Question 0.1 (Asymptotics): Suppose that \( Y \in \mathcal{Y} \) follows a generalized linear model (GLM) conditional on \( X \in \mathbb{R}^d \). That is, \( Y \mid X = x \) has density (with respect to a base measure \( \mu \)) parameterized by \( \theta \in \mathbb{R}^d \) defined by

\[
p_\theta(y \mid x) = \exp \left( T(y)x^T \theta - A(\theta; x) \right) \quad \text{where} \quad A(\theta; x) = \log \int \exp(T(y)x^T \theta) d\mu(y)
\]

(1)

and \( T : \mathcal{Y} \to \mathbb{R} \) is the sufficient statistic for \( Y \). You may assume that \( A(\theta; x) < \infty \) for all \( \theta \in \mathbb{R}^d \). Recall that for each \( x \), the function \( A(\theta; x) \) is convex in \( \theta \) and infinitely differentiable (in \( \theta \)) by standard exponential family results. Suppose you are given an i.i.d. sample from the model (1) with parameter \( \theta_0 \), where \( X_i \overset{iid}{\sim} P \) for some distribution \( P \) and \( Y_i \mid X_i \) follows the model (1). You choose the maximum likelihood estimator

\[
\hat{\theta}_n = \arg\max_{\theta} \left\{ \sum_{i=1}^n \log p_\theta(Y_i \mid X_i) \right\}.
\]

Assume the consistency result that \( \hat{\theta}_n \overset{P}{\to} \theta_0 \) (though you could prove this using the techniques you know from the class and Q4 from Problem Set 2), that \( \hat{\theta}_n \) exists for all suitably large \( n \), and that you may always swap the order of integration and differentiation. In addition, assume that \( \theta \mapsto \nabla^2 A(\theta; x) \) is \( H(x) \)-Lipschitz continuous, meaning that \( \| \nabla^2 A(\theta; x) - \nabla^2 A(\theta'; x) \|_{op} \leq H(x) \| \theta - \theta' \| \), where \( E[H(X)^2] < \infty \). Assuming that \( E[\nabla^2 A(\theta_0; X)] \) exists and is full rank, what is the asymptotic distribution of \( \sqrt{n}(\hat{\theta}_n - \theta_0) \)?

Answer: We perform a Taylor expansion. We have by standard exponential family calculations that

\[
\nabla A(\theta; x) = \mathbb{E}_\theta[T(Y) \mid X = x]x \quad \text{and} \quad \nabla^2 A(\theta; x) = xx^T \text{Var}_\theta(T(Y) \mid X = x).
\]

Let \( L_n(\theta) = \sum_{i=1}^n \log p_\theta(Y_i \mid X_i) \) be the (conditional) log likelihood. We then have by a Taylor expansion that

\[
0 = \nabla L_n(\hat{\theta}_n) = \nabla L_n(\theta_0) + \nabla^2 L_n(\theta_0)(\hat{\theta}_n - \theta_0) + E_n(\hat{\theta}_n, \theta_0)(\hat{\theta}_n - \theta_0)
\]

where \( E_n \) is an error matrix satisfying \( \| E_n(\hat{\theta}_n, \theta_0) \|_{op} \leq \frac{1}{n} \sum_{i=1}^n H(X_i)\| \hat{\theta}_n - \theta_0 \| \). That is, \( E_n = o_P(1) \). The law of large numbers implies that

\[
\nabla^2 L_n(\theta_0) \overset{a.s.}{\to} \mathbb{E}[\nabla^2 A(\theta_0; X)].
\]
Applying Slutsky’s theorem, we thus have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n} \left( \mathbb{E}[\nabla^2 A(\theta_0; X)] + o_P(1) \right)^{-1} \nabla L_n(\theta_0).$$

Of course, we have

$$n \nabla L_n(\theta_0) = \sum_{i=1}^{n} T(Y_i)X_i - \nabla A(\theta_0; X_i) = \sum_{i=1}^{n} (T(Y_i) - \mathbb{E}_{\theta_0}[T(Y) \mid X_i])X_i,$$

and so

$$\sqrt{n} \nabla L_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\text{Var}_{\theta_0}(T(Y) \mid X)XX^T]).$$

Noting that \( \nabla^2 A(\theta_0; x) = \text{Var}_{\theta_0}(T(Y) \mid X = x)xx^T \) gives the result that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\nabla^2 A(\theta_0; X)]^{-1}).$$
Question 0.2 (Asymptotics and efficiency): Consider the following classification problem, where we attempt to classify a vector $X \in \mathbb{R}^d$ as belonging to class $Y = 0$ or $Y = 1$. We have data drawn i.i.d. from a standard multivariate normal model, where

$$X \mid Y = 0 \sim \mathcal{N}\left(-\frac{1}{2} \theta_0, I_{d \times d}\right) \quad \text{and} \quad X \mid Y = 1 \sim \mathcal{N}\left(\frac{1}{2} \theta_0, I_{d \times d}\right),$$

(2)

where $I = I_{d \times d}$ is the $d \times d$ identity matrix and $\theta_0 \in \mathbb{R}^d$ is the unknown mean vector for each of the classes. We assume that $\mathbb{P}(Y = 1) = \frac{1}{2}$ and $\mathbb{P}(Y = 0) = \frac{1}{2}$. Assume you are given an i.i.d. sample $S := \{(X_i, Y_i)\}_{i=1}^n$ from the model (2), and would like to estimate $\theta_0$.

(a) In order to estimate $\theta_0$, the first obvious strategy is, given the sample $S$, to use the Gaussianity and estimate

$$\hat{\theta}_{G,n} := \frac{1}{N_1} \sum_{i: Y_i = 1} X_i - \frac{1}{N_0} \sum_{i: Y_i = 0} X_i,$$

where $N_y = \text{card}\{i \in [n] : Y_i = y\}$. Show that

$$\sqrt{n}(\hat{\theta}_{G,n} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_G)$$

and specify $\Sigma_G$.

(b) Under the model (2), give the conditional distribution

$$p_\theta(y \mid x) = \mathbb{P}(Y = y \mid X = x; \theta).$$

(c) Because you like classification, you decide to fit a logistic regression model, defining the logistic loss for a pair $(x, y)$ by

$$\ell(\theta; x, y) = \log(1 + \exp(x^T \theta)) - yx^T \theta.$$

Define the empirical risk and estimator

$$\hat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; X_i, Y_i) \quad \text{and} \quad \hat{\theta}_{LR,n} := \text{argmin}_{\theta} \hat{R}_n(\theta).$$

Show that

$$\sqrt{n}(\hat{\theta}_{LR,n} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{LR})$$

and specify $\Sigma_{LR}$. [Hint: Use Question 0.1]

(d) Show that if we take $\theta_0$ far from 0, that is, $\|\theta_0\| \to \infty$, we have $\text{tr}(\Sigma_{LR}) \to \infty$. Does $\text{tr}(\Sigma_G)$ have a limit as $\|\theta_0\| \to \infty$?

(e) Now consider the actual classification (log probability) risk, $R(\theta) = \mathbb{E}[\ell(\theta; X, Y)]$. Show that

$$n(R(\hat{\theta}_{G,n}) - R(\theta_0)) \xrightarrow{d} W_G \quad \text{and} \quad n(R(\hat{\theta}_{LR,n}) - R(\theta_0)) \xrightarrow{d} W_{LR}$$

for random variables $W_G$ and $W_{LR}$, and specify their distributions.

(f) With the random variables as above, what is $\sup_{\theta_0} \mathbb{E}[W_{LR}]$? Is $\inf_{\theta_0} \mathbb{E}[W_{LR}] > 0$?
(g) Suppose that instead of a Gaussian model, the data $X$ actually follows a different exponential family model, where $X$ has densities determined by $\theta = (\theta_0, \theta_1)$ with

$$p_\theta(x \mid y) = \exp \left( \theta^T y x - A(\theta_y) \right). \quad (3)$$

Show that the p.m.f. of $Y \in \{0, 1\}$ conditional on $X = x$ in this model has the sigmoid form

$$p(y \mid x) = \frac{e^{y\beta^T \phi(x)}}{1 + e^{\beta^T \phi(x)}}$$

for some function $\phi: \mathbb{R}^d \to \mathbb{R}^{d+1}$ and some $\beta \in \mathbb{R}^{d+1}$. Specify $\phi$, and specify $\beta$ in terms of $\theta$ and the prior probabilities $\pi(y) = \mathbb{P}(Y = y)$. Given that your data may come from model (3), which estimator $\hat{\theta}_{G,n}$ or $\hat{\theta}_{LR,n}$ do you prefer?

**Answer:**

(a) We have

$$\hat{\theta}_{G,n} - \theta_0 = \frac{1}{N_1} \sum_{i: Y_i = 1} \left( X_i - \frac{1}{2} \theta_0 \right) - \frac{1}{N_0} \sum_{i: Y_i = 0} \left( X_i + \frac{1}{2} \theta_0 \right).$$

By an elementary conditioning argument, the independence of the $X_i$ shows that $\hat{\theta}_{G,n}$ has identical distribution to

$$\hat{\theta}_{G,n} - \theta_0 \mid (N_0, N_1) \stackrel{d}{=} \frac{1}{\sqrt{N_1}} Z_1 - \frac{1}{\sqrt{N_0}} Z_0 \stackrel{d}{=} \left( \frac{1}{N_1} Z \right) + \left( \frac{1}{N_0} Z \right) \text{ where } Z, Z \sim \text{N}(0, I).$$

Multiplying by $\sqrt{n}$ and noting that $n/N_0 \xrightarrow{a.s.} 2$ and $n/N_1 \xrightarrow{a.s.} 2$, we have by Slutsky’s theorem that $\sqrt{n}(\hat{\theta}_{G,n} - \theta_0) \xrightarrow{d} \text{N}(0, 4I)$.

(b) Let $\pi(y)$ be the prior probability that $Y = y$. For simplicity let $y = 1$. We have

$$p_\theta(Y = 1 \mid x) = \frac{p_\theta(x \mid y) \pi(y)}{\pi(y)p_\theta(x \mid y) + \pi(1 - y)p_\theta(x \mid 1 - y)}$$

$$= \frac{\exp(-\frac{1}{2} \| x - \frac{1}{2} \theta \|^2)}{\exp(-\frac{1}{2} \| x - \frac{1}{2} \theta \|^2) + \exp(-\frac{1}{2} \| x + \frac{1}{2} \theta \|^2 + \log \frac{1-\pi}{\pi})}$$

$$= \frac{1}{1 + \exp(-x^T \theta + \log \frac{1-\pi}{\pi})}.$$

When $\pi = \frac{1}{2}$, this is simply the p.m.f.

$$p_\theta(y \mid x) = \frac{e^{y\theta^T x}}{1 + e^{\theta^T x}}.$$

(c) Evidently, this model is a GLM model by part (b). Moreover, we have that $T(y) = y$ and $A(\theta; x) = \log(1 + e^{\theta^T x})$, which satisfies $\nabla A(\theta; x) = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}} x = p_\theta(1 \mid x)x$ and $\nabla^2 A(\theta; x) = p(1 - p)x x^T$ for $p = p_\theta(y \mid x)$. Moreover, that $\theta_0$ is the population minimizer of the specified loss is also immediate from part (b). Thus, applying Question 6.1 we obtain

$$\sqrt{n}(\hat{\theta}_{LR,n} - \theta_0) \xrightarrow{d} \text{N}(0, \text{E}[p_{\theta_0}(Y \mid X)(1 - p_{\theta_0}(Y \mid X))XX^T]^{-1}).$$
(d) We show that as $\|\theta\| \to \infty$, we have $\mathbb{E}[p_0(1 - p_0)XX^T] \to 0$, where we reparametrized $\theta_0 \to \theta$. Because $(X, \theta)$ is rotationally symmetric, without loss of generality we may assume $\theta = t\theta_0$ for some $\theta_0$, where $t \in \mathbb{R}$ and $t \to \infty$. Indeed, for any $x$ such that $\theta_0^T x \neq 0$, we have
\[
\frac{e^{t\theta_0^T x}}{(1 + e^{t\theta_0^T x})^2} \to 0 \quad \text{as} \quad t \to \infty.
\]

As this is Lebesgue almost every $x$, the dominated convergence theorem gives that $\Sigma_{LR}^{-1} = \mathbb{E}[p_0(1 - p_0)XX^T] \to 0$ as $t \to \infty$. Taking inverses, we have the result that $\Sigma_{LR} \to \infty$ as $\|\theta_0\| \to \infty$. We have $\Sigma_G = 4I$ no matter what.

(e) We have that
\[
\nabla^2 R(\theta_0) = \Sigma_{LR}^{-1}
\]
by standard calculations (that is, it is the Fisher information). Thus a delta-method calculation gives
\[
n(\hat{R}_{LR,n} - R(\theta_0)) \xrightarrow{d} \frac{1}{2}(\Sigma_{LR}^{1/2}Z)\Sigma_{LR}^{-1}(\Sigma_{LR}^{1/2}Z) = \frac{1}{2}Z^T Z \quad \text{for} \quad Z \sim \mathcal{N}(0, I_{d\times d}).
\]

This is half of a $\chi_d^2$ distribution. On the other hand, for the Gaussian case, we have
\[
n(\hat{R}_{G,n} - R(\theta_0)) \xrightarrow{d} \frac{1}{2}(2Z)^T \Sigma_{LR}^{-1}(2Z) = \frac{1}{2}Z^T \left(\frac{1}{4}\Sigma_{LR}\right)^{-1}Z \quad \text{for} \quad Z \sim \mathcal{N}(0, I_{d\times d}).
\]

(f) We evidently have $\mathbb{E}[W_{LR}] = \frac{d}{2}$ and $\mathbb{E}[W_G] = 2\text{tr}(\Sigma_{LR}^{-1})$. By the previous question, we have $\mathbb{E}[W_G] \to 0$ as $\|\theta_0\| \to \infty$, because $\Sigma_{LR}^{-1} \to 0$. Moreover, at $\theta_0 = 0$ we have $\text{tr}(\Sigma_{LR}^{-1}) = \frac{d}{2}$ and so $\mathbb{E}[W_{LR}] = \mathbb{E}[W_G] = \frac{d}{2}$. Thus
\[
\sup_{\theta_0} \frac{\mathbb{E}[W_{LR}]}{\mathbb{E}[W_G]} = \lim_{\|\theta_0\| \to \infty} \frac{\mathbb{E}[W_{LR}]}{\mathbb{E}[W_G]} = \infty.
\]

That the infimum of the above quantity is positive is immediate from the fact that $\Sigma_{LR}$ is continuous in $\theta_0$ and so for any $r > 0$, $\{\theta_0 : \|\theta_0\| \leq r\}$ is compact, and $\inf_{\|\theta_0\| \leq r} \text{tr}(\Sigma_{LR}^{-1}) > 0$.

(g) By the Bayes rule, we have
\[
P(Y = 1|X = x) = \frac{\pi(1)p_0(x|1)}{\pi(1)p_0(x|1) + \pi(0)p_0(x|0)} = \frac{\pi(1)\exp(\theta_1^T x - A(\theta_1))}{\pi(1)\exp(\theta_1^T x - A(\theta_1)) + \pi(0)\exp(\theta_0^T x - A(\theta_0))} = \frac{\exp(\beta^T \phi(x))}{1 + \exp(\beta^T \phi(x))},
\]
where $\beta = [1, (\theta_1 - \theta_0)^T]^T$ and $\phi(x) = [A(\theta_0) - A(\theta_1) + \log \frac{\pi(1)}{\pi(0)}, x^T]^T \in \mathbb{R}^{d+1}$.

We prefer the logistic estimator, because the other is not even consistent when we don’t have a Gaussian mixture model.
**Question 0.3** (Moduli of continuity and non-asymptotic rates of convergence): Let $\theta \in \mathbb{R}^d$ and define

$$f(\theta) := \mathbb{E}[F(\theta; X)] = \int_X F(\theta; x) dP(x),$$

where $F(\cdot; x)$ is convex in its first argument (in $\theta$) for all $x \in \mathcal{X}$, and $P$ is a probability distribution. We assume $F(\cdot; \cdot)$ is integrable for all $\theta$. Recall that a function $h$ is convex if

$$h(t\theta + (1-t)\theta') \leq th(\theta) + (1-t)h(\theta') \text{ for all } \theta, \theta' \in \mathbb{R}^d, \ t \in [0,1].$$

Let $\theta_0 \in \text{argmin}_\theta f(\theta)$, and assume that $f$ satisfies the following $\nu$-strong convexity guarantee:

$$f(\theta) \geq f(\theta_0) + \frac{\nu}{2} \|\theta - \theta_0\|^2 \text{ for } \theta \text{ s.t. } \|\theta - \theta_0\| \leq \beta,$$

where $\beta > 0$ is some constant. We also assume that the instantaneous functions $F(\cdot; x)$ are $L$-Lipschitz with respect to the norm $\|\cdot\|$.

Let $X_1, \ldots, X_n$ be an i.i.d. sample according to $P$, and define $f_n(\theta) := \frac{1}{n} \sum_{i=1}^n F(\theta; X_i)$ and let

$$\hat{\theta}_n \in \text{argmin}_{\theta} f_n(\theta).$$

(a) Show that for any convex function $h$, if there is some $r > 0$ and a point $\theta_0$ such that $h(\theta) > h(\theta_0)$ for all $\theta$ such that $\|\theta - \theta_0\| = r$, then $h(\theta') > h(\theta_0)$ for all $\theta'$ with $\|\theta' - \theta_0\| > r$.

(b) Show that $f$ and $f_n$ are convex.

(c) Show that $\theta_0$ is unique.

(d) Let

$$\Delta(\theta, x) := [F(\theta; x) - f(\theta)] - [F(\theta_0; x) - f(\theta_0)].$$

Show that $\Delta(\theta, X)$ (i.e. the random version where $X \sim P$) is $4L^2 \|\theta - \theta_0\|^2$-sub-Gaussian.

(e) Show that for some constant $\sigma < \infty$, which may depend on the parameters of the problem (you should specify this dependence in your solution)

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta_0\| \geq \sigma \cdot \frac{1 + t}{\sqrt{n}}\right) \leq C \exp\left(-t^2\right)$$

for all $t \geq 0$, where $C < \infty$ is a numerical constant. [Hint: The quantity $\Delta_n(\theta) := \frac{1}{n} \sum_{i=1}^n \Delta(\theta, X_i)$ may be helpful, as may be the bounded differences inequality from HW6.]

**Answer:**

(a) Fix $\theta'$ such that $\|\theta' - \theta_0\| > r$. Then there is some $\theta \in [\theta_0, \theta']$ such that $\|\theta - \theta_0\| = r$, that is, there is a $t \in (0,1)$ with

$$\theta = t\theta_0 + (1-t)\theta', \text{ so } h(\theta) \leq th(\theta_0) + (1-t)h(\theta').$$

Rearranging by subtracting $h(\theta_0)$ from both sides yields $h(\theta) - h(\theta_0) \leq (1-t)(h(\theta') - h(\theta_0))$.

Noting that $h(\theta_0) < h(\theta)$ and that $t \in (0,1)$, we thus obtain

$$0 < h(\theta) - h(\theta_0) \leq (1-t)[h(\theta') - h(\theta_0)], \text{ or } h(\theta') > h(\theta_0).$$
(b) This is immediate: for any (positive) measure \( \mu \), including \( P \) and \( P_n \), we have
\[
\int F(t\theta + (1 - t)\theta'; x)d\mu(x) \leq \int tF(\theta; x) + (1 - t)F(\theta'; x)d\mu(x).
\]

(c) The uniqueness of \( \theta_0 \) is immediate by part (b) and (a), because \( f(\theta) \geq f(\theta_0) + \frac{\nu\beta}{2} > f(\theta_0) \) for all \( \theta \) with \( \|\theta - \theta_0\| = \beta \).

(d) We have that \( \mathbb{E}[\Delta(\theta, X)] = 0 \), and that
\[
|\Delta(\theta, x)| \leq |F(\theta; x) - F(\theta_0; x)| + |f(\theta) - f(\theta_0)| \leq 2L \|\theta - \theta_0\|,
\]
that is, \( \Delta \) is bounded by \( 2L \|\theta - \theta_0\| \). Using the standard result that a variable \( Z \in [a, b] \) is \( \frac{(b-a)^2}{4} \)-sub-Gaussian, we have that \( \Delta \) is \( 16L^2 \|\theta - \theta_0\|^2 / 4 = 4L^2 \|\theta - \theta_0\|^2 \) sub-Gaussian.

(e) Fix \( \delta \leq \beta \) and let \( \Theta_\delta = \{ \theta : \|\theta - \theta_0\| \leq \delta \} \). Suppose that \( \hat{\theta}_n \) is not within \( \delta \) of \( \theta_0 \), that is, \( \|\hat{\theta}_n - \theta_0\| \geq \delta \). Then by part (a), there must be some \( \theta_\delta \in \Theta_\delta \) such that \( f_n(\theta_\delta) \leq f_n(\theta_0) \). Then
\[
f_n(\theta_\delta) \leq f_n(\theta_0) = f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) + f(\theta_0) - f(\theta_\delta)
\]
\[
\leq f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) - \frac{\nu}{2} \|\theta_\delta - \theta_0\|^2.
\]
Rearranging, we have
\[
\frac{\nu}{2} \|\theta_\delta - \theta_0\|^2 \leq f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) - f_n(\theta_\delta) \leq |\Delta_n(\theta_\delta)| \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|.
\]
In particular, if we have that
\[
\|\hat{\theta}_n - \theta_0\| \geq \delta,
\]
then it must be the case that
\[
\frac{\nu}{2} \delta^2 \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|.
\]
(4)

Now, let us understand this last event (4). Let \( \Delta'_n \) be \( \Delta_n \) with the point \( x_i \) swapped for \( x'_i \). Then
\[
\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| - \sup_{\theta \in \Theta_\delta} |\Delta'_n(\theta)| \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta) - \Delta'_n(\theta)|
\]
\[
= \frac{1}{n} \sup_{\theta \in \Theta_\delta} \left| (F(\theta; x_i) - f(\theta)) - (F(\theta_0; x_i) - f(\theta_0)) - (F(\theta; x'_i) - f(\theta)) + (F(\theta_0); x'_i - f(\theta_0)) \right|
\]
\[
\leq \frac{1}{n} \sup_{\theta \in \Theta_\delta} \left| (F(\theta; x_i) - F(\theta_0; x_i)) + |F(\theta; x'_i) - F(\theta_0; x'_i)| \right| \leq \frac{2L}{n} \sup_{\theta \in \Theta_\delta} \|\theta - \theta_0\| = \frac{2L}{n} \delta.
\]
That is, \( \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \) satisfies bounded differences, and we have
\[
P \left( \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \geq \mathbb{E} \left[ \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \right] + t \right) \leq \exp \left( -\frac{nt^2}{2L^2 \delta^2} \right).
\]
Thus, we control the expected supremum of the errors \( \Delta_n(\theta) \) over the neighborhood \( \Theta_\delta \). We note by our standard symmetrization inequalities, and the fact that \( \theta \mapsto \sqrt{n} \Delta_n(\theta) \) is \( 4L^2 \|\theta - \theta_0\|^2 \)-sub-Gaussian process, that
\[
\mathbb{E} \left[ \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \right] \leq \frac{CL}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\Theta_\delta, \|\cdot\|, \epsilon)}d\epsilon,
\]
where $N$ denotes the covering numbers of $\Theta_\delta$ in norm $\|\cdot\|$ at radius $\epsilon$ as usual. But then we have $\log N(\Theta_\delta, \|\cdot\|, \epsilon) \leq d \log(1 + \frac{\delta}{\epsilon})$ for $\epsilon < \delta$, and 0 otherwise, so that

$$E \left[ \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \right] \leq \frac{CL\sqrt{d}}{\sqrt{n}} \int_0^\delta \sqrt{\log \left( 1 + \frac{\delta}{\epsilon} \right)} d\epsilon \leq C \frac{L\sqrt{d}\delta}{\sqrt{n}}.$$

That is, for some numerical constant $C$, we have

$$P \left( \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \geq C \frac{L\delta}{\sqrt{n}} (\sqrt{d} + t) \right) \leq e^{-t^2} \tag{5}$$

for all $t \geq 0$.

On the event that $\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \leq \frac{L\sqrt{d}}{\sqrt{n}} \delta + \frac{\sqrt{2}L}{\sqrt{n}} \delta t$, which occurs with probability at least $1 - e^{-t^2}$ by inequality (5), we have by inequality (4)

$$\delta^2 \leq C \frac{L}{\nu \sqrt{n}} \left( \sqrt{d} + t \right) \delta,$$

where $C < \infty$ is an absolute constant, or, setting $\sigma = C L \sqrt{d}/\nu \sqrt{n}$, that

$$\delta \leq \sigma (1 + t).$$

That is,

$$P \left( \|\hat{\theta}_n - \theta_0\| \leq C \frac{L}{\nu \sqrt{n}} \left( \sqrt{d} + t \right) \right) \geq 1 - e^{-t^2}.$$
**Question 0.4** (Non-asymptotic lower bounds on estimation): We wish to estimate the mean $\theta$ from some distribution $P$ belonging to a set $\mathcal{P}$ of distributions (each with at least two moments). Let $\theta(P) = \mathbb{E}_P[X]$ denote the mean of the distribution $P$, and consider the minimax absolute risk

$$
\mathcal{M}_n := \inf \sup_{\tilde{\theta}} \mathbb{E}_{P^n} \left[ \left| \tilde{\theta}(X_1, \ldots, X_n) - \theta(P) \right| \right],
$$

where the expectation is taken over the $n$ i.i.d. observations $X_i \overset{iid}{\sim} P$, and the infimum $\tilde{\theta}$ is taken over all estimators.

(a) Show that for any distributions $P_0, P_1 \in \mathcal{P}$, with $\theta_i = \theta(P_i)$ denoting their means, we have

$$
\mathcal{M}_n \geq \frac{1}{4} |\theta_0 - \theta_1| (1 - \|P_0^n - P_1^n\|_{TV}),
$$

where $P^n$ denotes the probability distribution of $n$ i.i.d. observations from $P$.

(b) Let $\beta \in (0, \infty)$. Suppose that $\mathcal{P}$ is the family of exponential distributions with scale (also mean) $\theta \in [\beta, 2\beta]$, is, $\mathcal{P} = \{\exp(\theta)\}_{\beta \leq \theta \leq 2\beta}$, where we recall that $X \sim \exp(\theta)$ if $X$ has density $f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ for $x \geq 0$ (and density $f(x) = 0$ otherwise). As a reminder, $\mathbb{E}_\theta[X] = \theta$ and $\text{Var}_\theta(X) = \theta^2$. Using the result of part (a), show that

$$
\frac{2\beta}{\sqrt{n}} \geq \mathcal{M}_n \geq c \frac{\beta}{\sqrt{n}},
$$

where $c > 0$ is a numerical constant.

**Answer:**

(a) Pick any $\theta_0, \theta_1$ (and associated $P_0 = P_0^n, P_1 = P_1^n$ for shorthand). For any estimator $\tilde{\theta}$ and any $\delta \geq 0$, we certainly have

$$
\sup_{P \in \mathcal{P}} \mathbb{E}_P[|\tilde{\theta} - \theta(P)|] \geq \frac{1}{2} \mathbb{E}_{P_0}[|\tilde{\theta} - \theta_0|] + \frac{1}{2} \mathbb{E}_{P_1}[|\tilde{\theta} - \theta_1|] \geq \frac{\delta}{2} \left( P_0(|\tilde{\theta} - \theta_0| \geq \delta) + P_1(|\tilde{\theta} - \theta_1| \geq \delta) \right).
$$

Now, set $\delta = \frac{|\theta_0 - \theta_1|}{2}$. Then we have that if $|\tilde{\theta} - \theta_0| < \delta$, it must be the case that $|\tilde{\theta} - \theta_1| \geq \delta$, which implies

$$
P_0(|\tilde{\theta} - \theta_0| \geq \delta) + P_1(|\tilde{\theta} - \theta_1| \geq \delta) = 1 - P_0(|\tilde{\theta} - \theta_0| < \delta) + P_1(|\tilde{\theta} - \theta_1| \geq \delta) \geq 1 - P_0(|\tilde{\theta} - \theta_0| \geq \delta) + P_1(|\tilde{\theta} - \theta_1| \geq \delta).
$$

But $\|P_0 - P_1\|_{TV} = \sup_A |P_0(A) - P_1(A)|$, so we have (for $\delta = \frac{|\theta_0 - \theta_1|}{2}$) that

$$
\sup_{P \in \mathcal{P}} \mathbb{E}_P[|\tilde{\theta} - \theta(P)|] \geq \frac{\delta}{2} (1 - \|P_0 - P_1\|_{TV}) = \frac{|\theta_0 - \theta_1|}{4} (1 - \|P_0 - P_1\|_{TV}),
$$

and replacing $P$ with $P^n$ gives the result.

(b) We use the fact, proved in HW8Q2, that

$$
\|P - Q\|_{TV}^2 \leq 2d_{hel}(P, Q)^2 = 2\left(1 - \int \sqrt{dP \cdot dQ} \right).
$$
Using that $\int \sqrt{dP} \sqrt{dQ} = 1 - d_{\text{hel}}(P,Q)^2$, when we have $n$-fold product distributions, this becomes

$$\|P^n - Q^n\|_{TV}^2 \leq 2d_{\text{hel}}(P^n, Q^n)^2 = 2 - 2 \left( \int \sqrt{dP} \sqrt{dQ} \right)^n = 2 - 2 \left( 1 - d_{\text{hel}}(P,Q)^2 \right)^n.$$ 

For exponential distributions with scale parameters $\alpha$ and $\beta$, that is, densities $\alpha^{-1}e^{-x/\alpha}$ and $\beta^{-1}e^{-x/\beta}$, respectively, we have

$$d_{\text{hel}}(P_0, P_1)^2 = 1 - \frac{1}{\sqrt{\alpha\beta}} \int_0^\infty \exp \left( -\frac{x(\alpha + \beta)}{2\beta\alpha} \right) dx$$

$$= 1 - \frac{2\sqrt{\alpha\beta}}{\alpha + \beta}.$$

Now, using the result of part (a), we have for any $\theta_0, \theta_1 \in \mathbb{R}_+$ that

$$M_n \geq \frac{\|\theta_0 - \theta_1\|}{4} \left( 1 - \sqrt{2} \sqrt{1 - (1 - d_{\text{hel}}(\theta_0, \theta_1)^2)^n} \right)$$

$$= \frac{\|\theta_0 - \theta_1\|}{4} \left( 1 - \sqrt{2} \sqrt{1 - \frac{2^n (\theta_0\theta_1)^{n/2}}{(\theta_0 + \theta_1)^n}} \right).$$

Now, set $\theta_1 = \theta_0(1 + 1/\sqrt{n})$, which yields

$$\frac{2^n (\theta_0\theta_1)^{n/2}}{(\theta_0 + \theta_1)^n} = \left( \frac{\sqrt{1 + n^{-\frac{1}{2}}} \sqrt{1}}{1 + \frac{1}{2} n^{-\frac{1}{2}}} \right)^n \downarrow e^{-\frac{1}{8}}$$

as $n \to \infty$. Thus we have

$$M_n \geq \frac{\theta_0}{4\sqrt{n}} \left( 1 - \sqrt{2} \sqrt{1 - e^{-\frac{1}{8}}} \right) \geq \frac{\theta_0}{8\sqrt{n}}.$$

For the upper bound, note that for any exponential distribution with scale parameter $\theta$ (and hence $\theta$ mean $\theta$) we have $\text{Var}_\theta(X) = \theta^2$. Thus the sample mean satisfies

$$E[\hat{\theta}_n - \theta]^2 \leq E[(\hat{\theta}_n - \theta)^2] = \sqrt{\frac{\text{Var}_\theta(X)}{n}} = \frac{\theta}{\sqrt{n}}.$$

This gives the result. 

$\square$