36-710, Fall 2019
Homework 5

Due Wed Nov 13, by 5:00pm in Alden's mailbox.

1. Let P and Q be two probability measures on some measurable space $\mathcal{X}$. Let X be a random variable taking values in $\mathcal{X}$ and suppose we want to test the null hypothesis that X comes from P versus the alternative that it comes from Q. Show that

$$\inf_{\phi} \mathbb{E}_P[\phi(X)] + \mathbb{E}Q[1 - \phi(X)] = 1 - d_{\mathrm{TV}}(P, Q),$$

   where the infimum is over all test functions $\phi\colon \mathcal{X} \to \{0, 1\}$.

2. Problem 15.10 from the textbook.

3. Exercise 2.2 from Tsybakov's book.

4. Let $\mathcal{P}_{d,\sigma} = \{N_d(\mu, \sigma^2 I_d), \mu = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d\}$ with some known σ and I_d the d-dimensional identity matrix (in fact we may take $\mathcal{P}_{d,\sigma}$ to be the larger class of sub-Gaussian random vectors with sub-Gaussian parameter no larger than σ^2). For any $P = N_d(\mu, I_d) \in \mathcal{P}_{d,\sigma}$, let $\theta(P) = \max_i \mu_i$ be the corresponding parameter, the magnitude of the largest mean value. We are interested in finding a minimax lower bound for this parameter, assuming the data consists of a d-dimensional vector $(X_1, \ldots, X_d)$ from a distribution on $\mathcal{P}_{d,\sigma}$ and using the loss $d(\theta, \theta') = |\theta - \theta'|$ (thus, $w(\cdot)$ is the identity function). Proceed as follows: let μ_0 be the zero d-dimensional vector and, for $i = 1, \ldots, d$, let μ_i be the d-dimensional vector with all 0 entries except for the one in the i-th coordinate, which is equal to $\sqrt{a\sigma^2 \log d}$, for some $a > 0$ to be set later during your calculations. For $i = 0, \ldots, d$, let $P_i = N_d(\mu_i, \sigma^2 I_d)$ and set $\overline{P} = \frac{1}{d}\sum_{i=1}^{d} P_i$. You will derive a lower bound based on the convex version of LeCam's two point argument by contrasting P_0 with the mixture $\overline{P}$.

   (a) Implement this strategy first using Pinsker's inequality to bound the total variation distance between P_0 and $\overline{P}$. Did the convex version of LeCam's two point argument bring any improvement over the vanilla version of LeCam's two point argument based on P_0 versus any P_i?

   (b) Now instead of the KL divergence use the χ^2 divergence (along with the fact that the squared of the total variation distance is bounded by the χ^2 divergence). You should now obtain a lower bound of the order $\sqrt{\sigma^2 \log d}$, which happens to be the minimax rate (no need to prove this, though you are ecnouraged to try).

5. **Reading Exercise (This one is going to be a pain, typographically)**. This is an old result that is relevant to modern adaptive data analysis. Here is the abstract set-up. Suppose we observe a random vector $(X_1, \ldots, X_d)$ from a distribution from the class $\mathcal{P}_{d,\sigma}$ described in the previous problem. Let $\mu = (\mu_1, \ldots, \mu_d)$ be the corresponding mean vector. Upon observing the sample, we compute $I = \mathrm{argmax}_{i=1,\ldots,d} X_i$. Notice that I is a random variable. Suppose now that we are interested in estimating (again, under the L_1 loss) the *random parameter* μ_I. This simple setting mimics a typical, though incorrect, in data analysis work-flow, where it often happens that the target for inference is not chosen before observing the data, but instead *based* on the observed data.

   A natural estimator of μ_I is $X_I = \max_i X_i$. An upper bound on its risk was recently derived in term of the mutual information between the sample and I (see the appendix to Chapter 15 in the book for basic information theoretic concepts):

- Russo, D. and You, J,. (2016). Controlling Bias in Adaptive Data Analysis Using Information Theory, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR 51:1232-1240.

In our settings, the resulting upper bound is order $O(\sqrt{\sigma^2 \log d})$. What about the lower bound? A lower bound based on Bayesian arguments was derived in this beautifully typeset paper:

- Sackrowitz, H. and Samuel-Cahn, E. (1986). Evaluating the chosen population: a Bayes and minimax approach, Lecture Notes–Monograph Series Volume 8, 386-399.

(Incidentally the authors show that X_I is actually not the minimax estimator).

- Extract the lower bound from the proof of claim (b) on page 396. You will find that this lower bound does not match the upper bound of $\sqrt{\sigma^2 \log d}$. The difference is only in the power of the $\log(d)$ term, which, in the present case, is significant.

As far as I know, determining the minimax rate for the estimation problem described above remains unsolved. It will be interesting but challenging project to close this gap. It is conjecture that the lower bound should also depend on the mutual information between X and I. Anyone interested in trying this out? This would be an interesting result in adaptive data analysis.