ORIE 6745: Causality and Learning for Intelligent Decision Making

Homework 1

1. Do any two of the following three:

   (a) Write out the complete inductive argument in the proof for McDiarmid's inequality (where I just put ... in the sequence of inequalities and noted "repeating this").

   (b) Write out the complete inductive argument in the proof for the Ledoux-Talegrand Lipschitz comparison lemma using the $(*)$ condition in Lec 3 (where I just put ... in the sequence of inequalities).

   (c) Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space and $\|v\|^2 = \langle v, v\rangle$. Write out the complete inductive proof for the parallelogram law we used in Lec 4: for all n and $v_1, \dots, v_n \in V$,

$$\frac{1}{2^n} \sum_{\sigma \in \{-1,+1\}^n} \left\| \sum_{i=1}^n \sigma_i v_i \right\|^2 = \sum_{i=1}^n \|v_i\|^2 .$$

2. Prove the following properties of the Rademacher complexity for $\mathcal{A} \subseteq \mathbb{R}^n$:

   (a) $\mathfrak{R}(\{\alpha + \beta a : a \in \mathcal{A}\}) = |\beta|\, \mathfrak{R}(\mathcal{A})$ for $\alpha \in \mathbb{R}^m, \beta \in \mathbb{R}$.

   (b) $\mathfrak{R}(\text{conv-hull}(\mathcal{A})) = \mathfrak{R}(\mathcal{A})$
   where $\text{conv-hull}(\mathcal{A}) = \left\{\sum_{j=1}^m \lambda_j a_j : m \in \mathbb{N}, a_1, \dots, a_m \in \mathcal{A}, \lambda \in \mathbb{R}^m, \lambda \geq 0, \sum_{j=1}^m \lambda_j = 1\right\}$.

   (c) $\mathfrak{R}(\text{symm-conv-hull}(\mathcal{A})) = \mathfrak{R}(\mathcal{A})$
   where $\text{symm-conv-hull}(\mathcal{A}) = \left\{\sum_{j=1}^m \lambda_j a_j : m \in \mathbb{N}, a_1, \dots, a_m \in \mathcal{A}, \lambda \in \mathbb{R}^m, \sum_{j=1}^m |\lambda_j| = 1\right\}$.

   (d) $\mathfrak{R}(\{a + b : a \in \mathcal{A}, b \in \mathcal{B}\}) = \mathfrak{R}(\mathcal{A}) + \mathfrak{R}(\mathcal{B})$.

   (e) If $\mathcal{F} \subseteq [\mathcal{X} \to \mathbb{R}]$ and $\mathcal{A}_n(\mathcal{F}) = \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\}$ then the above are also true for the functional complexity $\mathbb{E}\mathfrak{R}(\mathcal{A}_n(\mathcal{F}))$ when shifting, taking convex hulls of, and taking point-wise additions of function classes $\mathcal{F}$.

3. What is the VC dimension of $\{x \mapsto \mathbb{I}[a_1 \leq x_1 \leq b_1 \wedge \cdots \wedge a_d \leq x_d \leq b_d] : a_1, b_1, \dots, a_d, b_d \in \mathbb{R}\}$?

4. Consider a sequence of policy classes $\Pi_n \subseteq [\mathcal{X} \to \{\pm 1\}]$. Use Sauer's lemma to argue that

$$\sup_{x_1, \dots, x_n \in \mathcal{X}} |\{(\pi(x_1,), \dots, \pi(x_n)) : \pi \in \Pi_n\}|$$

   is polynomial in n if and only if $\text{VCdim}(\Pi_n)$ is polynomial in n.

   In all the following questions, $\mathcal{Z} = \{0, 1\}$.

5. For complete randomization show how to compute both $\text{Var}(\hat{\tau} \mid Y_{1:n}(0,1))$ and $\text{Var}(\hat{\tau})$ in terms of sample and population (co-)variances of potential outcomes. Discuss what are the differences between the two, qualitatively in terms of meaning and quantitatively in terms of value.

6. Consider the data shown below from a completely randomized experiment from this paper `http://diabetes.diabetesjournals.org/content/53/suppl_3/S215.full` comparing the body weight in grams of mice under a high-fat ($Z_1 = 1$) and regular diet ($Z_1 = 0$). Compute the mean difference and corresponding p-value for the sharp null. Try to repeat this using other statistics such as difference in medians and KS. Discuss the meaning of the results.

Diet (Z_i)	Weight (g) (Y_i)
0	21.51
0	28.14
0	24.04
0	23.45
0	23.68
0	19.79
0	28.4
0	20.98
0	22.51
0	20.1
0	26.91
0	26.25
1	25.71
1	26.37
1	22.8
1	25.34
1	24.97
1	28.14
1	29.58
1	30.92
1	34.02
1	21.9
1	31.53
1	20.73

7. Argue that the permutation test with either the usual mean differences statistics $\hat{\tau} = \frac{\sum_{i:Z_i=1} Y_i}{\sum_{i:Z_i=1} 1} - \frac{\sum_{i:Z_i=0} Y_i}{\sum_{i:Z_i=0} 1}$ or with the sum of treatment outcomes statistic $\hat{\tau}^{\text{total treatment outcomes}} = \sum_{i:Z_i=1} Y_i$ will always give the same p-value.

8. We know that even if the Fisherian sharp null is false, SATE can be zero and, correspondingly, mean differences takes small values (of order $1/\sqrt{n}$) centered at zero leading to low power of the permutation test. But what if we use the KS statistic? Specifically, if the permutation test with the KS statistic rejects the Fisherian sharp null no more than α fraction of the time for n arbitrarily large, can it be that the hypothesis is false nonetheless? Why or why not?

9. A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. To test her claim, Sir Ronald A Fisher takes $n = 8$ cups, prepares 4 with milk first ($Z_i = 1$) and 4 with tea first ($Z_i = 0$) and presents to the lady to taste. The lady declares $Y_i = 1$ if she thinks milk was added first and $Y_i = 0$ if tea first. The lady is aware of the design (4-4 split) and so will always declare $\sum_{i=1}^{8} Y_i = 4$ (note: units exchangeable but not iid). Using the permutation test, test the null hypothesis that the lady has no faculty of discrimination:
   (a) Argue that the Fisherian sharp null is the appropriate formalization of our null hypothesis.
   (b) What is the smallest significance of the test using $\alpha = 0.05$ as a p-value threshold?
   (c) What is the power of the test (probability of rejecting the null) if lady has perfect discriminatory powers ($Y_i(z) = z$)?
   (d) What is the power of the test (probability of rejecting the null) if lady gives each cup a score S_i and declares the top four scored cups to have milk first, where $S_i \sim \text{unif}[\min\{Z_i, 1-\gamma\}, \max\{Z_i, \gamma\}]$ for some $\gamma \in (0, 1)$.

10. Argue that if $\mathbb{E}[Y(z) \mid X] = \tau z + \beta^T X + \alpha$ then the $\hat{\tau}^{\text{OLS}}$ estimate, which regresses the factually observed Y on Z, X and takes the estimate as the coefficient on Z is unbiased for τ (note the difference between Z and z). Argue that under complete randomization the estimate is still consistent for PATE even if $\mathbb{E}[Y(z) \mid X]$ is not linear.
Now suppose $|Y(z)| \leq M$ and consider any sequence of function classes $\mathcal{F}_n \subseteq [\mathcal{X} \to [-M, M]]$ that satisfies $(f \in \mathcal{F}_n, \alpha \in \mathbb{R} \implies f + \alpha \in \mathcal{F}_n)$ and that has vanishing expected Rademacher complexity:

$$\mathbb{E}\mathfrak{R}\left(\{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}_n\}\right) \to 0.$$

Use the risk bounds we developed in Lecs 3–4 to prove that under complete randomization any estimate of the form

$$\hat{\tau}^{\text{ERM}} = \frac{1}{n}\sum_{i=1}^{n}(f_1(X_i) - f_0(X_i))$$

$$\text{for any} \quad (f_0, f_1) \in \operatorname*{argmin}_{f_0, f_1 \in \mathcal{F}_n} \sum_{i=1}^{n}(Y_i - f_{Z_i}(X_i))^2$$

is still consistent for PATE regardless of the form of $\mathbb{E}[Y(z) \mid X]$. Explain intuitively why this is and how it could break if randomization were violated or if the function classes had too high capacity (e.g., Rademacher complexities did not vanish).