Spring School "Structural Inference"

Exercises for lectures by Alexander Rakhlin

March 16, 2015

For $A \subseteq [-1,1]^n$ define Rademacher averages of A as

$$\mathfrak{R}(A) = \mathbb{E}_\epsilon \sup_{a \in A} \frac{1}{n} \sum_{t=1}^n \epsilon_t a_t$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. ± 1 Rademacher random variables.

Exercise 1 Prove that for any $r_1, \dots, r_n \in [0,1]$,

$$\mathbb{E} \sup_{a \in A} \sum_{t=1}^n \epsilon_t r_t a_t \leq \mathbb{E} \sup_{a \in A} \sum_{t=1}^n \epsilon_t a_t$$

Exercise 2 Define Gaussian averages of A as

$$G(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \sum_{t=1}^n \gamma_t a_t$$

where $\gamma_1, \dots, \gamma_n$ are independent $N(0,1)$. Show that

$$c\mathfrak{R}(A) \leq G(A) \leq C\sqrt{\log(n)}\mathfrak{R}(A)$$

and find explicit constants c, C.

Exercise 3 Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz. Prove that

$$\mathbb{E} \sup_{a \in A} \sum_{t=1}^n \epsilon_t \phi(a_t) \leq L\mathbb{E} \sup_{a \in A} \sum_{t=1}^n \epsilon_t a_t$$

Hint: condition on all but one ϵ_t, write out the two possibilities for ϵ_t, and combine the suprema. Make sure the argument does not leave any absolute values.

Exercise 4 Prove that for a finite collection $A \subset \mathbb{R}^n$ and any $c > 0$,

$$\mathbb{E} \max_{a \in A} \left\{ \sum_{t=1}^n \epsilon_t a_t - c a_t^2 \right\} \leq C \log |A|$$

Does C depend on the magnitude of vectors in A?

Hint: write out the moment-generating function and use $(e^{-x} + e^x)/2 \leq e^{x^2/2}$.

Exercise 5 We argued in the lecture that for a finite collection $A \subset [-1,1]^n$,

$$\mathbb{E} \max_{a \in A} \sum_{t=1}^{n} \epsilon_t a_t \le r\sqrt{2 \log N}, \quad r = \max_{a \in A} \|a\|_2$$

Now suppose B is a set of predictable processes with respect to $\{\mathcal{F}_t = \sigma(\epsilon_1, \ldots, \epsilon_t)\}_{t=0}^n$. That is, each $\boldsymbol{b} \in B$ is a sequence $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ where each $\boldsymbol{b}_t$ is $\mathcal{F}_{t-1}$-measurable. Prove that

$$\mathbb{E} \max_{\boldsymbol{b} \in B} \sum_{t=1}^{n} \epsilon_t \boldsymbol{b}_t \le r\sqrt{2 \log N}, \quad r = \max_{\epsilon \in \{\pm 1\}^n} \max_{\boldsymbol{b} \in B} \sqrt{\sum_{t=1}^{n} \boldsymbol{b}_t^2}.$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. Hint: Consider the moment generating function and peel off one term at a time, from n backwards to $t = 1$.

Exercise 6 Let W be a random variable with values in $\mathcal{A}$. Prove that for a measurable function $\Psi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$,

$$\mathbb{E}_W \sup_{b \in \mathcal{B}} \Psi(W, b) = \sup_{\gamma} \mathbb{E}_W \Psi(W, \gamma(W))$$

where the supremum ranges over all functions $\gamma : \mathcal{A} \to \mathcal{B}$. (Assume compactness or boundedness if needed to make the argument rigorous).

Exercise 7 Let $\epsilon_{1:n} \triangleq (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ be n i.i.d. Rademacher random variables. Use the previous exercise to conclude that for $\Psi : \mathcal{X}^n \times \{\pm 1\}^n \to \mathbb{R}$,

$$\sup_{x_1 \in \mathcal{X}} \mathbb{E}_{\epsilon_1} \ldots \sup_{x_n \in \mathcal{X}} \mathbb{E}_{\epsilon_n} \Psi(x_{1:n}, \epsilon_{1:n}) = \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_n} \mathbb{E}_{\epsilon_{1:n}} \Psi(\mathbf{x}_1, \mathbf{x}_2(\epsilon_1) \ldots, \mathbf{x}_n(\epsilon_{1:n-1}), \epsilon_{1:n})$$

where the last supremum is taken over functions $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$.

Exercise 8 Let Q be the set of distributions on some set $\mathcal{A}$ and P the set of distributions on $\mathcal{B}$. Under very general conditions on $\boldsymbol{\ell}, \mathcal{A}, \mathcal{B}$,

$$\min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \boldsymbol{\ell}(a, b) = \max_{p \in P} \min_{a \in \mathcal{A}} \mathbb{E}_{b \sim p} \boldsymbol{\ell}(a, b). \tag{1}$$

This is known as the minimax theorem. Note that the inner max/min can be taken at a pure strategy (delta distribution) because a linear function achieves its max/min at a corner of the probability simplex.

Prove the following: if $\boldsymbol{\ell}(a, b)$ is convex in a and $\mathcal{A}$ is a convex set, then the outer minimization

$$\min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \boldsymbol{\ell}(a, b) = \min_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \boldsymbol{\ell}(a, b)$$

is achieved at a pure strategy. We will use this result to restrict our attention to deterministic strategies.

Exercise 9 Let W be a random variable, and suppose that for any realization of W,

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} \{\boldsymbol{\ell}(a, b) + \Psi_t(b, W)\} \le \Psi_{t-1}(W)$$

Prove that

$$\inf_{q \in \Delta(\mathcal{A})} \sup_{b \in \mathcal{B}} \{\mathbb{E}_{a \sim q} \boldsymbol{\ell}(a, b) + \mathbb{E}_W \Psi_t(b, W)\} \le \mathbb{E}_W \Psi_{t-1}(W)$$

by exhibiting a strategy for the infimum. This statement will be useful for defining computationally-efficient random playout methods in Lecture #3.

Exercise 10 Consider the following online prediction problem, taking place over rounds $t = 1, \ldots, n$. On each round, we make a prediction $\widehat{y}_t \in [0, 1]$, observe an outcome $y_t \in \{0, 1\}$, and suffer the loss of $\boldsymbol{\ell}(\widehat{y}_t, y_t) = y_t + \widehat{y}_t - 2\widehat{y}_t \cdot y_t$. Take a potential function $\Phi : \{\pm 1\}^n \to \mathbb{R}$ with two properties: first, it is stable with respect to flip of any coordinate:

$$|\Phi(\ldots, -1, \ldots) - \Phi(\ldots, +1, \ldots)| \leq 1.$$

Second, $\mathbb{E}\Phi(b_1, \ldots, b_n) \geq n/2$ where b_i's are i.i.d. Bernoulli with bias $1/2$. Show that

$$\min_{\widehat{y}_t} \max_{y_t} \left\{ \boldsymbol{\ell}(\widehat{y}_t, y_t) + \mathbb{E}_{b_{t+1:n}} \Phi(y_1, \ldots, y_t, b_{t+1}, \ldots, b_n) \right\} \leq \mathbb{E}_{b_{t:n}} \Phi(y_1, \ldots, y_{t-1}, b_t, \ldots, b_n) + \frac{1}{2}$$

Conclude that there is a prediction strategy that guarantees

$$\sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{y}_t, y_t) \leq \Phi(y_1, \ldots, y_n) \tag{2}$$

for any sequence $y_1, \ldots, y_n$ of binary outcomes. Conversely, argue that if there is a function Φ that satisfies (2) for all sequences, then it must hold that $\mathbb{E}\Phi \geq n/2$.

Exercise 11 Write the loss function in the previous exercise as expected indicator loss under the randomized strategy with bias $\widehat{y}_t$. Use the previous exercise to argue that there must exist a randomized algorithm that predicts an arbitrary sequence of bits with the following strong guarantee:

the expected average number of mistakes (per n rounds) is at most the ***minimum*** *of proportion of* 1*'s and proportion of* 0*'s in the sequence, up to a* $O(1/\sqrt{n})$ *additive factor.*

That is, if the sequence, say, has 40% of 0's, then the method will only err roughly 40% of the time, even though the locations of 0's. The method is adaptive: it does not need to know any prior information about the sequence. This result might seem surprising, given that the sequence is not governed by any stochastic process that we can describe. (origin of this problem: T. Cover, 1960's)