Statistical Theory MATH-442
Spring 2025 Ho Yun & Yoav Zemel

Exercise sheet 3

Exercise 1 Suppose that $\Theta \subseteq \mathbb{R}^d$ and $X_1, \dots, X_n \overset{iid}{\sim} F_{\theta_0}$ for some $\theta_0 \in \Theta$, and

1. θ_0 is the unique maximiser of the **continuous** function ℓ.
2. For all M, $\sup_{\|\theta\| \le M} |\overline{\ell}_n(\theta) - \ell(\theta)| \xrightarrow{p-P_{\theta_0}} 0$.
3. For any $\epsilon > 0$ there exists $M_\epsilon < \infty$ such that $\sup_n \mathbb{P}_{\theta_0}(\|\widehat{\theta}_n^{MLE}\| > M_\epsilon) < \epsilon$.

Show that $\widehat{\theta}_n^{MLE} \xrightarrow{p-P_{\theta_0}} \theta_0$. **Hint:** first show an inequality of the form $P_{\theta_0}(\|\widehat{\theta}_n^{MLE} - \theta_0\| > \epsilon) \le 2\epsilon$ for all $\epsilon > 0$ and all $n \ge N_\epsilon$ large. Then show that this implies the convergence in probability.

Solution 1 Fix $\epsilon > 0$ and let M_ϵ as in 3. We may assume that $\|\theta_0\| \le M_\epsilon$ (otherwise increase M_ϵ). Since θ_0 is the unique maximiser of the continuous function ℓ on the compact set $\{\|\theta\| \le M_\epsilon\}$, we have

$$\delta = \delta(\epsilon) := \sup_{\|\theta\| \le M_\epsilon, \|\theta - \theta_0\| \ge \epsilon} \ell(\theta_0) - \ell(\theta) > 0.$$

Let A_n^ϵ be the event that $\|\widehat{\theta}_n^{MLE}\| \le M_\epsilon$ and $\sup_{\|\theta\| \le M_\epsilon} |\overline{\ell}_n(\theta) - \ell(\theta)| \le \delta/2$. For n large we have $P_{\theta_0}(A_n^\epsilon) > 1 - 2\epsilon$. When A_n^ϵ holds we have

$$\begin{aligned} 0 \le \ell(\theta_0) - \ell(\widehat{\theta}_n^{MLE}) = \ell(\theta_0) - \overline{\ell}_n(\widehat{\theta}_n^{MLE}) + \overline{\ell}_n(\widehat{\theta}_n^{MLE}) - \ell(\widehat{\theta}_n^{MLE}) &\le \ell(\theta_0) - \overline{\ell}_n(\theta_0) + \overline{\ell}_n(\widehat{\theta}_n^{MLE}) - \ell(\widehat{\theta}_n^{MLE}) \\ &\le 2 \sup_{\|\theta\| \le M_\epsilon} |\overline{\ell}_n(\theta) - \ell(\theta)| \le \delta. \end{aligned}$$

By definition of δ, this entails $\|\widehat{\theta}_n^{MLE} - \theta_0\| < \epsilon$. Therefore for n large

$$P_{\theta_0}(\|\widehat{\theta}_n^{MLE} - \theta_0\| > \epsilon) \le 1 - P_{\theta_0}(A_n^\epsilon) < 2\epsilon.$$

To show that this gives convergence in probability let $\rho > 0$ be arbitrary and let $\epsilon \le \rho$ be arbitrary. Then

$$\limsup_{n \to \infty} P_{\theta_0}(\|\widehat{\theta}_n^{MLE} - \theta_0\| > \rho) \le \limsup_{n \to \infty} P_{\theta_0}(\|\widehat{\theta}_n^{MLE} - \theta_0\| > \epsilon) \le 2\epsilon.$$

Since $\epsilon \le \rho$ is arbitrary we get $\limsup_{n \to \infty} P_{\theta_0}(\|\widehat{\theta}_n^{MLE} - \theta_0\| > \rho) = 0$, and since ρ is arbitrary this yields the convergence in probability.

Exercise 2 The second part of this question is not for the exam.

(a) **(equivariance of maximum likelihood estimators)**. Consider a model F_θ with $\theta \in \Theta$ and let $h : \Theta \to h(\Theta)$ be injective. Define $\phi = h(\theta)$ and consider the model $G_\phi = F_{h^{-1}(\phi)}$. Show that $\widehat{\phi}^{MLE} = h(\widehat{\theta}^{MLE})$.

(b) **(*invariance of maximum likelihood estimator with respect to the dominating measure)** Recall that f_θ is the Radon–Nikodym derivative of F_θ with respect to a σ-finite measure μ. Suppose that μ' is another measure that dominates μ, and that we replace f_θ by the $g_\theta = \partial dF_\theta / \partial \mu'$. Show that this does not change the maximum likelihood estimator. Deduce that if μ'' is another measure that dominates all the F_θ (but not necessarily μ), then the maximum likelihood estimators with respect to μ and with respect to μ'' are the same. **Hint:** recall that $g_\theta(x) = f_\theta(x)h(x)$, where $h(x) = \partial d\mu(x)/\partial d\mu'(x)$ is the Radon–Nikodym derivative and does not depend on θ, and $\mu(\{x : h(x) = 0\} = 0$.

Solution 2

(a) Maximising $\prod g(x_i; \phi) = \prod f(x_i, h^{-1}(\phi))$ with respect to ϕ is equivalent to maximising $\prod f(x_i, \theta)$ over θ, since h is bijective from Θ to $h(\Theta)$.

(b) Maximising $\prod g_\theta(x_i)$ is the same as maximising $\prod f_\theta(x_i)$ whenever all $h(x_i) > 0$. But this has probability one for any θ, because $\mu\{x : h(x) = 0\} = 0$ and μ dominating all the F_θ's implies that $P_\theta(h(X) = 0) = 0$ for all $\theta \in \Theta$. So with probability one, the maximisation problems are equivalent.

For μ'', we let $\mu' = \mu'' + \mu$ which dominates both μ and μ''. Then the likelihood problem with respect to μ' is equivalent to that with respect to μ as well as that with respect to μ''. Thus the latter two are equivalent.

Exercise 3 We say that a model $f(x;\theta)$ is a k-parameter exponential family in natural parametrisation if $\Theta \subseteq \mathbb{R}^k$ and

$$f(x;\theta) = \exp\left(\sum_{j=1}^{k} \theta_i T_j(x) - \gamma(\theta) + s(x)\right).$$

Assume that Θ has a nonempty interior and that the covariance matrix of $T(X) = (T_1(X), \ldots, T_k(X))$ is nonsingular for all $\theta \in int(\Theta)$.

Find the maximum likelihood estimator for θ based on a sample $X_1, \ldots, X_n$ from $f(x;\theta_0)$ with $\theta_0 \in int(\Theta)$ and show that it is consistent and asymptotically normal. You may use without proof that $\mathbb{E}_{\theta_0} T(X) = \nabla\gamma(\theta_0)$ and $var_{\theta_0} T(X) = \nabla^2\gamma(\theta_0)$. **Hint:** use the delta method and the inverse function theorem.

Solution 3 We have

$$\begin{aligned}
\ell_n(\theta) &= -n\gamma(\theta) + \sum_{i=1}^{n} s(x_i) + \sum_{i=1}^{n}\sum_{j=1}^{k} \theta_j T_j(x_i) \\
\ell_n'(\theta) &= -n\nabla\gamma(\theta) + \sum_{i=1}^{n} T(x_i) = 0 \iff \overline{T} = \nabla\gamma(\theta) \\
\nabla^2\ell_n(\theta) &= -n\nabla^2\gamma(\theta) = -n var_\theta T(X) \prec 0
\end{aligned}$$

where the last symbol means negative definite. Therefore ℓ_n is strictly concave. Similarly, ℓ is strictly concave and has a unique maximiser which we know is θ_0. Therefore the maximum likelihood estimator is consistent, so ℓ_n attains its unique maximum close to θ_0, and certainly in the interior of Θ. Thus $\widehat{\theta}_n^{MLE}$ is the unique solution of $\ell_n' = 0$, i.e., $(\nabla\gamma)^{-1}(\overline{T})$.

By the central limit theorem

$$\sqrt{n}(\overline{T} - \nabla\gamma(\theta_0)) = \sqrt{n}(\overline{T} - \mathbb{E}T(X)) \xrightarrow{d} N(0, var_{\theta_0}(T(X)) = N(0, \nabla^2\gamma(\theta_0)).$$

Since $\nabla(\nabla\gamma)(\theta_0) = \nabla^2\gamma(\theta_0) \prec 0$ is nonsingular, the inverse function theorem implies that the function $g(t) = (\nabla\gamma)^{-1}(t)$ is differentiable at $t = \nabla\gamma(\theta_0)$ with derivative $[\nabla^2(\nabla\gamma^{-1}(\nabla\gamma(\theta_0))]^{-1}$ at θ_0 and the delta method shows that

$$\sqrt{n}(\widehat{\theta}_n^{MLE} - \theta_0) \to [\nabla^2\gamma(\theta_0)]^{-1} N(0, \nabla^2\gamma(\theta_0)) = N(0, [\nabla^2\gamma(\theta_0)]^{-1})$$

Exercise 4

(a) Let G be any absolutely continuous distribution on the positive real line and let F_λ be the $\text{Exp}(\lambda)$ distribution. Find the value of $\lambda > 0$ that minimises the KL divergence between G and F_λ? Under which condition is λ unique?

(b) Now let G be an arbitrary distribution on $\mathbb{R}$. Find the values of μ and σ^2 that minimise $\text{KL}(G, N(\mu, \sigma^2))$. Under which conditions are these values unique? Can you generalise this to higher dimensions?

Solution 4

(a) Let g be the density of G. Then

$$KL(G, P_\lambda) = \int g \log \tfrac{g}{f_\lambda} = \int g \log g - \int g(x)[\log \lambda - \lambda x]dx = \int g \log g - \log \lambda + \lambda \mathbb{E}_G(X)$$

is minimised when $\lambda = 1/\mathbb{E}_G(X)$, provided all the expressions above are finite. Therefore, the conditions on G are that $\mathbb{E}_G(X) < \infty$ and $\mathbb{E}_G|\log g(X)|$ is finite.

(b) Here we have

$$KL(G, N(\mu, \sigma^2)) = \int g \log g - \int g \log \tfrac{1}{\sqrt{2\pi\sigma^2}} + \int g(x)(x-\mu)^2/2\sigma^2 \; dx = \int g \log g + \tfrac{1}{2} \log 2\pi\sigma^2 + \tfrac{\mathbb{E}_G(X-\mu)^2]}{2\sigma^2}$$

Assuming again $\int |g \log g| < \infty$, this is minimised $\mu = \mathbb{E}_G(X)$ and $\sigma^2 = var_G(X)$. The same holds in higher dimension, but the proof requires differentiation of matrix-valued function and that is beyond the scope of this course.

Exercise 5 Suppose that $S \sim$Exp(λ) and $C \sim$Exp(γ) are independent and define $T = \min(S, C)$ and $D = 1[T = S]$. Assume we have independent and identically distributed observations (T_i, D_i), $i = 1, \ldots, n$. Find the maximum likelihood estimator of the vector $(\lambda, \gamma)^\top$ and show that it is consistent and asymptotically normal. It is **not** required to compute the asymptotic covariance matrix.

Solution 5 To understand the joint distribution of (T, D) it suffices to compute for $t \geq 0$

$$\mathbb{P}_{\lambda,\gamma}(T \leq t, D = 1) = \mathbb{P}_{\lambda,\gamma}(S \leq t, S \leq C) = \int_0^t \lambda e^{-\lambda s} ds \int_s^\infty \gamma e^{-\gamma x} dx = \lambda \int_0^t e^{-\lambda s} e^{-\gamma s} ds$$

so that the joint density is $\lambda e^{-(\lambda+\gamma)t}$ at $(t, 1)$ and $\gamma e^{-(\lambda+\gamma)t}$ at $(t, 0)$. The log likelihood is

$$\ell_n(\lambda, \gamma) = \sum_{i=1}^n 1(D_i = 1) \log \lambda + 1(D_i = 0) \log \gamma - (\lambda+\gamma) T_i = n\overline{D}_n \log \lambda + n(1-\overline{D}_n) \log \gamma - n(\lambda+\gamma)\overline{T}_n.$$

This function is strictly concave, so the unique maximum likelihood estimator is $\widehat{\lambda}_n^{MLE} = \overline{D}_n / \overline{T}_n \in [0, \infty)$ and $\widehat{\gamma} = (1 - \overline{D}_n)/\overline{T}_n \in [0, \infty)$.

To understand the asymptotic distribution observe that

$$D \sim B\left(\tfrac{\lambda}{\lambda+\gamma}\right), \qquad T \sim Exp(\lambda + \gamma), \qquad cov_{\lambda,\gamma}(D, T) = 0$$

(in fact, D and T are independent), so that

$$\sqrt{n}\left(\begin{pmatrix}\overline{D}_n \\ \overline{T}_n\end{pmatrix} - \begin{pmatrix}\lambda \\ 1\end{pmatrix}/(\lambda+\gamma)\right) \xrightarrow{d} N(0, \Sigma), \qquad \Sigma = (\lambda+\gamma)^{-2} \begin{pmatrix}\lambda\gamma & \\ & 1\end{pmatrix}$$

Use the delta method with the function $g(x, y) = (x, 1 - x)/y$ defined on $[0, 1] \times (0, \infty)$ to obtain

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\lambda}_n^{MLE} \\ \widehat{\gamma}_n^{MLE}\end{pmatrix} - \begin{pmatrix}\lambda \\ \gamma\end{pmatrix}\right) \xrightarrow{d} N(0, J\Sigma J^\top), \qquad J = Dg(\lambda/(\lambda+\gamma), 1/(\lambda+\gamma))$$