,

$$\begin{aligned} w_k - \overline{y} &\le (1/2)^k (w_0 - \overline{y}) \\ \implies |w_k - \overline{y}| &\le (1/2)^k |w_0 - \overline{y}| \end{aligned}$$

Problem 4.2

For different values of w, we find

$$\begin{aligned} m\hat{L}'(w) = 6\hat{L}(w) = &= 6\frac{d}{dw}\frac{1}{6}\sum_{i=1}^{6} \ell_H(h, y) \\ &= \sum_{i=1}^{6} \begin{cases} w - y_i & |w - y_i| \le 1 \\ sgn(w - y_i) & |w - y_i| \ge 1 \end{cases} \end{aligned}$$

Taking $w_0 = -3$ with learning rate $h = 6$, we have the following three iterations of gradient descent,

$$\begin{aligned} w_1 &= w_0 - h\nabla_w\hat{L}(w_0) \\ &= -3 - 6\nabla_w\hat{L}(-3) \\ &= -3 - ((-3 - (-3)) + (-3 - (-2)) \\ &+ (sgn(-3 - (-0.3))) + (sgn(-3 - 0.4)) + (sgn(-3 - 1.5)) + (sgn(-3 - 4))) \\ &= -3 - (0 + 1 - 1 - 1 - 1 - 1) \\ &= -3 - (-3) \\ &= 0 \\ w_2 &= w_1 - h\nabla_w\hat{L}(w_1) \\ &= 0 - 6\nabla_w\hat{L}(0) \\ &= -(sgn(0 - (-3)) + sgn(0 - (-2)) \\ &+ (0 - (-0.3)) + (0 - 0.4) + (sgn(0 - 1.5)) + (sgn(0 - 4))) \\ &= -(1 + 1 + 0.3 - 0.4 - 1 - 1) \\ &= -(-0.1) \\ &= 0.1 \\ w_3 &= w_2 - h\nabla_w\hat{L}(w_2) \\ &= 0.1 - 6\nabla_w\hat{L}(0.1) \\ &= 0.1 - (sgn(0.1 - (-3)) + sgn(0.1 - (-2)) \\ &+ (0.1 - (-0.3)) + (0.1 - 0.4) + sgn(0.1 - 1.5)) + sgn(0.1 - 4)) \\ &= 0.1 - (1 + 1 + 0.4 - 0.3 - 1 - 1) \\ &= 0.1 - (0.1) \\ &= 0 \end{aligned}$$

Problem 4.3

As in the section, we consider (EL) with linear model $h_w(x) = w\cdot x$, linear data $y(x) = w^*\cdot x$ and quadratic loss $\ell(h_w(x), y) = (w\cdot x - w^*\cdot x)^2/2$, $\partial_h \ell(h_w(x), y) = h - y = w\cdot x - w^*\cdot x$. Then,

$$\begin{aligned}
\nabla \hat{L}(w) &= \frac{1}{m}\sum_{i=1}^{m} \partial_h \ell(h_w(x_i), y_i)\nabla_w h_w(x_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} \partial_h \ell(h_w(x_i), w^*\cdot x_i)(x_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} (w\cdot x_i - w^*\cdot x_i)(x_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} (w - w^*)\cdot x_i x_i \quad \square
\end{aligned}$$

Then as a matrix equation,

$$\begin{aligned}
\nabla \hat{L}(w) &= \frac{1}{m}\sum_{i=1}^{m} (w - w^*)\cdot x_i x_i \\
&= X^T X w - X^T X w^* \\
&= X^T X(w - w^*) \\
&= H(w - w^*)
\end{aligned}$$

where the expression for the coefficients is given by $H = X^T X$ since $H(w - w^*) = X^T X(w - w^*)$

In the case where $m = 4$ and $x_i = (1, i)$ for $i = 1, 2, 3, 4$ we have H as follows,

$$\begin{aligned}
H &= X^T X \\
&= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \\
&= \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}
\end{aligned}$$