Stats300b Étude 3 Solution

Due: Thursday, February 4 at 5:00pm on Gradescope.

Question 7.14 (Uniform convergence in a quantile regression problem)**:** Let pairs $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ and for a fixed $q \in (0,1)$ consider the "pinball" loss

$$\ell(\theta, z) = \ell(\theta, x, y) = q \left[\theta^T x - y\right]_+ + (1-q)\left[y - \theta^T x\right]_+ - q\left[-y\right]_+ - (1-q)\left[y\right]_+, \qquad (7.1)$$

where $[a]_+ = \max\{a, 0\}$ is the positive part of its argument. (One uses this loss to fit models that predict quantiles.) Define the population expectation $L(\theta) := \mathbb{E}_P[\ell(\theta, X, Y)]$.

(a) Show that if $\Theta \subset \mathbb{R}^d$ is compact and $\mathbb{E}[\|X\|] < \infty$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$, then

$$\sup_{\theta \in \Theta} |P_n \ell(\theta, X, Y) - L(\theta)| \xrightarrow{p} 0.$$

(b) Explain why we must normalize the losses (7.1) by subtracting $q[-y]_+ + (1-q)[y]_+$ to achieve the preceding convergence. (This should only take a sentence or two.)

Now, we derive asymptotics of the empirical minimizer $\widehat{\theta}_n$ of $L_n(\theta) := P_n\ell(\theta, X, Y)$, that is, $\widehat{\theta}_n \in \operatorname{argmin}_{\theta \in \Theta} L_n(\theta)$. You may use the following result:

Lemma 7.14.1 (Bertsekas [1])**.** *If $H : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is a convex function with $\mathbb{E}_P[|H(\theta, Z)|] < \infty$ and $\nabla_\theta H(\theta, x)$ exists for P-almost all x, then $h(\theta) := \mathbb{E}_P[H(\theta, Z)]$ is differentiable with gradient*

$$\nabla h(\theta) = \mathbb{E}_P[\nabla H(\theta, Z)] = \int \nabla H(\theta, z) dP(z) = \int_{z \in \mathcal{Z} : \nabla H(\theta, z) \text{ exists}} \nabla H(\theta, z) dP(z).$$

Assume that conditional on $X = x$, the random variable Y has cumulative distribution $F_x(\cdot)$ with continuous bounded positive density $f_x(\cdot)$ on $\mathbb{R}$. Assume additionally that $\operatorname{Cov}(X) \succ 0$, that is, X has full rank covariance with $\mathbb{E}[\|X\|_2^2] < \infty$, and that the population minimizer $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} L(\theta) \in \operatorname{int} \Theta$.

(c) Show that $\widehat{\theta}_n$ is consistent for $\theta^\star$. *Hint:* argue that the Hessian $\nabla^2 L(\theta)$ is positive definite in a neighborhood of $\theta^\star$. Then apply van der Vaart [2, Thm. 5.7]. You may assume you can exchange the order of expectation and differentiation in any integrals you desire. (It is possible to use dominated convergence to prove this valid in any case.)

(d) Show that

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \mathsf{N}\left(0, \nabla^2 L(\theta^\star)^{-1} \operatorname{Cov}(\nabla \ell(\theta^\star, X, Y)) \nabla^2 L(\theta^\star)^{-1}\right).$$

In addition, express the covariance $\operatorname{Cov}(\nabla \ell(\theta^\star, X, Y))$ and Hessian $\nabla^2 L(\theta^\star)$ in terms of expectations involving q and the random variables X, $f_X(\langle \theta^\star, X\rangle)$, and $F_X(\langle \theta^\star, X\rangle)$. *Hint:* you may use van der Vaart [2, Thm. 5.23] to show the claimed convergence.

(e) Suppose there exists $\theta_0 \in \operatorname{int} \Theta$ such that

$$F_x(\theta_0^T x) = 1 - q$$

for P-almost all x, and that the density $f_x(\theta_0^T x) = \rho > 0$ for P-almost all x. This would occur, for example, in the model

$$Y = \langle \beta^\star, X \rangle + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$$

so long as x includes the intercept term that $x_1 = 1$ (feel free to convince yourself of this!). Show that your result in part (d) simplifies to

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \mathsf{N}\left(0, \frac{q(1-q)}{\rho^2} \mathbb{E}[XX^T]^{-1}\right).$$

Answer:

(a) This we essentially did in class. Let $\epsilon > 0$ be arbitrary, and let $\mathcal{C} = \{\theta^1, \ldots, \theta^N\}$ be an ϵ-cover of Θ of size N for the ℓ_2-norm. Then define the bracketing functions

$$l_i(x, y) = q\left[\langle \theta^i, x\rangle - y\right]_+ + (1-q)\left[y - \langle \theta^i, x\rangle\right]_+ - q\left[-y\right]_+ - (1-q)\left[y\right]_+ - \epsilon \left\|x\right\|_2$$
$$u_i(x, y) = q\left[\langle \theta^i, x\rangle - y\right]_+ + (1-q)\left[y - \langle \theta^i, x\rangle\right]_+ - q\left[-y\right]_+ - (1-q)\left[y\right]_+ + \epsilon \left\|x\right\|_2 .$$

Then as $t \mapsto [t]_+$ is 1-Lipschitz, it is evident that for any $\theta \in \Theta$, there exists θ^i such that $\|\theta^i - \theta\|_2 \le \epsilon$, and for this i, we have

$$l_i(x, y) \le \ell(\theta, x, y) \le u_i(x, y) \ \text{ while } \ 0 \le u_i(x, y) - l_i(x, y) = 2\epsilon \left\|x\right\|_2 .$$

Evidently the class of functions has finite bracketing number, so van der Vaart [2, Thm. 19.4] gives the result.

(b) If we do not normalize the losses, consider the case that Y has Cauchy distribution while X is a point mass at 0. Then the expectation $L(\theta)$ is not defined as $\mathbb{E}[|Y|] = +\infty$.

(c) We compute the gradient and Hessian of the population loss near $\theta = \theta^\star$. We use Lemma 7.14.1, which gives

$$\nabla L(\theta) = q\mathbb{E}[\mathbf{1}\left\{\theta^T X - Y \ge 0\right\} X] - (1-q)\mathbb{E}[\mathbf{1}\left\{\theta^T X - Y \le 0\right\} X],$$

which follows by the assumption that Y has a density. Now, we note that

$$\mathbb{E}[\mathbf{1}\left\{\theta^T X - Y \ge 0\right\} X \mid X = x] = \mathbb{P}(Y \le \theta^T X \mid X = x) = F_x(\theta^T x)$$

and $\nabla_\theta F_x(\theta^T x) = f_x(\theta^T x) x$, and similarly,

$$\mathbb{P}(Y \ge \theta^T X \mid X = x) = 1 - F_x(\theta^T x),$$

so that

$$\nabla^2 L(\theta) = q\mathbb{E}[f_X(\theta^T X) X X^T] + (1-q)\mathbb{E}[f_X(\theta^T X) X X^T] = \mathbb{E}[f_X(\theta^T X) X X^T].$$

As the density f_X is assumed positive, we have

$$\nabla^2 L(\theta^\star) = \mathbb{E}[f_X(\langle \theta^\star, X\rangle) X X^T] \succ 0.$$

Now, we argue that consistency $\widehat{\theta}_n \xrightarrow{p} \theta^\star$ holds. That L is convex and f is bounded and continuous implies (e.g., by dominated convergence) that there is some $\lambda > 0$ such that $\nabla^2 L(\theta) \succeq \lambda I$ for all θ in a neighborhood of $\theta^\star$. In particular, for some $c > 0$ we have

$$L(\theta) \ge L(\theta^\star) + \frac{\lambda}{2} \min\{\|\theta - \theta^\star\|_2^2, c\left\|\theta - \theta^\star\right\|_2\}$$

by Question 2.5. Using the uniform convergence result that $\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \xrightarrow{p} 0$ allows us to apply van der Vaart [2, Thm. 5.7] immediately to give $\widehat{\theta}_n \xrightarrow{p} \theta^\star$.

(d) With consistency assured, we verify the conditions of [2, Thm. 5.23]. By construction and Lipschitz continuity we have $|\ell(\theta, x, y) - \ell(\theta', x, y)| \le \|\theta - \theta'\|_2 \|x\|_2$, and $\mathbb{E}[\|X\|_2^2] < \infty$. We have already shown the second-order Taylor expansion required in the theorem, that is, $\nabla^2 L(\theta) = \mathbb{E}[f_X(\theta^T X) X X^T] \succ 0$ in a neighborhood of $\theta^\star$. Thus, by the theorem,

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) = -\nabla^2 L(\theta^\star)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \ell(\theta^\star, X_i, Y_i) + o_P(1).$$

As

$$\mathbb{E}[\mathbf{1}\{Y \le \langle \theta^\star, X \rangle\} X X^T] = \mathbb{E}[F_X(\langle \theta^\star, X \rangle) X X^T] \text{ and}$$
$$\mathbb{E}[\mathbf{1}\{Y \ge \langle \theta^\star, X \rangle\} X X^T] = \mathbb{E}[(1 - F_X(\langle \theta^\star, X \rangle)) X X^T],$$

we have $\mathrm{Cov}(\nabla \ell(\theta^\star, X, Y)) = \mathbb{E}[(q^2 F_X(\langle \theta^\star, X \rangle) + (1-q)^2 (1 - F_X(\langle \theta^\star, X \rangle))) X X^T]$. This implies

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \mathsf{N}\left(0, \nabla^2 L(\theta^\star)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, X, Y)) \nabla^2 L(\theta^\star)^{-1}\right).$$

We have evidently written each of these in terms of q, $F_X(\langle \theta^\star, X \rangle)$, X, and $f_X(\langle \theta^\star, X \rangle)$ as desired.

(e) If there is a point θ_0 such that $F_X(\langle \theta_0, X \rangle) = (1 - q)$ with P-probability 1 over X, we may simplify our expressions. First, we must have $\theta^\star = \theta_0$, as $\mathbb{E}[\nabla \ell(\theta_0, X, Y)] = q(1-q)\mathbb{E}[X] - (1-q)q\mathbb{E}[X] = 0$. The covariance becomes

$$\mathrm{Cov}(\nabla \ell(\theta^\star, X, Y)) = \mathbb{E}[q^2(1-q) X X^T + (1-q)^2 q X X^T] = \mathbb{E}[q(1-q) X X^T],$$

while the second derivative becomes $\nabla^2 \ell(\theta^\star) = \rho \mathbb{E}[X X^T]$, giving the claimed result.

□

References

[1] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

[2] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.