Stat 300A Theory of Statistics

Homework 5: Solutions

Nikos Ignatiadis *Due on November 7, 2018*

- Solutions should be complete and concisely written. Please, use a separate sheet (or set of sheets) for each problem.
- We will be using Gradescope (`https://www.gradescope.com`) for homework submission (you should have received an invitation) - no paper homework will be accepted. Handwritten solutions are still fine though, just make a good quality scan and upload it to Gradescope.
- You are welcome to discuss problems with your colleagues, but should write and submit your own solution.

1: A function denoising problem

Let $\boldsymbol{\theta}$ be a discrete function sampled on a regular grid in $[0,1]$. Namely, for $n \in \mathbb{N}$, we let $\varepsilon = 1/n$, and

$$\boldsymbol{\theta} = \big(\theta(0), \theta(\varepsilon), \theta(2\varepsilon), \dots, \theta((n-1)\varepsilon)\big) \in \mathbb{R}^n \,. \tag{1}$$

We observe noisy measurements of this function $y_k = \theta(k\varepsilon) + z_k$, where $(z_k)_{k \leq n} \sim_{iid} \mathsf{N}(0, \sigma^2)$, and are interested in estimating $\boldsymbol{\theta}$ with respect to the normalized square loss $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2/n$.

We define the discrete derivative by letting $\Delta\theta(k\varepsilon) = [\theta((k+1)\varepsilon) - \theta(k\varepsilon)]/\varepsilon$ for $k \in \{0, \dots, n-2\}$, and $\Delta\theta((n-1)\varepsilon) = [\theta(0) - \theta((n-1)\varepsilon)]/\varepsilon$ (periodic boundary conditions). We consider the following parameter class

$$\Theta(R, n) = \Big\{\boldsymbol{\theta} : \ \sum_{k=0}^{n-1} \theta(k\varepsilon) = 0, \ \sum_{k=0}^{n-1} \varepsilon\big(\Delta\theta(k\varepsilon)\big)^2 \leq R\Big\} \,. \tag{2}$$

(*a*) Give an expression for the linear minimax risk $R_{\mathrm{L}}(\Theta(R, n))$.

[Hint: It might be convenient to use the discrete Fourier transform of $\boldsymbol{\theta}$.]

(*b*) Can you apply Pinsker's theorem and show that the linear minimax risk is close to the overall minimax risk $R_{\mathrm{M}}(\Theta(R, n))$? Justify your answer and state explicitly any eventual condition that you are imposing on R, n.

Solution

(*a*) Starting with this problem, we directly observe that we may write the constraint $\sum_{k=0}^{n-1} \varepsilon\big(\Delta\theta(k\varepsilon)\big)^2$ in Ellipsoidal form

$$\boldsymbol{\theta}^\top A \boldsymbol{\theta} \leq R/n$$

Here:

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & -1 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}$$

To apply Pinsker's result, we need to diagonalize A. To this end, first onsider the DFT matrix $(U_{kl})_{0\leq k,l\leq n-1}$ with $U_{kl} = \exp\left(\frac{-2\pi i kl}{n}\right)$. Furthermore, recall the following properties: $U^*U = nI_n$, so that $U/\sqrt{n}$ is unitary. We may check that $U/\sqrt{n}$ diagonalizes A with eigenvalues $2(1-\cos(2\pi j/n))$.

One way to see this is to use Parseval's identity for the DFT, as well as the Shift identity for the DFT (below we write $\boldsymbol{\theta}_{\cdot} = \boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\cdot+1} = (\theta_1, \dots, \theta_{n-1}, \theta_0))$ with $\theta_k = \theta(k\varepsilon)$). More concretely:

$$\begin{aligned}
n\|\boldsymbol{\theta}_{\cdot} - \boldsymbol{\theta}_{\cdot+1}\|^2 &= \|U\boldsymbol{\theta}_{\cdot} - U\boldsymbol{\theta}_{\cdot+1}\|^2 \;\; \text{Parseval} \\
&= \|U\boldsymbol{\theta}_{\cdot} - \exp(2i\pi\cdot/n)\cdot U\boldsymbol{\theta}_{\cdot}\|^2 \quad \text{(coordinatewise product, shift)} \\
&= \sum_{k=0}^{n-1} |1-\exp(2ik\pi/n)|^2 (U\boldsymbol{\theta}_{\cdot})_k^2 \\
&= \sum_{k=0}^{n-1} 2(1-\cos(2\pi k/n))(U\boldsymbol{\theta}_{\cdot})_k^2
\end{aligned}$$

Since the unitary matrix $U/\sqrt{n}$ diagonalizes A, we note that there must exist also an orthogonal (real) matrix O which diagonalizes A and has the same eigenvalues. Furthermore, note that 1st column and row of $U/\sqrt{n}$ just consists of entries $1/\sqrt{n}$, thus also the 1st row of O will consist of these entries. Thus upon mapping $\mathbf{y} \mapsto \tilde{\mathbf{y}} = O\mathbf{y}$, we observe that if we let $\tilde{\boldsymbol{\theta}} = O\boldsymbol{\theta}$, then $\tilde{\mathbf{y}} \sim \mathcal{N}(\tilde{\boldsymbol{\theta}}, \sigma^2)$. Furthermore the constraints turn into:

$$\tilde{\boldsymbol{\theta}}_0 = \sum_{i=0}^{n-1} \frac{1}{\sqrt{n}}\theta_i = 0$$

and

$$\sum_{k=0}^{n-1} 2(1-\cos(2\pi k/n))\tilde{\boldsymbol{\theta}}_k^2 \leq \frac{R}{n}$$

Furthermore, since $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 = \|O\hat{\boldsymbol{\theta}} - O\boldsymbol{\theta}\|_2^2$, we see that the transformed and the original estimation problems are equivalent and hence that their (linear) minimax risks must coincide. Also, since we know the first coordinate is 0, by a sufficiency argument we may discard the first observation $\tilde{Y}_0$ without loss of information and find ourselves in a $(n-1)$-dimensional Gaussian problem with the following Ellipsoidal form:

$$\tilde{\mathbf{Y}} \sim \mathcal{N}(\tilde{\boldsymbol{\theta}}, \sigma^2)$$

$$\tilde{\boldsymbol{\theta}} \in \tilde{\Theta} = \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{n-1} : \tilde{\boldsymbol{\theta}}^\top \tilde{A}\tilde{\boldsymbol{\theta}} \leq 1\}$$

Here $\tilde{A} = \mathrm{Diag}(\tilde{a}_1^2, \dots, \tilde{a}_{n-1}^2)$ and $\tilde{a}_j = \sqrt{\frac{2n}{R}(1-\cos(2\pi j/n))}$

We are finally ready to apply Theorem 4.1 from the notes to get (the notes gives us the linear minimax risk for the unnormalized loss so we further divide by n):

$$R_L(\theta) = \frac{1}{n}\inf_{\lambda\geq 0}\left\{\lambda^2 + \sigma^2\sum_{i=1}^{n-1}(1-\lambda\tilde{a}_j)_+^2\right\}$$

The minimum is achieved at the unique solution of:

$$\lambda = \sigma^2 \sum_{j=1}^{n-1} \tilde{a}_j (1 - \lambda \tilde{a}_j)_+$$

Let us now get a bit more insight into this expression, i.e. what is the minimax rate ignoring constants? We will write $\asymp$ to denote "rate equality", i.e. we will write $a_n \asymp b_n$ to mean $0 < \liminf a_n/b_n \le \limsup a_n/b_n < \infty$.

First let us note that (for j small enough so that the first order Taylor expansion of $1 - \cos(x) \approx x^2/2$ is accurate):

$$\tilde{a}_j \asymp \frac{j}{n^{1/2} R^{1/2}}$$

So with $\lambda := \lambda(k) \asymp \frac{n^{1/2} R^{1/2}}{k}$ we would get the equality:

$$\frac{n^{1/2} R^{1/2}}{k} \asymp \sigma^2 \sum_{j=1}^{k} \frac{j}{n^{1/2} R^{1/2}} \asymp \frac{\sigma^2}{n^{1/2} R^{1/2}} k^2$$

Solve for k to get:

$$k_*^3 \asymp \frac{nR}{\sigma^2}, \text{ i.e. } k_* \asymp \frac{n^{1/3} R^{1/3}}{\sigma^{2/3}}$$

So the optimal λ_* satisfies:

$$\lambda_* \asymp \sigma^{2/3} n^{1/6} R^{1/6}$$

Finally we get the affine minimax risk:

$$R_L(\Theta) \asymp \sigma^{4/3} R^{1/3} n^{-2/3}$$

In particular, we recover the rate for the nonparametric regression problem over first-order Sobolev ellipsoids (for fixed R).

(b) Directly applying Pinsker's theorem (Theorem 4.2), recalling that here we are dealing with a normalized loss, we get that for any $\varepsilon < 1/2$ we have (for a universal constant c_0) that:

$$R_M \le R_L \le (1 + c_0 \varepsilon) R_M + \frac{c_0}{n} \delta(\varepsilon)$$

Here:

$$\delta(\varepsilon) = \tilde{a}_{min}^{-2} \exp(-\Lambda_* \varepsilon^2 / 64)$$

$$\Lambda_* = \frac{\lambda_* / \sigma^2}{\max_{1 \le i \le (n-1)} \tilde{a}_i (1 - \lambda_* \tilde{a}_i)_+}$$

Note:

$$\tilde{a}_{min} \asymp \frac{1}{n^{1/2} R^{1/2}}$$

Hence we may bound the additive term as:

$$C_1 R$$

Note that if we can make the additive term $o(R_L)$,we will get $R_M/R_L \to 1$. One way to achieve this is (taking $\varepsilon \to 0$) to require that $R = o(R_L)$ or in other words $R = o(\sigma^{4/3} R^{1/3} n^{-2/3})$, i.e. $R = o(n^{-1}\sigma^2)$. For such shrinking radius R thus Pinsker gives that linear minimax and minimax risks are the same asymptotically.

Remark: Instead of considering a regime of shrinking radius, the same result also holds in a regime of $R >> n$, where the radius R increases at some appropriate rate compared to the sample size n. Both results are not that surprising given that we know that in the 1-dimensional bounded normal mean model in which $Z \sim \mathcal{N}(\mu, 1)$, $\mu \in [-\tau, \tau]$, the minimax risk and affine minimax risk are the same both in the regime where $\tau \to 0$ and $\tau \to \infty$.

2: A simple application of Le Cam's method

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable probability density function, and assume that there exists another density function $g : \mathbb{R}^d \to \mathbb{R}$, and a constant M such that, for all $\boldsymbol{x} \in \mathbb{R}^d$

$$\big\|\nabla f(\boldsymbol{x})\big\|_2 \le M\, g(\boldsymbol{x})\,. \tag{3}$$

We will denote by $\mathsf{P}_{\boldsymbol{\theta}}$ the probability distribution of $\boldsymbol{X} = \boldsymbol{\theta} + \boldsymbol{W}$ where $\boldsymbol{W} \sim f(\,\cdot\,)$ is noise with density f.

(*a*) Prove that, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,

$$\big\|\mathsf{P}_{\boldsymbol{\theta}_1} - \mathsf{P}_{\boldsymbol{\theta}_2}\big\|_{\mathrm{TV}} \le \frac{M}{2}\, \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2\,. \tag{4}$$

(*b*) Consider the problem of estimating $\boldsymbol{\theta} \in \Theta \equiv \mathbb{R}^d$ from data $\boldsymbol{X} \sim \mathsf{P}_{\boldsymbol{\theta}}$ under the square loss $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2$. Use the previous result to derive a lower bound on the minimax risk.

[Hint: It is sufficient to consider two priors Q_1, Q_2 given by Dirac's deltas.]

(*c*) Apply this lower bound to the case of Gaussian noise, namely to the case of f the density of the Gaussian distribution $\mathsf{N}(0, \sigma^2 \mathrm{I}_d)$. How does the result compare with the actual minimax risk?

Solution:

(*a*) We first note that $\mathsf{P}_{\boldsymbol{\theta}}$ has a density w.r.t. Lebesgue measure, namely $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{\theta})$ (i.e. we are dealing with a location family problem). Therefore:

$$
\begin{aligned}
\left\|\mathsf{P}_{\boldsymbol{\theta}_1} - \mathsf{P}_{\boldsymbol{\theta}_2}\right\|_{\mathrm{TV}} &= \frac{1}{2}\int_{\mathbb{R}^d} |f_{\boldsymbol{\theta}_1}(\boldsymbol{x}) - f_{\boldsymbol{\theta}_2}(\boldsymbol{x})| d\boldsymbol{x} \\
&= \frac{1}{2}\int_{\mathbb{R}^d} |f(\boldsymbol{x} - \boldsymbol{\theta}_1) - f(\boldsymbol{x} - \boldsymbol{\theta}_2)| d\boldsymbol{x} \\
&= \frac{1}{2}\int_{\mathbb{R}^d} |\int_0^1 \frac{d}{dt} f(\boldsymbol{x} - \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) dt| d\boldsymbol{x} \\
&= \frac{1}{2}\int_{\mathbb{R}^d} |\int_0^1 \nabla f(\boldsymbol{x} - \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2))^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) dt| d\boldsymbol{x} \\
&\leq \frac{1}{2}\int_{\mathbb{R}^d} \int_0^1 \|\nabla f(\boldsymbol{x} - \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2))\| \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\| dt d\boldsymbol{x} \\
&\leq \frac{\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|}{2}\int_{\mathbb{R}^d} \int_0^1 M g(\boldsymbol{x} - \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)) dt d\boldsymbol{x} \\
&= \frac{M}{2}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\| && \text{(by Fubini's theorem)}
\end{aligned}
$$

(b) We will directly apply Le Cam's Lemma. To this end, first note that for any $a \in \mathbb{R}^d$ we have that:

$$
||a - \boldsymbol{\theta}_1||^2 + ||a - \boldsymbol{\theta}_2||^2 \geq \frac{1}{2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2
$$

In other words we may take $d(\theta_1, \theta_2) = \frac{1}{2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2$. We want this to be $\geq 2\delta$.
Hence let us set $\delta = \frac{1}{4}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2$, where we will choose these parameters later.
Then:

$$
1 - \left\|\mathsf{P}_{\boldsymbol{\theta}_1} - \mathsf{P}_{\boldsymbol{\theta}_2}\right\|_{\mathrm{TV}} \geq 1 - \frac{M}{2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_2
$$

Le Cam gives the lower bound:

$$
\geq \frac{||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2}{8}\left(1 - \frac{M}{2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_2\right)
$$

Plugging in $||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| = \frac{4}{3M}$ we get the lower bound $\frac{2}{27M^2}$.

(c)

$$
\|\nabla f(x)\| = \frac{\|x\|}{\sigma^2(2\pi\sigma^2)^{d/2}} e^{-1/(2\sigma^2)\|x\|^2}
$$

The r.h.s. has finite integral. Letting $Z \sim \mathsf{N}(0, I_d)$, the desired bound holds with

$$
M^{-1} = \int \frac{\|x\|}{\sigma^2(2\pi\sigma^2)^{d/2}} e^{-1/(2\sigma^2)\|x\|^2} dx = \frac{1}{\sigma}\mathbb{E}[\|Z\|] \tag{5}
$$

We know $\mathbb{E}\|Z\| \approx \sqrt{d}$, so plugging this into the expression from the previous part gives:

$$
R_{\mathrm{B}}(Q) \geq \frac{2\sigma^2(\mathbb{E}[\|Z\|])^2}{27} \approx \frac{2\sigma^2}{27d}
$$

The minimax risk in the problem is $R_{\mathrm{M}} = \sigma^2 d$, so our argument recovers the correct dependence in σ^2 but not in d.

3: Some properties of distances between distributions

(a) Let $\mathsf{P} = \mathsf{P}_1 \times \mathsf{P}_2 \times \cdots \times \mathsf{P}_n$ and $\mathsf{Q} = \mathsf{Q}_1 \times \mathsf{Q}_2 \times \cdots \times \mathsf{Q}_n$ be two product-form distributions (where, for each $i \le n$, P_i, Q_i are probability measures on the same space $\mathcal{X}_i$). Show that

$$\left\|\mathsf{P} - \mathsf{Q}\right\|_{\mathrm{TV}} \le \sum_{i=1}^{n} \left\|\mathsf{P}_i - \mathsf{Q}_i\right\|_{\mathrm{TV}} . \tag{6}$$

[Hint: Start with $n = 2$. It is fine to assume that the $\mathcal{X}_i$'s are finite sets.]

(b) Prove that there cannot be a reverse Pinsker inequality. Namely, there does not exist any function $f : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ with $f(t) > 0$ for $t > 0$ such that, for any two distributions P, Q.

$$D(\mathsf{P}\|\mathsf{Q}) \le f(\|\mathsf{P} - \mathsf{Q}\|_{\mathrm{TV}}) . \tag{7}$$

(c) Assume that P and Q are probability distributions over a finite set $\mathcal{X}$, with probability mass functions p, q, and assume $\mathsf{q}(x) \ge \mathsf{q}_{\min} > 0$ for all $x \in \mathcal{X}$. Prove that there exists $g : \mathbb{R}_{\ge 0} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ with $g(t, s) > 0$ for $t, s > 0$ such that, for any two probability mass functions p, q, we have

$$D(\mathsf{P}\|\mathsf{Q}) \le g(\|\mathsf{P} - \mathsf{Q}\|_{\mathrm{TV}}, \mathsf{q}_{\min}) . \tag{8}$$

We would like the function g to be such that $\lim_{z\to 0} g(z; \mathsf{q}_{\min}) = 0$ for any $\mathsf{q}_{\min} > 0$. Give an explicit expression for the function g.

[Hint: Write $D(\mathsf{P}\|\mathsf{Q}) = \mathbb{E}_{\mathsf{Q}}(X \log X - X + 1)$, for $X = \frac{\mathrm{d}P}{\mathrm{d}Q}$.]

Solution:

(a) Consider the case where $X_1 \in \mathcal{X}_1, X_2 \in \mathcal{X}_2$ where $\mathcal{X}_i$ are finite sets. We will show the result in the case where $n = 2$, the general case follows by induction.

$$\begin{aligned}
\|P - Q\|_{TV} &= \frac{1}{2} \sum_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} |p_1(x_1)p_c(x_2) - q_1(x_1)q_2(x_2)| \\
&= \frac{1}{2} \sum_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} |(p_1(x_1) - q_1(x_1))p_2(x_2) + (q_2(x_2) - p_2(x_2))q_1(x_1)| \\
&\le \frac{1}{2} \sum_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} |p_1(x_1) - q_1(x_1)|p_2(x_2) + |q_2(x_2) - p_2(x_2)|q_1(x_1) \\
&= \|P_1 - Q_1\|_{TV} + \|P_2 - Q_2\|_{TV}
\end{aligned}$$

(b) To show this it suffices to argue that for any $v > 0$, there exist P, Q with $||P - Q||_{TV} = v$ but $D(P||Q) = \infty$. Consider $\mathcal{X} = \{1, 2, 3\}$. Let $P = v\delta_1 + (1 - v)\delta_2$ and $Q = v\delta_3 + (1 - v)\delta_2$ so that $\|P - Q\|_{TV} = v$. But $D(P||Q) = \infty$ because $Q(1) = 0$ and hence $\sum_{x\in\mathcal{X}} P(x) \log(\frac{P(x)}{Q(x)}) = \infty$.

(c) With $X = \frac{dP}{dQ}$, and using the hint, we write the KL divergence as

$$
\begin{aligned}
D(P||Q) &= \mathbb{E}_Q(X \log X - X + 1) \\
&\leq \mathbb{E}_Q(X(X-1) - X + 1) \\
&= \mathbb{E}_Q(X^2) - 2\mathbb{E}_Q(X) + 1 \\
&= \mathbb{E}_Q(X^2) - 1 \\
&= \sum_{x \in \mathcal{X}} \frac{p(x_i)^2}{q(x_i)^2} q(x_i) - 1 \\
&= \sum_{x \in \mathcal{X}} \frac{(p(x_i) - q(x_i))^2}{q(x_i)} \\
&\leq \frac{1}{\mathsf{q}_{\min}} \sum_{x \in \mathcal{X}} (p(x_i) - q(x_i))^2 \\
&\leq \frac{1}{\mathsf{q}_{\min}} \left(\sum_{x \in \mathcal{X}} |p(x_i) - q(x_i)| \right)^2 \\
&= \frac{4||P - Q||_{TV}^2}{\mathsf{q}_{\min}}
\end{aligned}
$$

Thus we may choose $g(t, s) = \frac{4t^2}{s}$ for $t, s > 0$.

References