CS F320: Foundations of Data Science

COMPREHENSIVE EXAMINATION (40 points, 180 minutes)
May 20, 2 PM − 5 PM
IC: Snehanshu Saha

Handwritten/printed notes and calculator are allowed. Laptop, mobile phone and any other form of electronic gadget is NOT allowed. Any violation will be interpreted as unfair means and disciplinary action will be taken. Clear and concise arguments backed by unambiguous mathematics will be considered for FULL CREDIT.

1. (8 points)
   Recall,

   (a) A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be $\mu-$strongly convex for $\mu \geq 0$ if $\forall x, y \in \mathbb{R}^n, t \in [0, 1]$ we have that

   $$f(ty - (1-t)x) \leq tf(y) + (1-t)f(x) - \frac{\mu}{2}t(1-t)||x-y||_2^2$$

   (b) A differentiable function is $\mu-$strongly convex iff $\forall x, y \in \mathbb{R}^n$

   $$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}||x-y||_2^2$$

   (c) A twice differentiable function is $\mu-$strongly convex iff $\forall x \in \mathbb{R}^n$

   $$z^T\nabla^2 f(x) \geq \mu||z||_2^2$$

   (d) if $f : \mathbb{R}^n \to \mathbb{R}$ is smooth and $\mu-$strongly convex for $\mu > 0$ then $\forall x^* \in X^*(f)$ and $x^* \in \mathbb{R}^n$

   $$\frac{1}{2\mu}||\nabla f(x)||_2^2 \geq f(x) - f^* \geq \frac{\mu}{2}||x - x^*||_2^2$$

   Consider x^* as a singleton optima or a set of optimal points in the gradient landscape. Let us consider a strongly convex gradient descent approach. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $L-$smooth, $\mu-$strongly convex function for $\mu > 0$. Then for $x_0 \in \mathbb{R}^n$, let $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \forall k \geq 0$. Then, show that $f(x_k) - f^* \leq (1 - \frac{\mu}{L})^k(f(x_0) - f*)$ and consequently, we require $\frac{L}{\mu}log(\frac{f(x_0)-f^*}{\epsilon})$ iterations to find an $\epsilon-$optimal point.
   (You need to first show that $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}||\nabla f(x_k)||_2^2$ and use the fact that $1 + x \leq e^x$)

2. (5 points)
   Recall the definition of $L-$smoothness. Let $f : \mathbb{R}^n \to \mathbb{R}$ be $L-$smooth. Then $\forall x, y \in \mathbb{R}^n$ show that

$$f(y) - (f(x) + \nabla f(x)^T (y - x)) \leq \frac{L}{2}||x - y||_2^2$$

3. (7 points)
   Recall, if f is strongly convex with Lipschitz gradient L, then $\forall x, y \in \mathbb{R}^n$, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}||x - y||_2^2 + \frac{1}{\mu + L}||\nabla f(x) - \nabla f(y)||_2^2$$

   Suppose, $0 < h \leq \frac{2}{\mu+L}$ and f is strongly convex with Lipschitz gradient. Then show that the gradient descent algorithm satisfies

$$||x_k - x^*||^2 \leq (1 - \frac{2h\mu L}{\mu + L})^k ||x_0 - x^*||^2$$

4. (7 points)
   Let us continue from problem (3). If $h = \frac{2}{\mu+L}$, then show that

$$||x_k - x^*|| \leq (\frac{\kappa - 1}{\kappa + 1})^k ||x_0 - x^*||$$

   and

$$f(x_k) - f^* \leq \frac{L}{2}(\frac{\kappa - 1}{\kappa + 1})^{2k} ||x_0 - x^*||^2$$

   where κ is the condition number.

5. (4 points) [Computing Lipschitz constant for the regression problem]
   Recall, a function defined on Ω is Lipschitz if

$$||f(x) - f(y)|| \leq L||x - y||, \forall x, y \in \Omega$$

   .
   Equivalently, $|\frac{f(x)-f(y)}{x-y}| \leq L$ where L is the Lipschitz constant. Also, recall that, elsewhere, we conjectured that finding the Lipschitz constant is particularly useful as it helps accelerate (convex) gradient descent. Now, recall that the least square cost function

$$g(w) = \frac{1}{2m}\sum_{i=1}^{m}(x^{(i)}w - y^{(i)})^2$$

   for linear regression.
   Show that its Lipschitz constant is bounded by

$$\frac{k}{m}||x^T x|| + \frac{1}{m}||y^T x||$$

   where $||w||, ||v|| \leq k$ and m is the number of training examples. Recall, we conjectured that the inverse of the Lipschitz constant is the learning rate which helps in faster convergence of the gradient descent process.

6. (4 points) [Convex everywhere smooth cost functions in regression] Convexity of the loss is a desirable property because convex functions have a unique global minima and are much more well behaved (convexity implies local lipschitzness). The latter is significant in ensuring a predictable, smooth trajectory of the optimizer while navigating through the generated loss landscape.
Show that $logcosh(x)$ is convex i.e. $J = \sum_{i=1}^{m} logcosh(y_i - v^T x_i)$ is convex. This establishes $logcosh(x)$ as a viable alternative to the least square cost function for regression!

7. (3+2 points)

   (a) Use SVD to show the following identity for any $n \times d$ matrix D

   $$(\lambda I_d + D^T D)^{-1} D^T = D^T (\lambda I_n + DD^T)^{-1}$$

   Note, Σ is an $n \times d$ diagonal matrix and $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$ are diagonal matrices with $\sigma_i i^2$ on the diagonal.

   (b) Let D be a $n \times d$ matrix and $\overline{y}$ be a $n \times 1$ column vector containing the dependent variables of linear regression. The **Tikhonov** regularization solution to the linear regression predicts the dependent variables of a test instance $\overline{z}$: using the following equation:

   $$\text{Prediction}(\overline{z}) = \overline{zw} = \overline{z}(D^T D + \lambda I)^{-1} D^T \overline{y}$$

   Here, the vectors $\overline{z}$ and $\overline{w}$ are treated as $1 \times d$ and $d \times n$ matrices respectively. Show using the result of $7(a)$, how you can write the prediction as given above purely in terms of similarities between training points or between $\overline{z}$ and training points. Note that $\overline{z}D$ is a row vector containing the dot product between the ______ (train?) and ______ (test?) instances and DD^T contains similarities between pairs of training instances.