Solution to Exercises on Gradient Descent and Stochastic Gradient Descent
(Problems are from previous years' exams)
CS-526 Learning Theory

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. Since you have so much data the mean field model is likely going to predict your performance correctly. The NTK applies even for a fixed amount of data but it only applies in the setting of vanishing learning rate (and the width of the network should be large). Since you have so much data the basic generalization bound will also tell you that if you chose according to the empirical mean you will likely choose close to an optimal hypothesis. But you still need to argue that the stochastic gradient will in fact do well in this scenario.

3. The main problem is that there is no easy and efficient way to compute the solution of the associated differential equation. In fact, solving such types of differential equations is typically done by running stochastic gradient descent! :-) So this framework can be used to discuss convergence and other theoretical questions but currently cannot be used to predict the performance or to optimize the parameters of the system.

4. (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.

5. 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.

A Conservation Law For Neural Networks

1.

$$\begin{aligned}\frac{d}{dt}(\frac{1}{2}\|\dot{\boldsymbol{w}}(t)\|^2 + L(\boldsymbol{w}(t))) &= \langle \dot{\boldsymbol{w}}(t), \ddot{\boldsymbol{w}}(t)\rangle + \langle \nabla L(\boldsymbol{w}(t)), \dot{\boldsymbol{w}}(t)\rangle \\ &= \langle \underbrace{\ddot{\boldsymbol{w}}(t) + \nabla L(\boldsymbol{w}(t))}_{=0}, \dot{\boldsymbol{w}}(t)\rangle = 0\end{aligned}$$

2.

$$\begin{aligned}\frac{d}{dt}A &= \dot{\boldsymbol{w}}\dot{\boldsymbol{w}}^T + \boldsymbol{w}\ddot{\boldsymbol{w}}^T - \ddot{\boldsymbol{w}}\boldsymbol{w}^T - \dot{\boldsymbol{w}}\dot{\boldsymbol{w}}^T \\ &= -\boldsymbol{w}(\nabla f(\|\boldsymbol{w}\|))^T + \nabla f(\|\boldsymbol{w}\|)\boldsymbol{w}^T \\ &= -\frac{\boldsymbol{w}\boldsymbol{w}^T}{\|w\|}f'(\|\boldsymbol{w}\|) + \frac{\boldsymbol{w}\boldsymbol{w}^T}{\|w\|}f'(\|\boldsymbol{w}\|) = 0,\end{aligned}$$

where in the last line we use $\nabla\|\boldsymbol{w}\| = \frac{\boldsymbol{w}}{\|w\|}$ and so $\nabla f(\|\boldsymbol{w}\|) = \frac{\boldsymbol{w}}{\|w\|}f'(\|\boldsymbol{w}\|)$.

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

$$F(X) = \sum_{j=1}^{n} \lambda_j \mathrm{Tr} X^T u_j u_j^T X = \sum_{j=1}^{n} \lambda_j \mathrm{Tr}(X^T u_j)(X^T u_j)^T$$
$$= \sum_{j=1}^{n} \lambda_j (X^T u_j)^T (X^T u_j) = \sum_{j=1}^{n} \lambda_j \|X^T u_j\|^2 \geq 0$$

   since $\lambda_j > 0$ for all j.

2. We find

$$f^{''}(s) = 2\mathrm{Tr} X^T B X + 2\mathrm{Tr} Y^T B Y - \mathrm{Tr} X^T B Y - \mathrm{Tr} Y^T B X$$
$$= 2\mathrm{Tr}(X-Y)^T B (X-Y) \geq 0$$

   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

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