Stat 300A Theory of Statistics

Homework 3

Andrea Montanari *Due on October 17, 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: Convex compact parameter space

Let $\mathscr{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a statistical model with $\Theta \subseteq \mathbb{R}^d$ a convex compact set, and Θ is not a singleton (Θ contains at least two points). Let $\Theta^\varepsilon = \{\boldsymbol{\theta} : \ d(\boldsymbol{\theta}, \Theta) \le \varepsilon\}$, where $d(\boldsymbol{\theta}, \Theta) \equiv \inf\{\boldsymbol{v} \in \Theta : \|\boldsymbol{v} - \boldsymbol{\theta}\|_2\}$. Assume the estimator $\hat{\boldsymbol{\theta}}$ to take values in $\mathbb{R}^d$ (i.e. the decision space is $\mathcal{A} = \mathbb{R}^d$).

(a) Consider the case of square loss $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2$. Assume that (for some $\varepsilon, \delta > 0$) $\mathsf{P}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}(\boldsymbol{X}) \notin \Theta^\varepsilon) > \delta$ for all $\boldsymbol{\theta} \in \Theta$. Prove that $\hat{\boldsymbol{\theta}}(\,\cdot\,)$ cannot be minimax optimal.

(b) Keeping to the square loss, consider now the linear model $\mathsf{P}_{\boldsymbol{\theta}} = \mathsf{N}(\boldsymbol{D}\boldsymbol{\theta}, \sigma^2 \mathrm{I}_n)$, where $\boldsymbol{D} \in \mathbb{R}^{n \times d}$ is a known design matrix, of rank d, and $\sigma^2 > 0$ is known noise variance. Prove that no affine estimator (i.e. no etimator of the form $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{M}\boldsymbol{y} + \boldsymbol{\theta}_0$) can be minimax optimal.

(c) Produce a counter-example showing that the conclusion at point (a) does no longer hold if Θ is not convex.

(d) Consider the case $d = 1$, $\Theta = [\theta_{\min}, \theta_{\max}]$, and assume that L is continuous, with $a \mapsto L(a, \theta)$ is strictly decreasing for $a < \theta$, and strictly increasing for $a > \theta$. Assume that (for some $\varepsilon, \delta > 0$) $\mathsf{P}_\theta(\hat{\theta}(\boldsymbol{X}) \notin \Theta^\varepsilon) > \delta$ for all $\theta \in \Theta$, and that the risk function $\theta \mapsto R(\hat{\theta}; \theta)$ is continuous. Prove that $\hat{\theta}$ cannot be minimax optimal.

What can you conclude if $a \mapsto L(a, \theta)$ is decreasing (but not necessarily strictly decreasing) for $a < \theta$ and increasing (but not necessarily strictly decreasing) for $a > \theta$.

2: On the minimax estimator of a binomial parameter

Let $X \sim \mathsf{P}_\theta = \mathrm{Binom}(n, \theta)$, where $\theta \in \Theta = [0, 1]$, and we consider the square loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. Recall that a minimax estimator is given by

$$\hat{\theta}_{\mathrm{MM}}(X) = \frac{\sqrt{n}}{1 + \sqrt{n}} \cdot \frac{X}{n} + \frac{1}{1 + \sqrt{n}} \cdot \frac{1}{2}\,. \tag{1}$$

We know already that this is Bayes optimal with respect to the prior distribution $\mathsf{Q} = \mathrm{Beta}(\sqrt{n}/2, \sqrt{n}/2)$.

(a) Consider the case $n = 1$. Construct a two points prior $\mathsf{Q} = q\delta_{\theta_1} + (1 - q)\delta_{\theta_2}$ whose Bayes optimal estimator coincides with $\hat{\theta}_{\mathrm{MM}}$.

(*b*) Show that, for any n, there exists a prior supported on m number of points for some integer m, whose Bayes estimators coincides with $\hat{\theta}_{\rm MM}$.

[You can assume that the linear system $\sum_{i=0}^m q_i(i/m)^k = \int \theta^k \mathsf{Q}(\mathrm{d}\theta)$, $k \in \{0,\dots,n+1\}$ has a solution $\mathbf{q} = (q_0,\dots,q_m) \geq 0$ for m large enough. (Here $\mathsf{Q} = \mathrm{Beta}(\sqrt{n}/2, \sqrt{n}/2)$.)]

3: Minimax estimation of sparse vectors

Let $\Theta \subseteq \mathbb{R}^d$ and consider estimation with a loss $L : \mathcal{A}\times\mathbb{R}^d \to \mathbb{R}_{\geq 0}$ upper bounded by L_0: $\sup_{a\in\mathcal{A},\boldsymbol{\theta}\in\Theta} L(a,\boldsymbol{\theta}) \leq L_0$.

(*a*) Prove that, for any probability distribution Q on $\mathbb{R}^d$,

$$R_{\rm M}(\Theta) \geq R_{\rm B}(\mathsf{Q}) - L_0\, \mathsf{Q}(\Theta^c)\,, \tag{2}$$

where $\mathsf{Q}(\Theta^c) = \int_{\Theta^c} \mathsf{Q}(\mathrm{d}\boldsymbol{\theta})$ is the probability of Θ^c under Q, and $R_{\rm B}(\mathsf{Q}) = \int_{\mathbb{R}^d} R(A;\boldsymbol{\theta})\,\mathsf{Q}(\mathrm{d}\boldsymbol{\theta})$. (Here we assume that $\mathsf{P}_{\boldsymbol{\theta}}$ is not only defined for $\boldsymbol{\theta} \in \Theta$, but for any $\boldsymbol{\theta} \in \mathbb{R}^d$.)

Given two integers $1 \leq k \leq d$, and a real number $M \geq 0$, define the set of vectors

$$\Theta(d,k;M) = \Big\{\boldsymbol{\theta} \in \{0,+M,-M\}^d :\ \|\boldsymbol{\theta}\|_0 \leq k\Big\}\,, \tag{3}$$

where $\|\boldsymbol{\theta}\|_0 = |\mathrm{supp}(\boldsymbol{\theta})|$ is the number of non-zero entries in $\boldsymbol{\theta}$. We we are interested in the minimax error for the Gaussian location model with this parameters space $\mathscr{P} = \{\mathsf{P}_{\boldsymbol{\theta}} :\ \boldsymbol{\theta} \in \Theta(d,k;M)\}$, action space $\mathbb{R}^d$, and square loss $L(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2$. We will denote this minimax risk by $R_{\rm M}(d,k;M)$.

(*b*) Prove that, in determining the minimax error, we can restrict ourselves to estimators that take values in $\mathcal{A} = B^d(\mathbf{0};M\sqrt{k}) = \{\boldsymbol{\theta} \in \mathbb{R}^d :\ \|\boldsymbol{\theta}\|_2 \leq M\sqrt{k}\}$. Further, we can replace the square loss by $\tilde{L}(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) = \min(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2; 4M^2k)$

(*c*) Prove that there exists a least favorable prior Q_*, and that it can be taken of the form

$$\mathsf{Q}_* = \sum_{\ell=0}^k p_\ell \mathsf{Q}_\ell \tag{4}$$

where $p = (p_\ell)_{0\leq\ell\leq k}$ is a probability distribution over $\{0,1,\dots,k\}$, and Q_ℓ is the uniform distribution over vectors in $\boldsymbol{\theta} \in \Theta(d,k;M)$ with $\|\boldsymbol{\theta}\|_0 = \ell$.

[Hint: Note that this claim is equivalent to $\mathsf{Q}_*(\{\boldsymbol{\theta}_1\}) = \mathsf{Q}_*(\{\boldsymbol{\theta}_2\})$, for any $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \Theta(d,k;M)$ with $\|\boldsymbol{\theta}_1\|_0 = \|\boldsymbol{\theta}_2\|_0$.]

Computing the Bayes risk for the prior Q_* described above is somewhat intricate. We thus consider a simpler prior $\mathsf{Q}_{M,\varepsilon}$. Under $\mathsf{Q}_{M,\varepsilon}$ the coordinates of $\boldsymbol{\theta}$ are independent with $\mathsf{Q}_{M,\varepsilon}(\{\theta_i = M\}) = \mathsf{Q}_{M,\varepsilon}(\{\theta_i = -M\}) = \varepsilon/2$, and $\mathsf{Q}_{M,\varepsilon}(\{\theta_i = 0\}) = 1-\varepsilon$. Equivalently $\mathsf{Q}_{M,\varepsilon} = \mathsf{q}_{M,\varepsilon} \times \cdots \times \mathsf{q}_{M,\varepsilon}$, where $\mathsf{q}_{M,\varepsilon}$ is the three points distribution $\mathsf{q}_{M,\varepsilon} = (1-\varepsilon)\delta_0 + (\varepsilon/2)\delta_M + (\varepsilon/2)\delta_{-M}$.

(*d*) Prove that

$$R_{\rm M}(d,k;M) \geq \tilde{R}_{\rm B}(\mathsf{Q}_{M,\varepsilon}) - 4M^2k\,\mathbb{P}\Big(\mathrm{Binom}(d,\varepsilon) > k\Big)\,. \tag{5}$$

where $\tilde{R}_{\rm B}$ is the Bayes risk for the loss function $\tilde{L}$.

Setting $\varepsilon = (k/d)(1-\eta)$, it is possible to show (for instance by Bernstein inequality [BLM13]) that

$$\mathbb{P}\Big(\text{Binom}(d,\varepsilon) > k\Big) \le e^{-k\eta^2/4}\,. \tag{6}$$

Let $R_{\rm B}$ denote the Bayes risk for the square loss. It is also possible to show that

$$\tilde{R}_{\rm B}(\mathsf{Q}_{M,\varepsilon}) \ge R_{\rm B}(\mathsf{Q}_{M,\varepsilon}) - (M^2+1)\, o_\eta(k)\,, \tag{7}$$

where $o_\eta(k)$ is a quantity such that $\lim_{k\to\infty} o_\eta(k)/k = 0$ for any $\eta > 0$.

(e) Prove that the above implies implies

$$R_{\rm M}(d,k;M) \ge d\, R_{\rm B}(\mathsf{q}_{M,\varepsilon}) - (M^2+1) o_\eta(k)\,. \tag{8}$$

where $R_{\rm B}(\mathsf{q}_{M,\varepsilon})$ is the Bayes risk for the one-dimensional problem of estimating $\theta \sim \mathsf{q}_{M,\varepsilon}$ from $X = \theta + Z$, $Z \sim \mathsf{N}(0,1)$.

Optional

This question will not be graded and is mainly food for thought:

- Continuing from te previous problem, what is the behavior of $R_{\rm B}(\mathsf{q}_{M,\varepsilon})$ with ε and M? What are the consequences for $R_{\rm M}(d,k;M)$? Of particular interest is the regime $\varepsilon \ll 1$ (corresponding to $k \ll d$).

References

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.