2017 USC Marshall Statistics PhD Screening Exam
Part I: in class (8:30AM-1:00PM, June 5, 2017)

Instructions:

- This open-book exam is 270 minutes long. Write your answers on the exam sheets. Make sure to hand in all pages.
- There are five questions, each counts 20 points and is expected to be finished in 54 minutes.
- Read the questions carefully. You must show your work to get full credit. Give specific references if you are using any result from class. If you cannot finish a problem within the given exam period, explain your ideas and list the key steps of your proposed solution.

1. Let $X_1, X_2, \ldots, X_k$ be independent standard normal random variables and $\gamma_1(t), \ldots, \gamma_k(t)$ infinitely differentiable functions of a real variable defined on a closed, bounded interval, such that $\sum_{i=1}^k \gamma_i^2(t) = 1$ for all t. Let $Z(t) = \sum_{i=1}^k \gamma_i(t) X_i$. Let $\dot{Z}(t), \ddot{Z}(t)$, etc, denote first, second, etc derivatives of $Z(t)$ with respect to t.

   (a) Show that $\mathrm{Cov}(Z(t), \dot{Z}(t)) = 0$.

(b) Evaluate $\mathbb{E}(Z(t)|\ddot{Z}(t))$ in terms of $\ddot{Z}(t)$ and expressions of the form

$$\sum_{i=1}^{k}(\gamma_i(t))^a(\delta^m\gamma_i(t)/\delta t^m)^b,$$

for some a, b, m values.

Continue your answers for Question 1 if needed.

2. This question has two parts.

   (a) Let $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma)$, where σ is known. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Consider the test: reject H_0 if $\bar{X} > c$. Derive the power function $\beta(\mu)$. When μ increases, how will $\beta(\mu)$ change?

(b) i. Explain in your own words, what is p-value?

ii. Explain in your own words, how do we interpret confidence intervals?

3. Let $(e_1, \ldots, e_n)$ be the natural basis for $\mathbb{R}^n$ and $\langle \theta, x \rangle = \theta_1 x_1 + \cdots + \theta_n x_n$ the scalar product on $\mathbb{R}^n$. Define

$$L(\theta) = 1 + \exp(\theta_1) + \cdots + \exp(\theta_n)$$

and fix $m \in \mathbb{R}^n$ such that $m_j > 0$ for all $j = 1, \ldots,$ and such that

$$m_0 = 1 - m_1 - \cdots - m_n > 0.$$

(a) For $\lambda > 0$, show that

$$I(m, \lambda) = \int_{\mathbb{R}^n} \exp(\lambda \langle \theta, m \rangle) \{L(\theta)\}^{-\lambda} \, d\theta = \frac{\Gamma(\lambda m_0)\Gamma(\lambda m_1) \cdots \Gamma(\lambda m_n)}{\Gamma(\lambda)}$$

Hint: Write $u_j = \exp(\theta_j)$, prove the result for $n = 1$, and thereafter do an induction on n.

(b) Compute $\lim_{\lambda\to 0} \lambda^n I(m,\lambda)$. *Hint*: Since $\lim_{x\to 0} x\Gamma(x) = 1$.

(c) Compute $\lim_{\lambda\to\infty} \lambda^{n/2} I(m,\lambda)$. *Hint*: Use Sirling formula for $x \to \infty$.

(d) Consider the set of F of probabilities on $\mathbb{R}^n$ of the form:

$$p_0\delta_0 + \sum_{j=1}^{n} p_j \delta_{e_j} \ ,$$

where $p_j > 0$ for $j = 0, 1, \ldots, n$. Here δ_e denotes unit point mass at e. Show that F is an exponential family and find its natural parameterization. Describe the conugate prior as explicitly as you can (What is the domian of the conjugate prior parameters? What is the normalizaing constant?).

4. Let f be m strongly convex and L-lipschitz gradient, and define $\kappa = \frac{L}{m}$. Show that Gradient descent, $x_{k+1} = x_k - \alpha \nabla f(x_k)$, reaches ϵ accuracy in $O(\kappa \log \frac{1}{\epsilon})$ steps for the following 3 measures of convergence : $f(x_k) - f(x^*)$, $\|\nabla f(x_k)\|^2$, and $\|x_k - x^*\|^2$. Clearly state your choice of step-size α.

   Hint: First prove the result for either $\|x_k - x^*\|^2$ or $f(x_k) - f(x^*)$, then use lipschitz gradient/strong convexity to convert to the other measures of convergence.

Continue your answers for Question 4 if needed.

5. Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$ is an n-dimensional vector of response, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is an $n \times p$ deterministic design matrix of p covariates $\mathbf{x}_j$'s, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional vector of regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is an n-dimensional vector of noise. Assume that the true regression coefficient vector $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})^T$ is sparse with s nonzero components and the number of variables p may exceed the number of observations n.

Part a

Explain why the model may no longer be identifiable when $p > n$ if we do not make any sparsity constraint on the parameter vector $\boldsymbol{\beta}$. Show some details when necessary.

Part b

One way to enforce the model identifiability is assuming that the true regression coefficient vector $\boldsymbol{\beta}_0$ is sufficiently sparse with $s \leq n$ or even $s = o(n)$. Provide some justification why such a $\boldsymbol{\beta}_0$ can be unique in this scenario. For example, you may work with the special case of Gaussian design, in which each row of $\mathbf{X}$ is an independent copy of $N(\mathbf{0}, I_p)$, and show that model identifiability would hold with large probability provided that s is sufficiently small compared to n. Some intuition would be sufficient and you are not required to show all details.

Part c

Let us now assume the uniqueness of $\boldsymbol{\beta}_0$ and we aim at recovering the support $A_0 = \operatorname{supp}(\boldsymbol{\beta}_0)$ of $\boldsymbol{\beta}_0$, that is, the true underlying sparse model. One computationally attractive method is the Lasso which is the L_1-penalized least squares

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ (2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\},$$

where $\lambda \geq 0$ is a regularization parameter. Denote by $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ the minimizer to the above problem. It is appealing for the Lasso estimator to have the model selection consistency, meaning that the support of $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ is identical to A_0 with asymptotic probability one. To investigate such a property, we may first work with the case when $\operatorname{supp}(\widehat{\boldsymbol{\beta}}_{\text{Lasso}}) = A_0$ for a fixed choice of regularization parameter λ. Write down some details for this special case and see what kind of condition needs to be imposed on the deterministic design matrix $\mathbf{X}$. Provide some statistical interpretation for the condition you end up with.

Continue your answers for part (c) if needed.

Part d

Let us assume that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ for some $\sigma > 0$. What is the typical choice of the regularization parameter λ for the Lasso? It is appealing to make the choice of the regularization parameter free of the error standard deviation σ which is generally unknown in practice. Provide some brief thoughts on how we may achieve such a goal.