STAT 581: Advanced Theory of Statistical Inference I **Autumn 2026**

Lecture 3: M-estimation and the maximum likelihood estimator

Instructor: Yen-Chi Chen

⊙ We thank previous instructors: Jon Wellner, Alex Luedtke, Fang Han, and Andrea Rotnitzky.

⊙ Some of this lecture notes are based on the following book:

> [van der Vaart] Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

In particular, Chapter 5 and 7 are useful references.

3.1 Maximum likelihood inference: classical conditions

We start with the classical conditions for the maximum likelihood estimator (MLE). In the next sections, we will discuss other popular examples of M-estimators.

Let $X_1, \ldots, X_n$ be IID from some unknown distribution F. In parametric modeling, we assume that F belongs to a specific family of distributions, indexed by a (often multivariate) parameter $\theta \in \Theta \subset \mathbb{R}^d$. We assume that distributions in this family have a known probability density function (PDF) or probability mass function (PMF), which we denote by $p(x;\theta)$.

For example, for a Gaussian model, $\theta = (\mu, \sigma^2)$ and the PDF is:

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The goal is to estimate the parameter θ from the observed data. Under the likelihood model, the goal is to estimate the underlying parameter θ.

3.1.1 The likelihood function

Given the observed data X, the likelihood function $L(\theta|X)$ is defined as the PDF/PMF evaluated at X, but viewed as a function of the parameter θ.

$$L(\theta|X) = p(X;\theta)$$

The maximum likelihood principle states that we should choose the parameter θ that maximizes the likelihood of observing the data we have. The estimator that achieves this is the *maximum likelihood estimator (MLE)*.

$$\hat{\theta} = \mathsf{argmax}_\theta L(\theta|X)$$

For IID data, the joint PDF is the product of the individual PDFs, so the likelihood is:

$$L(\theta|X_1,\ldots,X_n) = \prod_{i=1}^n p(X_i;\theta)$$

Maximizing the likelihood is equivalent to maximizing its logarithm, which is often mathematically simpler. With this, we define the log-likelihood function to be $\ell(\theta|x) = \log L(\theta|x)$. For IID data, the (total) log-likelihood is:

$$\ell_n(\theta) = \sum_{i=1}^{n} \ell(\theta|X_i) = \sum_{i=1}^{n} \log p(X_i;\theta)$$

The MLE $\hat{\theta}_n$ can then be defined as the maximizer of $\ell_n(\theta)$:

$$\hat{\theta}_n = \mathsf{argmax}_\theta \ell_n(\theta) = \mathsf{argmax}_\theta \bar{\ell}_n(\theta),$$

where $\bar{\ell}_n(\theta) = \frac{1}{n}\ell_n(\theta)$.

In most cases, the maximizer satisfies the first-order condition, i.e., it occurs at zero gradient location. In the case of likelihood function, we define the score function to be

$$S(\ell|x) = \nabla_\theta \ell(\theta|x), \qquad \bar{S}(\theta) = \mathbb{E}[S(\ell|X_1)] = \nabla_\theta \bar{\ell}(\theta), \qquad \bar{S}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(\ell|X_i).$$

We say the MLE solves the score equation if

$$\bar{S}_n(\hat{\theta}_n) = 0, \qquad \bar{S}(\theta^*) = 0.$$

For many common parametric models, the MLE does solve the score equation.

Those *nice* M-estimators often involve utilizing the gradient condition (gradient equals 0) to obtain an estimator. So sometimes they are also called Z-estimators. Thus, a number of literature just interchangeably uses the two terms.

3.1.2 Asymptotic theory of MLE

A key result, which holds even if the true data-generating process is not in the parametric family (i.e., the model is mis-specified), is the asymptotic normality of the MLE. Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, \Sigma^*)$$

for some vector θ^* and covariance matrix Σ^*.

The parameter θ^* is the *population MLE*, defined as the value that maximizes the expected log-likelihood:

$$\theta^* = \mathsf{argmax}_\theta \bar{\ell}(\theta),$$

where

$$\bar{\ell}(\theta) = \mathbb{E}[\ell(\theta|X_1)] = E[\log p(X_1;\theta)] = \int p(x) \log p(x;\theta) dx,$$

where $p(x)$ is the true density of the data. To see why θ^* should be the target of $\hat{\theta}_n$, we first note that for each θ, the law of large numbers implies

$$\bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta|X_i) \xrightarrow{P} \mathbb{E}[\ell(\theta|X_1)] = \bar{\ell}(\theta).$$

Therefore, it is reasonable to view $\theta^* = \mathsf{argmax}_\theta \bar{\ell}(\theta)$ as the target of $\hat{\theta}_n = \mathsf{argmax}_\theta \bar{\ell}_n(\theta)$.

To formally state the asymptotic theory of MLE, we also need to define the Hessian matrices:

$$\bar{H}(\theta) = \nabla_\theta \bar{S}(\theta) = \nabla_\theta \nabla_\theta \bar{\ell}(\theta)$$
$$\bar{H}_n(\theta) = \nabla_\theta \bar{S}_n(\theta) = \nabla_\theta \nabla_\theta \bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \nabla_\theta \ell(\theta|X_i).$$

Theorem 3.1 *Assume the following conditions:*

(M1) The parameter space Θ is compact and θ^* lies in the interior of Θ.

(M2) The MLEs $(\hat{\theta}_n, \theta^*)$ solves the corresponding score equations and are unique.

(M3) All eigenvalues of $\bar{H}(\theta^*)$ are away from 0, i.e., $\bar{H}(\theta^*)$ is invertible.

(M4) There exists a function $\Lambda(x)$ such that

$$\sup_{\theta\in\Theta} \max_{j_1,j_2,j_3} \left| \frac{\partial^3}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}} \ell(\theta|x) \right| \leq \Lambda(x)$$

and $\mathbb{E}[|\Lambda(x)|] < \infty$.

Then we have

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, \Sigma^*),$$

where the asymptotic covariance matrix is

$$\Sigma^* = \bar{H}^{-1}(\theta^*)\mathbb{E}[S(\theta^*|X_1)S(\theta^*|X_1)^T]\bar{H}^{-1}(\theta^*).$$

Conditions (M1) is the common assumption on the parameter space. Note that the compact parameter space is an important requirement with condition (M4). (M2) is a very mild condition that holds for most MLE. (M3) requires the maximizer is well-defined; since we are maximizing the likelihood function, this will implies that all eigenvalues are negative at $\theta = \theta^*$. Assumption (M4) is a critical assumption for ensuring the remainder terms in Taylor expansion is small (via Taylor remainder theorem). Note that (M4) can be relaxed but here we assume this stronger form to make the proof easier and also, it will ensure uniform convergence of log-likelihood, score, and Hessian, which will be useful later.

Warning. Sometimes people only use a second-order derivative in (M4) and apply the mean-value theorem. This idea does NOT work for multivariate θ. The primary reason is that there is NO mean-value theorem for vector-valued functions. A high level idea is that for each coordinate, we do have a mean value theorem. But the location where the mean-value occurs differ from one coordinate to the other. So there is no single point that the mean-value theorem works jointly.

Proof: By (M2), the MLEs solve the score equations

$$\bar{S}_n(\hat{\theta}_n) = 0, \qquad \bar{S}(\theta^*) = 0.$$

Now we consider the quantity:

$$\bar{S}_n(\theta^*) - \bar{S}(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} S(\theta^*|X_i) - \mathbb{E}[S(\theta^*|X_i)],$$

which has a sample average form. By multivariate central limit theorem, we know that

$$\sqrt{n}(\bar{S}_n(\theta^*) - \bar{S}(\theta^*)) \xrightarrow{d} N(0, \mathbb{E}[S(\theta^*|X_i)S(\theta^*|X_i)^T]). \tag{3.1}$$

Thus, this motivates us to investigate the quantity $\bar{S}_n(\theta^*) - \bar{S}(\theta^*)$.

Using the score equation,

$$\bar{S}(\theta^*) = 0 = \bar{S}_n(\hat{\theta}_n),$$

we have

$$\begin{aligned} \bar{S}_n(\theta^*) - \bar{S}(\theta^*) &= \bar{S}_n(\theta^*) - \bar{S}_n(\hat{\theta}_n) \\ &= -[\bar{S}_n(\hat{\theta}_n) - \bar{S}_n(\theta^*)] \\ &= -\nabla_\theta \bar{S}_n(\theta^*)(\hat{\theta}_n - \theta^*) + R_n, \end{aligned} \tag{3.2}$$

where $R_n \in \mathbb{R}^d$ is the Taylor remainder, which has an integral form that the j-th element is

$$R_{n,j} = \int_{t=0}^{t=1} (\hat{\theta}_n - \theta^*)^T \underbrace{[\nabla_\theta \nabla_\theta \bar{S}_{n,j}((1-t)\theta^* + t\hat{\theta}_n)]}_{\Psi_{n,j}} (\hat{\theta}_n - \theta^*) dt$$

such that $\bar{S}_{n,j}(\theta)$ is the j-th element of $\bar{S}_n(\theta)$. Using the upper bound in (M4), every element in the matrix $\Psi_{n,j}$ is bounded by $\frac{1}{n}\sum_{i=1}^b \Lambda(X_i)$, and (M4) requires $\mathbb{E}[|\Lambda(X_1)|] < \infty$, so the strong law of large numbers applies and thus, we conclude that

$$R_n = O_P(\|\hat{\theta}_n - \theta^*\|^2),$$

which is negligible compare to the other quantity.

Thus, we conclude that

$$\bar{S}_n(\theta^*) - \bar{S}(\theta^*) = -\nabla_\theta \bar{S}_n(\theta^*)(\hat{\theta}_n - \theta^*) + O_P(\|\hat{\theta}_n - \theta^*\|^2).$$

By the law of large numbers, the matrix $\nabla_\theta \bar{S}_n(\theta^*)$ has a limit

$$\nabla_\theta \bar{S}_n(\theta^*) = \bar{H}_n(\theta^*) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \nabla_\theta \ell(\theta|X_i) \xrightarrow{P} \bar{H}(\theta^*).$$

By (M3), the matrix $\bar{H}(\theta^*)$ is invertible, so

$$[\nabla_\theta \bar{S}_n(\theta^*)]^{-1} \xrightarrow{P} \bar{H}^{-1}(\theta^*).$$

Combining this result with equation (3.1) and using Slutsky's theorem, we conclude that

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta^*) &= [\nabla_\theta \bar{S}_n(\theta^*)]^{-1} \sqrt{n}[\bar{S}_n(\theta^*) - \bar{S}(\theta^*)] + o_P(1) \\ &\xrightarrow{d} N(0, \bar{H}^{-1}(\theta^*)\mathbb{E}[S(\theta^*|X_1)S(\theta^*|X_1)^T]\bar{H}^{-1}(\theta^*)). \end{aligned}$$

■

Remark 3.2 *Here are some important remarks.*

- **Sandwich estimator.** *There is a simple estimator of the underlying covariance matrix via the plug-in approach:*

$$\hat{\Sigma}^* = \bar{H}_n^{-1}(\hat{\theta}_n) \left[\frac{1}{n}\sum_{i=1}^n S(\hat{\theta}_n|X_i)S(\hat{\theta}_n|X_i)^T\right] \bar{H}_n^{-1}(\hat{\theta}_n).$$

  This estimator is also known as the sandwich estimator.

- **Fisher's information.** *The matrix* $I(\theta^*) \equiv \mathbb{E}[S(\theta^*|X_1)S(\theta^*|X_1)^T]$ *is called* Fisher's information matrix. *So the sandwich estimator uses a plug-in for the Hessian matrix and the Fisher's information matrix.*

- **Bootstrap covariance estimator.** *In case we do not want to use sandwich estimator, we can use the empirical bootstrap to estimate the covariance matrix: we generate* $X_1^*, \cdots, X_n^*$ *by sampling with replacement from* $X_1, \cdots, X_n$ *and compute the MLE using* $X_1^*, \cdots, X_n^*$*, denoted as* $\hat{\theta}_n^*$*. Repeat this process* B *times, leading to* $\hat{\theta}_n^{*(1)}, \cdots, \hat{\theta}_n^{*(B)}$*. We use the sample covariance matrix of these* B *bootstrap MLEs as the estimator of the covariance matrix.*

- **Model correctness.** *We do NOT assume the model is correct. When the model is correct, i.e., there exists* $\theta_0 \in \Theta$ *that generates our data, then we have two additional results:*

  - *The population MLE* $\theta^* = \theta_0$*.*
  - *The Hessian matrix* $\bar{H}(\theta^*) = -\mathbb{E}[S(\theta^*|X_1)S(\theta^*|X_1)^T]$*, so the asymptotic covariance matrix* $\Sigma^* = \bar{H}(\theta^*) = -I(\theta^*)$*, which is also the Fisher's information matrix.*

- **Mean-value theorem.** *While we cannot directly use the mean-value theorem to deal with the Taylor expansion, it is still possible to use it to relax assumptions (M4). The trick is: we apply the mean value theorem to each element of the vector* $\bar{S}_n(\hat{\theta}_n) - \bar{S}_n(\theta^*)$*. For the* j*-th element, we have*

$$\bar{S}_{n,j}(\hat{\theta}_n) - \bar{S}_{n,j}(\theta^*) \in \mathbb{R},$$

where $\bar{S}_{n,j}(\theta) = \frac{1}{n}\sum_{i=1}^{\frac{\partial}{\partial\theta_j}} \ell(\theta|X_i)$*. The mean value theorem implies that there exists* $\tilde{\theta}_{n,j}$ *lies between* $\hat{\theta}_n$ *and* θ^* *such that*

$$\bar{S}_{n,j}(\hat{\theta}_n) - \bar{S}_{n,j}(\theta^*) = [\nabla_\theta \bar{S}_{n,j}(\tilde{\theta}_{n,j})]^T(\hat{\theta}_n - \theta^*).$$

Now we define the matrix $B_n \in \mathbb{R}^{d\times d}$ *such that the* j*-th row of* B_n *is* $[\nabla_\theta \bar{S}_{n,j}(\tilde{\theta}_{n,j})]^T$*. Then we can still have*

$$B_n \xrightarrow{P} \bar{H}(\theta^*)$$

without assuming third-order derivative is upper bounded because $\tilde{\theta}_{n,j} \xrightarrow{P} \theta^*$ *for each* j*.*

3.2 Examples of M-estimators

Finding the estimator by maximizing or minimizing a criterion is a very common procedure in both Statistics and Machine Learning. In Machine Learning, this occurs in the **empirical risk minimization (ERM)**, where our estimator is the minimizer of the empirical risk

$$\hat{\theta}_n = \mathsf{argmin}_\theta \quad \bar{R}_n(\theta), \qquad \bar{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta, X_i), \tag{3.3}$$

where $\mathcal{L}(\theta, X_i)$ is the loss of the model when parameter is θ and observation X_i.

Clearly, if we set the loss function to be the negative log-likelihood function, i.e., $\mathcal{L}(\theta, X_i) = -\ell(\theta|X_i)$, then the MLE is the ERM estimator. Using our analysis in the MLE, we expect the ERM estimator converges to the *population risk minimizer*:

$$\theta^* = \mathsf{argmin}_\theta \quad \bar{R}(\theta), \qquad \bar{R}(\theta) = \mathbb{E}\left[\mathcal{L}(\theta, X_1)\right]. \tag{3.4}$$

The population risk $\bar{R}(\theta)$ is often interpreted as the expected loss of making a prediction on a new observation.

The asymptotic theory of $\hat{\theta}_n$ toward θ^* in Theorem 3.1 applies to any of these ERM estimators as long as (M1-4) hold.

Here are some examples of the ERM problems.

Example 3.3 (Least square regression) *Consider a regression problem where we want to predict Y using X. Our prediction can be written as a function $m_\theta(x)$, indexed by the parameter θ. The linear model is the case where we assume $m_\theta(x) = \theta^T x$. The least square approach estimates θ by*

$$\hat{\theta}_{LS} = \mathsf{argmin} \sum_{i=1}^{n} (Y_i - m_\theta(X_i))^2 = \mathsf{argmin} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_\theta(X_i))^2,$$

which is the ERM with loss function $\mathcal{L}(\theta, X_i) = (Y_i - m_\theta(X_i))^2$.

Thus, the population least square parameter is $\theta^*_{LS} = \mathsf{argmin}_\theta \mathbb{E}\left[(Y_1 - m_\theta(X_1))^2\right]$ and the asymptotic normality in Theorem 3.1 applies.

Example 3.4 (Logistic regression) *When $Y \in \{0, 1\}$, the regression problem is related to the binary classification problem. A popular approach in this scenario is the logistic regression model, where we model the log-odds*

$$\log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = f_\theta(x).$$

In the simplest form of the logistic regression, $f_\theta(x) = \theta^T x$ is the linear model. The log-odds model implies the following probability model:

$$P(Y = 1|X = x) = \frac{e^{f_\theta(x)}}{1 + e^{f_\theta(x)}} \equiv \phi(x; \theta).$$

The maximum likelihood principle can be applied to this case, leading to the following estimator

$$\begin{aligned}
\hat{\theta}_n &= \mathsf{argmax}_\theta \sum_{i=1}^{n} Y_i \log \phi(X_i; \theta) + (1 - Y_i) \log(1 - \phi(X_i; \theta)) \\
&= \mathsf{argmax}_\theta \sum_{i=1}^{n} Y_i f_\theta(X_i) - \log[1 + e^{f_\theta(X_i)}] \\
&= \mathsf{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} -Y_i f_\theta(X_i) + \log[1 + e^{f_\theta(X_i)}].
\end{aligned}$$

Again, this is the ERM estimator and the population quantity $\hat{\theta}_n$ is converging to is

$$\theta^* = \mathsf{argmin}_\theta \mathbb{E}\left[-Y_1 f_\theta(X_1) + \log[1 + e^{f_\theta(X_1)}]\right].$$

Theorem 3.1 and assumptions (M1-4) imply the asymptotic normality of $\hat{\theta}_n - \theta^*$.

Example 3.5 (Classification) *Suppose $Y \in \{0, 1, \cdots, K\}$ be a class label and X is our feature vector. A classifier makes a prediction about the label from a given feature vector x, so it can be written as $c(x)$ and when the classifier is determined by a set of parameter θ, we write it as $c_\theta(x)$. The classification problem is often done by introducing a loss function $L(y_1, y_2)$ that measures the amount of loss incurred when the true label is y_2 but our predicted label is y_1. A common loss function for classification is the $0 - 1$ loss where*

$L(y_1, y_2) = I(y_1 \neq y_2)$. Namely, we lose a value of 1 if we are making a mistake in the prediction and do not lose anything if we are correct.

The classifier is often trained by minimizing the prediction error. Since classifiers are now parameterized by θ, training a classifier is equivalent to estimating/learning the underlying parameter θ. The training is often done by the ERM:

$$\hat{\theta}_n = \mathsf{argmin}_\theta \sum_{i=1}^n L(c_\theta(X_i), Y_i) = \mathsf{argmin}_\theta \frac{1}{n} \sum_{i=1}^n L(c_\theta(X_i), Y_i).$$

By ERM and the above analysis, it is clearly that the population quantity corresponding to $\hat{\theta}_n$ is

$$\theta^* = \mathsf{argmin}_\theta \mathbb{E}[L(c_\theta(X_1), Y_1)].$$

Example 3.6 (Mode estimation with kernel density estimator) *Now we consider a slightly different problem in nonparametric estimation. Suppose our data $X_1, \cdots, X_n \sim p_0$, where p_0 is an unknown PDF. Our goal is to estimate $m_0 = \mathsf{argmax}_x p_0(x)$, the mode of p_0.*

Intuitively, a nonparametric method to estimating m_0 is via a plug-in estimate, where we first estimate the PDF $\hat{p}$ and then construct our mode estimator as $\hat{m}_0 = \mathsf{argmax}_x \hat{p}(x)$. Now suppose we use the kernel density estimator (KDE), where $\hat{p} = \hat{p}_h$ is

$$\hat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right),$$

where $h > 0$ is the smoothing bandwidth that controls the amount of smoothing and $K(\cdot) \geq 0$ is a kernel function such as a Gaussian. In this case, the mode estimator is

$$\hat{m}_h = \mathsf{argmax}_x \hat{p}_h(x),$$

which corresponds to estimating the mode of a smoothed density:

$$m_h^* = \mathsf{argmax}_x \bar{p}_h(x), \qquad \bar{p}_h(x) = \mathbb{E}\left[\frac{1}{h^d} K\left(\frac{X_1 - x}{h}\right)\right].$$

When $h \to 0$, one can show that

$$\bar{p}_h(x) - p_0(x) = O(h^2)$$

under conventional assumptions.

The ERM theory (Theorem 3.1) shows that $\hat{m}_h$ has asymptotic normality for estimating $\bar{m}_h$ when h is fixed. When $h \to 0$, we may modify the derivation in Theorem 3.1 and obtain

$$\sqrt{nh^{d+2}}(\hat{m}_h - \bar{m}_h) \xrightarrow{d} N(0, \Sigma^*)$$

for some covariance matrix Σ^*.

3.3 General theory for M-estimators

In the previous section, we have seen four key conditions (M1-4). Some of these conditions are not necessary. In this section, we will discuss how to relax these conditions.

Recall that in the general M-estimation setup, we have IID observations $X_1, \cdots, X_n$ from an unknown distribution and our goal is to learn a parameter $\theta \in \Theta \subset \mathbb{R}^d$ that maximizes the empirical criterion function

$$\hat{\theta}_n = \mathsf{argmax}_\theta \frac{1}{n} \sum_{i=1}^n \ell(\theta | X_i). \tag{3.5}$$

The population maximizer is

$$\theta^* = \mathsf{argmax}_\theta \, \mathbb{E}[\ell(\theta | X_1)]. \tag{3.6}$$

For simplicity, we write

$$\hat{\ell}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta | X_i)$$

to be the empirical criterion function (its negative value can be viewed as the empirical risk) and

$$\bar{\ell}(\theta) = \mathbb{E}[\ell(\theta | X_1)]$$

to be the population criterion function.

3.3.1 Consistency of M-estimator

If our goal is just to show that the M-estimator is consistent, i.e.,

$$\hat{\theta}_n - \theta^* = o_P(1),$$

then clearly conditions for asymptotic normality (M1-4) is more than enough.

Here we introduce a useful theorem.

Theorem 3.7 (Consistency of M-estimator: Theorem 5.7 in van der Vaart) *Consider estimator $\tilde{\theta}_n$ (not necessarily M-estimator). Assume that following:*

- **(C1) Uniform error bound:** $\sup_{\theta \in \Theta} \left| \hat{\ell}_n(\theta) - \bar{\ell}(\theta) \right| = o_P(1)$
- **(C2) Well-defined maximizer:** *for any* $\epsilon > 0$, $\sup_{\theta : \|\theta - \theta^*\| > \epsilon} \bar{\ell}(\theta) < \bar{\ell}(\theta^*)$.
- **(C3) Accuracy of $\tilde{\theta}_n$:** $\hat{\ell}_n(\tilde{\theta}_n) \geq \hat{\ell}_n(\theta^*) + o_P(1)$,.

Then we have $\tilde{\theta}_n - \theta^* = o_P(1)$.

Theorem 3.8 applies to a broader class of estimator that is not limited to the M-estimator. Note that only the condition (C3) is a requirement on the estimator. The M-estimator automatically satisfies this (and can drop the $o_P(1)$). Condition (C2) is a reasonable condition otherwise the population maximizer θ^* is not well-defined. Therefore, for an M-estimator, the only condition we need to verify is (C1). This would involve something called *Glivenko-Cantelli* class and empirical process theory, which you will learn more in STAT 582-583. Generally speaking, for common parametric models such as normal, exponential, Binomial, and Poisson, all these conditions hold.

As we have mentioned previously, sometimes we write the estimator in the form of Z-estimator, i.e., estimators are defined via solving the estimating equations

$$(\hat{\theta}_n, \theta^*) : \hat{S}_n(\hat{\theta}_n) \equiv \frac{1}{n} \sum_{i=1}^n s(\hat{\theta}_n | X_i) = 0, \qquad \bar{S}(\theta^*) \equiv \mathbb{E}[s(\theta^* | X_1)] = 0.$$

For Z-estimators, we also have a similar result as Theorem 3.8.

Theorem 3.8 (Consistency of Z-estimator: Theorem 5.9 in van der Vaart) *Consider estimator $\tilde{\theta}_n$ (not necessarily M-estimator). Assume that following:*

- **(Z1) Uniform error bound:** $\sup_{\theta\in\Theta}\left|\hat{S}_n(\theta)-\bar{S}(\theta)\right| = o_P(1)$
- **(Z2) Well-defined maximizer:** *for any* $\epsilon>0$, $\inf_{\theta:\|\theta-\theta^*\|>\epsilon}\bar{S}(\theta) > 0 = \bar{S}(\theta^*)$.
- **(Z3) Accuracy of $\tilde{\theta}_n$:** $\hat{S}_n(\tilde{\theta}_n) = o_P(1)$.

Then we have $\tilde{\theta}_n - \theta^* = o_P(1)$.

3.3.2 Asymptotic Normality of M-estimator

In Theorem 3.1, we have seen the asymptotic normality of the M-estimator. In our derivation, we do remark that the conditions are sufficient (and convenient) conditions that works for most models. Now we relax these conditions and cast it in a more abstract way.

Theorem 3.9 (Asymptotic normality of M-estimator: Theorem 5.23 in van der Vaart) *Assume the following:*

- **(V1) Smoothness of criterion around θ^*.** *For any θ in an open subset around θ^*,*
  - **(V1-1):** *$\ell(\theta|x)$ is differentiable at θ^* for P-almost every X and with a derivative $s(\theta^*|x) = \nabla\ell(\theta^*|x)$,*
  - **(V1-2):** *for any θ_1,θ_2 in this open subset, there exists $\eta(x)$ such that $|\ell(\theta_1|x)-\ell(\theta_2|x)| \le \eta(x)\|\theta_1-\theta_2\|$ and $\mathbb{E}(\eta^2(X)) < \infty$. The quantity $\eta(x)$ behaves like the Lipschitz constant of the criterion function.*
- **(V2) Second-order Taylor expansion:** *The population criterion $\bar{\ell}(\theta)$ admits a second-order Taylor expansion at $\theta=\theta_0$ in the sense that*

$$\bar{\ell}(\theta) = \bar{\ell}(\theta^*) + \frac{1}{2}(\theta-\theta^*)^T\bar{H}(\theta^*)(\theta-\theta^*) + o(\|\theta-\theta^*\|^2),$$

  where the Hessian matrix $\bar{H}(\theta^*)$ is non-singular.
- **(V3) Consistent estimator:** *The M-estimator $\hat{\theta}_n - \theta^* = o_P(1)$.*

Then we have

$$\sqrt{n}(\hat{\theta}_n-\theta^*) = \bar{H}^{-1}(\theta^*)\frac{1}{\sqrt{n}}\sum_{i=1}^n s(\theta^*|X_i) + o_P(1).$$

Example 3.10 (Asymptotic normality of sample median) *Now we consider an unconventional example that Theorem 3.1 fails while we can still apply Theorem 3.9: the sample median. Here we assume that the data consists of IID univariate random variables $X_1,\cdots,X_n \sim F$, where F admits a PDF p that has a positive density at the neighborhood of the population median $F^{-1}(0.5)$.*

Consider the criterion function

$$\ell(\theta|x) = -|x-\theta|.$$

The sample criterion function is

$$\hat{\ell}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}|X_i - \theta|.$$

You can easily verify that the maximizer of $\hat{\ell}_n(\theta)$ is the sample median $\hat{\theta}_n = \mathsf{Median}(X_1, \cdots, X_n)$. Similarly, the population criterion is $\ell(\theta) = -\mathbb{E}|X_1 - \theta|$ and $\theta^* = F^{-1}(0.5)$ is the population median.

Clearly, we cannot apply Theorem 3.1 since $\ell(\theta|x)$ is not differentiable at $\theta = x$. However, condition (V1) in Theorem 3.9 is still valid since this single point has 0 probability to occur. It is clear that the derivative $s(\theta^*|x) = \nabla\ell(\theta^*|x) = -\mathsf{sign}(x - \theta^*)$. For $\ell(\theta|x) = -|x - \theta|$, it is clearly that $|\ell(\theta_1|x) - \ell(\theta_2|x)| \le 1 \cdot |\theta_1 - \theta_2|$. Thus, both (V1-1), (V1-2) hold.

To verify (V2), a direct computation shows

$$\begin{aligned}\bar{\ell}(\theta) &= -\mathbb{E}|X_1 - \theta| \\ &= -\mathbb{E}[(X_1 - \theta)I(X_1 < \theta)] - \mathbb{E}[(\theta - X_1)I(X_1 > \theta)] \\ &= -\int_{-\infty}^{\theta}(x - \theta)p(x)dx - \int_{\theta}^{\infty}(\theta - x)p(x)dx.\end{aligned}$$

Using integration by parts and the fact that $\lim_{x\to-\infty} xF(x) = 0$, the first term becomes

$$-\int_{-\infty}^{\theta}(x - \theta)p(x)dx = -[(x - \theta)F(x)]_{-\infty}^{\theta} + \int_{-\infty}^{\theta}F(x)dx = \int_{-\infty}^{\theta}F(x)dx.$$

Similarly, the second term can be obtained using $p(x) = -\frac{d}{dx}(1 - F(x))$,

$$-\int_{\theta}^{\infty}(\theta - x)p(x)dx = -[(\theta - x)(1 - F(x))]_{\theta}^{\infty} + \int_{\theta}^{\infty}(1 - F(x))dx = \int_{\theta}^{\infty}(1 - F(x))dx.$$

Therefore, we conclude that

$$\bar{\ell}(\theta) = \int_{-\infty}^{\theta}F(x)dx + \int_{\theta}^{\infty}(1 - F(x))dx.$$

Clearly, the two derivative of it is:

$$\bar{S}(\theta) = \frac{d}{d\theta}\bar{\ell}(\theta) = F(\theta) - (1 - F(\theta)) = 2F(\theta) - 1, \qquad \bar{H}(\theta) = \frac{d}{d\theta}\bar{S}(\theta) = 2p(\theta).$$

So the Hessian is $\bar{H}(\theta^*) = 2p(\theta^*) > 0$ when the PDF has a positive density at the population median.

You can verify on your own about the consistency of sample median (V3) so we skip it here.

With the above results, we conclude that the sample median $\hat{\theta}_n = \mathsf{Median}(X_1, \cdots, X_n)$ is an asymptotic normal estimator of the population median θ^* in the sense that

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N\left(0, \frac{1}{4p^2(\theta^*)}\right).$$

While we may use this normality to construct a confidence interval, it will require a density estimator, which generally requires further smoothness conditions and we typically cannot estimate it at a good convergence rate. To bypass this, we would recommend to use the bootstrap method, which avoids the need of a density estimation.

Example 3.11 (Non-linear regression) *Now we consider a non-linear regression problem, where our data are IID*

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

from the model

$$Y_i = m_{\theta^*}(X_i) + \epsilon_i, \qquad (X_i, \epsilon_i) \sim q_X(x) q_E(e),$$

where q_X and q_E are some PDFs (we assume X_i, ϵ_i are independent).

The model class $m_\theta(x)$ is something we know but can be complex such as $m_\theta(x) = \theta_1 + \theta_2 e^{\theta_3 (x-\theta_4)^2}$. Our goal is to estimate θ^* using our data. Note that in this problem, we assume that the model is correctly specified.

A conventional approach is to estimate θ^* via the least square method:

$$\hat{\theta}_n = \mathsf{argmin}_\theta \, \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_\theta(X_i))^2,$$

so our criterion function $\ell(\theta | x, y) = -(y - m_\theta(x))^2$.

The gradient condition (V1) will require placing conditions on

$$s(\theta | x, y) = \nabla \ell(\theta | x, y) = (y - m_\theta(x)) \nabla_\theta m_\theta(x)$$

around θ^* to ensure the smoothness of $s(\theta | x, y)$. Note that you can see that an easy way to ensure this holds is to assume both X, Y are supported on a compact set (so they are bounded) and restrict the parameter space Θ to be a compact set and assume θ^* lie in the interior of the parameter space.

The second-order condition (V2) will put constraints on the $\bar{\ell}(\theta) = -\mathbb{E}[(Y_1 - m_\theta(X_1))^2]$. A direct computation shows that

$$\begin{aligned} \bar{\ell}(\theta) &= -\mathbb{E}[(Y_1 - m_\theta(X_1))^2] \\ &= -\mathbb{E}((m_{\theta^*}(X_1) - m_\theta(X_1) + \epsilon_1)^2) \\ &= -\mathbb{E}((m_{\theta^*}(X_1) - m_\theta(X_1))^2) + \mathbb{E}(\epsilon_1^2). \end{aligned}$$

We can drop the last term $\mathbb{E}(\epsilon_1^2)$ since it does not involve θ.

Now if the model $m_\theta(x)$ is smooth in θ at $\theta = \theta^*$, we would expect a Taylor expansion

$$m_\theta(x) - m_{\theta^*}(x) \approx (\theta - \theta^*)^T g(\theta^*, x), \tag{3.7}$$

where $g(\theta, x) = \nabla_\theta m_\theta(x)$.

Then we immediately have

$$\bar{\ell}(\theta) \approx -(\theta - \theta^*)^T \mathbb{E}(g(\theta^*, X_1) g(\theta^*, X_1)^T)(\theta - \theta^*) + c,$$

where $c = \mathbb{E}(\epsilon_1^2)$ is just a constant. Thus, the 'Hessian' is $\bar{H}(\theta^*) = \mathbb{E}(g(\theta^*, X_1) g(\theta^*, X_1)^T)$. Note that in the above example, the matrix $\bar{H}(\theta^*)$ is not the Hessian matrix since it only involves the first-order derivative. The matrix $I(\theta^*) \equiv \mathbb{E}(g(\theta^*, X_1) g(\theta^*, X_1)^T)$ is the Fisher's information matrix.

To apply Theorem 3.9, we just need to place conditions on the model $m_\theta(x)$ with respect to θ to ensure the neighborhood condition in (V1) as well as the approximation in in equation (3.7) (from condition (V2)) holds.

3.3.3 Influence function

In Theorem 3.9, we have seen the asymptotic linearity:

$$\hat{\theta}_n = \theta^* + \frac{1}{n}\sum_{i=1}^{n} \underbrace{\bar{H}^{-1}(\theta^*)s(\theta^*|X_i)}_{=f^*(X_i)} + o_P(1/\sqrt{n}).$$

Ignoring the smaller order term, this shows that the estimator $\hat{\theta}$ is the population parameter θ^* plus a linear noise average over n IID observations. So the function $f^*(x) = \bar{H}^{-1}(\theta^*)s(\theta^*|x)$ is called the *asymptotic influence function* since it measures the influence of an observation at x.

To see how $f^*(X_k)$ measures the influence of X_k, we consider a leave-one-out scenario that $\hat{\theta}_{n,-k}$ is the M-estimator without k-th observation. Then we have

$$\begin{aligned}
\hat{\theta}_n - \hat{\theta}_{n,-k} &\approx \frac{1}{n}\sum_{i=1}^{n} f^*(X_i) - \frac{1}{n-1}\sum_{i\neq k}^{n} f^*(X_i) \\
&\approx f^*(X_k) - \frac{1}{n(n-1)}\sum_{i\neq k}^{n} f^*(X_i) \\
&\approx f^*(X_k).
\end{aligned}$$

3.4 Quadratic mean differentiability

Theorem 3.9 relaxes the classical third-order derivative conditions from Theorem 3.1 to a second-order derivative condition. But in the non-linear regression (Example 3.11), we have seen an interesting phenomenon: it seems to be possible to only require a first-order derivative condition when the model is correctly specified. In this section, we will use the concept of *quadratic mean differentiability* to formalize this.

We use the notation $p_\theta(x) = p(x;\theta)$ do denote the PDF under parameter θ. In this way, we can easily say p_θ is a statistical model. A statistical model p_θ is **quadratic mean differentiable (QMD)** if there exists a vector $s(\theta|x) \in \mathbb{R}^d$ such that

$$(QMD) \qquad \int \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2}h^T s(\theta|x)\sqrt{p_\theta(x)}\right]^2 dx = o(\|h\|^2) \tag{3.8}$$

for any $h \to 0$. When the parameter does not change the support of p_θ, equation (3.8) can be equivalently written as

$$\int \left[\sqrt{\frac{p_{\theta+h}(x)}{p_\theta(x)}} - 1 - \frac{1}{2}h^T s(\theta|x)\right]^2 p_\theta(x)dx = o(\|h\|^2). \tag{3.9}$$

How should we think of the vector $s(\theta|x)$? A simple way is to view it as an approximation of $\sqrt{p_{\theta+h}(x)} -$

$\sqrt{p_\theta(x)}$ when $h \to 0$. By Taylor approximation,

$$\begin{aligned}\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} &\approx h^T \nabla_\theta \sqrt{p_\theta(x)} \\ &= h^T \left[\frac{1}{\sqrt{p_\theta(x)}} \nabla_\theta \sqrt{p_\theta(x)}\right] \cdot \sqrt{p_\theta(x)} \\ &= h^T \left[\frac{1}{2p_\theta(x)} \nabla_\theta p_\theta(x)\right] \cdot \sqrt{p_\theta(x)} \\ &= \frac{1}{2} h^T \nabla_\theta \log p_\theta(x) \cdot \sqrt{p_\theta(x)} \\ &= \frac{1}{2} h^T s(\theta|x) \cdot \sqrt{p_\theta(x)},\end{aligned}$$

so

$$s(\theta|x) = \nabla_\theta \log p(x;\theta)$$

is the usual *score function*.

The key to QMD is that the model is smooth in the sense that for any sequence of vectors $h \to 0$, the limiting direction is always the score vector $s(\theta|x)$. While most smooth models such as Gaussian, exponential, Poisson, Binomial are QMD, not all models are QMD and here is a famous counterexample.

Example 3.12 (Uniform distribution: not QMD) *Consider a univariate uniform distribution over* $[0,\theta]$. *We will show that it is not QMD.*

For $X \sim \mathsf{Uni}[0,\theta]$, *its PDF is*

$$p_\theta(x) = \frac{1}{\theta} I(0 \le x \le \theta).$$

A simple way to check QMD is to use the fact that equation (3.8) *requires*

$$\int \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 dx = O(h^2)$$

when we do not include the linear term. This is because equation (3.8) *implies*

$$\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} = \frac{1}{2} h^T s(\theta|x) \sqrt{p_\theta(x)} + o(h),$$

where the $o(h)$ *term has to remains* $o(h^2)$ *after square integral.*

For the uniform distribution, we have

$$\begin{aligned}\int \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 dx &= \int_0^{\theta+h} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 dx \\ &= \int_0^{\theta} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 dx + \int_\theta^{\theta+h} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 dx.\end{aligned}$$

You can easily verify that the first integral is fine

$$\int_0^{\theta} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 dx = O(h^2).$$

However, using the fact $\sqrt{p_\theta(x)} = 0$ *for* $x \in [\theta, \theta + h]$ *(WLOG, we assume* $h > 0$*) that the second integral becomes*

$$\int_\theta^{\theta+h} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 dx = \int_\theta^{\theta+h} p_{\theta+h}(x)dx = \frac{h}{\theta+h} = O(h).$$

The second integral is at the order of $O(h)$*, not* $O(h^2)$*! Therefore, the uniform distribution is not QMD.*

For a smooth model in the QMD, we have the following useful theorem about the local behavior of the model.

Theorem 3.13 (Likelihood ratio; Theorem 7.2. in van der Vaart) *Suppose the parameter space* Θ *is an open subset of* $\mathbb{R}^d$ *and the model* p_θ *is QMD at* θ*. Then we have*

- **Score equation.** $\mathbb{E}(s(\theta|X)) = 0$ *when* $X \sim p_\theta$*.*
- **Fisher's information matrix.** *The Fisher's information matrix* $I(\theta) = \mathbb{E}(s(\theta|X)s(\theta|X)^T)$ *exists when* $X \sim p_\theta$*.*
- **Local approximation to the likelihood ratio.** *For every converging sequence* $h_n \to h \in \mathbb{R}^d$*,*

$$\log \prod_{i=1}^n \frac{p_{\theta+h_n/\sqrt{n}}(X_i)}{p_\theta(X_i)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T s(\theta|X_i) - \frac{1}{2} h^T I(\theta) h + o_P(1),$$

  where $o_P(1)$ *term is under the model* p_θ*.*

Theorem 3.13 further implies the following asymptotic normality condition when the likelihood function is Lipschitz.

Theorem 3.14 (Asymptotic normality; Theorem 5.39 in van der Vaart) *Suppose the parameter space* Θ *is an open subset of* $\mathbb{R}^d$ *and assume the following conditions:*

(Q1) *The model* p_θ *is QMD at* θ^**.*

(Q2) *The Fisher's information matrix* $I(\theta^*) = \mathbb{E}(s(\theta^*|X)s(\theta^*|X)^T)$ *is non-singular when* $X \sim p_{\theta^*}$*.*

(Q3) *There exists smooth* $\eta(x)$ *such that* $|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \le \eta(x)\|\theta_1 - \theta_2\|$ *in the neighborhood of* θ^* *and* $\int \eta^2(x) p_{\theta^*}(x)dx < \infty$*.*

(Q4) *The MLE* $\hat{\theta}_n$ *is consistent.*

Then

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \bar{I}^{-1}(\theta^*) \frac{1}{\sqrt{n}} \sum_{i=1}^n s(\theta^*|X_i) + o_P(1).$$

Compared to Theorem 3.9, the major difference is that we replace the second-order Taylor expansion condition (V2) with the QMD condition (Q1). Moreover, the Hessian matrix in the influence function is replaced by Fisher's information matrix.

Proof: This theorem is proved by verifying conditions in Theorem 3.9.

We will focus on showing the main condition (V2) that QMD implies that the population log-likelihood function admits a second-order Taylor expansion.

One way to prove this is via the likelihood ratio of Theorem 3.13. But here we use an alternative way to prove it by directly expand the QMD.

Let $X \sim p_{\theta^*}$. To verify (V2), we essentially need to show that for any $\epsilon \to 0$ and $h \in \mathbb{R}^d$,

$$\sup_{h:\|h\|=1} \left| \underbrace{\mathbb{E}(\log p_{\theta^*+\epsilon h}(X))}_{=\bar{\ell}(\theta^*+\epsilon h)} - \underbrace{\mathbb{E}(\log p_{\theta^*}(X))}_{=\bar{\ell}(\theta^*)} - \frac{\epsilon^2}{2} h^T H(\theta^*) h \right| = o(\epsilon^2) \tag{3.10}$$

for some non-singular matrix $H(\theta^*)$.

So a key quantity to consider is the difference between the first two terms:

$$\Delta^*(\epsilon, h) \equiv \bar{\ell}(\theta^* + \epsilon h) - \bar{\ell}(\theta^*) = \mathbb{E}\left[\log\left(\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}\right)\right] = \mathbb{E}\left[2\log\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}}\right]$$

We know that as $\epsilon \to 0$, the ratio $\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}}$ should be approaching 1. So we utilize the following fact about $\log(1+\gamma)$:

$$\log(1+\gamma) = \gamma - \frac{1}{2}\gamma + o(\gamma^2).$$

And we choose $\gamma = \sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}} - 1$, which leads to

$$2\log\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}} = \underbrace{2\left(\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}} - 1\right)}_{(A)} - \underbrace{\left(\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}} - 1\right)^2}_{=(B)} + \underbrace{o\left(\left\|\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}} - 1\right\|^2\right)}_{(C)}.$$

Thus, $\Delta^*(\epsilon, h) = \mathbb{E}_{X\sim p_{\theta^*}}[(A) + (B) + (C)]$.

Clearly, term (C) is the smaller-order term so we ignore. In fact arguing how (C) can be dropped is a bit more involved. See the proof of Theorem 5.39 of [van der Vaart]. We focus on terms (A) and (B).

The expectation of (A) under $X \sim p_{\theta^*}$ is

$$\begin{aligned}
\mathbb{E}\left(2\left(\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}} - 1\right)\right) &= 2\int \left[\sqrt{p_{\theta^*+\epsilon h}(x)} - \sqrt{p_{\theta^*}(x)}\right]\sqrt{p_{\theta^*}(x)}dx \\
&= 2\int \sqrt{p_{\theta^*+\epsilon h}(x)p_{\theta^*}(x)} - 2 \\
&= -\int \left[\sqrt{p_{\theta^*+\epsilon h}(x)} - \sqrt{p_{\theta^*}(x)}\right]^2 dx \\
&= -H^2(p_{\theta^*+\epsilon h}, p_{\theta^*}),
\end{aligned}$$

where H^2 is called the Hellinger distance.

The expectation of (B) turns out to be the same Hellinger distance:

$$\mathbb{E}\left(\sqrt{\frac{p_{\theta^*+\epsilon h}(X)}{p_{\theta^*}(X)}} - 1\right)^2 = \int \left[\sqrt{p_{\theta^*+\epsilon h}(x)} - \sqrt{p_{\theta^*}(x)}\right]^2 dx = H^2(p_{\theta^*+\epsilon h}, p_{\theta^*}).$$

As a result, we have shown that

$$\Delta^*(\epsilon, h) \equiv \bar{\ell}(\theta^* + \epsilon h) - \bar{\ell}(\theta^*) = -2H^2(p_{\theta^*+\epsilon h}, p_{\theta^*}). \tag{3.11}$$

Recall the QMD requires

$$\int \left[\underbrace{\sqrt{p_{\theta+\epsilon h}(x)} - \sqrt{p_\theta(x)}}_{=f_1(x)} - \underbrace{\frac{\epsilon}{2} h^T s(\theta|x) \sqrt{p_\theta(x)}}_{f_2(x)} \right]^2 dx = o(\epsilon^2),$$

which means that the L_2 distance between the two functions f_1 and f_2 is bounded by

$$\|f_1 - f_2\|_\mu = o(\epsilon), \tag{3.12}$$

where the norm $\|\cdot\|_\mu$ means $\|f\|_\mu = \sqrt{\int f^2(x)dx}$.

One may notice that

$$\|f_1\|_\mu^2 = \int \left(\sqrt{p_{\theta+\epsilon h}(x)} - \sqrt{p_\theta(x)} \right)^2 dx = H^2(p_{\theta^*+\epsilon h}, p_{\theta^*}).$$

Here is the last trick that we will use called reverse triangle inequality: $|\|f\|_\mu - \|g\|_\mu| \le \|f - g\|_\mu$. With the reverse triangle inequality along with equation (3.12), we have

$$\|f_1\|_\mu = \|f_2\|_\mu + o(\epsilon),$$

which implies

$$H(p_{\theta^*+\epsilon h}, p_{\theta^*}) = \sqrt{\int \left(\sqrt{p_{\theta^*+\epsilon h}(x)} - \sqrt{p_{\theta^*}(x)} \right)^2 dx} = \frac{\epsilon}{2} \sqrt{h^T \int s(\theta^*|x) s^T(\theta^*|x) p_{\theta^*}(x) dx h} + o(\epsilon^2).$$

Taking squares of both sides leads to

$$H^2(p_{\theta^*+\epsilon h}, p_{\theta^*}) = \frac{\epsilon^2}{4} h^T \underbrace{\int s(\theta^*|x) s^T(\theta^*|x) p_{\theta^*}(x) dx}_{=I(\theta^*)} h + o(\epsilon^2).$$

By equation (3.11) and the above result, we conclude that

$$\Delta^*(\epsilon, h) \equiv \bar{\ell}(\theta^* + \epsilon h) - \bar{\ell}(\theta^*) = -2H^2(p_{\theta^*+\epsilon h}, p_{\theta^*}) + o(\epsilon^2) = -\frac{\epsilon^2}{2} h^T I(\theta^*) h + o(\epsilon^2),$$

which implies equation (3.10). So we have verified condition (V2).

■

Remark 3.15 *Here we point out some key remarks that were not highlighted in conventional textbooks.*

- **Model correctness.** *Both Theorem 3.13 and 3.14 require the parametric model to be* correct. *Many conditions require $X \sim p_{\theta^*}$. This essentially means that we are considering the scenario where the model is correct. Note that sometimes people assume that $X \sim p_{\theta_0}$ for some parameter θ_0 in the interior of the parameter space. This will imply that the population maximizer $\theta^* = \theta_0$ since the MLE minimizes the Kullback-Leiber divergence.*

- **Relaxing the derivative conditions: yes and no.** *From Theorem 3.1 to Theorem 3.9 to Theorem 3.13, we do see the reduction of the derivative conditions:*
  - *Classical regime: we need third-order condition (M4) in Theorem 3.1.*
  - *Modern regime: we need second-order condition (V2) in Theorem 3.9.*
  - *Correct model regime: we only need first-order condition (Q1) in Theorem 3.14.*

  While it is true that both (V2) and (Q1) relax the classical third-order condition (M4), (Q1) does not remove the second order condition (V2). In fact, QMD implies (V2)! What Theorem 3.13 shows: under QMD (Q1) and correct model (and other conditions), the Hessian is exactly the Fisher's information matrix. So the QMD (Q1) offers a sufficient condition for (V2) to work.

- **Hellinger distance.** *In the proof of Theorem 3.13, we have see the Hellinger distance for two PDFs p, q as*

$$H^2(p,q) = \frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx = 1 - \int \sqrt{p(x)q(x)}dx.$$

  When p, q are PMFs, their Hellinger distance is

$$H^2(p,q) = \frac{1}{2}\sum_x \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 = 1 - \sum_x \sqrt{p(x)q(x)}.$$

  A simple way to understand Hellinger distance is to view the square root of a PDF or PMF as a vector on a sphere. This is because once we take a square root of a PDF or a PMF, say $g(x) = \sqrt{p(x)}$, we immediately have $\int g(x)^2 dx = 1$ or $\sum_x g(x)^2 = 1$. In particular, for the case of a PMF $p(x)$ that taking values on $x = 1, 2, \cdots, k$, if we consider the vector $v \in \mathbb{R}^k$ such that $v_j = \sqrt{p(j)}$, then $\|v\|^2 = 1 = \sum_x p(x)$. So the square root of a probability vector v is a vector on the sphere. In this regard, the Hellinger distance is essentially the Euclidean distance between two vectors on a unit sphere. Once we view $\sqrt{p(x)} = f(x)$ as a vector on a unit sphere (for continuous case, this corresponds to a infinite dimensional vector), the QMD condition is essentially a smoothness condition that we need $f_\theta(x) = \sqrt{p_\theta(x)}$ to change smoothly as we vary θ under the Hellinger distance.

3.5 Failure of MLE (Ritov-Wasserman)

Before we end the lecture, we discuss an interesting example highlighting the failure of MLE (and inconsistency of Bayesian estimator). Suppose our data is triplet $(X_1, R_1, Y_1), \cdots, (X_n, R_n, Y_n)$ from the following distribution:

- $\theta \in [0,1]^B$ is unknown vector with B is a huge number compare to n.
- $\xi \in [0,1]^B$ is a known vector with each $0 < \delta \le \xi_j \le 1 - \delta < 1$ for some δ.
- $X_i \sim \mathsf{Uni}\{1, 2, \cdots, B\}$.
- $R_i \sim \mathsf{Ber}(\xi_{X_i})$,
- If $R_i = 1$, $Y_i \sim \mathsf{Ber}(\theta_{X_i})$. If $R_i = 0$, Y_i is missing.

Our goal is to estimate the average of all θ_j, i..e, we want

$$\psi = \frac{1}{B}\sum_{b=1}^{B} \theta_b.$$

The joint likelihood function of a single observation (X_i, R_i, Y_i) is

$$\begin{aligned} L(\theta|X_i, R_i, Y_i) &= p(X_i)p(R_i|X_i)p(Y_i|X_i, R_i) \\ &= \frac{1}{B}\xi_{B_i}^{R_i}(1-\xi_{B_i})^{1-R_i}\theta_{X_i}^{R_iY_i}(1-\theta_{X_i})^{R_i(1-Y_i)} \\ &\propto \theta_{X_i}^{R_iY_i}(1-\theta_{X_i})^{R_i(1-Y_i)}. \end{aligned}$$

Thus, the log-likelihood over n observations is

$$\sum_{i=1}^{n} R_iY_i\log\theta_{X_i} + R_i(1-Y_i)\log(1-\theta_{X_i}) = \sum_{b=1}^{B} n_{b,1,1}\log\theta_b + n_{b,1,0}\log(1-\theta_b),$$

where $n_{b,1,1} = \sum_{i=1}^{n} I(X_i = b, R_i = 1, Y_i = 1)$ and $n_{n,1,0} = \sum_{i=1}^{n} I(X_i = b, R_i = 1, Y_i = 0)$. Thus, the MLE of θ_b is

$$\hat{\theta}_b = \frac{n_{b,1,1}}{n_{b,1,1} + n_{b,1,0}}.$$

However, if B is much larger than n (high-dimensional setting), we are very likely to have $n_{b,1,1} = n_{b,1,0} = 0$, so the MLE is not well-defined! So we cannot estimate ψ.

We may consider a Bayes estimator by placing a prior on θ and compute the posterior, which leads to a posterior distribution of ψ. While the posterior distribution of ψ can be computed, it will be very similar to the prior distribution since most entries are missing, so the most posterior distribution of θ_j is identical to its prior.

In this case, both MLE and Bayes estimator do not lead to a consistent estimator. You may be curious if we can use a penalized MLE (maximizing the log-likelihood function with a penalty/regularization term) for this case, this approach also does not work since you can write the penalized MLE as the Bayes rule under suitable priors.

Though it may seem to be impossible to estimate ψ, you can show that the inverse probability weighting (Horvitz-Thompson) estimator still work:

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{\xi_{X_i}}.$$

You can easily show that $\hat{\psi}$ is a consistent estimator and we even have asymptotic normality of $\sqrt{n}(\hat{\psi} - \psi)$.

Remark 3.16 *Here are some remarks about this example.*

- **Individual parameter versus overall parameter.** *Here you see an interesting example that we do not have consistency of individual parameter θ_j but the overall parameter $\psi = \frac{1}{B}\sum_{b=1}^{B}\theta_b$ can still be consistently estimated. If our goal is to infer individual parameter, no method will work since we do not have observations from many θ_b.*
- **Survey sample.** *While this problem may seem artificial, it could occur easily in survey sample. The missing data problems and binary outcome models are common scenarios in survey sample. The covariate X can be viewed as a numerical indicator of the* characteristic *of the individual that comes from demographic variables. In this case, B is the number of unique characteristics in the population of interest and ξ_j is the survey weight for individual with characteristic $X_i = j$.*
- **High-dimensional statistics.** *The setup of this example is rather simple: we are mostly using Bernoulli random variables. But it highlights the complexity of high-dimensional problem: we may need to place a strong modeling assumption otherwise we may not be able to estimate the parameter of interest. You can easily show that if we assume that parameters $\theta_b = \bar{\theta}$, then the MLE works.*

A Computational learning: gradient descent

In the previous sections, we have established key statistical learning theory of an M-estimator. Now we will investigate the computational perspective about this estimator. We will come back the conditions (M1-4) since these curvature conditions offer a simple way to show that numerically, we can compute the estimator quickly.

For simplicity, we assume that our estimator is from ERM

$$\hat{\theta}_n = \mathsf{argmin}_\theta \bar{R}_n(\theta), \qquad \bar{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta, X_i).$$

Numerically, a popular approach to compute $\hat{\theta}_n$ is the **gradient descent** (GD) method.

Starting with an initial guess $\theta^{(0)}$, the GD creates a sequence of points $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \cdots$ via the following procedure:

$$\theta^{(t+1)} = \theta^{(t)} - \gamma \nabla \bar{R}_n(\theta^{(t)}), \tag{3.13}$$

where $\gamma > 0$ is a stepsize constant. Namely, the sequence of points are generated by moving the current point toward the descending direction of the current gradient. In the case of likelihood inference, the gradient $\nabla \bar{R}_n(\theta) = -\bar{S}_n(\theta)$ is the empirical score function. So clearly, the MLE occurs at a stationary point.

The GD is a very common procedure in convex optimization. Here we will focus on the behavior of GD under smoothness conditions related to (M1-4) in Theorem 3.1. To this end, we will introduce two smoothness conditions.

A.1 L-smooth and M-strongly convex

L-smooth. A smooth function $f : \mathbb{R}^d \to \mathbb{R}$ is called *L-smooth* if

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|.$$

When f is twice-differentiable, the L-smoothness can be achieved by requiring all eigenvalues of $\nabla\nabla f(x)$ is bounded by L for all x. In view of Assumptions (M1-4) in Theorem 3.1, Assumptions (M1) and (M4) imply that the population log-likelihood function $\bar{\ell}(\theta)$ is L-smooth.

Convex. Convexity is another important property for optimization. Intuitively, a convex function is a function that curves upward like a U- or V-shape. Formally, a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if

$$\alpha f(x) + (1-\alpha) f(y) \ge f(\alpha x + (1-\alpha) y)$$

for any $x, y \in \mathbb{R}^d$ and $\alpha \in [0, 1]$. The convexity is often used in the Jensen's inequality that for a random vector $X \in \mathbb{R}^d$ and a convex function f, we have

$$\mathbb{E}[f(X)] \ge f(\mathbb{E}(X)).$$

A useful example of convex function is the absolute value function in univariate $f(x) = |x|$ when $x \in \mathbb{R}^d$. In the multivariate case, the L_1 norm $f(x) = \sum_{j=1}^{d} |x_j| = \|x\|_1$ is also convex. This result is particularly important in High-dimensional statistics because the L_1 norm is used very frequently in penalized estimator. The fact that it is convex allows the computation to be done in a quick way. Generally speaking, the GD converges very fast when the objective function is convex.

M-strongly convex. In our ERM, we are considering nice Hessian matrix (invertible around the maximizer/minimizer), which corresponds to an even stronger concept than the convexity: strongly convexity. A

function $f : \mathbb{R}^d \to \mathbb{R}$ is called *M-strongly convex* if $f(x) - \frac{M}{2}\|x - x^*\|^2$ is convex and $x^* = \mathsf{argmin}_x f(x)$. For a twice-differentiable function, another way to think about M-strongly convex is that all eigenvalues of the Hessian matrix $\nabla\nabla f(x)$ are greater than or equal M for all x. Assumptions (M3) and (M4) in Theorem 3.1, imply that locally around θ^*, the population log-likelihood $\bar{\ell}(\theta)$ is strongly convex.

When comparing L-smoothness and M-strongly convexity together, we see that:

- L-smoothness: upper bound on the curvature, which implies

$$f(y) - f(x) \leq (y - x)^T \nabla f(x) + \frac{L}{2}\|x - y\|^2. \tag{3.14}$$

- M-strongly convexity: lower bound on the curvature, which implies

$$f(y) - f(x) \geq (y - x)^T \nabla f(x) + \frac{M}{2}\|x - y\|^2. \tag{3.15}$$

The inequalities in equations (3.14) and (3.15) can be viewed as performing a Taylor expansion to the second-order. The upper and lower bounds on the eigenvalues of Hessian matrix control the shape of the objective function.

A.2 Convergence rate of the gradient descent

With the concept of L-smoothness and M-strongly convex, we can obtain the algorithmic convergence rate of the GD.

Theorem 3.17 *Suppose the objective function $f(\theta) \equiv \bar{R}_n(\theta)$ is L-smooth and M-strongly convex. Then we have*

$$\|\theta^{(t)} - \hat{\theta}_n\|^2 \leq (1 - \gamma M)^t \|\theta^{(0)} - \hat{\theta}_n\|^2$$

when the stepsize $\gamma < \min\left\{\frac{1}{M}, \frac{1}{L}\right\}$.

Theorem 3.17 shows that the GD procedure converges geometrically to the MLE. This convergence rate is called linear convergence in optimization literature (the log of the convergence rate is linear in terms of the number of iterations).

While Theorem 3.17 states that the GD converges under appropriate smoothness assumptions, these smoothness assumptions are on our empirical risk function (a random/sample-based quantity). Ideally, we do not want to place smoothness conditions on the estimators and instead, we would prefer to put conditions on the population quantity or the underlying distribution. We will investigate how Theorem 3.17 can be applied under the conventional MLE assumptions (M1-4).

Proof:

A direct expansion shows that

$$\begin{aligned}
\|\theta^{(t+1)} - \hat{\theta}_n\|^2 &= \|\theta^{(t)} - \gamma\nabla f(\theta^{(t)}) - \hat{\theta}_n\|^2 \\
&= \|\theta^{(t)} - \hat{\theta}_n\|^2 - 2\gamma(\theta^{(t)} - \hat{\theta}_n)^T \nabla f(\theta^{(t)}) + \gamma^2 \|\nabla f(\theta^{(t)})\|^2 \\
&= \|\theta^{(t)} - \hat{\theta}_n\|^2 + 2\gamma(\hat{\theta}_n - \theta^{(t)})^T \nabla f(\theta^{(t)}) + \gamma^2 \|\nabla f(\theta^{(t)})\|^2
\end{aligned} \tag{3.16}$$

The middle term $2\gamma(\hat{\theta}_n - \theta^{(t)})^T \nabla f(\theta^{(t)})$ has a useful upper bound from equation (3.15), where

$$(y - x)^T \nabla f(x) \leq f(y) - f(x) - \frac{M}{2}\|x - y\|^2.$$

Choosing $y = \hat{\theta}_n$ and $x = \theta^{(t)}$ leads to

$$2\gamma(\hat{\theta}_n - \theta^{(t)})^T \nabla f(\theta^{(t)}) \leq 2\gamma(f(\hat{\theta}_n) - f(\theta^{(t)})) - M\gamma\|\theta^{(t)} - \hat{\theta}_n\|^2.$$

Therefore, equation (3.16) has an upper bound

$$\|\theta^{(t+1)} - \hat{\theta}_n\|^2 \leq (1 - \gamma M)\|\theta^{(t)} - \hat{\theta}_n\|^2 + 2\gamma(f(\hat{\theta}_n) - f(\theta^{(t)})) + \gamma^2\|\nabla f(\theta^{(t)})\|^2. \tag{3.17}$$

Since $\hat{\theta}_n$ is the minimizer of $f(\theta)$, we have

$$f(\theta) \geq f(\hat{\theta}_n).$$

Thus,

$$f\left(\theta - \frac{1}{L}\nabla f(\theta)\right) \geq f(\hat{\theta}_n)$$

for any θ. Moreover, we minus $f(\theta)$ in both sides, which leads to

$$f\left(\theta - \frac{1}{L}\nabla f(\theta)\right) - f(\theta) \geq f(\hat{\theta}_n) - f(\theta). \tag{3.18}$$

Recall the L-smoothness property in equation (3.14):

$$f(y) - f(x) \leq (y - x)^T \nabla f(x) + \frac{L}{2}\|x - y\|^2.$$

Choosing $y = \theta - \frac{1}{L}\nabla f(\theta)$ and $x = \theta$, the left-hand-side of equation (3.18) is upper bounded by

$$f\left(\theta - \frac{1}{L}\nabla f(\theta)\right) - f(\theta) \leq -\frac{1}{L}\|\nabla f(\theta)\|^2 + \frac{L}{2}\left\|\frac{1}{L}\nabla f(\theta)\right\|^2 = -\frac{1}{2L}\|\nabla f(\theta)\|^2.$$

Putting this back to equation (3.18), we conclude that

$$f(\hat{\theta}_n) - f(\theta) \leq -\frac{1}{2L}\|\nabla f(\theta)\|^2$$

and applying this to equation (3.17), we obtain

$$\begin{aligned}
\|\theta^{(t+1)} - \hat{\theta}_n\|^2 &\leq (1 - \gamma M)\|\theta^{(t)} - \hat{\theta}_n\|^2 + 2\gamma(f(\hat{\theta}_n) - f(\theta^{(t)})) + \gamma^2\|\nabla f(\theta^{(t)})\|^2 \\
&\leq (1 - \gamma M)\|\theta^{(t)} - \hat{\theta}_n\|^2 - \frac{\gamma}{L}\|\nabla f(\theta)\|^2 + \gamma^2\|\nabla f(\theta)\|^2 \\
&= (1 - \gamma M)\|\theta^{(t)} - \hat{\theta}_n\|^2 - \frac{\gamma}{L}(1 - \gamma L)\|\nabla f(\theta)\|^2 \\
&\leq (1 - \gamma M)\|\theta^{(t)} - \hat{\theta}_n\|^2
\end{aligned} \tag{3.19}$$

when $\gamma < \frac{1}{L}$. Note that to ensure equation (3.19) is contracting, we also need $\gamma < \frac{1}{M}$, which is the other requirement of the stepsize γ.

By telescoping equation (3.19), we conclude that

$$\|\theta^{(t)} - \hat{\theta}_n\|^2 \leq (1 - \gamma M)^t\|\theta^{(0)} - \hat{\theta}_n\|^2$$

when $\gamma < \min\left\{\frac{1}{M}, \frac{1}{L}\right\}$, which completes the proof. ■

B Bridging statistical and computational learning

While Theorem 3.17 shows that the GD is a fast algorithm to numerically compute the estimator, the assumptions are directly imposed on the empirical risk function $\bar{R}_n(\theta)$. In statistics, we often want to impose conditions on the population quantity such as assumptions (M1-4) in Theorem 3.1. Thus, we want to understand what computational learning theory we can obtain under assumptions (M1-4).

Challenge of bridging the two learning theories. While assumptions (M1) and (M4) imply that the population risk $\bar{R}(\theta)$ is L-smooth for some L, assumptions (M1-4) does not require $\bar{R}(\theta)$ to be strongly convex. In fact, $\bar{R}(\theta)$ may not even be a convex function and could have multiple local maxima. The MLE theory still applies when there are multiple local maxima.

B.1 Local strongly convex of the population risk

Having said this, the eigenvalue condition in (M3) and the smoothness of Hessian matrix from (M4) imply that $\bar{R}(\theta)$ is *locally strongly convex*.

Lemma 3.18 *Under assumption (M1-4), there exists a radius $\zeta_1 > 0$ such that $\bar{R}(\theta) = -\bar{\ell}(\theta)$ is strongly convex within $B(\theta^*, \zeta_1) \subset \Theta$.*

Proof:

Let

$$\lambda^*_{\min} = \lambda_{\min}(\nabla_\theta \nabla_\theta \bar{R}(\theta^*))$$

be the smallest eigenvalue at $\theta = \theta^*$. By assumption (M3), the Hessian $\nabla_\theta \nabla_\theta \bar{R}(\theta^*)$ is invertible, so $\lambda^*_{\min} > 0$.

The compact support condition of (M1) and the bounded third-order derivative condition in (M4) implies that the Hessian matrix

$$\bar{H}_R(\theta) = \nabla_\theta \nabla_\theta \bar{R}(\theta)$$

is smooth in the sense that there exists a constant $\phi_3 > 0$ such that

$$\|\bar{H}_R(\theta_1) - \bar{H}_R(\theta_2)\|_2 \le \phi_3 \|\theta_1 - \theta_2\|_2.$$

This is useful because the Weyl's theorem[1] show that for two symmetric matrices A, B,

$$|\lambda_{\min}(A) - \lambda_{\min}(B)| \le \|A - B\|_2.$$

Thus, for any point θ, its smallest eigenvalue

$$\begin{aligned}\lambda_{\min}(H_R(\theta)) &\ge \lambda_{\min}(H_R(\theta^*)) - |\lambda_{\min}(H_R(\theta)) - \lambda_{\min}(H_R(\theta^*))| \\ &\ge \lambda^*_{\min} - \phi_3 \|\theta - \theta^*\|_2.\end{aligned}$$

Therefore, for any θ such that $\|\theta - \theta^*\|_2 < \frac{\lambda^*_{\min}}{\phi_3}$, we have

$$\lambda_{\min}(H_R(\theta)) \ge \lambda^*_{\min} - \phi_3 \|\theta - \theta^*\|_2 > 0, \tag{3.20}$$

which means that the function $\bar{R}(\theta)$ is strongly convex.

As a result, we can choose $\zeta_1 = \frac{\lambda^*_{\min}}{\phi_3}$ and the result follows.

[1] see, e.g., https://en.wikipedia.org/wiki/Weyl%27s_inequality

■

The nice part of Lemma 3.18 is that the objective function $\bar{R}(\theta)$ is locally strongly convex.

Note that if we want to obtain a precise constant the strongly convex, we will need to pick ζ_1 cleverly. For instance, based on equation (3.20), we may choose

$$\zeta_1 = \frac{\lambda^*_{\min}}{2\phi_3} \Longrightarrow \lambda_{\min}(H_R(\theta)) \geq \frac{1}{2}\lambda^*_{\min}. \tag{3.21}$$

With this choice, $\bar{R}(\theta)$ is M-strongly convex within $\theta \in B(\theta^*, \zeta_1)$ with $M = \frac{1}{2}\lambda^*_{\min}$. For the L-smoothness, assumption (M4) implies that there exists a finite constant

$$h_{\max} = \sup_{\theta\in\Theta} \|\bar{H}_R(\theta)\|_2 < \infty. \tag{3.22}$$

Then clearly, we have

$$\|\nabla_\theta \bar{R}(\theta_1) - \nabla_\theta \bar{R}(\theta_2)\|_2 \leq h_{\max}\|\theta_1 - \theta_2\|_2.$$

So the function $\bar{R}(\theta)$ is L-smooth with $L = h_{\max}$.

Thus, the gradient descent method with objective function being $\bar{R}(\theta)$ converges linearly if our initial point $\theta^{(0)} \in B\left(\theta^*, \frac{\lambda^*_{\min}}{2\phi_3}\right)$ and we choose the stepsize

$$\gamma < \min\left\{\frac{2}{\lambda^*_{\min}}, \frac{1}{h_{\max}}\right\}.$$

One important thing to keep in mind for a locally convex function is that the GD is NOT guaranteed to discover the global minimum. It could get stuck at a local minimum. The local convexity only implies that the GD converges under a good initialization. How to find a good initialization remains an open question.

B.2 Transferring the smoothness to the empirical risk

While the analysis in the previous section shows that applying GD on $\bar{R}(\theta)$ with a good initialization converges quickly, our actual application of GD is on the empirical risk/sample log-likelihood $\bar{R}_n(\theta)$. Thus, we need to investigate if the L-smoothness and strongly convexity holds on $\bar{R}_n$ for an area around the minimizer $\hat{\theta}_n$.

In Statistics, we generally do not want to assume conditions on the data since such conditions are either true or false given a set of observations and since the data is random, so there will be a 'probability' on those conditions being true. Therefore, we want to use the conventional assumptions (M1-4) and investigate if we can show that $\bar{R}_n$ is locally strongly convex with a good probability.

To transfer the smoothness of population risk $\bar{R}(\theta)$ to the empirical risk $\bar{R}_n(\theta)$, we will utilize the following result.

Lemma 3.19 *Let $f_\theta : \mathbb{R}^p \to \mathbb{R}$ be a function indexed by $\theta \in \Theta \subset \mathbb{R}^d$ and Θ is a compact set. Suppose*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq q(x)\|\theta_1 - \theta_2\|_2 \tag{3.23}$$

such that $\mathbb{E}|q(X)| < \infty$. Then

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n f_\theta(X_i) - \mathbb{E}[f_\theta(X_i)]\right| \xrightarrow{P} 0.$$

Lemma 3.19 follows from Example 19.7 and Theorem 19.4 of [van der Vaart]. The details of the proof would require some techniques from empirical process theory so we omit it.

Results in Lemma 3.19 are known as Glivenko-Cantelli (GC) theory for the function class $\{f_\theta : \theta \in \Theta\}$.

Lemma 3.19 is particularly useful in our case because assumption (M4) requires the existence of a absolutely integrable function $\Lambda(x)$ for the third-order derivative:

$$\sup_{\theta\in\Theta} \max_{j_1,j_2,j_3} \left| \frac{\partial^3}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}} \ell(\theta|x) \right| \leq \Lambda(x).$$

Under the compact parameter space (M1), this implies a similar result for the lower-order derivatives. Namely, there exists $\Lambda_1(x), \Lambda_2(x)$ such that $\mathbb{E}|\Lambda_k(x)| < \infty$ and

$$\begin{aligned} \sup_{\theta\in\Theta} \max_{j_1,j_2} \left| \frac{\partial^2}{\partial\theta_{j_1}\partial\theta_{j_2}} \ell(\theta|x) \right| &\leq \Lambda_2(x), \\ \sup_{\theta\in\Theta} \max_{j_1} \left| \frac{\partial}{\partial\theta_{j_1}} \ell(\theta|x) \right| &\leq \Lambda_1(x). \end{aligned}$$

The uniform bound on the derivative imply the Lipschitz condition in equation (3.23). Thus, Lemma 3.19 implies the following uniform convergence:

$$\begin{aligned} \sup_{\theta\in\Theta} |\bar{R}_n(\theta) - \bar{R}(\theta)| &\xrightarrow{P} 0, \\ \sup_{\theta\in\Theta} \|\nabla\bar{R}_n(\theta) - \nabla\bar{R}(\theta)\|_{\max} &\xrightarrow{P} 0, \\ \sup_{\theta\in\Theta} \left\| \underbrace{\nabla\nabla\bar{R}_n(\theta)}_{=\bar{H}_{R,n}(\theta)} - \underbrace{\nabla\nabla\bar{R}(\theta)}_{=\bar{H}_R(\theta)} \right\|_{\max} &\xrightarrow{P} 0. \end{aligned} \tag{3.24}$$

With the above result, we can formally state the algorithmic convergence on the empirical risk.

Theorem 3.20 (Convergence of gradient descent for sample MLE) *Suppose we apply the gradient descent on the empirical risk $\bar{R}_n(\theta)$. Assume conditions (M1-4) for $\bar{\ell}(\theta) = -\bar{R}(\theta)$. There exists a constant ζ_0 and a threshold of stepsize γ_0 such that if our initialization $\theta^{(0)} \in B(\hat{\theta}_n, \zeta_0)$ and stepsize $\gamma < \gamma_0$, then with a probability tending to 1, there is a constant $\rho_\gamma \in (0,1)$ depending on γ such that*

$$\|\theta^{(t)} - \hat{\theta}_n\|^2 \leq \rho_\gamma^t \|\theta^{(0)} - \hat{\theta}_n\|^2.$$

The constants in Theorem 3.20 can be chosen to be

$$\zeta_0 = \frac{1}{2}\zeta_1 = \frac{\lambda^*_{\min}}{4\phi_3}, \quad \gamma_0 = \min\left\{ \frac{1}{2h_{\max}}, \frac{4}{\lambda^*_{\min}} \right\}$$

and $\rho_\gamma = 1 - \gamma \cdot \frac{\lambda^*_{\min}}{4}$, where $\lambda^*_{\min} = \lambda_{\min}(\bar{H}_R(\theta^*))$ and $h_{\max} = \sup_{\theta\in\Theta} \|\bar{H}_R(\theta)\|_2$ and ϕ_3 depends on the third-order derivative of $\bar{R}(\theta)$. All these constants are non-random and only depends on the population distribution.

Proof:

Given Theorem 3.17 and Lemma 3.18, we only need to show that $\bar{R}_n(\theta)$ is both L^*-smooth and M^*-strongly convex within $B(\hat{\theta}_n, \zeta_0)$ for some L^*, M^*, ζ_0.

L-smoothness. The L-smoothness of $\bar{R}(\theta)$ comes from equation (3.22), where the parameter $L = h_{\max} = \sup_\theta \|\bar{H}(\theta)\|$. Thus, the empirical risk $\bar{R}_n(\theta)$ is also L-smooth with

$$L = \sup_\theta \|\bar{H}_{R,n}(\theta)\|_2.$$

However, this quantity is random quantity (maximal of the sample Hessian), so we cannot directly use it for our stepsize threshold (γ_0), which is a non-random quantity. Using the uniform bound in equation (3.24) and assumptions (M1) and (M4), we can easily upper bound it by

$$\sup_\theta \|\bar{H}_{R,n}(\theta)\|_2 \leq 2 \sup_\theta \|\bar{H}_R(\theta)\|_2 = 2h_{\max}.$$

Let

$$E_{1,n} = \left\{ \sup_\theta \|\bar{H}_{R,n}(\theta)\|_2 \leq 2h_{\max} \right\}$$

be such event and it holds with a probability

$$P(E_{1,n}) = P\left(\sup_\theta \|\bar{H}_{R,n}(\theta)\|_2 \leq 2h_{\max} \right) \to 1.$$

Thus, we will proceed with saying $\bar{R}_n(\theta)$ is L^*-smooth with $L^* = 2h_{\max}$.

M-strongly convex and ζ_0. The strongly convex comes from the eigenvalue conditions. But here is a caveat, we are considering regions around $\hat{\theta}_n$, not θ^*. To make the analysis easier, we utilize the fact that $\hat{\theta}_n \xrightarrow{P} \theta^*$ by Theorem 3.1.

We consider the following event

$$E_{2,n} = \left\{ \|\hat{\theta}_n - \theta^*\| \leq \frac{1}{2}\zeta_1 \right\},$$

where $\zeta_1 = \frac{\lambda^*_{\min}}{2\phi_3}$. Clearly,

$$P(E_{2,n}) \to 1,$$

and under $E_{2,n}$, the ball

$$B\left(\hat{\theta}, \frac{1}{2}\zeta_1\right) \subset B(\theta^*, \zeta_1),$$

so we choose

$$\zeta_0 = \frac{1}{2}\zeta_1 = \frac{\lambda^*_{\min}}{4\phi_3}, \tag{3.25}$$

which implies $B\left(\hat{\theta}, \zeta_0\right) \subset B(\theta^*, \zeta_1)$. By equation (3.21), this implies that

$$\lambda_{\min}(\bar{H}_R(\theta)) \geq \frac{1}{2}\lambda^*_{\max}.$$

Namely, the eigenvalues of the population risk $\bar{H}_R(\theta)$ are bounded from below.

We then use the uniform bound in equation (3.24) again such that

$$\|\bar{H}_{R,n}(\theta) - H_R(\theta)\|_2 \xrightarrow{P} 0.$$

Consider the event

$$E_{3,n} = \left\{ \|\bar{H}_{R,n}(\theta) - H_R(\theta)\|_2 \leq \frac{1}{4}\lambda^*_{\max} \right\}.$$

Clearly, $P(E_{3,n}) \to 1$ and under $E_{3,n}$, the minimal eigenvalue

$$\begin{aligned}
\lambda_{\min}(\bar{H}_{R,n}(\theta)) &\geq \lambda_{\min}(\bar{H}_R(\theta)) - |\lambda_{\min}(\bar{H}_{R,n}(\theta)) - \lambda_{\min}(\bar{H}_R(\theta))| \\
&\geq \lambda_{\min}(\bar{H}_R(\theta)) - \frac{1}{4}\lambda^*_{\max} \\
&\geq \frac{1}{2}\lambda^*_{\min} - \frac{1}{4}\lambda^*_{\max} \\
&= \frac{1}{4}\lambda^*_{\min}
\end{aligned}$$

for any point $\theta \in B(\hat{\theta}_n, \zeta_0)$.

As a result, under events $E_{2,n}$ and $E_{3,n}$, all eigenvalues of $\bar{H}_{R,n}(\theta)$ are above $\frac{1}{4}\lambda^*_{\min}$ for any $\theta \in B(\hat{\theta}_n, \zeta_0)$. Namely, the function $\bar{R}_n(\theta)$ is M^*-strongly convex with $M^* = \frac{1}{4}\lambda^*_{\min}$ when $\theta \in B(\hat{\theta}_n, \zeta_0)$.

By Theorem 3.17, we conclude that

$$\|\theta^{(t)} - \hat{\theta}_n\|^2 \leq (1 - M^*\gamma)^t \|\theta^{(0)} - \hat{\theta}_n\|^2,$$

when $\theta^{(0)} \in B(\hat{\theta}_n, \zeta_0)$ and

$$\gamma < \gamma_0 = \min\left\{ \frac{1}{M^*}, \frac{1}{L^*} \right\} = \min\left\{ \frac{1}{2h_{\max}}, \frac{4}{\lambda^*_{\min}} \right\}.$$

This result holds when events $E_{1,n}, E_{2,n}, E_{3,n}$ holds, which has a probability

$$\begin{aligned}
P(E_{1,n} \cap E_{2,n} \cap E_{3,n}) &= 1 - P(E^C_{1,n} \cup E^C_{2,n} \cup E^C_{3,n}) \\
&\geq 1 - (1 - P(E_{1,n})) - (1 - P(E_{2,n})) - (1 - P(E_{2,n})) \\
&\to 1.
\end{aligned}$$

Note that we can also get a bound on how fast the probability converges to 1 using concentration bounds.

■