Solutions to Homework 6 (graded)
CS-526 Learning Theory

Exercise 1

1. $f(x) = \max_{1\leq i\leq m} f_i(x)$ where $f_i(x) = a_i^T x + b_i$ is convex differentiable with gradient $\nabla f_i(x) = a_i$. By Claim 14.6, it follows that $\forall x : a_j \in \partial f(x)$ where $j \in \arg\max_i f_i(x)$.

2. $f(x) = \max_{1\leq i\leq m} f_i(x)$ where $f_i(x) = |a_i^T x + b_i|$ is convex subdifferentiable. Fix x, let $j \in \arg\max_i f_i(x)$ and choose $v \in \partial f_j(x)$ as follows:

$$v = \begin{cases} -a_j & \text{if } a_j^T x + b_j < 0\,, \\ 0 & \text{if } a_i^T x + b_i = 0\,, \\ +a_j & \text{if } a_j^T x + b_j > 0\,. \end{cases}$$

A straightforward generalization of Claim 14.6 shows that v is a subgradient of f at x.

3. Note that the sup is really a maximum as $t \mapsto p(t,x)$ is a continuous function on a compact. Hence $f(x) = \max_{t\in[0,1]} p(t,x)$ and $\forall t \in [0,1] : \nabla_x p(t,x) = [1, t, \ldots, t^{n-1}]^T \in \mathbb{R}^n$. A straightforward generalization of Claim 14.6 shows that $[1, t(x), \ldots, t(x)^{n-1}]^T \in \partial f(x)$, where $t(x) \in \arg\max_{t\in[0,1]} p(t,x)$.

Exercise 2

Fix $\mathbf{w}$, $\mathbf{u}$. The function f is λ-strongly convex, so for all $\alpha \in [0,1]$ we have:

$$\begin{aligned} & f((1-\alpha)\mathbf{w} + \alpha\mathbf{u}) \leq (1-\alpha)f(\mathbf{w}) + \alpha f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 \\ \Leftrightarrow \quad & f(\mathbf{w} + \alpha(\mathbf{u}-\mathbf{w})) - f(\mathbf{w}) \leq \alpha\Big(f(\mathbf{u}) - f(\mathbf{w}) - \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\Big) \qquad (1) \end{aligned}$$

Let $\mathbf{v} \in \partial f(\mathbf{w})$. Then, $\forall \alpha \in [0,1] : f(\mathbf{w} + \alpha(\mathbf{u}-\mathbf{w})) \geq f(\mathbf{w}) + \langle \alpha(\mathbf{u}-\mathbf{w}), \mathbf{v}\rangle$. Combining this inequality and (1) gives:

$$\begin{aligned} & \langle \alpha(\mathbf{u}-\mathbf{w}), \mathbf{v}\rangle \leq \alpha\Big(f(\mathbf{u}) - f(\mathbf{w}) - \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\Big) \\ \Leftrightarrow \quad & \langle \mathbf{u}-\mathbf{w}, \mathbf{v}\rangle \leq f(\mathbf{u}) - f(\mathbf{w}) - \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 \\ \Leftrightarrow \quad & \langle \mathbf{w}-\mathbf{u}, \mathbf{v}\rangle \geq f(\mathbf{w}) - f(\mathbf{u}) + \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 \end{aligned}$$

Taking the limit $\alpha \to 0+$ ends the proof: $\langle \mathbf{w}-\mathbf{u}, \mathbf{v}\rangle \geq f(\mathbf{w}) - f(\mathbf{u}) + \frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2$.

Exercise 3

a) Fix $A, B \in \mathcal{S}_n^+$ and $\alpha \in [0,1]$. Let $\mathbf{e} \in \mathbb{R}^n$ a unit-norm eigenvector of $\alpha A + (1-\alpha)B$ associated to the maximum eigenvalue, i.e., $(\alpha A + (1-\alpha)B)\mathbf{e} = \lambda_{\max}(\alpha A + (1-\alpha)B)\mathbf{e}$ and $\|\mathbf{e}\| = 1$. We have:

$$\begin{aligned} f(\alpha A + (1-\alpha)B) = \mathbf{e}^T(\alpha A + (1-\alpha)B)\mathbf{e} &= \alpha \mathbf{e}^T A \mathbf{e} + (1-\alpha)\mathbf{e}^T B \mathbf{e} \\ &\leq \alpha \lambda_{\max}(A) + (1-\alpha)\lambda_{\max}(B) \\ &= \alpha f(A) + (1-\alpha) f(B)\,. \end{aligned}$$

This shows that f is convex.
b) Let $A \in \mathcal{S}_n^+$. A subgradient of f at A is a matrix $V \in \mathbb{R}^{n \times n}$ that satisfies:

$$\forall B \in \mathcal{S}_n^+ : f(B) \geq f(A) + \mathrm{Tr}\big((B-A)^T V\big)\,.$$

Consider any $\mathbf{e} \in \mathbb{R}^n$ which is a unit-norm eigenvector of A associated to the maximum eigenvalue, i.e., $A\mathbf{e} = \lambda_{\max}(A)\mathbf{e}$ and $\|\mathbf{e}\| = 1$. Then for all $B \in \mathcal{S}_n^+$:

$$\begin{aligned} f(A) = \lambda_{\max}(A) = \mathbf{e}^T A \mathbf{e} = \mathbf{e}^T B \mathbf{e} + \mathbf{e}^T(A-B)\mathbf{e} &\leq \lambda_{\max}(B) + \mathbf{e}^T(A-B)\mathbf{e} \\ &= f(B) + \mathrm{Tr}(\mathbf{e}^T(A-B)\mathbf{e}) \\ &= f(B) + \mathrm{Tr}((A-B)^T \mathbf{e}\mathbf{e}^T)\,. \end{aligned}$$

In the last equality we used that $(A-B)^T = A - B$ and that the trace is preserved by cyclic permutations. We see that $\mathbf{e}\mathbf{e}^T$ satisfies the definition of a subgradient: $\mathbf{e}\mathbf{e}^T \in \partial f(A)$.

Exercise 4

a) Assume that A has the singular value decomposition $U\Lambda V^T$. Plugging this into the expression $I - \alpha A^T A$ we see that $I - \alpha A^T A$ has the singular value decomposition $V\Lambda' V^T$, where Λ' is of dimension $n \times n$ and has the singular values $1 - \alpha\sigma_i^2$. For the given choice of α all these singular values are non-negative and the largest is $1 - \alpha\sigma_{\min}^2(A) = 1 - \frac{\sigma_{\min}^2(A)}{\sigma_{\max}^2(A)}$.
b) We get

$$\nabla f(\mathbf{x}) = A^T(A\mathbf{x} - \mathbf{b}) = A^T A(\mathbf{x} - \mathbf{x}^*),$$

where we used the fact that A has full column rank so that $A\mathbf{x}^* = b$. Hence GD can be rewritten as

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha A^T A(\mathbf{x}^t - \mathbf{x}^*). \tag{2}$$

c) Subtracting $\mathbf{x}^*$ from both sides of (2) gives

$$\mathbf{x}^{t+1} - \mathbf{x}^* = \mathbf{x}^t - \mathbf{x}^* - \alpha A^T A(\mathbf{x}^t - \mathbf{x}^*) = (I - \alpha A^T A)(\mathbf{x}^t - \mathbf{x}^*).$$

By taking norms we obtain

$$\begin{aligned} ||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 &\leq \sigma_{\max}(I - \alpha A^T A)||\mathbf{x}^t - \mathbf{x}^*||_2 \\ &= (1 - \alpha\sigma_{min}(A)^2)||\mathbf{x}^t - \mathbf{x}^*||_2. \end{aligned}$$

Exercise 5

1. From the Taylor's theorem, we have

$$L(\beta') = L(\beta) + (\beta' - \beta)^T \nabla L(\beta) + \frac{1}{2}(\beta' - \beta)^T \text{Hessian}(L)(\xi\beta + (1-\xi)\beta)(\beta' - \beta),$$

for some $\xi \in [0, 1]$. We can rewrite L as

$$L(\beta) = \frac{1}{n}\|X\beta - y\|^2,$$

which gives $\text{Hessian}(L)(\beta) = \frac{2}{n}X^T X$. Hence, we have

$$(\beta' - \beta)^T \text{Hessian}(L)(\xi\beta + (1-\xi)\beta)(\beta' - \beta) \geq \frac{2\mu}{n}\|\beta' - \beta\|^2,$$

giving the desired result.

2. Expanding $\|\beta_{k+1} - \beta^*\|^2$, we have

$$\begin{aligned}\|\beta_{k+1} - \beta^*\|^2 &= \|\beta_k - \eta\nabla\ell(\beta_k, x_{i_k}, y_{i_k}) - \beta^*\|^2 \\ &= \|\beta_k - \beta^*\|^2 - 2\eta\langle\beta_k - \beta^*, \nabla\ell(\beta_k, x_{i_k}, y_{i_k})\rangle + \eta^2\|\nabla\ell(\beta_k, x_{i_k}, y_{i_k})\|^2\end{aligned}$$

Let $\mathbb{E}_k$ denote expectation conditioned on the randomness till step k. We have

$$\begin{aligned}\mathbb{E}_k\|\beta_{k+1} - \beta^*\|^2 &= \|\beta_k - \beta^*\|^2 - 2\eta\langle\beta_k - \beta^*, \mathbb{E}_k\nabla\ell(\beta_k, x_{i_k}, y_{i_k})\rangle + \eta^2\mathbb{E}_k\|\nabla\ell(\beta_k, x_{i_k}, y_{i_k})\|^2 \\ &= \|\beta_k - \beta^*\|^2 - 2\eta\langle\beta_k - \beta^*, \nabla L(\beta_k)\rangle + 4\eta^2\mathbb{E}_k[(x_{i_k}^T\beta_k - y_{i_k})^2\|x_{i_k}\|^2]\end{aligned}$$

The result from the previous question with $\beta = \beta_k, \beta' = \beta^*$ gives

$$\langle\beta_k - \beta^*, \nabla L(\beta_k)\rangle \geq L(\beta_k) + \frac{\mu}{n}\|\beta_k - \beta^*\|^2.$$

Hence,

$$\mathbb{E}_k\|\beta_{k+1} - \beta^*\|^2 \leq \|\beta_k - \beta^*\|^2 - 2\eta L(\beta_k) - \frac{2\eta\mu}{n}\|\beta_k - \beta^*\|^2 + 4\eta^2\mathbb{E}_k[(x_{i_k}^T\beta_k - y_{i_k})^2\|x_{i_k}\|^2]$$

Using the fact that $\|x_{i_k}\|^2 = 1$, we get

$$\begin{aligned}\mathbb{E}_k\|\beta_{k+1} - \beta^*\|^2 &\leq \|\beta_k - \beta^*\|^2 - 2\eta L(\beta_k) - \frac{2\eta\mu}{n}\|\beta_k - \beta^*\|^2 + 4\eta^2 L(\beta_k) \\ &= \left(1 - \frac{2\eta\mu}{n}\right)\|\beta_k - \beta^*\|^2 - 2(\eta - 2\eta^2)L(\beta_k), \\ &\leq \left(1 - \frac{2\eta\mu}{n}\right)\|\beta_k - \beta^*\|^2, \quad \text{for } \eta \leq \frac{1}{2}.\end{aligned}$$

Taking expectation, we get

$$\mathbb{E}\|\beta_{k+1} - \beta^*\|^2 \leq \left(1 - \frac{2\eta\mu}{n}\right)\mathbb{E}\|\beta_k - \beta^*\|^2.$$

Now, recursively applying the above result, we get

$$\mathbb{E}\|\beta_k - \beta^*\|^2 \leq \left(1 - \frac{2\eta\mu}{n}\right)^k \mathbb{E}\|\beta_0 - \beta^*\|^2.$$

3. Here the convergence is guaranteed for the iterates (β_k) itself without averaging, and is exponentially fast. The convergence result derived in class is for function value at the average of iterates, and the convergence is polynomial in the number of iterations.