3 Asymptotic Statistics

In this section, we discuss the asymptotic properties of the parametric models. We will start with describing the supervised learning setup which will be the focus of next few lectures.

3.1 Supervised learning setting

We assume that we observed n pairs of feature/response pairs $(x_i, y_i) \sim p(x, y)$ for $i = 1, 2, ..., n$ where $x \in \mathcal{X} \subset \mathbb{R}^d$ and $y \in \mathcal{Y} \subset \mathbb{R}$ (which could be a real number or discrete class label). Data pairs are i.i.d. and underlying joint distribution $\sim p(x, y)$ is unknown to us. Our goal is to learn some function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ using the observed data that will help us predict y_i given features x_i, i.e., $y_i \approx \hat{f}(x_i)$.

We will need to define a measure to evaluate the quality of learned function $\hat{f}$.

- **Loss:** For this, we choose a loss function $\ell(y, f(x)) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. For example, a commonly used loss function is the squared error loss function $\ell(y, f(x)) = (y - f(x))^2$, or another one is the absolute value of the error $\ell(y, f(x)) = |y - f(x)|$. Loss function evaluates the error on only one sample.
- **Risk:** However, we would like to measure the error on average which is why we define the risk $R : \mathcal{F} \to \mathbb{R}_+$ of this function to be $R(f) = \mathbb{E}[\ell(y, f(x))]$. Hereby, the expectation will be implicitly over all random variables inside brackets, and $\mathcal{F}$ denotes the set of functions. The risk is a function of f and it also depends on the joint density $p(x, y)$, and loss ℓ.
- **Goal (revised):** Find $f \in \mathcal{F}$ such that $R(f)$ is small (to be revised again).

Example. [Bias-Variance Decomposition (first step)] We choose the loss as the squared error loss, $\ell(y, f(x)) = (y - f(x))^2$ and write the risk as

$$\begin{aligned}
R(f) &= \mathbb{E}[(y - f(x))^2] \\
&= \mathbb{E}[\mathbb{E}[(y - f(x))^2 | x]] \quad \text{(Law of iterated expectations)} \\
&= \mathbb{E}\Big[\mathbb{E}[(y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - f(x))^2 | x]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[(y - \mathbb{E}[y|x])^2 | x] + \mathbb{E}[(f(x) - \mathbb{E}[y|x])^2 | x] + 2\underbrace{\mathbb{E}[(y - \mathbb{E}[y|x])(\mathbb{E}[y|x] - f(x))|x]}_{=0}\Big] \\
&= \underbrace{\mathbb{E}[\mathrm{Var}(y|x)]}_{\text{Irreducible error}} + \mathbb{E}[(f(x) - \mathbb{E}(y|x))^2] \quad = \text{Irreducible error + Variance + Bias}^2
\end{aligned}$$

Since the irreducible error is not a function of f, the lower bound on the risk of f is obtained when $f_*(x) = \mathbb{E}[y|x]$. This is attainable if $f_* \in \mathcal{F}$. That is,

$$\inf_{f \in \mathcal{F}} R(f) = \mathbb{E}[\mathrm{Var}(y|x)] + \inf_{f \in \mathcal{F}} E[(f(x) - E(y|x))^2].$$

We will return to bias-variance decomposition later.

3.1.1 Parametric Models

When we are searching for a function f satisfying $y_i \approx f(x_i)$ for $i = 1, ..., n$, we need to restrict ourselves to a specific set of functions $\mathcal{F}$ to avoid overfitting. Otherwise, we can simply choose any function satisfying $y_i = f(x_i)$ for $\forall i$.

In this subsection, we focus our attention on a parametric function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where f_θ is a function (or hypothesis) and Θ is the parameter space.

Example. Consider the set of linear functions that have weights constrained in a ball of radius λ,

$$\mathcal{F} = \{f_\theta(x) = \langle x, \theta \rangle \ : \ \|\theta\|_2 \leq \lambda\}$$

Notice that the parameter space is given by $\Theta = \{\theta \ : \ \|\theta\|_2 \leq \lambda\}$.

In the case of parametric models, it is generally redundant to write the function f_θ, instead we will simply use the parameter θ to describe it. For example,

$$\ell(y, f_\theta(x)) \triangleq \ell((y, x), \theta) \quad \text{and} \quad R(f_\theta) \triangleq R(\theta),$$

is more compact and conveys the same information for parametric function classes.

We would like to minimize the population risk $R(\theta)$, that is, we want

$$\theta_* \in \arg\min_{\theta \in \Theta} R(\theta) = \mathbb{E}[\ell((x, y), \theta)]$$

for $(x, y) \sim p$. But we don't have access to the joint density p, therefore we cannot minimize this objective. Instead, what we can estimate the population risk with the empirical risk using our n training samples. The empirical risk is just a sample mean estimator for the population risk and given as

$$\hat{\theta} \in \arg\min_{\theta} \hat{R}(\theta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), \theta).$$

Notice the hat in $\hat{R}$ and $\hat{\theta}$ which indicates that these are estimators that depend on data. These quantities are random variables (or vectors) whereas $R(\theta)$ and θ_* are deterministic values ($R(\hat{\theta})$ is also random).

- The quantity $\hat{R}(\hat{\theta})$ is the **training error**.
- The quantity $R(\hat{\theta})$ is simply **test error**. It is worth noting that in machine learning courses, we define test error as an estimator to this quantity.

Notice that when n is large, we expect to have $R(\theta) \approx \hat{R}(\theta)$; thus, it would makes sense to have the minimizers of these functions close together $\hat{\theta} \approx \theta_*$.

The following quantity will be used repeatedly as a notion of generalization error.

Definition 12 (Excess risk)**.** *We define the excess risk of an estimator $\hat{\theta}$ as the distance between the test error and the minimum achievable error*

$$\textit{Excess risk} = R(\hat{\theta}) - R(\theta_*).$$

3.2 MLE Framework

In the MLE framework, we assume that data pairs are sampled in the following hierarchical way

$$\begin{aligned} y_i | x_i \sim & p_{\theta_*}(y|x) \\ x_i \sim & p(x) \end{aligned}$$

where θ_* is the true but unknown parameter. However, we make the very strong assumption that the parametric form $p_\theta(x)$ is known. This is like assuming that we know a random variable z is Gaussian $z \sim \mathcal{N}(\theta_*, 1)$ but we don't know the value of θ_*.

The MLE is motivated as a finding "the most likely" parameter. If we translate this to our framework, we simply choose a loss function that is the negative of the log-likelihood, i.e.

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

We give two examples below.

Example. Parametric distribution is normal with mean $\langle x, \theta\rangle$ and some variance σ^2 (which doesn't matter). That is

$$\begin{aligned} y|x &\sim \mathcal{N}(\langle x,\theta\rangle, \sigma^2) \\ p_\theta(y|x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\langle x,\theta\rangle)^2}{2\sigma^2} \right\} \\ \ell((x,y),\theta) &= -\log p_\theta(y|x) = (y-\langle x,\theta\rangle)^2 + \text{ const.} \end{aligned}$$

which is the squared error loss yielding linear regression.

Example. Parametric distribution is Bernoulli with mean $\sigma(\langle x,\theta\rangle)$ where σ, in this case, is the sigmoid function. That is

$$\begin{aligned} y|x &\sim \text{Ber}(\sigma(\langle x,\theta\rangle)) \\ p_\theta(y|x) &= \sigma(\langle x,\theta\rangle)^y (1-\sigma(\langle x,\theta\rangle))^{1-y} \\ \ell((x,y),\theta) &= -\log p_\theta(y|x) = -y\log(\sigma(\langle x,\theta\rangle)) - (1-y)\log(1-\sigma(\langle x,\theta\rangle)) \end{aligned}$$

which is the cross-entropy loss yielding logistic regression. Both of above settings belong to large class of regression models called generalized linear models (GLMs). They are obtained by modeling the natural parameter in exponential families with a linear function of feature vector. As seen above, Gaussian leads to linear regression whereas Bernoulli leads to logistic regression.

MLE problem: We observe n data point: $(y_i, x_i) \sim p_{\theta_*}(y|x)p(x)$, $i = 1, \ldots, n$. Our goal here is to estimate the true parameter θ_*, by minimizing the empirical risk:

$$\hat{\theta} = \arg\min_\theta \;\; \hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell((x_i,y_i),\theta) = -\frac{1}{n}\sum_{i=1}^n \log p_\theta(y_i|x_i).$$

Let's see how this is related to population risk minimizer. We start investigating by writing out the gradient and the Hessian of $R(\theta)$.

- $\nabla R(\theta_*) = \mathbb{E}[-\nabla \log p_{\theta_*}(y|x)] = 0$. Therefore, the true parameter is a critical point of the population risk.
- $\nabla^2 R(\theta_*) = \mathbb{E}[-\nabla^2 \log p_{\theta_*}(y|x)] = E[\nabla \log p_{\theta_*}(y|x) \nabla \log p_{\theta_*}(y|x)^T] = \mathcal{I}_{\theta_*} \succeq 0$. This still doesn't prove that θ_* is a local minimum. Note that the Hessian of the risk must be positive semi-definite (PSD) since zz^T is PSD as $u^T zz^T u = (u^T z)^2$.

In what follows, for simplicity we assume that $\mathcal{I}_{\theta_*} \succ 0$ which clearly implies that true parameter θ_* is a local minimum. Actually, if we assume identifiability of our parametric family, that is $\theta \neq \theta' \implies p_\theta \neq p_{\theta'}$, then θ_* can be shown to be a unique global minimum.

3.3 Asymptotics of MLE

First, we need a few definitions.

Definition 13 (Convergence of random variables)**.**

(a) ***Convergence in probability****: We write $\hat{\theta}_n \xrightarrow{p} \theta_*$, if for every $\epsilon > 0$ we have*

$$\mathbb{P}(|\hat{\theta}_n - \theta_*| > \epsilon) \to 0 \quad as \quad n \to \infty.$$

(b) ***Convergence in distribution****: We write $\hat{\theta}_n \xrightarrow{d} \theta_*$, if X_n and X have CDFs $F_n(x)$ and $F(x)$, respectively and for every continuity point of $F(x)$, we have $\lim_{n\to\infty} F_n(x) = F(x)$. This is also called weak convergence as it is a very weak notion of convergence. The letter d in the symbol $\xrightarrow{d}$ is to specify that the convergence is in distribution, and it should not be confused be the dimension d.*

(c) ***Consistency****: We say $\hat{\theta}_n$ is a consistent estimator for θ_* if $\hat{\theta}_n \xrightarrow{p} \theta_*$.*

In our asymptotic setting, we fix the dimension d and let number of samples $n \to \infty$. We drop the subscript n to ease the notation.

3.3.1 Asymptotic normality

The following theorem is charactering the asymptotics of the MLE.

Theorem 14 (Asymptotics of MLE)**.** *Assume that $\hat{\theta}$ is consistent for θ_*, and the Fisher information satisfies $\mathcal{I}_{\theta_*} \succ 0$, and that $\sup_\theta \|\nabla^3 \log p_\theta\|_{op} < B$. Then,*

1. $\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\theta_*}^{-1})$.

2. $n(R(\hat{\theta}) - R(\theta_*)) \xrightarrow{d} \frac{1}{2}\chi^2_d$.

Remark. We make two important remarks about the above theorem.

1. The first result is giving us the asymptotic distribution of the MLE. We observe that the variance of this distribution is the inverse Fisher information which validates its name: larger the Fisher information is, lower the variance of this distributions. Therefore, the estimator gives more information about the true parameter.

   It is worth noting that these types of distributional results are useful in constructing confidence intervals; hence quantifying uncertainty in models.

2. The second item above is the asymptotic distribution of the excess risk. Since χ^2_d is a random variable with mean d and variance $2d$, the right hand side is roughly of order d/n. That is,

   $$R(\hat{\theta}) - R(\theta_*) \approx \mathcal{O}\left(\frac{d}{n}\right).$$

   The excess risk gets worse with increased dimension buy gets better with increased number of samples. We should emphasize that this is an asymptotic rate and it is quite fast compared to the non-asymptotic rates that we will obtain in the future lectures. It is also worth noting that this is an equality rather than a upper bound.

Proof sketch.

We start by proving the first item, the normality of the MLE. The distribution of excess risk will follow. Our proof outline is 1- we apply Taylor's theorem, 2- identify a term that is an iid sum which converges to a Gaussian random variable by central limit theorem (CLT), 3- show that the other quantities converge in probability to deterministic quantities. We finally apply the Slutsky's theorem to conclude the proof.

Lemma 15 (Slutsky's Theorem)**.** *For a sequence of random variables* $\{x_n, y_n, z_n\}_{n\in\mathbb{N}}$ *satisfying* $x_n \xrightarrow{d} x$, $y_n \xrightarrow{p} a$ *and* $z_n \xrightarrow{p} b$ *where* a, b *are constants, then we have* $x_n y_n + z_n \xrightarrow{d} ax + b$.

We omit the proof as it is a simple application of the continuous mapping theorem.

We notice that $\nabla R(\theta_*) = \nabla \hat{R}(\hat{\theta}) = 0$. We expand the latter using the Taylor's theorem around the true parameter θ_*.

$$\nabla \hat{R}(\hat{\theta}) = \nabla \hat{R}(\theta_*) + \nabla^2 \hat{R}(\theta_*)(\hat{\theta} - \theta_*) + \frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})[\hat{\theta} - \theta_*, \hat{\theta} - \theta_*].$$

Above, the last term is a tensor in $\nabla^3 \hat{R}(\bar{\theta}) \in \mathbb{R}^{d\times d\times d}$, when multiplied by a vector (e.g. $\hat{\theta} - \theta_*$) it reduces to a $d \times d$ matrix. Also, $\bar{\theta}$ is chosen somewhere on the line of and between $\hat{\theta}$ and θ_* (it is worth noting that mean value theorem doesn't hold for vector valued functions which can be easily fixed by using the integral form Taylor's theorem).

We notice that the left hand side is zero. Rearranging terms, we get

$$-\nabla \hat{R}(\theta_*) = [\nabla^2 \hat{R}(\theta_*) + \frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})(\hat{\theta} - \theta_*)](\hat{\theta} - \theta_*) \tag{3.1}$$

Multiplying both sides with $\sqrt{n}$, we obtain

$$\underbrace{-\sqrt{n}\nabla \hat{R}(\theta_*)}_{\text{iid sum}/\sqrt{n}} = [\underbrace{\nabla^2 \hat{R}(\theta_*)}_{\text{iid sum}/n} + \underbrace{\frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})(\hat{\theta} - \theta_*)]}_{\xrightarrow{p} 0} \underbrace{\sqrt{n}(\hat{\theta} - \theta_*)}_{\text{of interest}}$$

We observe that the left hand side of (3.1) is a iid sum divided by $\sqrt{n}$. By the CLT, we obtain

$$-\sqrt{n}\nabla \hat{R}(\theta_*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla \log p_{\theta_*}(y_i|x_i) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}(\nabla \log p_{\theta_*}(y_i|x_i))).$$

Here, the expected value is 0 since $\mathbb{E}[\nabla \log p_{\theta_*}(y_i|x_i)] = 0$, and $\mathrm{Cov}(\nabla \log p_{\theta_*}(y_i|x_i)) = \mathcal{I}_{\theta_*}$.

For the first term on the right hand side of (3.1), we have another iid sum but this time divided by n. We use law of large numbers (LLN) to obtain

$$\nabla^2 \hat{R}(\theta_*) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \log p_{\theta_*}(y_i|x_i) \xrightarrow{p} \nabla^2 R(\theta_*) = \mathcal{I}_{\theta_*}.$$

The second term on the right hand side of (3.1) converges to 0 in probability by the consistency assumption. Therefore, multiplying both sides with $\mathcal{I}_{\theta_*}^{-1}$, we obtain that

$$\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{I}_{\theta_*}^{-1}\mathcal{N}(0, \mathcal{I}_{\theta_*}). \tag{3.2}$$

We proceed by using a very useful property of Gaussian random vectors.

Lemma 16. *Let $z \sim \mathcal{N}(\mu, \Sigma)$ be a d-dimensional Gaussian random vector. Then for a matrix $A \in \mathbb{R}^{l \times d}$ we have $Az \sim \mathcal{N}(A\mu, A\Sigma A^\top)$.*

Using the above lemma together with (3.2) and obtain

$$\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\theta_*}^{-1}). \tag{3.3}$$

This concludes the proof of the first part. For the proof of second part, we again use Taylor's theorem and write

$$R(\hat{\theta}) - R(\theta_*) = \Big\langle \nabla R(\theta_*), \hat{\theta} - \theta_* \Big\rangle + \frac{1}{2}\Big\langle \nabla^2 R(\theta_*)(\hat{\theta} - \theta_*), \hat{\theta} - \theta_* \Big\rangle + \frac{1}{6}\nabla^3 \hat{R}(\bar{\theta})[\hat{\theta} - \theta_*, \hat{\theta} - \theta_*, \hat{\theta} - \theta_*],$$

where again the first term on the right hand side disappears, and $\bar{\theta}$ is in between θ_* and $\hat{\theta}$ (this time without any issue since R is real-valued). Multiplying both sides with n and rearranging, we obtain

$$n\{R(\hat{\theta}) - R(\theta_*)\} = \frac{1}{2}\Big\langle \sqrt{n}(\hat{\theta} - \theta_*), \{\nabla^2 R(\theta_*) + \frac{1}{3}\nabla^3 R(\bar{\theta})[\hat{\theta} - \theta_*]\}\sqrt{n}(\hat{\theta} - \theta_*) \Big\rangle.$$

Using the previous result (3.3), we know that $\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} z$ where $z \sim \mathcal{N}(0, \mathcal{I}_{\theta_*}^{-1})$, and the term multiplying $\nabla^3 R$ vanishes due to consistency. Therefore, as $n \to \infty$, the right hand side converges in distribution to

$$n\{R(\hat{\theta}) - R(\theta_*)\} \xrightarrow{d} \frac{1}{2}\langle z, \mathcal{I}_{\theta_*} z \rangle.$$

We use Lemma 16 to deduce that

$$\frac{1}{2}\langle z, \mathcal{I}_{\theta_*} z \rangle = \frac{1}{2}\langle \mathcal{I}_{\theta_*}^{1/2} z, \mathcal{I}_{\theta_*}^{1/2} z \rangle = \frac{1}{2}\|\tilde{z}\|^2 \sim \frac{1}{2}\chi_d^2$$

where $\tilde{z} \sim \mathcal{N}(0, I)$ with I denoting the identity matrix. This concludes the proof of the second statement.

□

It is important to identify the contribution of each assumption. It is obvious that the CLT follows from the iid average structure of the MLE problem (also there for many learning tasks). The bounded third derivative is needed to control higher-order terms. Lastly, consistency is needed to kill the third-order term which reduces everything to a quadratic problem in the asymptotic limit.

3.3.2 Consistency

It turns out that the consistency assumption is actually true for the MLE, under certain assumptions (Note that the below assumptions are stronger than what is in fact needed).

Theorem 17 (MLE is consistent)**.** *Assume that the following assumptions are satisfied.*

(a) **Uniform convergence:** *The empirical process satisfies* $\sup_{\theta \in \Theta} |\hat{R}(\theta) - R(\theta)| \xrightarrow{p} 0$.

(b) **Identifiability:** *For every* $\epsilon > 0$, $\inf_{\theta : \|\theta - \theta_*\| \geq \epsilon} R(\theta) > R(\theta_*)$.

(c) **Compactness***: Θ is non-empty and compact.*

Then, $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \hat{R}(\theta)$ *is consistent.*

Remark. The first assumption above is a very strong notion of convergence and it will be quite handy when we talk about generalization. The second assumption simple means that we can identify the function has a unique minimizer θ_* and around that point, R grows. The last assumption is only needed to ensure that θ_* and $\hat{\theta}$ belong to the set Θ.

Proof. By the compactness assumption, we have $\hat{\theta}, \theta_* \in \Theta$. Next, notice that since $\hat{\theta}$ minimizes $\hat{R}$ in Θ, we have $\hat{R}(\hat{\theta}) \leq \hat{R}(\theta_*)$. We can write

$$\begin{aligned}
\hat{R}(\hat{\theta}) \leq & \hat{R}(\theta_*) \\
= & \hat{R}(\theta_*) - R(\theta_*) + R(\theta_*) \\
\leq & \sup_{\theta \in \Theta} \left| \hat{R}(\theta) - R(\theta) \right| + R(\theta_*) \xrightarrow{p} R(\theta_*) \quad \text{by assump (a).}
\end{aligned} \tag{3.4}$$

Also, since θ_* minimizes R, we write

$$\begin{aligned}
0 \leq R(\hat{\theta}) - R(\theta_*) \leq & R(\hat{\theta}) - \hat{R}(\hat{\theta}) \quad \text{as } n \to \infty \text{ by (3.4)}, \\
\leq & \sup_{\theta \in \Theta} \left| \hat{R}(\theta) - R(\theta) \right| \xrightarrow{p} 0, \quad \text{by assump (a).}
\end{aligned} \tag{3.5}$$

Notice that we squeezed the excess risk between zeros. So for every $\epsilon > 0$, the following holds for the events

$$\left\{ \| \hat{\theta} - \theta_* \| \geq \epsilon \right\} \underset{\text{by assumption (b)}}{\subseteq} \left\{ R(\hat{\theta}) - R(\theta_*) > \delta_\epsilon \right\}$$

Probability of the right hand side above goes to 0 as we let $n \to \infty$ due to (3.5).

□