Stat 300A Theory of Statistics

Homework 1

Andrea Montanari *Due on October 3, 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: Properties of exponential families

Recall that an exponential family in canonical form is a class of probability measures on $\mathbb{R}^n$, taking the form

$$\mathsf{P}_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})}\, \exp\big\{\langle \boldsymbol{\theta}, \boldsymbol{T}(\boldsymbol{x})\rangle\big\}\, \nu(\mathrm{d}\boldsymbol{x})\,, \tag{1}$$

where $\nu(\mathrm{d}\boldsymbol{x})$ is a reference measure on $\mathbb{R}^n$. For the purpose of this problem, you can assume that $\mathsf{P}_{\boldsymbol{\theta}}$ has a density with respect to the Lebesgue measure on $\mathbb{R}^n$, which therefore can be written as

$$\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})}\, \exp\big\{\langle \boldsymbol{\theta}, \boldsymbol{T}(\boldsymbol{x})\rangle\big\}\, h(\boldsymbol{x})\,, \tag{2}$$

where $h : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a measurable function. Alternatively, you can assume that $\mathsf{P}_{\boldsymbol{\theta}}$ is supported on $\mathbb{Z}^d$, with probability mass function of the form (2). Recall that the log partition function is defined as $\phi(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$, which is finite for $\boldsymbol{\theta} \in \Theta_N$ (the natural parameter space). (In the following, you are not required to justify the exchange of order of derivative and integrals.)

(a) Prove that Θ_N is convex and $\phi : \Theta_N \to \mathbb{R}$ is a convex function.

(b) Prove the following identities hold for $\boldsymbol{\theta} \in \Theta_N^\circ$ (the interior of Θ_N)

$$\frac{\partial \phi}{\partial \theta_i}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\{T_i(\boldsymbol{X})\}\,, \tag{3}$$

$$\frac{\partial^2 \phi}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}\{T_i(\boldsymbol{X}); T_j(\boldsymbol{X})\}\,. \tag{4}$$

(c) Assume that $\boldsymbol{x} \mapsto h(\boldsymbol{x})$, $\boldsymbol{x} \mapsto \boldsymbol{T}(\boldsymbol{x})$ are differentiable. Prove that the following identity (Stein's identity) hold for any differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ such that both sides make sense:

$$\mathsf{E}_{\boldsymbol{\theta}}\left\{\left[\frac{1}{h(\boldsymbol{x})}\frac{\partial h}{\partial x_i}(\boldsymbol{x}) + \langle \boldsymbol{\theta}, \frac{\partial \boldsymbol{T}}{\partial x_i}(\boldsymbol{x})\rangle\right] g(\boldsymbol{x})\right\} + \mathsf{E}_{\boldsymbol{\theta}}\left\{\frac{\partial g}{\partial x_i}(\boldsymbol{x})\right\} = 0 \tag{5}$$

(d) Assume that p is a multivariate Gaussian density, namely

$$\mathsf{p}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\mathrm{det}(\boldsymbol{\Sigma})^{1/2}}\, \exp\left\{-\frac{1}{2}\langle (\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\rangle\right\}\,. \tag{6}$$

Show that Stein's identity in this case reduces to

$$\mathbb{E}\big\{(\boldsymbol{x}-\boldsymbol{\mu})\, g(\boldsymbol{x})\big\} = \boldsymbol{\Sigma}\mathbb{E}\big\{\nabla g(\boldsymbol{x})\big\}\,. \tag{7}$$

2: Exercises on sufficient statistics

(a) Consider a statistical model composed of k probability distributions. Namely

$$\mathscr{P} = \left\{\mathsf{p}_1, \mathsf{p}_2, \dots, \mathsf{p}_k\right\}, \tag{8}$$

where p_ℓ are densities on $\mathbb{R}^n$ (we identify the probability distribution with its density).

Show that there exists a set of $k-1$ sufficient statistics.

(b) Let $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R}^2$ with $\theta_1 < \theta_2$ and define $\mathsf{P}_{\boldsymbol{\theta}}$ to be the distribution of n i.i.d. random variables $X_1, \dots, X_n \sim \mathsf{Unif}([\theta_1, \theta_2])$. Let $x_{\min} = \min(x_1, \dots, x_n)$, and $x_{\max} = \max(x_1, \dots, x_n)$. Prove that $(x_{\min}, x_{\max})$ is a sufficient statistics for the model $\mathscr{P} = (\mathsf{P}_{\boldsymbol{\theta}})$.

(c) Consider the Gaussian linear model. Namely, for a fixed design matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, we have $\mathsf{P}_{\boldsymbol{\theta}} = \mathsf{N}(\boldsymbol{A}\boldsymbol{\theta}, \sigma^2 \mathrm{I}_n)$, $\boldsymbol{\theta} \in \Theta = \mathbb{R}^d$. Show that there exists a sufficient statistic of dimension d.

3: Optimal linear estimation in heteroscedastic Gaussian model

Assume $\sigma_1, \dots, \sigma_d > 0$ to be known, and consider the statistical model $\mathsf{P}_\theta = \mathsf{N}(\theta\mathbf{1}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$, and $\theta \in \Theta = \mathbb{R}$ (with $\mathbf{1}$ denoting the all-ones vector). In other words, $X_i = \theta + \sigma_i G_i$ where $(G_i)_{i\le d} \sim_{iid} \mathsf{N}(0,1)$. (Here $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \sum_{i=1}^m u_i v_i$ denotes the usual scalar product of $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$.)

(a) Show that there exists a sufficient statistics of the form $\ell(\boldsymbol{x}) = \langle \boldsymbol{c}, \boldsymbol{x}\rangle$, where $\boldsymbol{c} \in \mathbb{R}^d$, and determine the vector $\boldsymbol{c}$.

(b) Using the result at the previous point, determine the optimal linear estimator $\hat{\theta}(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x}\rangle$, with respect to the square loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. Here optimality is to be understood in minimax sense, that is we want to minimize $R_{\mathrm{M}}(\hat{\theta})$ among all linear estimators.

(c) Generalize the above to the case of general correlated Gaussian noise (i.e. $\boldsymbol{\Sigma}$ not necessarily diagonal, but strictly positive definite).