Lecture-18: Minimax risk for GLM

1 Minimax risk of GLM with non-quadratic loss

Lemma 1.1. *Let $Z \sim \mathcal{N}(0,1)$. Then for $1 \leqslant q < \infty$, we have $\min_{y\in\mathbb{R}} \mathbb{E}\,|y+Z|^q = \mathbb{E}\,|Z|^q$.*

Proof. We fix $a \in \mathbb{R}_+, y \in \mathbb{R}$, and denote the distribution of random variable Z by $F_Z : \mathbb{R} \to [0,1]$. From the symmetry of F_Z around 0 and unimodality at 0, we observe that

$$P\{|y+Z| \leqslant a\} = F_Z(a-y) - F_Z(-a-y) \leqslant F_Z(a) - F_Z(-a) = P\{|Z| \leqslant a\}.$$

The equality is achieved at $y = 0$, and the result follows from the following observation

$$\mathbb{E}\,|y+Z|^q = \int_{x\in\mathbb{R}_+} P\left\{|y+Z|^q > x\right\} dx \geqslant \int_{x\in\mathbb{R}_+} P\left\{|Z|^q > x\right\} dx = \mathbb{E}\,|Z|^q.$$

□

Theorem 1.2. *Consider the statistical decision theory simple setting under unstructured Gaussian location model on $\Theta = \mathcal{X} = \mathcal{Y} = \mathcal{Y}' = \mathbb{R}^d$, and* i.i.d. *observation sample $X : \Omega \to \mathcal{X}^m$ with common distribution $\mathcal{N}(\theta, I_d)$. Denoting $Z \sim \mathcal{N}(0, I_d)$, for $1 \leqslant q < \infty$, we have*

$$\inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}[\left\|\theta - \hat\theta\right\|_q^q \mid \theta] = m^{-q/2}\mathbb{E}\,\|Z\|_q^q.$$

Proof. Denoting estimate $\hat\theta \triangleq \hat T(X)$, we observe that the loss function $\ell : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_+$ defined for all $\theta, \hat\theta \in \Theta$ as $\ell(\theta,\hat\theta) \triangleq \left\|\theta - \hat\theta\right\|_q^q = \sum_{i=1}^d \left|\theta_i - \hat\theta_i\right|^q$ is separable. We further note that $\mathcal{N}(\theta, I_d)$ is a product distribution. Thus the experiment is a d-fold tensor product of the one-dimensional version. From tensorization minimax theorem, the minimax risk for this experiment lies between the aggregate minimax and the aggregate worst case Bayes risk for $d = 1$.

For $d = 1$, the upper bound is achieved by the sample mean $\bar X \triangleq \frac{1}{m}\sum_{i=1}^m X_i$, which is distributed according to $\mathcal{N}(\theta, \frac{1}{m})$ and is a sufficient statistic for θ. For the lower bound, we consider a Gaussian prior $\pi = \mathcal{N}(0, s)$ for which the posterior distribution is also Gaussian $P_{\theta|X} = \mathcal{N}(\mathbb{E}[\theta \mid X], (1/s + m)^{-1})$ for conditional mean $\mathbb{E}[\theta \mid X] = sm/(1+sm)$. From Lemma 1.1, it follows that the Bayes estimator is simply the conditional mean $\mathbb{E}[\theta \mid X]$, and hence the Bayes risk is

$$R^*_\pi = \mathbb{E}\,|\theta - \mathbb{E}[\theta \mid X]|^q = (1/s + m)^{-q/2}\mathbb{E}\,|Z|^q.$$

Taking limit as $s \to \infty$ proves the matching lower bound. □

2 Log-concavity, Anderson's lemma, and exact minimax risk in GLM

Computing the exact minimax risk is frequently difficult especially in high dimensions. Nevertheless, for the special case of unconstrained GLM, the minimax risk is known exactly in arbitrary dimensions for a large collection of loss functions. We have previously seen in Theorem 1.2 that this is possible for loss functions of the form $\ell : (\theta,\hat\theta) \mapsto \left\|\theta - \hat\theta\right\|_q^q$. Examining the proof of this result, we note that the major limitation is that it only applies to separable loss functions, so that tensorization allows us to reduce the problem to one dimension. This does not apply to (and actually fails) for non-separable loss, since tenosrization minimax theorem, if applicable, dictates the risk to grow linearly with the dimension, which is not always the case. We next discuss a more general result that goes beyond separable losses.

Definition 2.1. A function $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is called *bowl-shaped* if its sublevel set $K_c \triangleq \left\{x \in \mathbb{R}^d : \rho(x) \leqslant c\right\}$ is convex and even symmetric i.e. $K_c = -K_c$ for all $c \in \mathbb{R}$.

Theorem 2.2. *Consider the statistical decision theory simple setting for unstructured GLM with $\Theta = \mathcal{X} = \mathcal{Y} = \mathcal{Y}' \triangleq \mathbb{R}^d$, i.i.d. observation sample $X : \Omega \to \mathcal{X}^m$ with the common distribution $\mathcal{N}(\theta, I_d)$, and the loss function be $\ell(\theta, \hat{\theta}) = \rho(\theta - \hat{\theta})$ where $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is bowl-shapm:ed and lower-semicontinuous. Let $Z \sim \mathcal{N}(0, I_d)$, then the minimax risk is given by*

$$R^* \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}[\rho(\theta - \hat{\theta}) \mid \theta] = \mathbb{E}\rho\Big(\frac{Z}{\sqrt{m}}\Big).$$

Furthermore, the upper bound is attained by $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$.

Proof. We show upper bound by taking estimator as the sample mean, and show the lower bound by showing that the Bayes estimator is conditional mean for Gaussian prior.

(a) **Upper bound.** For the estimator $\hat{\theta} \triangleq \hat{T}(X) \triangleq \bar{X}$, the distribution of $(\theta - \hat{\theta})$ and $\frac{Z}{\sqrt{m}}$ are identical, and $\mathbb{E}[\rho(\theta - \hat{\theta}) \mid \theta] = \mathbb{E}\rho\Big(\frac{Z}{\sqrt{m}}\Big)$. for all $\theta \in \Theta$.

(b) **Lower bound.** We lower bound the minimax risk R^* by the Bayes risk R^*_π with the prior $\pi = \mathcal{N}(0, sI_d)$. We take the estimate $\hat{\theta}^* \triangleq \mathbb{E}[\theta \mid X]$. Under the Gaussian prior π and estimate $\hat{\theta}^* = \mathbb{E}[\theta \mid X]$, we observe that $\theta - \hat{\theta}^* \sim \mathcal{N}(0, \frac{s}{1+sm})$ which is identical to the distribution of $\sqrt{\frac{s}{1+sm}}Z$. From Anderson's Lemma A.7, we obtain that for bowl shaped functions ρ,

$$\mathbb{E}\rho\Big(\frac{Z}{\sqrt{1/s + m}}\Big) = \mathbb{E}_\pi \rho(\theta - \hat{\theta}^*) = \inf_{\hat{\theta}} \mathbb{E}_\pi \rho(\theta - \hat{\theta}^* + \hat{\theta}^* - \hat{\theta}) = R^*_\pi.$$

Since ρ is lower semicontinuous, sending $s \to \infty$ and applying Fatou's lemma, we obtain $R^* \geqslant \lim_{s\to\infty} R^*_\pi(s) = \lim_{s\to\infty} \mathbb{E}\rho(\frac{Z}{\sqrt{1/s+m}}) \geqslant \mathbb{E}\rho(\frac{Z}{\sqrt{m}})$.

□

Corollary 2.3. *Consider a map $\rho : \mathbb{R}^d \to \mathbb{R}_+$ defined as $\rho(x) \triangleq \|x\|^q$ for some $q > 0$ and arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$. Then $R^* = m^{-q/2}\mathbb{E}\|Z\|^q$.*

Proof. It follows from Theorem 2.2 applied to bowl shaped loss function $\big\|\theta - \hat{\theta}\big\|^q$. □

Example 2.4 (Applications of Corollary 2.3). Consider the statistical decision theory simple setting where $\Theta = \mathcal{Y} = \mathcal{Y}'$, and the loss function $\ell : \theta \times \hat{\theta} \mapsto \ell(\theta, \hat{\theta}) \triangleq \rho(\theta - \hat{\theta})$ is defined in terms of bowl-shaped loss functions $\rho : \mathbb{R}^d \times \mathbb{R}_+$ for all $x \in \mathbb{R}^d$.

- For $\Theta \subseteq \mathbb{R}^d$ and $\rho(x) \triangleq \|x\|_2^2$, the minimax risk is $R^* \asymp \frac{1}{m}\mathbb{E}\|Z\|^2 = \frac{d}{m}$.
- For $\Theta \subseteq \mathbb{R}^d$ and $\rho(x) \triangleq \|x\|_\infty$, we have $\mathbb{E}\|Z\|_\infty \asymp \sqrt{\ln d}$ and the minimax risk is $R^* = \sqrt{\frac{d}{m}}$.
- For $\Theta \subseteq \mathbb{R}^{d\times d}$ and $\rho(\theta) = \|\theta\|_{\text{op}}$ denote the operator norm that is the maximum singular value. In this case, $\mathbb{E}\|Z\|_{\text{op}} \asymp \sqrt{d}$ and so minimax risk is $R^* = \sqrt{\frac{d}{m}}$.
- For $\Theta \subseteq \mathbb{R}^{d\times d}$ and $\rho(\theta) = \|\theta\|_F$, the minimax risk $R^* \asymp \frac{d}{\sqrt{m}}$.

Remark 1. We can also phrase the result of Corollary 2.3 in terms of the sample complexity $m^*(\epsilon)$ as defined before. For example, for $q = 2$ we have $m^*(\epsilon) = \Big\lceil \frac{1}{\epsilon}\mathbb{E}\|Z\|_2^2 \Big\rceil$. The above examples show that the scaling of sample complexity $m^*(\epsilon)$ with dimension depends on the loss function and the "rule of thumb" that the sampling complexity is proportional to the number of parameters need not always hold. Finally, for the sake of high-probability as opposed to average risk bound, consider $\rho(\theta - \hat{\theta}) = \mathbb{1}_{\{\theta - \hat{\theta} > \epsilon\}}$, which is lower semicontinuous and bowl-shaped. Then the exact expression $R^* = P\big\{\|Z\| \geqslant \epsilon\sqrt{m}\big\}$. This result is stronger since the sample mean is optimal simultaneously for all ϵ, so that integrating over ϵ recovers the result in Corollary 2.3.

A Log-concavity

Definition A.1. A measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ is said to be log-concave if for any $A, B \in \mathcal{B}(\mathbb{R}^d)$ and $\lambda \in [0,1]$, we have

$$\mu(\lambda A + (1-\lambda)B) \geqslant \mu(A)^\lambda \mu(B)^{1-\lambda}.$$

Theorem A.2. *Let $\mathcal{X} \triangleq \mathbb{R}^d$ and $\mu \in \mathcal{M}(\mathcal{X})$ that has a density $f \triangleq \frac{d\mu}{d\,\mathrm{vol}} \in \mathbb{R}_+^{\mathcal{X}}$ with respect to Lebesgue measure $\mathrm{vol} \in \mathcal{M}(\mathcal{X})$. Then, μ is log-concave iff f is log-concave.*

Example A.3 (Lebesgue measure). Let $\mu = \mathrm{vol}$ be the Lebesgue measure on $\mathbb{R}^d$,which satisfies Theorem A.2 for $f \equiv 1$. It follows that for any $\lambda \in [0,1]$,

$$\mathrm{vol}(\lambda A + (1-\lambda)B) \geqslant \mathrm{vol}(A)^\lambda \,\mathrm{vol}(B)^{1-\lambda}.$$

Example A.4 (Guassian measure). Let $\mu \triangleq \mathcal{N}(0,\Sigma)$ with positive semidefinite covariance matrix $\Sigma \succ 0$. Then, it has a log-concave density f, since $\ln f(x) = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det(\Sigma) - \frac{1}{2}x^T\Sigma^{-1}x$ is concave in x.

Theorem A.5 (Brunn-Minkowski). *Let $d \in \mathbb{N}$ and $\mathcal{X} \triangleq \mathbb{R}^d$. Then for any $A, B \in \mathcal{B}(\mathcal{X}), \lambda \in [0,1]$, we get*

$$\mathrm{vol}(A+B)^{1/d} \geqslant \mathrm{vol}(A)^{1/d} + \mathrm{vol}(B)^{1/d}.$$

Proof. Let $A, B \in \mathcal{B}(\mathcal{X})$, and define two unit volume sets $A' \triangleq \mathrm{vol}(A)^{-1/d}A$ and $B' \triangleq \mathrm{vol}(B)^{-1/d}B$. Taking $\lambda \triangleq \frac{\mathrm{vol}(A)^{1/d}}{\mathrm{vol}(A)^{1/d}+\mathrm{vol}(B)^{1/d}}$ and $A', B' \in \mathcal{B}(\mathcal{X})$ in Example A.3, we obtain

$$\frac{\mathrm{vol}(A+B)}{(\mathrm{vol}(A)^{1/d} + \mathrm{vol}(B)^{1/d})^d} = \mathrm{vol}(\lambda A' + (1-\lambda)B') \geqslant \mathrm{vol}(A')^\lambda \,\mathrm{vol}(B')^{1-\lambda} = 1.$$

□

Lemma A.6. *Let $K \subseteq \mathbb{R}^d$ be an even symmetric convex set and $X \sim \mathcal{N}(0,\Sigma)$. Then $\max_{y\in\mathbb{R}^d} P\{X + y \in K\} = P\{X \in K\}$.*

Proof. From Example A.4 and Theorem A.2, it follows that distribution of X is log-concave. Let $y \in \mathbb{R}^d$, then we observe that $\frac{1}{2}(K+y) + \frac{1}{2}(K-y) = K$ from the convexity of K. Applying the log-concavity of distribution of X for $\lambda = \frac{1}{2}$ and measurable sets $A \triangleq K + y, B \triangleq K - y$, we obtain

$$P\{X \in K\} = P\left\{X \in \frac{1}{2}(K+y) + \frac{1}{2}(K-y)\right\} \geqslant \sqrt{P\{X \in K+y\}\,P\{X \in K-y\}}.$$

The equality is obtained for $y = 0$. From even symmetry of K, we have $K = -K$, and hence $\{X \in K - y\} = \{X \in -K - y\}$. Since X has an even symmetric distribution, we obtain $P\{X \in -K - y\} = P\{X \in K + y\}$. It follows that $P\{X \in K\} \geqslant P\{X \in K + y\}$, with the equality at $y = 0$, and the result follows. □

Lemma A.7 (Anderson). *Let $X \sim \mathcal{N}(0,\Sigma)$ for some positive definite $\Sigma \succ 0$ and $\rho : \mathbb{R}^d \to \mathbb{R}_+$ a bowl-shaped loss function. Then, $\min_{y\in\mathbb{R}^d} \mathbb{E}\rho(y+X) = \mathbb{E}\rho(X)$.*

Proof. Denote the sublevel set $K_c \triangleq \left\{x \in \mathbb{R}^d : \rho(x) \leqslant c\right\}$ for each $c \in \mathbb{R}$. Since ρ is bowl- shaped, K_c is convex and even symmetric. For $y \in \mathbb{R}^d$, we write the mean

$$\mathbb{E}\rho(y+X) = \int_{x\in\mathbb{R}_+} P\{\rho(y+X) > x\}\,dx = \int_{x\in\mathbb{R}_+} P\{X + y \notin K_x\}\,dx.$$

From Lemma A.6, we have $\min_{y\in\mathbb{R}^d} P\{X + y \notin K_x\} = P\{X \notin K_x\}$, and it follows that

$$\min_{y\in\mathbb{R}^d} \mathbb{E}\rho(y+X) = \int_{x\in\mathbb{R}_+} P\{X \notin K_x\}\,dx = \int_{x\in\mathbb{R}_+} P\{\rho(X) > x\}\,dx = \mathbb{E}\rho(X).$$

□