Computational Learning Theory Comprehensive Examination (35 points)
27/12/2022—— 9 AM-12 PM—— CS F453
Instructor: Snehanshu Saha Name: ______________________

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. Read all directions carefully and write your answers in the space provided.

ANSWER ALL QUESTIONS

NOTE: Use supplementary pages provided to do the scratch work.

1. (10 points) Mark the following statements as True (**T**) or False (**T**). No explanation is necessary. Each question carries 1 point.
   (i) The update rule for the least mean squares $h_\theta(x) = \theta^T x$ is $\theta_k \leftarrow \theta_k + (h_\theta(x^i) - y^i)$
   (ii) The update rule for perceptron $h_\theta(x) = \text{sign}(\theta^T x)$ is $\theta_k \leftarrow \theta_k + \lambda(h_\theta(x^i) - y^i)x_k^i$
   (iii) Say, a linear regression model with the mean square loss fits a data set with reasonably good accuracy. We add 3 outlier points to the data. The loss function function needs to be changed to the mean absolute error loss.
   (iv) the Least Mean Square training rule does not perform a gradient descent to minimize the cost/error function
   (v) Lasso can be interpreted as least-squares linear regression where weights are regularized with the L_1 norm
   (vi) Perceptron can achieve zero training error on any linearly separable dataset
   (vii) A typical $k-$Means clustering algorithm minimizes a loss function over k clusters, sample points $x_1, ..., x_n$, and centers $\mu_1, ..., \mu_k$. This is a typical batch gradient descent framework for updating the cluster means.
   (viii) For all real differentiable functions $f : R^n \to R$, with at least one local minimum and given any initial point $x \in R^n$, there exists a learning rate sequence such that the gradient descent algorithm converges to a local minimum of f.
   (ix) Saha has a magical learning algorithm which returns the true labelling function regardless of the training set. He claims his algorithm has low bias, since its predictions are always correct, but high variance, since its predictions are quite different for different datapoints. Is Bob correct about bias?
   (x) Is Saha correct about variance?

2. (5 points) Consider the two functions: $f_a(x) = \max[0, (x-a)^T A(x-a))], f_b(x) = 2\max_i|x_i - a_i|$ with $x \in R^n, a \in R^n$ and positive semidefinite $A \in R^{n*n}$. Prove or disprove that $f_a(x), f_b(x)$ are convex in x.

   Hint: Use convexity preserving operations.

3. (5 points) The weight update in SGD: $w_{i+1} \leftarrow w_i - \eta_i \nabla_w f(w_i)$ may be thought of as a discretization to the first order ODE: $w'(t) = -\nabla_w f(w_i)$. The minimizer of the SGD is therefore conceived as a stable equilibrium of the ODE. That is, the minimum, w^* can be thought of as a fixed point to the iterates $w_{i+1} = G(\eta_i, w_i) \equiv w^* = G(\eta_i, w^*)$. Assume G to be locally Lipschitz on R^n. The stable equilibrium/fixed point w^* is guaranteed if we can construct a Lyapunov function, E suitably and show that such an E satisfies the following properties.
   (1) E is continuous
   (2) $E(w_i) = 0$ iff $w_i = W^*$
   (3) $E(w_i) > 0$ iff $w_i \neq W^*$ and
   (4) $E(w_{i+1}) \leq E(w_i) \forall i \in N$

Verify that $E(w) = \frac{1}{2}||w - w^*||^2$ is one such Lyapunov function. Assume L_2 norm.

4. (5 points) Consider the loss function $l(w, (x, y)) = log(1+exp(yw^T x))$, Assume x is bounded by $||x|| \leq B$, and $y \in [1, 1]$. Show that l is both Lipschitz bounded and convex and a smooth convex loss function (as a function of w, for every y, x). Calculate that parameters of Lipschitzness and smoothness.

5. (5 points) The gradient descent update rule is given by

$$w_{i+1} := w_i - \alpha \cdot \nabla_w f$$

where f is the loss function. When the learning rate, α, is too small, then convergence takes a long time. However, when the learning rate is too large, the solution diverges.

For a function f, the Lipschitz constant is given by $\max |\nabla_{\mathbf{w}} f|$. Therefore, by setting $\alpha = \frac{1}{L}$, we have $\Delta \mathbf{w} \leq 1$, constraining the change in the weights. This makes it optimal to set the learning rate to the reciprocal of the Lipschitz constant.

Show that, for the Least Square Cost function, the Lipschitz constant is given by the right hand side of the inequality:

$$\boxed{\frac{\|g(\mathbf{w}) - g(\mathbf{v})\|}{\|\mathbf{w} - \mathbf{v}\|} \leq \frac{k}{m} \left\|\mathbf{X}\mathbf{X}^T\right\| + \frac{1}{m} \left\|\mathbf{y}^T \mathbf{X}\right\|}$$

6. (5 points) Show that Convex-Lipschitz-bounded problems are PAC learnable,