EE660 Homework 3

Assigned: 10/14/2025, Due: 10/28/2025

Instructions: You may collaborate with others on this problem set, but each student must independently write their own solutions. We highly encourage you to use LaTeX to typeset your solutions. We will accept handwritten assignments, however if the solution to a problem is too illegible for the grader to read, then they may use their discretion and consider the problem incomplete. Solutions are due by 11:59pm Pacific Time on the due date, and are only to be submitted on DEN. Do not email the course staff with your assignment.

Template: The link `https://www.overleaf.com/read/hjgknqqhryqy` contains a basic LaTeX template that you may use. Note, however, that you are not required to use this template.

1. Problem 1

Let $\{\mathscr{H}_i\}_{i=1}^k$ be k hypothesis classes mapping X to $\{\pm 1\}$. Put $d := \max_{i=1,\dots,k} \mathrm{VCdim}(\mathscr{H}_i)$, and suppose that d is finite. Show that there exists universal constants c_0, c_1 such that:

$$\mathrm{VCdim}(\cup_{i=1}^k \mathscr{H}_i) \leqslant c_0 d + c_1 \log k.$$

Hint: use the Sauer-Shelah lemma and the following fact (cf. Proposition D.2 in the lecture notes). Let b, c be positive reals satisfying $bc \geqslant 1$. Then for every $n > 0$ we have that:

$$n \geqslant 2b \log(2bc) \text{ implies } n \geqslant b \log(cn).$$

2. Problem 2

Show that:

$$\mathrm{VCdim}(\{x \mapsto \mathrm{sgn}(1 - \|x - \theta\|) \mid \theta \in \mathbb{R}^d\}) \leqslant d + 2.$$

You may use the following fact without proof: Suppose that $\mathscr{F}$ is a d-dimensional vector space of functions mapping $\mathsf{X} \mapsto \mathbb{R}$. Then,

$$\mathrm{VCdim}(\{x \mapsto \mathrm{sgn}(f(x)) \mid f \in \mathscr{F}\}) \leqslant d.$$

3. Problem 3

In this problem, we will study stochastic gradient descent (SGD) for ERM problems. First, recall the empirical risk over $\theta \in \mathbb{R}^p$:

$$\hat{L}_n[\theta] = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta), \quad \ell_i(\theta) := \ell(f_\theta(x_i), y_i).$$

We consider a family of SGD estimators given a function $g : \mathbb{R}^p \times \Omega \mapsto \mathbb{R}^p$,

$$\theta_{t+1} = \theta_t - \eta_t g(\theta_t, \omega_t),$$

where at each iteration, ω_t is drawn from a fixed distribution $\mathcal{D}_\Omega$ over Ω independently across time. We will consider the setting where $\ell \geqslant 0$ and each $\ell_i(\theta)$ for $i \in [n]$ is L-smooth over $\theta \in \mathbb{R}^p$, i.e.,

$$\ell_i(\bar{\theta}) \leqslant \ell_i(\theta) + \langle \nabla \ell_i(\theta), \bar{\theta} - \theta \rangle + \frac{L}{2}\|\bar{\theta} - \theta\|^2, \quad \forall \theta, \bar{\theta} \in \mathbb{R}^p.$$

(a) Suppose that for every fixed $\theta \in \mathbb{R}^p$, we have:

$$\mathbb{E}_{\omega \sim \mathcal{D}_\Omega}[g(\theta, \omega)] = \nabla_\theta \hat{L}_n(\theta), \quad \mathrm{tr}(\mathrm{Cov}_{\omega \sim \mathcal{D}_\Omega}(g(\theta, \omega))) \leqslant B^2,$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, and $\mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\mathsf{T}]$ for a random vector X. Show that if $\eta_t \leqslant 1/L$ for all $t \in \mathbb{N}$, then we have:

$$\mathbb{E}[\hat{L}_n[\theta_{t+1}]] \leqslant \mathbb{E}\left[\hat{L}_n[\theta_t] - \frac{\eta_t}{2}\|\nabla_\theta \hat{L}_n[\theta_t]\|^2 + \frac{LB^2\eta_t^2}{2}\right], \quad t \in \mathbb{N}.$$

(b) Use part (a) to conclude that:

$$\min_{t \in \{0,\ldots,T-1\}} \mathbb{E}\|\nabla_\theta \hat{L}_n[\theta_t]\|^2 \leqslant \frac{2\hat{L}_n[\theta_0] + LB^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}.$$

(c) We now consider *mini-batched* SGD. Consider the specific choice of $g^k(\theta, \omega)$ for $k \in \mathbb{N}_+$:

$$g^k(\theta, \omega) := \frac{1}{k}\sum_{j=1}^{k} \nabla \ell_{i_j}(\theta), \quad \omega = \{i_j\}_{j=1}^k,$$

where each index $i_j \sim \mathrm{Unif}([n])$ is drawn independently across the batch, i.e., $i_{j_1} \perp i_{j_2}$ for $j_1 \neq j_2$. Suppose the following variance bound holds for all $\theta \in \mathbb{R}^p$:

$$\mathbb{E}_{\zeta \sim \mathrm{Unif}([n])}\|\nabla \ell_\zeta(\theta) - \hat{L}_n(\theta)\|^2 \leqslant B_1^2.$$

Use part (b) to derive the following bound for mini-batch SGD:

$$\min_{t \in \{0,\ldots,T-1\}} \mathbb{E}\|\nabla_\theta \hat{L}_n[\theta_t]\|^2 \leqslant \frac{2\hat{L}_n[\theta_0] + LB_1^2/k \cdot \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}.$$

Discuss using both this bound, and taking into account computational considerations, the trade-offs of small vs. large batch-size SGD.

4. Problem 4

In this problem, we will study gradient descent on overparameterized least-squares problems. Consider the following least-squares loss:

$$\ell(x) := \frac{1}{2}\|Ax - b\|^2,$$

where $x \in \mathbb{R}^d$, $b \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n \times d}$. We will consider the overparameterized regime when $d > n$, and where $\operatorname{rank}(A) = n$. Consider the gradient descent update with step size $\eta > 0$:

$$x_{t+1} = x_t - \eta \nabla \ell(x_t), \quad x_0 = 0.$$

In this problem, we will show that, despite there being an infinite number of global minima for $\ell(x)$, gradient descent always converges to a particular solution.

(a) Show that $x_t \in \operatorname{Span}(A^\mathsf{T})$ for all $t \in \mathbb{N}$.

(b) Show there exists a unique solution $x_\star$ such that $Ax_\star = b$ and $x_\star \in \operatorname{Span}(A^\mathsf{T})$.

(c) Define κ to be $\kappa := \frac{\lambda_{\max}(AA^\mathsf{T})}{\lambda_{\min}(AA^\mathsf{T})}$, where $\lambda_{\max}(M)$ (resp. $\lambda_{\min}(M)$) denotes the maximum (resp. minimum) eigenvalue of M. Show that if we set $\eta = 1/\lambda_{\max}(AA^\mathsf{T})$,

$$\ell(x_t) \leqslant (1 - 1/\kappa)^t \ell(x_0).$$

Hint: Since $\ell(\cdot)$ is a quadratic function, its second-order Taylor expansion is exact. That is,

$$\ell(y) = \ell(x) + \langle \nabla \ell(x), y - x \rangle + \frac{1}{2}\langle y - x, \nabla^2 \ell(x)(y - x) \rangle.$$

(d) Use part (c) to show that for any $\varepsilon > 0$,

$$T \geqslant 2\kappa \log\left(\frac{\|b\|}{\varepsilon}\right) \implies \|Ax_T - b\| \leqslant \varepsilon.$$

You will receive full credit for correctly showing the above implication for any valid universal constant $c > 0$.

Hint: Use the inequality $1 + x \leqslant \exp(x)$ for any $x \in \mathbb{R}$.