An introduction on basic learning theories of M-estimation

Yen-Chi Chen
University of Washington
October 6, 2025

This note is revised from my lecture notes on UW STAT 535 Statistical Machine Learning. We will cover the basic form of statistical and computational learning theories for common M-estimators (maximal estimators) that arises from maximal likelihood principle, least square regression, and empirical risk minimization. In particular, we will show that conventional assumptions for asymptotic normality of an M-estimator (statistical learning) also leads to a linear convergence of a gradient descent/ascent algorithm with a suitable initialization (computational learning).

Contents

1 Statistical learning: likelihood inference

We discuss the statistical learning theory for M-estimator using the maximum likelihood estimator (MLE) as a motivating example. 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.

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).

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

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

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

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

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

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

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,

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

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 $\widehat{\theta}_n$, 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 $\widehat{\theta}_n = \mathsf{argmax}_\theta \bar{\ell}_n(\theta)$.

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 1 *Assume the following conditions:*

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

(M2) The MLEs $(\widehat{\theta}_n, \theta^*)$ 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

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

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

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

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

$$\sqrt{n}(\bar{S}_n(\theta^*) - \bar{S}(\theta^*)) \xrightarrow{d} N(0, \mathbb{E}[S(\theta^*|X_i)S(\theta^*|X_i)^T]). \tag{1}$$

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

Using the score equation,

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

we have

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

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

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

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

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

which is negligible compare to the other quantity.

Thus, we conclude that

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

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 (1) and using Slutsky's theorem, we conclude that

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

□

1.3 Remarks

Here are some important remarks.

- **Sanwich estimator.** There is a simple estimator of the underlying covariance matrix via the plug-in approach:
$$\widehat{\Sigma}^* = \bar{H}_n^{-1}(\widehat{\theta}_n)\left[\frac{1}{n}\sum_{i=1}^{n} S(\widehat{\theta}_n|X_i)S(\widehat{\theta}_n|X_i)^T\right]\bar{H}_n^{-1}(\widehat{\theta}_n).$$
This estimator is also known as the sandwich estimator.

- **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 $\widehat{\theta}_n^*$. Repeat this process B times, leading to $\widehat{\theta}_n^{*(1)}, \cdots, \widehat{\theta}_n^{*(B)}$. 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 called 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 $\bar{S}_n(\widehat{\theta}_n) - \bar{S}_n(\theta^*)$. For the j-th element, we have
$$\bar{S}_{n,j}(\widehat{\theta}_n) - \bar{S}_{n,j}(\theta^*) \in \mathbb{R},$$
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 $\widetilde{\theta}_{n,j}$ lies between $\widehat{\theta}_n$ and θ^* such that
$$\bar{S}_{n,j}(\widehat{\theta}_n) - \bar{S}_{n,j}(\theta^*) = [\nabla_\theta \bar{S}_{n,j}(\widetilde{\theta}_{n,j})]^T(\widehat{\theta}_n - \theta^*).$$
Now we define the matrix $B_n \in \mathbb{R}^{d\times d}$ such that the j-th row of B_n is $[\nabla_\theta \bar{S}_{n,j}(\widetilde{\theta}_{n,j})]^T$. Then we can still have
$$B_n \xrightarrow{P} \bar{H}(\theta^*)$$
without assuming third-order derivative is upper bounded because $\widetilde{\theta}_{n,j} \xrightarrow{P} \theta^*$ for each j.

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
$$\widehat{\theta}_n = \mathsf{argmin}_\theta \quad \bar{R}_n(\theta), \qquad \bar{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta, X_i), \tag{2}$$
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*:

$$\theta^* = \mathsf{argmin}_\theta \quad \bar{R}(\theta), \qquad \bar{R}(\theta) = \mathbb{E}\left[\mathcal{L}(\theta, X_1)\right]. \tag{3}$$

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 $\widehat{\theta}_n$ toward θ^* in Theorem 1 applies to any of these ERM estimators as long as (M1-4) hold.

Here are some examples of the ERM problems.

2.1 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

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

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 1 applies.

2.2 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

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

Again, this is the ERM estimator and the population quantity $\widehat{\theta}_n$ 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 1 and assumptions (M1-4) imply the asymptotic normality of $\widehat{\theta}_n - \theta^*$.

2.3 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:

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

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

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

2.4 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 $\widehat{p}$ and then construct our mode estimator as $\widehat{m}_0 = \mathsf{argmax}_x \widehat{p}(x)$. Now suppose we use the kernel density estimator (KDE), where $\widehat{p} = \widehat{p}_h$ is

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

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

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

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 1) shows that $\widehat{m}_h$ has asymptotic normality for estimating $\bar{m}_h$ when h is fixed. When $h \to 0$, we may modify the derivation in Theorem 1 and obtain

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

for some covariance matrix Σ^*.

3 Computational learning: gradient descent

In Section 1 and Theorem 1, we have developed basic statistical learning result of an M-estimator. Now we will investigate the computational perspective about this estimator.

For simplicity, we assume that our estimator is from ERM

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

Numerically, a popular approach to compute $\widehat{\theta}_n$ 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:

$$\theta^{(t+1)} = \theta^{(t)} - \gamma \nabla \bar{R}_n(\theta^{(t)}), \tag{4}$$

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 1. To this end, we will introduce two smoothness conditions.

3.1 L-smooth and M-strongly convex

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

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

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 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

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

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

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

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 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

$$f(y) - f(x) \leq (y-x)^T \nabla f(x) + \frac{L}{2}\|x - y\|^2. \tag{5}$$

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

$$f(y) - f(x) \geq (y-x)^T \nabla f(x) + \frac{M}{2}\|x - y\|^2. \tag{6}$$

The inequalities in equations (5) and (6) 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.

3.2 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 2 *Suppose the objective function $f(\theta) \equiv \bar{R}_n(\theta)$ is L-smooth and M-strongly convex. Then we have*

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

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

Theorem 2 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 2 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 2 can be applied under the conventional MLE assumptions (M1-4).

Proof.

A direct expansion shows that

$$\begin{aligned}
\|\theta^{(t+1)} - \widehat{\theta}_n\|^2 &= \|\theta^{(t)} - \gamma \nabla f(\theta^{(t)}) - \widehat{\theta}_n\|^2 \\
&= \|\theta^{(t)} - \widehat{\theta}_n\|^2 - 2\gamma(\theta^{(t)} - \widehat{\theta}_n)^T \nabla f(\theta^{(t)}) + \gamma^2 \|\nabla f(\theta^{(t)})\|^2 \\
&= \|\theta^{(t)} - \widehat{\theta}_n\|^2 + 2\gamma(\widehat{\theta}_n - \theta^{(t)})^T \nabla f(\theta^{(t)}) + \gamma^2 \|\nabla f(\theta^{(t)})\|^2
\end{aligned} \tag{7}$$

The middle term $2\gamma(\widehat{\theta}_n - \theta^{(t)})^T \nabla f(\theta^{(t)})$ has a useful upper bound from equation (6), where

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

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

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

Therefore, equation (7) has an upper bound

$$\|\theta^{(t+1)} - \widehat{\theta}_n\|^2 \leq (1 - \gamma M)\|\theta^{(t)} - \widehat{\theta}_n\|^2 + 2\gamma(f(\widehat{\theta}_n) - f(\theta^{(t)})) + \gamma^2 \|\nabla f(\theta^{(t)})\|^2. \tag{8}$$

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

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

Thus,

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

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

$$f\left(\theta - \frac{1}{L}\nabla f(\theta)\right) - f(\theta) \geq f(\widehat{\theta}_n) - f(\theta). \tag{9}$$

Recall the L-smoothness property in equation (5):

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

Choosing $y = \theta - \frac{1}{L}\nabla f(\theta)$ and $x = \theta$, the left-hand-side of equation (9) 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 (9), we conclude that

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

and applying this to equation (8), we obtain

$$\begin{aligned}
\|\theta^{(t+1)} - \widehat{\theta}_n\|^2 &\leq (1-\gamma M)\|\theta^{(t)} - \widehat{\theta}_n\|^2 + 2\gamma(f(\widehat{\theta}_n) - f(\theta^{(t)})) + \gamma^2\|\nabla f(\theta^{(t)})\|^2 \\
&\leq (1-\gamma M)\|\theta^{(t)} - \widehat{\theta}_n\|^2 - \frac{\gamma}{L}\|\nabla f(\theta)\|^2 + \gamma^2\|\nabla f(\theta)\|^2 \\
&= (1-\gamma M)\|\theta^{(t)} - \widehat{\theta}_n\|^2 - \frac{\gamma}{L}(1-\gamma L)\|\nabla f(\theta)\|^2 \\
&\leq (1-\gamma M)\|\theta^{(t)} - \widehat{\theta}_n\|^2
\end{aligned} \tag{10}$$

when $\gamma < \frac{1}{L}$. Note that to ensure equation (10) is contracting, we also need $\gamma < \frac{1}{M}$, which is the other requirement of the stepsize γ.

By telescoping equation (10), we conclude that

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

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

4 Bridging statistical and computational learning

While Theorem 2 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 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.

4.1 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 *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

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

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

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

Thus, for any point θ, its smallest eigenvalue

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

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

$$\lambda_{\min}(H_R(\theta)) \geq \lambda^*_{\min} - \phi_3 \|\theta - \theta^*\|_2 > 0, \tag{11}$$

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.

□

The nice part of Lemma 3 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 (11), we may choose

$$\zeta_1 = \frac{\lambda^*_{\min}}{2\phi_3} \Longrightarrow \lambda_{\min}(H_R(\theta)) \geq \frac{1}{2}\lambda^*_{\min}. \tag{12}$$

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

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

$$h_{\max} = \sup_{\theta \in \Theta} \|\bar{H}_R(\theta)\|_2 < \infty. \tag{13}$$

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.

4.2 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 $\widehat{\theta}_n$.

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 4 *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*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq q(x)\|\theta_1 - \theta_2\|_2 \tag{14}$$

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 4 follows from Example 19.7 and Theorem 19.4 of the following book:

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

The details of the proof would require some techniques from empirical process theory so we omit it.

Results in Lemma 4 are known as Glivenko-Cantelli (GC) theory for the function class $\{f_\theta : \theta \in \Theta\}$.

Lemma 4 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 (14). Thus, Lemma 4 implies the following uniform convergence:

$$\begin{aligned} \sup_{\theta\in\Theta} |\bar{R}_n(\theta) - \bar{R}(\theta)| &\xrightarrow{P} 0, \\ \sup_{\theta\in\Theta} \|\nabla\bar{R}_n(\theta) - \nabla\bar{R}(\theta)\|_{\max} &\xrightarrow{P} 0, \\ \sup_{\theta\in\Theta} \left\| \underbrace{\nabla\nabla\bar{R}_n(\theta)}_{=\bar{H}_{R,n}(\theta)} - \underbrace{\nabla\nabla\bar{R}(\theta)}_{=\bar{H}_R(\theta)} \right\|_{\max} &\xrightarrow{P} 0. \end{aligned} \tag{15}$$

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

Theorem 5 (Convergence of gradient descent) *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 $\theta^{(0)} \in B(\widehat{\theta}_n, \zeta_0)$ 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*

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

The constants in Theorem 5 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 2 and Lemma 3, we only need to show that $\bar{R}_n(\theta)$ is both L^*-smooth and M^*-strongly convex within $B(\widehat{\theta}_n, \zeta_0)$ for some L^*, M^*, ζ_0.

***L*-smoothness.** The L-smoothness of $\bar{R}(\theta)$ comes from equation (13), 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 (15) 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 $\widehat{\theta}_n$, not θ^*. To make the analysis easier, we utilize the fact that $\widehat{\theta}_n \xrightarrow{P} \theta^*$ by Theorem 1.

We consider the following event

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

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

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

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

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

so we choose

$$\zeta_0 = \frac{1}{2}\zeta_1 = \frac{\lambda^*_{\min}}{4\phi_3}, \tag{16}$$

which implies $B\left(\widehat{\theta}, \zeta_0\right) \subset B(\theta^*, \zeta_1)$. By equation (12), 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 (15) 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 $\theta \in B(\widehat{\theta}_n, \zeta_0)$.

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 $\theta \in B(\widehat{\theta}_n, \zeta_0)$. Namely, the function $\bar{R}_n(\theta)$ is M^*-strongly convex with $M^* = \frac{1}{4}\lambda^*_{\min}$ when $\theta \in B(\widehat{\theta}_n, \zeta_0)$.

By Theorem 2, we conclude that

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

when $\theta^{(0)} \in B(\widehat{\theta}_n, \zeta_0)$ 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.

□