CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma — Lecture 2
Scribe: Nian Si, Honglin Yuan — Sept. 26th, 2018

1 Review and Overview

This lecture is about the proof of asymptotics of Maximum likelihood estimator (MLE).

Let $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})$ be some training data, and $\{p_\theta(y|x) : \theta \in \mathbf{R}^p\}$ be the parametrized family of distributions. In this lecture we consider the following "well-specified" case, in which θ^* is the ground-truth parameter, i.e., for any fixed x, $y|x$ is distributed as $p_{\theta^*}(y|x)$.

For MLE, the loss function is defined as

$$l((x,y),\theta) := -\log p_\theta(y|x).$$

The expected loss of a parameter θ is defined as

$$L(\theta) := \mathbf{E}_x \mathbf{E}_{y|x\sim p_{\theta^*}}[l((x,y),\theta)],$$

and the training loss of θ is the empirical average

$$\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l((x^{(i)}, y^{(i)}), \theta).$$

The MLE takes the estimator $\hat{\theta}$ that minimizes the training loss $\hat{L}(\theta)$, i.e.,

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in\mathbf{R}^p} \hat{L}(\theta). \tag{1}$$

The main goal of this lecture is to show the following theorem regarding the asymptotic behavior of MLE. The results are helpful for us to get an intuition of the order of suboptimality.

Theorem 1. *Assuming that $\hat{\theta} \xrightarrow{d} \theta^*$ and certain regularity conditions hold* [1]*, we have*

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, Q^{-1}) \tag{2}$$

$$n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}\chi^2(p) \tag{3}$$

where Q is the Fisher information matrix

$$Q = \mathbf{E}_x \mathbf{E}_{y|x\sim p_{\theta^*}}[\nabla_\theta \log p_{\theta^*}(y|x) \nabla_\theta \log p_{\theta^*}(y|x)^T],$$

and $\chi^2(p)$ is the chi-square distribution, i.e, $\chi^2(p) = \sum_{i=1}^{p} X_i^2$, $X_i \sim N(0,1)$, X_i i.i.d..

[1] See `https://en.wikipedia.org/wiki/Maximum_likelihood_estimation` for details

2 Proof of Asymototics of MLE

Before giving a proof of the theorem, we first show several lemmas.

The first lemma deals with the linear transformation of multivariate Gaussian vector and its relationship to chi-square distribution.

Lemma 1. *(a) If* $Z \sim N(0, \Sigma^{-1})$*, we have* $AZ \sim N(0, A\Sigma^{-1}A^T)$*..*

(b) If $Z \sim N(0, \Sigma^{-1})$ *and* $Z \in \mathbf{R}^p$*, we have* $Z^T \Sigma Z \sim \chi^2(p)$*.*

Proof. The proof of (a) follows directly from the characteristic function of multivariate Gaussian. For (b), we note that

$$\Sigma^{1/2} Z \sim N(0, \Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}) = N(0, I).$$

Then,

$$Z^T \Sigma Z = \left\| Z^T \Sigma Z \right\|_2^2 \sim \chi^2(p).$$

□

Lemma 2. *(a)* $Q = \nabla^2 L(\theta^*)$*.*

(b) $\nabla^2 L(\theta^*) \succeq 0$*, i.e.,* $\nabla^2 L(\theta^*)$ *is positive semidefinite.*

(c) $\nabla L(\theta^*) = 0$*.*

Proof. First, we prove (c). Let $z = y|x$. Since $p_\theta(z)$ is a density, we have $\int p_\theta(z)dz = 1$. By taking derivatives of both sides, we have

$$\int \nabla p_\theta(z)dz = \nabla \left(\int p_\theta(z)dz \right) = 0.$$

Similarly, we have

$$\int \nabla^2 p_\theta(z)dz = 0. \tag{4}$$

Then, since $\nabla p_\theta(z) = [\nabla \log(p_\theta(z))]p_\theta(z)$, we have

$$\mathbf{E}_{z\sim p_\theta}[\nabla \log(p_\theta(z))] = \int \nabla \log(p_\theta(z))p_\theta(z)dz = 0.$$

By plugging θ^* in, we have

$$\nabla L(\theta^*) = -\mathbf{E}_x \mathbf{E}_{z\sim p_{\theta^*}}[\nabla \log p_{\theta^*}(y|x)] = 0.$$

This gives (c).

Then, we turn to prove (a). Note that

$$\nabla^2 L(\theta) = \mathbf{E}_x \mathbf{E}_{y|x}\left[\nabla^2 l((x,y),\theta)\right] = \mathbf{E}_x \mathbf{E}_{y|x}\left[-\nabla^2 \log p_\theta(y|x)\right] = \mathbf{E}_x \mathbf{E}_z\left[-\nabla^2 \log p_\theta(z)\right].$$

Recall that we define $z = y|x$. For the inner expectation, we have

$$\mathbf{E}_z\left[-\nabla^2 \log p_\theta(z)\right] = \mathbf{E}_z\left[-\nabla\left(\frac{\nabla p_\theta(z)}{p_\theta(z)}\right)\right] = \mathbf{E}_z\left[\frac{\nabla p_\theta(z)\nabla p_\theta(z)^T}{p_\theta(z)^2} - \frac{\nabla^2 p_\theta(z)}{p_\theta(z)}\right].$$

From (4), we have

$$\mathbf{E}_z\left[\frac{\nabla^2 p_\theta(z)}{p_\theta(z)}\right] = \int \frac{\nabla^2 p_\theta(z)}{p_\theta(z)} p_\theta(z) dz = 0.$$

Also,

$$\mathbf{E}_z\left[\frac{\nabla p_\theta(z)\nabla p_\theta(z)^T}{p_\theta(z)^2}\right] = \mathbf{E}_z[\nabla_\theta \log p_{\theta^*}(z)\nabla_\theta \log p_{\theta^*}(z)^T].$$

Hence $\nabla^2 L(\theta^*) = Q$. (b) follows directly from the fact that $Q \succeq 0$. □

Corollary 1. *If $\nabla^2 L(\theta^*)$ is of full rank, then $\nabla^2 L(\theta^*)$ is positive definite, θ^* is a local minimizer of $L(\theta)$.*

We now give the proof of theorem 1.

Proof. Recall that we defined that

$$\nabla \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla l((x^{(i)}, y^{(i)}), \theta).$$

By doing Taylor expansion of $\nabla \hat{L}(\theta)$ at θ^*, we have

$$\nabla \hat{L}(\theta) = \nabla \hat{L}(\theta^*) + \nabla^2 \hat{L}(\theta^*)(\theta - \theta^*) + O(\|\theta - \theta^*\|_2^2).$$

Particularly, for $\theta = \hat{\theta}$, we have

$$\nabla \hat{L}(\hat{\theta}) = \nabla \hat{L}(\theta^*) + \nabla^2 \hat{L}(\theta^*)(\hat{\theta} - \theta^*) + O(\left\|\hat{\theta} - \theta^*\right\|_2^2).$$

And by definition of minimizer (1) , $\nabla \hat{L}(\hat{\theta}) = 0$. Then,

$$\hat{\theta} - \theta^* = -\left(\nabla^2 \hat{L}(\theta^*)\right)^{-1} \nabla \hat{L}(\theta^*) + O(\left\|\hat{\theta} - \theta^*\right\|_2^2). \tag{5}$$

Now, by CLT, we have

$$\sqrt{n}\nabla \hat{L}(\theta^*) = \sqrt{n}\left(\nabla \hat{L}(\theta^*) - \nabla L(\theta^*)\right) \xrightarrow{d} N(0, \mathrm{Cov}\left(\nabla l((x, y), \theta^*)\right)).$$

Substituting into (5), we obtain

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) = -\left(\nabla^2 \hat{L}(\theta^*)\right)^{-1} \sqrt{n}\nabla \hat{L}(\theta^*) + \text{higher order term}.$$

By Slutsky's theorem, we have

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \nabla^2 L(\theta^*)^{-1} N(0, \mathrm{Cov}\left(\nabla l((x, y), \theta)\right)).$$

Since $Q = \nabla^2 L(\theta^*) = \mathrm{Cov}\left(\nabla l((x, y), \theta)\right)$ and (a) of lemma 1, we have

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} N(0, Q^{-1}).$$

For the second part of theorem, by Taylor expansion at θ^*, we have

$$L(\hat{\theta}) - L(\theta^*) = \nabla L(\theta^*)^T(\hat{\theta} - \theta^*) + \frac{1}{2}(\hat{\theta} - \theta^*)^T \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) + o(\left\|\hat{\theta} - \theta^*\right\|_2^2).$$

Since $\nabla L(\theta^*) = 0$, we have

$$L(\hat{\theta}) - L(\theta^*) = \frac{1}{2}(\hat{\theta} - \theta^*)^T \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) + o(\left\|\hat{\theta} - \theta^*\right\|_2^2),$$

by the first part of theorem, $Q = \nabla^2 L(\theta^*)$ and (b) of lemma 1, we have

$$n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}\chi^2(p),$$

completing the proof. □

3 Consistency

Our main result (Theorem 1) assumes consistency of the MLE $\hat{\theta}$, which can be shown under certain additional assumptions on the problem. Here we state one such consistency result, which we hope reveals the type of assumptions one need to have consistency.

Theorem 2 (Theorem 5.7, [**?**])**.** *Suppose MLE is performed in $\Theta \subset \mathbf{R}^p$, i.e. $\theta = \mathrm{argmin}_{\theta \in \Theta} \hat{L}(\theta)$. Assume that*

1. *(Uniform convergence) We have*

$$\sup_{\theta \in \Theta} |\hat{L}(\theta) - L(\theta)| \xrightarrow{d} 0;$$

2. *(Identifiability) For every $\varepsilon > 0$,*

$$\inf_{\theta \in \Theta,\ \|\theta - \theta^*\| \geq \varepsilon} L(\theta) > L(\theta^\star).$$

Then the MLE has consistency: $\hat{\theta} \xrightarrow{d} \theta^\star$.

Remark 1. The positive definiteness of the Hessian $\nabla^2 L(\theta^*)$ guarantees that identifiability holds *locally* (in a neighborhood of θ^*), but does not imply identifiability because of a lack of *global* information.

To give a counter-example, suppose θ^* is the global minimizer of L and the Hessian is positive definite, but there exists a sequence of θ_n such that $\|\theta_n - \theta^*\| = n$ but $L(\theta_n) = L(\theta^*) + 1/n$, then the identifiability is violated: the inf is not strictly greater than but equal to $L(\theta^*)$. The reason is that the Hessian does not reveal information about L outside an infinitesimal neighborhood of θ^*.

One way to exclude such adversarial case is to assume convexity: when L is convex, a local strong growth implies global growth and thus identifiability.