ECE 598MR: Statistical Learning Theory

Maxim Raginsky

Homework 3

Assigned November 4, 2014; due November 18, 2014

---

1. In this problem and the next one, we will prove a simple comparison inequality for Gaussian random vectors, which lies at the basis of more sophisticated comparison tools like Slepian's lemma. First, we need to establish some preliminary results.

   (a) **Integration by parts for univariate Gaussians.** Let γ be a zero-mean Gaussian random variable with variance σ^2. Let $F : \mathbb{R} \to \mathbb{R}$ be a differentiable function that behaves sufficiently well at infinity in the sense that $\lim_{|u|\to\infty} F(u)e^{-u^2} = 0$. Prove that

   $$\mathbb{E}[\gamma F(\gamma)] = \sigma^2 \mathbb{E}[F'(\gamma)]. \tag{1}$$

   (b) **Integration by parts for multivariate Gaussians.** Now let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ be a vector of zero-mean Gaussian random variables, not necessarily independent. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function that has similar moderate growth at infinity in every coordinate. Prove that, for any $1 \le i \le n$,

   $$\mathbb{E}[\gamma_i F(\boldsymbol{\gamma})] = \sum_{j=1}^{n} \mathbb{E}[\gamma_i \gamma_j] \mathbb{E}\left[\frac{\partial F}{\partial \gamma_j}(\boldsymbol{\gamma})\right] \tag{2}$$

   Hint: For each fixed i, define a new random vector $\tilde{\boldsymbol{\gamma}}$ by

   $$\tilde{\gamma}_j = \begin{cases} \gamma_i, & \text{if } j = i \\ \gamma_j - \gamma_i \frac{\mathbb{E}[\gamma_i \gamma_j]}{\mathbb{E}[\gamma_i^2]}, & \text{if } j \neq i. \end{cases}$$

   Prove that $\tilde{\boldsymbol{\gamma}}$ is a Gaussian random vector, and that $\tilde{\gamma}_i$ is independent of each $\tilde{\gamma}_j$. Now express F in terms of $\tilde{\boldsymbol{\gamma}}$ and apply the univariate formula (1).

   (c) **Softmax.** For a fixed parameter $\beta > 0$, define the *softmax* function $F_\beta : \mathbb{R}^n \to \mathbb{R}$ by

   $$F_\beta(\boldsymbol{x}) \triangleq \frac{1}{\beta} \log \sum_{i=1}^{n} e^{\beta x_i}. \tag{3}$$

   Prove that

   $$\max_i x_i \le F_\beta(\boldsymbol{x}) \le \max_i x_i + \frac{\log n}{\beta}.$$

   What happens in the limit as $\beta \to \infty$?

(d) **More softmax.** For a given vector $\boldsymbol{x} \in \mathbb{R}^n$, let $p_i(\boldsymbol{x}) \triangleq \frac{\partial F_\beta}{\partial x_i}(\boldsymbol{x})$, $1 \le i \le n$. Prove that $(p_1(\boldsymbol{x}), \ldots, p_n(\boldsymbol{x}))$ is a probability distribution on the set $\{1, \ldots, n\}$, and express the second partial derivatives

$$\frac{\partial^2 F_\beta}{\partial x_i \partial x_j}(\boldsymbol{x}), \qquad 1 \le i, j \le n$$

in terms of these probabilities.

2. **A Gaussian comparison inequality.** In this problem, you will prove the following result: let $\boldsymbol{V} = (V_1, \ldots, V_n)$ and $\boldsymbol{W} = (W_1, \ldots, W_n)$ be two zero-mean Gaussian random vectors. We can assume, without loss of generality, that $\boldsymbol{W}$ and $\boldsymbol{V}$ are independent; however, their components are not assumed to be independent. Let $\sigma^{\boldsymbol{V}}_{ij} = \mathbb{E}[(V_i - V_j)^2]$ for all $1 \le i, j \le n$, and define $\sigma^{\boldsymbol{W}}_{ij}$ in the same way. Suppose that

$$\sigma^{\boldsymbol{V}}_{ij} \le \sigma^{\boldsymbol{W}}_{ij}, \qquad 1 \le i, j \le n. \tag{4}$$

Then

$$\mathbb{E}\left[\max_i V_i\right] \le \mathbb{E}\left[\max_i W_i\right]. \tag{5}$$

(a) For $t \in [0,1]$, define the Gaussian random vector $\boldsymbol{Z}_t = \sqrt{1-t}\boldsymbol{V} + \sqrt{t}\boldsymbol{W}$. Fix a parameter $\beta > 0$, and consider the function $\varphi(t) \triangleq \mathbb{E}[F_\beta(\boldsymbol{Z}_t)]$, where F_β is the softmax function defined in (3). Note that $\varphi(0) = \mathbb{E}[F_\beta(\boldsymbol{V})]$ and $\varphi(1) = \mathbb{E}[F_\beta(\boldsymbol{W})]$. Prove that

$$\varphi'(t) = \frac{1}{2} \sum_{1 \le i,j \le n} \left(\kappa^{\boldsymbol{W}}_{ij} - \kappa^{\boldsymbol{V}}_{ij}\right) \mathbb{E}\left[\frac{\partial^2 F_\beta}{\partial x_i \partial x_j}(\boldsymbol{Z}_t)\right], \tag{6}$$

where $\kappa^{\boldsymbol{V}}_{ij} \triangleq \mathbb{E}[V_i V_j]$, and $\kappa^{\boldsymbol{W}}_{ij}$ is defined in the same way.

Hint: Use the multivariate Gaussian integration-by-parts formula twice.

(b) For each t, consider the *random* vector $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{n,t})$, where

$$p_{i,t} \triangleq \frac{\partial F_\beta}{\partial x_i}(\boldsymbol{Z}_t).$$

Use (6) and the results from Problem 1(d), to prove that

$$\varphi'(t) = \frac{\beta}{4} \sum_{1 \le i,j \le n} \left(\sigma^{\boldsymbol{W}}_{ij} - \sigma^{\boldsymbol{V}}_{ij}\right) \mathbb{E}[p_{i,t} p_{j,t}]. \tag{7}$$

Hint: How are the quantities $\sigma^{\boldsymbol{V}}_{ij}$ related to $\kappa^{\boldsymbol{V}}_{ij}$'s?

(c) Conclude from (7) that if $\boldsymbol{V}$ and $\boldsymbol{W}$ satisfy (4), then $\varphi'(t) \ge 0$ for all $t \in [0,1]$, i.e., the function φ is increasing. Use this fact to show that

$$\mathbb{E}[F_\beta(\boldsymbol{V})] \le \mathbb{E}[F_\beta(\boldsymbol{W})], \qquad \forall \beta > 0. \tag{8}$$

(d) Finally, derive (5) from (8).

3. **Some examples of the Gaussian comparison technique.**

   (a) Let $\boldsymbol{W} = (W_1, \ldots, W_n)$ be a zero-mean Gaussian random vector, where $n \geq 2$. Prove that

   $$\mathbb{E}\left[\max_i W_i\right] \geq \frac{\sigma_*}{2}\sqrt{\log n},$$

   where

   $$\sigma_* \triangleq \min_{i \neq j} \sqrt{\mathbb{E}[(W_i - W_j)^2]}.$$

   Hint: Consider comparing $\boldsymbol{W}$ with the Gaussian vector $\boldsymbol{V} = \frac{\sigma_*}{\sqrt{2}}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ are i.i.d. zero-mean, unit-variance Gaussian random variables. You may also find the following fact useful: if $\gamma_1, \ldots, \gamma_n$ are i.i.d. $N(0,1)$ random variables (where $n \geq 2$), then

   $$\mathbb{E}\left[\max_i \gamma_i\right] \geq \sqrt{\frac{\log n}{2}}.$$

   (b) Let $\boldsymbol{Y}$ be an $n \times m$ random matrix, whose entries Y_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$, are i.i.d. $N(0,1)$ random variables. The *spectral norm* of Y is defined as

   $$\|\boldsymbol{Y}\| \triangleq \sup_{v \in S^{n-1}} \sup_{w \in S^{m-1}} \langle v, \boldsymbol{Y} w \rangle$$

   where $S^{n-1} \triangleq \left\{v \in \mathbb{R}^n : \|v\|_2 = 1\right\}$ is the ℓ_2 unit sphere in $\mathbb{R}^n$. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n), \boldsymbol{\gamma}' = (\gamma'_1, \ldots, \gamma'_m)$ be two independent vectors of i.i.d. $N(0,1)$ random variables. Define two Gaussian processes indexed by pairs $(v, w) \in S^{n-1} \times S^{m-1}$:

   $$Y_{v,w} \triangleq \langle v, \boldsymbol{Y} w \rangle, \qquad Z_{v,w} \triangleq \langle v, \boldsymbol{\gamma} \rangle + \langle w, \boldsymbol{\gamma}' \rangle.$$

   Prove that

   $$\mathbb{E}\left[\left(Y_{v,w} - Y_{v',w'}\right)^2\right] \leq \mathbb{E}\left[\left(Z_{v,w} - Z_{v',w'}\right)^2\right]$$

   for all $v, v' \in S^{n-1}$ and all $w, w' \in S^{m-1}$. Now use this fact, together with Slepian's lemma, to prove that

   $$\mathbb{E}\|\boldsymbol{Y}\| \leq \sqrt{n} + \sqrt{m}.$$

4. **Intrinsic limitations of learning to predict.** In our analysis of regression with quadratic loss, we have focused on the ERM algorithm and developed high-probability bounds on its excess loss. In this problem, we will see that there are certain intrinsic limitations any learning algorithm will face even in the realizable case when $Y = f(X)$ (with probability one) and the function f is a member of the chosen hypothesis class $\mathscr{F}$.

   Let μ be the marginal probability distribution of X, and for each $f \in \mathscr{F}$ let $Y^f = f(X)$. Let $\mathbb{P}_f$ denote the joint distribution of (X, Y^f). That is, under $\mathbb{P}^f$ we have

   $$\mathbb{P}_f(A \times B) = \int_A \mu(\mathrm{d}x) \mathbf{1}_{\{f(x) \in B\}}$$

for all measurable sets $A \subset \mathsf{X}$ and all $B \subset \mathbb{R}$. Consider a learning algorithm $\mathscr{A}_n$ that receives a sequence of i.i.d. training samples $Z_i^f = (X_i, Y_i^f)$, $1 \le i \le n$, drawn from $\mathbb{P}_f$, where $f \in \mathscr{F}$ is unknown. Consider also the following *random* subset of $\mathscr{F}$:

$$\mathscr{V}_n(f) \triangleq \left\{h \in \mathscr{F} : h(X_i) = f(X_i),\, 1 \le i \le n\right\}.$$

This set, called the *version space*, consists of all functions $h \in \mathscr{F}$ that agree with the unknown target function f on the training data. Let $D_n(f)$ denote the *diameter* of the version space in $L^2(\mu)$ norm:

$$D_n(f) \triangleq \sup_{h,h' \in \mathscr{V}_n(f)} \|h - h'\|_{L^2(\mu)} \equiv \sup_{h,h' \in \mathscr{V}_n(f)} \left(\int_{\mathsf{X}} \left|h(x) - h'(x)\right|^2 \mu(\mathrm{d}x)\right)^{1/2}.$$

Note that $D_n(f)$ is a random variable, since it depends on the training data. Our goal is to prove that, no matter how sophisticated $\mathscr{A}_n$ is, it cannot attain better performance than a constant multiple of $D_n^2(f)$.

(a) Suppose that $\mathscr{A}_n$ is the ERM algorithm: upon receiving the training data $Z^n = (Z_1, \ldots, Z_n)$ with $Z_i = (X_i, Y_i)$, $1 \le i \le n$, it outputs

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

Prove that if Z^n are i.i.d. samples from $\mathbb{P}_{f^*}$ for some $f^* \in \mathscr{F}$, then

$$L(\widehat{f}_n) \equiv \int_{\mathsf{X}} \left(\widehat{f}_n(x) - f^*(x)\right)^2 \mu(\mathrm{d}x) \le D_n^2(f^*).$$

(b) Now we will prove the following converse result: for an arbitrary learning algorithm $\mathscr{A}_n$, there exists at least one $f \in \mathscr{F}$, such that

$$\mathbb{P}_f^n\left(L(\tilde{f}_n) \ge \frac{D_n^2(f)}{16}\right) \ge \frac{1}{2}, \tag{9}$$

where $\tilde{f}_n = \mathscr{A}_n\left(Z^{f,n}\right)$ is the output of $\mathscr{A}_n$ given training data $Z^n = Z^{f,n}$ drawn i.i.d. from $\mathbb{P}_f$. We will prove this in several steps.

i. Given $f \in \mathscr{F}$, consider the version space $\mathscr{V}_n(f)$ and let $h_{0,f}, h_{1,f} \in \mathscr{V}_n(f)$ be such that $\|h_{0,f} - h_{1,f}\|_{L^2(\mu)} = D_n(f)$. Let ε be a Bernoulli(1/2) random variable independent of X^n, and define the random function

$$h_f \triangleq (1-\varepsilon) h_{0,f} + \varepsilon h_{1,f}.$$

That is, if $\varepsilon = 0$, then $h_f = h_{0,f}$; if $\varepsilon = 1$, then $h_f = h_{1,f}$. Prove that, for any realization of ε, $D_n(f) = D_n(h_f)$.

ii. Prove that, for any realization of ε,

$$\sup_{f \in \mathscr{F}} \mathbb{P}_f^n\left(\left\|\mathscr{A}_n(Z^{n,f}) - f\right\|_{L^2(\mu)} \ge \frac{D_n(f)}{4}\right) \ge \sup_{f \in \mathscr{F}} \mu^n\left(\left\|\mathscr{A}_n(Z^{n,h_f}) - h_f\right\|_{L^2(\mu)} \ge \frac{D_n(f)}{4}\right). \tag{10}$$

iii. Let Π_n denote the quantity on the right-hand side of (10). Note that Π_n is a random variable that depends on ε. Prove that

$$\mathbb{E}_\varepsilon \Pi_n \geq \frac{1}{2} \sup_{f \in \mathscr{F}} \left(\mu^n(A_{0,f}) + \mu^n(A_{1,f}) \right), \tag{11}$$

where, for $b \in \{0,1\}$, we have defined the event

$$A_{b,f} \triangleq \left\{ \left\| \mathscr{A}_n(Z^{n,h_{b,f}}) - h_{b,f} \right\|_{L^2(\mu)} \geq \frac{D_n(f)}{4} \right\}.$$

iv. Prove that the union of the events $A_{0,f}$ and $A_{1,f}$ occurs with μ-probability one, and conclude from this and from (11) that $\mathbb{E}_\varepsilon \Pi_n \geq 1/2$.

Hint: Use the fact $\|h_{0,f} - h_{1,f}\|_{L^2(\mu)} = D_n(f)$, and that both $h_{0,f}$ and $h_{1,f}$ are in the version space $\mathscr{V}_n$, and therefore the function output by the learning algorithm $\mathscr{A}_n$ upon seeing the training data

$$(X_1, h_{0,f}(X_1)), \ldots, (X_n, h_{0,f}(X_n))$$

is the same as the function output by $\mathscr{A}_n$ upon seeing the training data

$$(X_1, h_{1,f}(X_1)), \ldots, (X_n, h_{1,f}(X_n))$$

with the *same* i.i.d. input sequence $X_1, \ldots, X_n \sim \mu$.

v. Finally, use all of the above to prove that there exists at least one $f \in \mathscr{F}$, such that (9) holds true.

The moral of the story is: even if there is no noise in the data, the best performance of any learning algorithm is controlled by the richness of the function class $\mathscr{F}$. In particular, if $\mathscr{F}$ is very rich, the version space is likely to be large (as measured by the $L^2(\mu)$ norm) because there will be many functions that can match the target function on a given sample. This limitation is there even if we design our algorithm with full knowledge that the target function f is in our hypothesis class, and even if we know the marginal distribution μ of X ahead of time.