# ECE 543: Statistical Learning Theory

Maxim Raginsky

## Homework 5

Assigned April 12, 2018; due April 19, 2018

---

**Required reading:** lecture notes of Hajek and Raginsky, Chapter 12

---

1. Consider the online learning problem with $\mathcal{F} = \mathsf{Z} = [-1,1]$ and $\ell(f,z) = 1 + fz$.

   (a) What does the gradient descent algorithm reduce to for this example?

   (b) Express $\min_{f^* \in \mathcal{F}} J_T(f^*, z^T)$ in terms of $z^T = (z_1, \ldots, z_T)$. Here, $J_T(f^*, z^T) = \sum_{t=1}^T \ell(f^*, z_t)$..

   (c) Suppose an online algorithm $\tilde{A}$ (i.e. an algorithm of the form $(\tilde{f}_t = \tilde{A}(\tilde{f}_1, \ldots, \tilde{f}_{t-1}, z_1, \ldots, z_{t-1}))$) minimizes $\max_{z^T \in \mathsf{Z}^T} J_T((f_t), z^T)$ over all online algorithms. Is the sequence $(\tilde{f}_1, \ldots, \tilde{f}_T)$ produced by $\tilde{A}$ uniquely determined? (This part shows that there is a difference between minimizing maximum loss, and minimizing maximum regret against all fixed strategies.)

   (d) Suppose for this part that the sequence $Z^T = (Z_1, \ldots, Z_T)$ is a Rademacher sequence (i.e., the $Z_t$'s are iid, each equally likely to be $\pm 1$.). Show that

   $$\lim_{T \to \infty} \frac{\mathbf{E}\left[\min_{f^* \in \mathcal{F}} J_T(f^*, Z^T)\right] - T}{\sqrt{T}} = -c,$$

   and identify the constant $c > 0$. (Hint: Apply the central limit theorem.) In contrast, find $\mathbf{E} J_T((f_t), Z^T)$ for $(f_t)$ produced by an arbitrary online algorithm. Finally, explain why, for any $\varepsilon > 0$, $\sup_{z^T} R_T((f_t), z^T) \geq (1-\varepsilon)c\sqrt{T}$ for all sufficiently large $T$ and any online algorithm.

2. Continue to consider the setting of Problem 1, with $\mathcal{F} = \mathsf{Z} = [-1,1]$ and $\ell(f,z) = 1 + fz$.

   (a) Consider the projected GD algorithm run with step size $\alpha_t = \frac{1}{\sqrt{t}}$ for $t \geq 1$ and initial state $f_1 = 0$. Suppose $T$ is even and $z_1 = \ldots = z_{T/2} = -1$ and $z_{T/2+1}, \ldots, z_T = 1$. Show that there is a finite constant $c > 0$ such that $R_T((f_t), z^T) \leq -T + c\sqrt{T}$ for all $T$ sufficiently large. (Large negative maximum regret means the algorithm has much smaller loss than any fixed $f^*$.)

   (b) A conclusion of problem 1(d) is that for any online algorithm, there is a constant $c' > 0$ so that $\max_{z^T \in \mathsf{Z}^T} R_T((f_t), z^T) \geq c'\sqrt{T}$ for all $T$ sufficiently large. The proof uses randomization and is thus nonconstructive. To gain more insight into this result, prove this same result for the projected gradient descent algorithm with fixed step size $\alpha$ with $\alpha$ by using deterministic sequences $z^T$. For convenience you can assume $T$ is even and consider only $0 < \alpha \leq 1$. (Hint: The choice of $z^T$ depends on $\alpha$. The entire range $\alpha > 0$ can be covered by two choices of $z^T$.)

3. Continue to consider the setting of Problem 2, with $\mathcal{F} = \mathsf{Z} = [-1,1]$ and $\ell(f,z) = 1 + fz$. Write $f_s^t$ to denote $(f_s,\ldots,f_t)$, and $f^T$ to denote $(f_1,\ldots,f_T)$, and define $z_s^t$ similarly. Let $A$ represent an arbitrary online algorithm of the learner, and $B$ represent an arbitrary online algorithm of the adversary. That is, $A$ determines $f_1$ and it has mappings of the form $f_t = A(f_1^{t-1}, z_1^{t-1})$ for $2 \le t \le T$, and $B$ has mappings of the form $z_t = B(f_1^t, z_1^{t-1})$ for $1 \le t \le T$. Together, a choice of $A$ and $B$ determine $f^T$ and $z^T$ uniquely. The goal of this problem is to determine the min max regret, $R_T^* = \min_A \max_B R_T(f^T, z^T)$.

   (a) Show by induction on $k$ that for $0 \le k \le T$

$$R_T^* = \min_A \max_B \sum_{t=1}^{T-k} f_t z_t + V_k\left(\sum_{t=1}^{T-k} z_t\right), \tag{1}$$

   where $V_0(s) = |s|$ for all $s \in \mathbb{R}$, and

$$V_{k+1}(s) = \min_{f\in[-1,1]}\left(\max_{z\in[-1,1]} fz + V_k(s+z)\right). \tag{2}$$

   In particular, $R_T^* = V_T(0)$. (Hint: These are equations of dynamic programming, working backwards from the end of a problem by induction. First check the base case, $k = 0$. For the induction step, separate out the terms $z_{T-k}$ and $f_{T-k}$ on the righthand side of (1).

   (b) Show $V_k(s) = \mathbf{E}|s + Z_1 + \ldots + Z_k|$ for $0 \le k \le T$, where $Z_1,\ldots,Z_T$ are independent Rademacher random variables. In particular, $R_T^* = \mathbf{E}|Z_1 + \ldots + Z_T|$. How does this compare with the lower bound on maximum regret you found in Problem 1? (Hint: Let $V_k(s) = \mathbf{E}|s + Z_1 + \ldots + Z_k|$ and show by induction it indeed satisfies (2). The base case for $k = 0$ is true by definition.)