ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 3: Minimax risk of GLM and four extensions

Lecturer: Yihong Wu *Scribe: Ashok Vardhan, Jan 28, 2016 [Ed. Mar 24]*

To end this chapter, we discuss several extension of the Gaussian local model (GLM) to illustrate the following concepts such as *tensor product of experiments* and *sample complexity.*

Recall the scalar GLM we discussed in the last lecture, where $X = \theta + Z$, where $\theta \in \mathbb{R}, Z \sim \mathcal{N}(0, \sigma^2)$ and the loss function is quadratic $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Then we have

$$R^* = \sigma^2. \tag{3.1}$$

This follows from

- Lower bound: If $\theta \sim \mathcal{N}(0, \sigma_0^2)$, we know $R_\pi = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}$. Letting $\sigma_0 \to \infty$ yields $R^* \geq \sigma^2$.
- Upper bound: Let the estimator be $\hat{\theta} = X$. Thus $R_\theta(\hat{\theta}) = \sigma^2$ for all θ. Hence $R^* \leq \sigma^2$.

3.0.1 Multivariate version and tensor product of experiments

We observe $X = \theta + Z$, where $\theta \in \mathbb{R}^p, Z \sim \mathcal{N}(0, \sigma^2 I_p)$ and the loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$. Then

$$R^* = p\sigma^2. \tag{3.2}$$

This can be obtained using similar argument to the univariate case:

- Lower bound: $\theta \sim \mathcal{N}(0, \sigma_0^2 I_p)$ and $\sigma_0 \to \infty$.
- Upper bound: take $\hat{\theta} = X$.

The multivariate GLM can be viewed as a tensor product of the univariate GLM, and their minimax risks satisfy a general relationship. We discuss this notion below:

Minimax risk for tensor product of the experiment Given statistical experiments $\mathcal{P}_i = \{P_{\theta_i} : \theta_i \in \Theta_i\}$ and the corresponding loss function ℓ_i, for $i \in [p]$, consider their tensor product, which is the following statistical experiment:

$$\mathcal{P} = \left\{ P_\theta = \prod_{i=1}^p P_{\theta_i} : \theta = \{\theta_1, \dots, \theta_p\} \in \mathbf{\Theta} \triangleq \prod_{i=1}^p \Theta_i \right\},$$

$$X = (X_1, \dots, X_p) \text{ where } X_i \overset{\text{ind}}{\sim} P_{\theta_i},$$

$$\ell(\theta, \hat{\theta}) = \sum_{i=1}^n \ell_i(\theta_i, \hat{\theta}_i), \forall \theta, \hat{\theta} \in \mathbf{\Theta}.$$

Then the minimax risk of the tensor product experiment is related to the minimax risk $R^*(\mathcal{P}_i)$ and worst-case Bayes risks $R_B^*(\mathcal{P}_i) \triangleq \sup_\pi R_\pi(\mathcal{P}_i)$ of individual experiments as follows:[1]

[1] Here the minimax risk is defined allowing randomized procedures.

Theorem 3.1 (Minimax risk of tensor product)**.**

$$\sum_{i=1}^{p} R_B^*(\mathcal{P}_i) \le R^*(\mathcal{P}) \le \sum_{i=1}^{p} R^*(\mathcal{P}_i). \tag{3.3}$$

Consequently, if minimax theorem holds for each experiment, i.e., $R^*(\mathcal{P}_i) = R_B^*(\mathcal{P}_i)$, we have

$$R^*(\mathcal{P}) = \sum_{i=1}^{p} R^*(\mathcal{P}_i). \tag{3.4}$$

Proof. The right inequality simply follows by separately estimating θ_i based on X_i, namely, $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$. For the left inequality, consider a product prior $\pi = \prod_{i=1}^{p} \pi_i$. Then X_i's are independent. For any $\hat{\theta}_i = \hat{\theta}_i(X_1, \ldots, X_p, U_i)$, where U_i is independent external randomness, we can rewrite $\hat{\theta}_i = \hat{\theta}_i(X_i, \tilde{U}_i)$, where $\tilde{U}_i = (X_{\backslash i}, U_i) \perp\!\!\!\perp X_i$ serves as randomization. Therefore the Bayes risk of $\hat{\theta}_i$ satisfies: $\mathbb{E}[\ell(\theta_i, \hat{\theta}_i)] \ge R_{\pi_i}^*$. Summing over i and taking suprema over priors π_i's yields the left inequality of (3.5). □

Remark 3.1 (Minimax risk of tensor product < sum of minimax risks)**.** The right inequality of (3.5) can be strict. This might appear surprising since X_i only carries information about θ_i and it is intuitive to estimate θ_i based solely on X_i. Nevertheless, the following is a counterexample:

Consider $X = \theta Z$, where $\theta \in \mathbb{N}$, $Z \sim \text{Bern}(\frac{1}{2})$. The estimator $\hat{\theta}$ takes values in $\mathbb{N}$ as well and the loss function is $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} < \theta\}$, i.e., whoever guesses the greater number wins. The minimax risk for this experiment is equal to $\mathbb{P}\left[Z = 0\right] = \frac{1}{2}$. To see this, note that if $Z = 0$, then all information about θ is erased. Therefore for any (randomized) estimator $P_{\hat{\theta}|X}$, the risk is lower bounded by $R_\theta(\hat{\theta}) = \mathbb{P}[\hat{\theta} < \theta] \ge \mathbb{P}[\hat{\theta} < \theta, Z = 0] = \frac{1}{2}\mathbb{P}[\hat{\theta} < \theta | X = 0]$. Therefore sending $\theta \to \infty$ yields $\sup_\theta R_\theta(\hat{\theta}) \ge \frac{1}{2}$. This is achievable by $\hat{\theta} = X$. Clearly, this is a case where minimax theorem does not hold, which is very similar to the trivial example given in the last lecture.

Next consider the tensor product of two copies of this experiment. We show that the minimax risk is strictly less than one. For $i = 1, 2$, let $X_i = \theta_i Z_i$, where $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{1}{2})$. Consider the following estimator $\hat{\theta}_1 = \hat{\theta}_2 = X_1 \vee X_2$. Then for any $\theta_1, \theta_2 \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}[\ell(\theta, \hat{\theta})] &= \mathbb{P}[\hat{\theta}_1 < \theta_1] + \mathbb{P}[\hat{\theta}_2 < \theta_2] = \mathbb{P}[Z_1 = 0, Z_2 < \theta_1/\theta_2] + \mathbb{P}[Z_2 = 0, Z_1 < \theta_2/\theta_1] \\ &= \frac{1}{2}(\mathbb{P}[Z_2 < \theta_1/\theta_2] + \mathbb{P}[Z_1 < \theta_2/\theta_1]) \le \frac{3}{4}.\end{aligned}$$

Remark 3.2 (Non-uniqueness of minimax estimator)**.** In general, minimax risk achieving strategies need not be unique. For instance, consider Example 3.1 where $\hat{\theta} = X$ is the maximum likelihood estimator as well as the minimax. On the other hand, the risk of the James-Stein estimator

$$\hat{\theta}_{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X$$

dominates that of MLE everywhere (see Fig. 3.1). Therefore $\hat{\theta}_{JS}$ also achieves $R^* = p$ for $p \ge 3$.

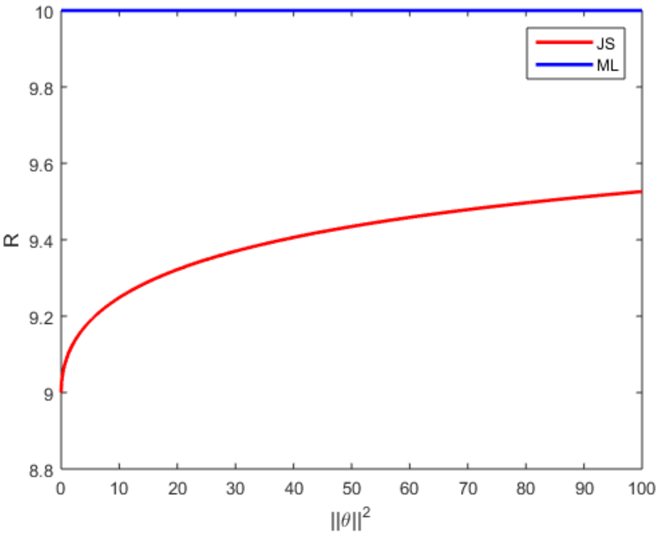

Figure 3.1: Risks of MLE and JS estimators for $p = 10$.

3.0.2 Multiple samples and sample complexity

We now consider a variant of GLM where we observe $X = (X_1, \ldots, X_n)$ where $X_i = \theta + Z_i, Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I_p), \theta \in \mathbb{R}^p$. In this case, we have

$$R^* = \frac{p\sigma^2}{n}. \tag{3.5}$$

To see this, note that for the case of i.i.d. Gaussian random variables, $\bar{X}$ is a sufficient statistic of X for θ, because the joint pdf $p_{X_1,\ldots,X_n|\theta}$ is of the form $h\left(X\right) g_\theta(\bar{X})$, and hence by Fisher's factorization criterion, $\theta \to \bar{X} \to (X_1, \ldots, X_n)$. Therefore the model reduces to $\bar{X} \sim \mathcal{N}(\theta, \frac{\sigma^2}{n} I_p)$, which is the single-sample multivariate case and the minimax risk follows from (3.3).

Sample complexity Given the experiment $\{P_\theta : \theta \in \Theta\}$, consider the experiment

$$\mathcal{P}_n = \left\{ P_\theta^{\otimes n} : \theta \in \Theta \right\}.$$

Note this is not the tensor product of the given experiment because all samples are generated by a common parameter. It is easy to see that $n \mapsto R^* \left(\mathcal{P}_n\right)$ is decreasing since we can always discard samples. Typically, $R^* \left(\mathcal{P}_n\right) \to 0$ as $n \to \infty$. Thus it is natural to consider how fast $R^*(\mathcal{P}_n)$ decreases with n (convergence rate). Equivalently, one can ask what is the minimum number of samples to attain a prescribed error ϵ even in the worst case. This motivates the following definition.

Definition 3.1 (Sample complexity)**.** Given an error margin $\epsilon > 0$, we define the *sample complexity* of the statistical model as

$$n^*(\epsilon) \triangleq \min \left\{ n \in \mathbb{N} : R^* \left(\mathcal{P}_n\right) \leq \epsilon \right\}.$$

In machine learning and related fields, it is customary and useful to consider high-probability bound instead of average risk bound and it is useful to define the sample complexity to be the minimum number of samples required to achieve a prescribed loss with high confidence. In other words, given $\epsilon > 0$ and $0 < \delta < 1$, the sample complexity $n^*(\epsilon, \delta)$ is the smallest n such that there exists $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ satisfying

$$\mathbb{P}_\theta(\ell(\theta, \hat{\theta}) \le \epsilon) \ge 1 - \delta, \quad \forall \theta \in \Theta.$$

This is in fact just a special case of Definition 3.1 with the loss function ℓ replaced by $\mathbf{1}\{\ell(\theta, \hat{\theta}) \ge \epsilon\}$.

Remark 3.3. For the multi-sample GLM with unit variance, we know that $R^* = \frac{p}{n}$. Hence the sample complexity is given by $n^*(\epsilon) = \lceil \frac{p}{\epsilon} \rceil$. Here we notice that the sample complexity grows linearly with the dimension p. This is the common wisdom that "the sample size need to scale at least proportionally to the number of parameters", also known as "counting the degrees of freedom". Indeed in high dimensions we typically expect the sample complexity to grow with the ambient dimension. However, such claim of linear growth should be taken with a grain of salt because it highly depends on what loss function and what is target we are estimating. For example, consider the matrix case $\theta \in \mathbb{R}^{p \times p}$ and let ϵ be a small constant. Then

- For quadratic loss, namely, $\|\theta - \hat{\theta}\|_F^2$, then we have $R^* = \frac{p^2}{n}$ and hence $n^*(\epsilon) = \Theta(p^2)$.
- If the loss function is $\|\theta - \hat{\theta}\|_{op}^2$, then we have $R^* \asymp \frac{p}{n}$ and hence $n^*(\epsilon) = \Theta(p)$.
- If we only want to estimate the scalar functional $\|\theta\|_{\ell_\infty}$, then $n^*(\epsilon) = \Theta(\sqrt{\log p})$.

3.0.3 Nonparametric extension

The result we obtained on the minimax risk of GLM can be in fact generalized to the following nonparametric setting. Consider the class of distributions (which need have density) on the real line with bounded variance:

- Model: $\mathcal{P} = \{P \in \mathcal{M}(\mathbb{R}) : \mathrm{var}_P \le 1\}$, where var_P denotes the variance of the distribution P.
- Data: $X = (X_1, \ldots, X_n) \overset{iid}{\sim} P$ for some $P \in \mathcal{P}$.
- Objective: We wish to estimate $\theta(P)$ where $\theta(P)$ = mean of the distribution P.
- Loss function: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ for $\theta, \hat{\theta} \in \mathbb{R}$.

Then the minimax risk is

$$R^*(\mathcal{P}) = \frac{1}{n}.$$

Proof. Restricting the analysis to the subcollection of Gaussian distributions $\mathcal{P}_G = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$, we know that $R^*(\mathcal{P}_G) = \frac{1}{n}$. Hence $R^*(\mathcal{P}) \ge \frac{1}{n}$. On the other hand, for the estimator $\hat{\theta} = \bar{X}$,

$$R_\theta(\hat{\theta}) = \mathbb{E}[(\theta(P) - \hat{\theta})^2] = \mathbb{E}[(\theta(P) - \bar{X})^2] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n (\theta(P) - X_i)^2\right] \le \frac{1}{n}.$$

Hence $\sup_{P \in \mathcal{P}} R_\theta(\hat{\theta}) \le \frac{1}{n}$ and $R^*(\mathcal{P}) \le \frac{1}{n}$. Thus $R^*(\mathcal{P}) = \frac{1}{n}$. □

3.0.4 Non-quadratic loss

One can also consider non-quadratic loss functions such as $\|\theta - \hat{\theta}\|_1$ when $\theta \in \mathbb{R}^p$ or $\|\theta - \hat{\theta}\|_{op}$ when $\theta \in \mathbb{R}^{p \times p}$, etc., where R^* will no longer be given by (3.6). We will prove the following result later in the course.

Theorem 3.2. *For the Gaussian location model where $X = (X_1, \ldots, X_n) \overset{iid}{\sim} \mathcal{N}(\theta, I_p)$ and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$ for some arbitrary norm $\|\cdot\|$, one has*

$$R^* = \frac{\mathbb{E}[\|Z\|^2]}{n}.$$

Thus (3.6) can be seen as a direct consequence of this theorem. In this case, the sample complexity $n^*(\epsilon)$ scales as $\frac{\mathbb{E}[\|Z\|^2]}{\epsilon}$, depending on the norm.

References

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:229–318, 1967.