Solution 7
CS-526 Learning Theory

1. Short problems

1. **A**, **C** and **D**. The first sum is the classical cross entropy loss in a logistic regression problem. We can check that this first sum is convex (nonnegative second derivative) and Lipschitzian (bounded first derivative). These properties remain when summing the regularization term.

2. (a) For $x \geq 0$, $f''(x) = 2 - 2.5\cos(x)$, which is negative for some values of $x \geq 0$. Hence the function is not convex.
   (b) f is not differentiable at $x = 0$ due to the term $|x|$.
   (c) This is a little tricky. The function has a derivative everywhere except at 0 where it has a subderivative. But it is not convex and hence not subdifferentiable everywhere.

3. We have $\nabla\|x\|^\alpha = \alpha\|x\|^{(\alpha-1)}\frac{x}{\|x\|}$. Therefore $\nabla h_\alpha(x) = \alpha\|x\|^{(\alpha-1)}\frac{x}{\|x\|}g'(\|x\|^\alpha)$ and

$$\|\nabla h_\alpha(x)\| = \alpha\|x\|^{(\alpha-1)}|g'(\|x\|^\alpha)| \leq \alpha\rho\|x\|^{(\alpha-1)}$$

   So $h_{\alpha=1}$ is a Lipschitz function with constant ρ. For $\alpha > 1$ the equality shows that $\|\nabla h_\alpha(x)\|$ is not bounded so we dont have a Lipschitz function. For $\alpha < 1$ $\|\nabla h_\alpha(x)\|$ is unbounded when $x \to 0$ unless we assume that g vanishes fast enough at the origin so we dont have a Lipschitz constant.

4. The function $|x|^3$ is convex (but not strictly convex) and this can be seen by computing the second derivative. Also $|x|$ is convex. The sum of convex fuunctions is convex therefore the whole function is convex. For $x \neq 0$ the subgradient is just the derivative $3ax^2 + b\text{sgn}(x)$. For $x = 0$ the subgradient is $[-b, +b]$.

5. Small calculation shows that (a) is true.

2. Gradient Descent for Positive Semi-definite Matrices

1. Use the spectral decomposition $B = \sum_{j=1}^n \lambda_j u_j u_j^T$ and since B is positive definite all $\lambda_j > 0$ (and we can take eigenvectors with real components). Then

$$\begin{aligned} F(X) &= \sum_{j=1}^n \lambda_j \text{Tr} X^T u_j u_j^T X = \sum_{j=1}^n \lambda_j \text{Tr}(X^T u_j)(X^T u_j)^T \\ &= \sum_{j=1}^n \lambda_j (X^T u_j)^T (X^T u_j) = \sum_{j=1}^n \lambda_j \|X^T u_j\|^2 \geq 0 \end{aligned}$$

   since $\lambda_j > 0$ for all j.

2. We find

$$\begin{aligned} f''(s) &= 2\mathrm{Tr}X^T BX + 2\mathrm{Tr}Y^T BY - \mathrm{Tr}X^T BY - \mathrm{Tr}Y^T BX \\ &= 2\mathrm{Tr}(X-Y)^T B(X-Y) \geq 0 \end{aligned}$$

Thus f is convex. Since $f(s) = f((1-s).0 + s.1)$ we have $f(s) \leq (1-s)f(0) + sf(1)$. This inequality reads

$$F((sX + (1-s)Y) \leq sF(X) + (1-s)F(Y)$$

3. The gradient of $F(X)$ is the matrix

$$\nabla_X F(X) = BX$$

This can be computed using components $\frac{\partial}{\partial X_{ij}} F(X)$. Since F is convex it is above its tangent and this shows (see class)

$$F(Y) - F(X) \geq \langle \nabla_X F(X), Y - X \rangle = \mathrm{Tr}(BX)^T (Y - X)$$

Note the last result can also be found working with components.

The function is not Lipschitz because the gradient BX is not bounded (locally it is Lipschitz but we did not talk about this in class).

4. For L the gradient is $\nabla L(X) = BX + AX - A$. The gradient descent algorithm is as follows: initialize with X_1 and for $t = 1, \cdots, T$ do

$$X_{t+1} = X_t - \eta(BX_t + AX_t - A)$$

Summing over $t = 1, \cdots, T$ we get

$$\frac{1}{T}(X_{T+1} - X_1) = -\eta((B+A)\frac{1}{T}\sum_{t=1}^{T} X_t - A)$$

Since we assume $\|X_t\| \leq M$ uniformly in t, we can use $\|X_1\| \leq M$ and $\|X_{T+1}\| \leq M$ to get

$$\|\frac{1}{T}\sum_{t=1}^{T} X_t - (B+A)^{-1}A\| \leq \frac{2M}{\eta T}\|(B+A)^{-1}\|$$

3. Variant of standard gradient descent; forward and backward schemes

1. The backward Euler scheme is

$$x^{t+1} = x^t + S^{-1}\nabla f(x^{t+1}).$$

2. The first term f is convex. The second term is strictly convex because S is positive definite (with $\lambda_{min} > 0$). Thus the sum is strictly convex.

   Since f is differentiable we can differentiate the gradient of the quantity in the bracket in order to find the argmin:

$$\nabla f(x) + \eta^{-1} S(x - x^t) = 0$$

   which implies the backward Euler scheme:

$$x^{t+1} = x^t - \eta S^{-1} \nabla f(x^{t+1})$$

3. Let $S^{-1} = U^T \Lambda^{-1} U$ with U an orthogonal matrix, and $\Lambda = \text{Diag}\,(\lambda_1 \cdots \lambda_d)$. With $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x^t$, we have

$$\begin{aligned}
f(\bar{x}) - f\left(x^*\right) &\leqslant \frac{1}{T} \sum_{t=1}^{T} \left(f\left(x^t\right) - f\left(x^*\right) \right) \qquad \text{convexity} \\
&\leqslant \frac{1}{T} \sum_{t=1}^{T} \left\langle \nabla f\left(x^t\right), x^t - x^* \right\rangle \qquad \text{convexity} \\
&= \frac{1}{T} \sum_{t=1}^{T} \left\langle U \nabla f\left(x^t\right), U x^t - U x^* \right\rangle \\
&= \sum_{k=1}^{d} \frac{1}{T} \sum_{t=1}^{T} (U \nabla f)_k (x^t) \left(U\left(x^t - x^*\right) \right)_k \\
&= \sum_{k=1}^{d} \frac{\lambda_k}{\eta T} \sum_{t=1}^{T} \left(\frac{\eta}{\lambda_k} \right) (U \nabla f)_k (x^t) \left(U\left(x^t - x^*\right) \right)_k \\
= \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \sum_{t=1}^{T} \Bigg\{ &- \left(\left(U\left(x^t - x^*\right) \right)_k - \frac{\eta}{\lambda_k} (U \nabla f)_k (x^t) \right)^2 + \left(U\left(x^t - x^*\right) \right)_k^2 + \frac{\eta^2}{\lambda_k^2} \left(U \nabla_f \right)_k (x^t)^2 \Bigg\}
\end{aligned}$$

   Now, from the backward equation we have:

$$\begin{aligned}
& x^{t+1} = x^t - \eta U^T \Lambda^{-1} U \nabla (x^t) \\
& \Rightarrow U x^{t+1} = U x^t - \eta \Lambda^{-1} U \nabla f(x^t) \\
& \left(U x^{t+1} \right)_k = \left(U x^t \right)_k - \frac{\eta}{\lambda_k} (U \nabla f)_k (x^t)
\end{aligned}$$

From which we get

$$
\begin{aligned}
f(\bar{x}) - f\left(x^*\right) &\leq \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \sum_{t=1}^{T} \left\{ -\left(U\left(x^{t+1} - x^*\right)\right)_k^2 + \left(U\left(x^t - x^*\right)\right)_k^2 + \frac{\eta^2}{\lambda_k^2} \left(U\nabla_f\right)_k \left(x^t\right)^2 \right\} \\
&= \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \left[\left(U\left(x^1 - x^*\right)\right)_k^2 - \left(U\left(x^{T+1} - x^*\right)\right)_k^2 \right] + \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \sum_{t=1}^{T} \frac{\eta^2}{\lambda_k^2} (U\nabla f)_k (x^t)^2 \\
&\leq \frac{\lambda_{\max}}{2\eta T} \sum_{k=1}^{d} \left(U\left(x^1 - x^*\right)\right)_k^2 + \frac{\eta}{2T\lambda_{\min}} \sum_{t=1}^{T} \|U\nabla f\|^2 \\
&= \frac{\lambda_{\max}}{2\eta T} \left\|U\left(x^1 - x^*\right)\right\|^2 + \frac{\eta}{2\lambda_{\min}} \|\nabla f\|^2 \\
&\leq \frac{\lambda_{\max}}{2\eta T} R^2 + \frac{\eta}{2\lambda_{\min}} \rho^2
\end{aligned}
$$

where we used that $x^1 = 0$ and $\|x^*\|^2 \leq R^2$ (by assumption) in the last inequality. Set

$$
\eta^2 = \frac{\lambda_{\max}\lambda_{\min} R^2}{\rho^2 T}
$$

Then, we find:

$$
\begin{aligned}
f(\bar{x}) - f\left(x^*\right) &\leq \frac{\lambda_{\max} R^2 \rho \sqrt{T}}{2\sqrt{\lambda_{\max}\lambda_{\min}} RT} + \frac{\sqrt{\lambda_{\max}\lambda_{\min}} R}{\rho\sqrt{T}} \frac{\rho^2}{2\lambda_{\min}} \\
&= \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{2\sqrt{T}} + \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{2\sqrt{T}} \\
&= \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{\sqrt{T}}
\end{aligned}
$$