Problem Set 3: More minimax, maximum entropy, redundancy

Stats 311/EE 377

Due: Tuesday, 11/11/2014

Question 1: In this question, we study the question of whether adaptivity can give better estimation performance for linear regression problems. That is, for \( i = 1, \ldots, n \), assume that we observe variables \( Y_i \) in the usual linear regression setup,

\[
Y_i = \langle X_i, \theta \rangle + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2),
\]

where \( \theta \in \mathbb{R}^d \) is unknown. But now, based on observing \( Y_{1:i-1} = \{Y_1, \ldots, Y_{i-1}\} \), we allow an adaptive choice of the next predictor variables \( X_i \in \mathbb{R}^d \). Let \( \mathcal{L}_{\text{ada}}^n(F^2) \) denote the family of linear regression problems under this adaptive setting (with \( n \) observations) where we constrain the Frobenius norm of the data matrix \( X^\top = [X_1 \cdots X_n] \), \( X \in \mathbb{R}^{n \times d} \), to have bound \( \|X\|_2^\text{Fr} = \sum_{i=1}^n \|X_i\|_2^2 \leq F^2 \). We use Assouad’s method to show that the minimax mean-squared error satisfies the following bound:

\[
M(\mathcal{L}_{\text{ada}}^n(F^2), \|\cdot\|_2) := \inf_\hat{\theta} \sup_{\theta \in \mathbb{R}^d} E[\|\hat{\theta} - \theta\|_2^2] \geq \frac{d\sigma^2}{n} \cdot \frac{1}{4 \frac{1}{dn} F^2}. \tag{2}
\]

Here the infimum is taken over all adaptive procedures satisfying \( \|X\|_2^\text{Fr} \leq F^2 \).

Justify the following steps in the proof of inequality (2):

(i) Fix \( \delta > 0 \). Let \( v \in \{-1, 1\}^d \), and for each such \( v \), define \( \theta_v = \delta v \). Also let \( P^n_v \) denote the joint distribution (over all adaptively chosen \( X_i \)) of the observed variables \( Y_1, \ldots, Y_n \), and define \( P^n_{+j} = \frac{1}{2^{d-1}} \sum_{v:v_j=1} P^n_v \) and \( P^n_{-j} = \frac{1}{2^{d-1}} \sum_{v:v_j=-1} P^n_v \), so that \( P^n_{\pm j} \) denotes the distribution of the \( Y_i \) when \( v \in \{-1, 1\}^d \) is chosen uniformly at random but conditioned on \( v_j = \pm 1 \). Then

\[
M(\mathcal{L}_{\text{ada}}^n(F^2), \|\cdot\|_2) \geq \frac{\delta^2}{2} \sum_{j=1}^d \left[ 1 - \|P^n_{+j} - P^n_{-j}\|_{\text{TV}} \right].
\]

(ii) We have

\[
\frac{\delta^2}{2} \sum_{j=1}^d \left[ 1 - \|P^n_{+j} - P^n_{-j}\|_{\text{TV}} \right] \geq \frac{\delta^2 d}{2} \left[ 1 - \left( \frac{1}{d} \sum_{j=1}^d \|P^n_{+j} - P^n_{-j}\|_{\text{TV}}^2 \right)^{\frac{1}{2}} \right].
\]

(iii) Let \( P_{+j}^{(i)} \) be the distribution of the random variable \( Y_i \) conditioned on \( v_j = +1 \) (with the other coordinates of \( v \) chosen uniformly at random), and let \( P_{+j}^{(i)}(\cdot \mid y_{1:i-1}, x_{1:i-1}) \) denote the
distribution of $Y_i$ conditioned on $v_j = +1$, $Y_i^{i-1} = y_i^{i-1}$, and the first $i - 1$ vectors $X_i^{i-1} = x_i^{i-1}$. (And similarly for $P_{-i}^{j}$.) Justify
\[
\|P_{+j}^n - P_{-j}^n\|_{TV}^2 \leq \frac{1}{2} D_{KL}( P_{+j}^n \| P_{-j}^n ) \\
\leq \frac{1}{2} \sum_{i=1}^n D_{KL}( P_{+i}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) \| P_{-i}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) ) dP_{+i}^{i-1}(y_i^{i-1}, x_i^{i-1}),
\]
where $dP_{+i}^{i-1}(y_i^{i-1}, x_i^{i-1})$ denotes the joint density of $Y_i^{i-1}$ and $X_i^{i-1}$.

(iv) Let $Q^{(i)}$ denote the (adaptive) distribution choosing the vector $X_i$, so that $Q^{(i)}(\cdot \mid y_i^{i-1}, x_i^{i-1})$ denotes the distribution of $X_i$ conditional on $Y_i^{i-1} = y_i^{i-1}$ and $X_i^{i-1} = x_i^{i-1}$. Let $P_{v,+}^i$ be the distribution $P_e$ with the $j$th coordinate of $v$ forced to be $v_j = 1$ (and similarly for $P_{v,-}^i$). Justify
\[
D_{KL}( P_{v,+}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) \| P_{v,-}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) ) \leq \frac{1}{2d} \sum_{v \in \{-1,1\}^d} D_{KL}( P_{v,+}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) \| P_{v,-}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) ) dQ^{(i)}(x_i \mid y_i^{i-1}, x_i^{i-1}).
\]

(v) Thus we have (here the expectation is taken over the procedure choosing the $X_i$)
\[
\sum_{j=1}^d D_{KL}( P_{+j}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) \| P_{-j}^{(i)}( \cdot \mid y_i^{i-1}, x_i^{i-1}) ) \leq \frac{\delta^2}{2\sigma^2} \mathbb{E} \left[ \|X_i\|_2^2 \mid Y_i^{i-1} = y_i^{i-1}, X_i^{i-1} = x_i^{i-1} \right].
\]

(vi) We have
\[
\sum_{j=1}^d \|P_{+j}^n - P_{-j}^n\|_{TV}^2 \leq \frac{\delta^2}{4\sigma^2} \mathbb{E}[\|X\|_2^2],
\]
where the final expectation is over $V$ drawn uniformly in $\{-1,1\}^d$ and all $Y_i, X_i$.

(vii) Show how to choose $\delta$ appropriately to conclude the minimax bound (2).

**Question 2:** Suppose under the setting of Question 1 that we may no longer be adaptive, meaning that the matrix $X \in \mathbb{R}^{n \times d}$ must be chosen ahead of time (without seeing any data). Assuming $n \geq d$, is it possible to attain (within a constant factor) the risk (2)? If so, give an example construction, if not, explain why not.

**Question 3:** Prove that the log determinant function is concave over the positive semidefinite matrices. That is, show that for $X, Y \in \mathbb{R}^{d \times d}$ satisfying $X \succeq 0$ and $Y \succeq 0$, we have
\[
\log \det(\lambda X + (1 - \lambda) Y) \geq \lambda \log \det(X) + (1 - \lambda) \log \det(Y)
\]
for any $\lambda \in [0, 1]$. *Hint: think about log-partition functions.*

**Question 4 (Maximum entropy):** Consider the following optimization problem over symmetric positive semidefinite matrices in $\mathbb{R}^{d \times d}$:
\[
\text{maximize} \quad \log \det(\Sigma) \quad \text{subject to} \quad \Sigma_{ij} = \sigma_{ij}
\]

2
where \( \sigma_{ij} \) are specified only for indices \( i, j \in S \) (but we know that \( \sigma_{ij} = \sigma_{ji} \) and \( (i, i) \in S \) for all \( i \)). Let \( \Sigma^* \) denote the solution to this problem, assuming there is a positive definite matrix \( \Sigma \) satisfying \( \Sigma_{ij} = \sigma_{ij} \) for \( i, j \in S \). Show that for each unobserved pair \( (i, j) \notin S \), the \( (i, j) \) entry \( [\Sigma^*]_{ij} \) of the inverse \( \Sigma^* \) is 0. Hint: The distribution maximizing the entropy \( H(X) = -\int p(x) \log p(x) dx \) subject to \( \mathbb{E}[X_i X_j] = \sigma_{ij} \) has Gaussian density of the form \( p(x) = \exp(\sum_{(i,j)\in S} \lambda_{ij} x_i x_j - \Lambda_0) \).

**Question 6** (Strong versions of redundancy): In this question, we consider expected losses under the Bernoulli distribution. Assume that \( X_i \overset{i.i.d.}{\sim} \text{Bernoulli}(p) \), meaning that \( X_i = 1 \) with probability \( p \) and \( X_i = 0 \) with probability \( 1 - p \). We consider four different loss functions, and their associated expected regret, for measuring the accuracy of our predictions of such \( X_i \). For each of the four choices below, we prove expected regret bounds on

\[
\text{Red}_n(\hat{\theta}, P, \ell) := \sum_{i=1}^{n} \mathbb{E}_P[\ell(\hat{\theta}(X_{1:i-1}^{i-1}), X_i)] - \inf_{\theta} \sum_{i=1}^{n} \mathbb{E}_P[\ell(\theta, X_i)],
\]

where \( \hat{\theta} \) is a predictor based on \( X_1, \ldots, X_{i-1} \) at time \( i \). Define \( S_i = \sum_{j=1}^{i} X_j \) to be the partial sum up to time \( i \). For each of parts (ii)–(iv), at time \( i \) use the predictor

\[
\hat{\theta}_i = \hat{\theta}(X_{1:i-1}^{i-1}) = \frac{S_{i-1} + \frac{1}{2}}{i}.
\]

(a) Loss function: \( \ell(\theta, x) = \frac{1}{2}(x - \theta)^2 \). Show that \( \text{Red}_n(\hat{\theta}, P, \ell) \leq C \cdot \log n \) where \( C \) is a constant.

(b) Loss function: \( \ell(\theta, x) = x \log \frac{1}{\theta} + (1 - x) \log \frac{1}{1 - \theta} \), the usual log loss for predicting probabilities. Show that \( \text{Red}_n(\hat{\theta}, P, \ell) \leq C \cdot \log n \) whenever the true probability \( p \in (0, 1) \), where \( C \) is a constant. Hint: Note that there exists a prior \( \pi \) for which \( \hat{\theta} \) is a Bayes strategy. What is this prior?

(c) Loss function: \( \ell(\theta, x) = |x - \theta| \). Show that \( \text{Red}_n(\hat{\theta}, P, \ell) \geq c \cdot n \), where \( c > 0 \) is a constant, whenever the true probability \( p \notin \{0, \frac{1}{2}, 1\} \).

(d) Extra credit: Show that there is a numerical constant \( c > 0 \) such that for any procedure \( \hat{\theta} \), the worst-case redundancy \( \sup_{p \in [0,1]} \text{Red}_n(\hat{\theta}, \text{Bernoulli}(p), \ell) \geq c \sqrt{n} \) for the absolute loss \( \ell \) in part (d). Give a strategy attaining this redundancy.

**Question 5** (Minimax redundancy and different loss functions): In this question, we consider expected losses under the Bernoulli distribution. Assume that \( X_i \overset{i.i.d.}{\sim} \text{Bernoulli}(p) \), meaning that \( X_i = 1 \) with probability \( p \) and \( X_i = 0 \) with probability \( 1 - p \). We consider four different loss functions, and their associated expected regret, for measuring the accuracy of our predictions of such \( X_i \). For each of the four choices below, we prove expected regret bounds on

\[
\text{Red}_n(\hat{\theta}, P, \ell) := \sum_{i=1}^{n} \mathbb{E}_P[\ell(\hat{\theta}(X_{1:i-1}^{i-1}), X_i)] - \inf_{\theta} \sum_{i=1}^{n} \mathbb{E}_P[\ell(\theta, X_i)],
\]

where \( \hat{\theta} \) is a predictor based on \( X_1, \ldots, X_{i-1} \) at time \( i \). Define \( S_i = \sum_{j=1}^{i} X_j \) to be the partial sum up to time \( i \). For each of parts (ii)–(iv), at time \( i \) use the predictor

\[
\hat{\theta}_i = \hat{\theta}(X_{1:i-1}^{i-1}) = \frac{S_{i-1} + \frac{1}{2}}{i}.
\]

(a) Loss function: \( \ell(\theta, x) = \frac{1}{2}(x - \theta)^2 \). Show that \( \text{Red}_n(\hat{\theta}, P, \ell) \leq C \cdot \log n \) where \( C \) is a constant.

(b) Loss function: \( \ell(\theta, x) = x \log \frac{1}{\theta} + (1 - x) \log \frac{1}{1 - \theta} \), the usual log loss for predicting probabilities. Show that \( \text{Red}_n(\hat{\theta}, P, \ell) \leq C \cdot \log n \) whenever the true probability \( p \in (0, 1) \), where \( C \) is a constant. Hint: Note that there exists a prior \( \pi \) for which \( \hat{\theta} \) is a Bayes strategy. What is this prior?

(c) Loss function: \( \ell(\theta, x) = |x - \theta| \). Show that \( \text{Red}_n(\hat{\theta}, P, \ell) \geq c \cdot n \), where \( c > 0 \) is a constant, whenever the true probability \( p \notin \{0, \frac{1}{2}, 1\} \).

(d) Extra credit: Show that there is a numerical constant \( c > 0 \) such that for any procedure \( \hat{\theta} \), the worst-case redundancy \( \sup_{p \in [0,1]} \text{Red}_n(\hat{\theta}, \text{Bernoulli}(p), \ell) \geq c \sqrt{n} \) for the absolute loss \( \ell \) in part (d). Give a strategy attaining this redundancy.
Consider any distribution $Q$ on the set $\mathcal{X}$ and let $\epsilon \in [0, 1]$, and define the set of points $\theta$ where $Q$ is $\epsilon$-better than the worst case redundancy as

$$B_\epsilon := \{\theta \in \Theta : D_{\text{KL}}(P_\theta | Q) \leq (1 - \epsilon)C^*_n\}.$$  

(a) Show that for any prior $\pi$, we have

$$\pi(B_\epsilon) \leq \frac{\log 2 + C^*_n - I_{\pi}(T; X^n_i)}{\epsilon C^*_n}.$$  

As an aside, note this implies that if $\pi_i$ is a sequence of priors tending to $\sup_{\pi} I_{\pi}(T; X^n_i)$ and the redundancy $C^*_n \to \infty$, then so long as $C^*_n - I_{\pi_i}(T; X^n_i) \ll \epsilon C^*_n$, we have $\pi_i(B_\epsilon) \approx 0$.

(b) Assume that $\pi$ attains the supremum in the definition of $C^*_n$. Show that

$$\pi(B_\epsilon) \leq O(1) \cdot \exp(-\epsilon C^*_n).$$

**Hint:** Introduce the random variable $Z$ to be 1 if the random variable $T \in B_\epsilon$ and 0 otherwise, then use that $Z \to T \to X^n_i$ forms a Markov chain, and expand the mutual information. For part (b), the inequality $\frac{1 - x}{x} \log \frac{1 - x}{1 - x} \leq 1$ for all $x \in [0, 1]$ may be useful.

**This final question is extra credit and quite challenging. Have fun!**

**Question 7** (A generalized version of Fano’s inequality; cf. Proposition 2.6 in the notes): Let $\mathcal{V}$ and $\hat{\mathcal{V}}$ be arbitrary sets, and suppose that $\pi$ is a (prior) probability measure on $\mathcal{V}$, where $V$ is distributed according to $\pi$. Let $V \to X \to \hat{V}$ be Markov chain, where $V$ takes values in $\mathcal{V}$ and $\hat{V}$ takes values in $\hat{\mathcal{V}}$. Let $N \subset \mathcal{V} \times \hat{\mathcal{V}}$ denote a measurable subset of $\mathcal{V} \times \hat{\mathcal{V}}$ (a collection of neighborhoods), and for any $\hat{v} \in \hat{\mathcal{V}}$, denote the slice

$$N_{\hat{v}} := \{v \in \mathcal{V} : (v, \hat{v}) \in N\}.$$  

That is, $N$ denotes the neighborhoods of points $v$ for which we do not consider a prediction $\hat{v}$ for $v$ to be an error, and the slices $N_{\hat{v}}$ index the neighborhoods. Define the “volume” constants

$$p_{\text{max}} := \sup_{\hat{v}} \pi(V \in N_{\hat{v}}) \quad \text{and} \quad p_{\text{min}} := \inf_{\hat{v}} \pi(V \in N_{\hat{v}}).$$

Define the error probability $P_{\text{err}} = \mathbb{P}[(V, \hat{V}) \notin N]$ and entropy $h_2(p) = -p \log p - (1 - p) \log (1 - p)$.

(a) Prove that for any Markov chain $V \to X \to \hat{V}$, we have

$$h_2(P_{\text{err}}) + P_{\text{err}} \log \frac{1 - p_{\text{min}}}{p_{\text{max}}} \geq \log \frac{1}{p_{\text{max}}} - I(V; \hat{V}).$$

(b) Conclude from inequality (5) that

$$\mathbb{P}[(V, \hat{V}) \notin N] \geq 1 - \frac{I(V; X) + \log 2}{\inf_{\hat{v}} \log \frac{1}{\pi(V_{\hat{v}})}}.$$  

(c) Now we give a version explicitly using distances. Let $\mathcal{V} \subset \mathbb{R}^d$ and define $N = \{(v, v') : \|v - v'\| \leq \delta\}$ to be the points within $\delta$ of one another. Let $B_v$ denote the $\|\cdot\|$-ball of radius 1 centered at $v$. Conclude that for any prior $\pi$ on $\mathbb{R}^d$ that

$$\mathbb{P}\left(\|V - \hat{V}\|_2 \geq \delta\right) \geq 1 - \frac{I(V; X) + \log 2}{\log \sup_{\pi} \|\delta B_v\|}.$$