

Lecture 13: Maximum Likelihood Estimation

1 Summary of Lecture 12

In the last lecture we derived a risk (MSE) bound for regression problems; i.e., select an $f \in \mathcal{F}$ so that $E[(f(X) - Y)^2] - E[(f^*(X) - Y)^2]$ is small, where $f^*(x) = E[Y|X = x]$. The result is summarized below.

Theorem 1 (Complexity Regularization with Squared Error Loss) Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = [-b/2, b/2]$, $\{X_i, Y_i\}_{i=1}^n$ iid, P_{XY} unknown, $\mathcal{F} = \{\text{collection of candidate functions}\}$,

$$f : \mathbb{R}^d \rightarrow \mathcal{Y}, \quad R(f) = E[(f(X) - Y)^2].$$

Let $c(f)$, $f \in \mathcal{F}$, be positive numbers satisfying $\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1$, and select a function from \mathcal{F} according to

$$\hat{f}_n = \arg \min \left\{ \hat{R}_n(f) + \frac{1}{\epsilon} \frac{c(f) \log 2}{n} \right\},$$

with $\epsilon \leq \frac{3}{5b^2}$ and $\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2$. Then,

$$E[R(\hat{f}_n)] - R(f^*) \leq \left(\frac{1 + \alpha}{1 - \alpha} \right) \min_{f \in \mathcal{F}} \left\{ R(f) - R(f^*) + \frac{1}{\epsilon} \frac{c(f) \log 2}{n} \right\} + O(n^{-1})$$

where $\alpha = \frac{cb^2}{1 - 2b^2\epsilon/3}$

2 Maximum Likelihood Estimation

The focus of this lecture is to consider another approach to learning based on maximum likelihood estimation. Consider the classical signal plus noise model:

$$Y_i = f\left(\frac{i}{n}\right) + W_i, \quad i = 1, \dots, n$$

where W_i are iid zero-mean noises. Furthermore, assume that $W_i \sim p(w)$ for some known density $p(w)$. Then

$$Y_i \sim p\left(y - f\left(\frac{i}{n}\right)\right) \equiv p_{f_i}(y)$$

since $Y_i - f\left(\frac{i}{n}\right) = W_i$.

A very common and useful loss function to consider is

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (-\log p_{f_i}(Y_i)).$$

Minimizing \hat{R}_n with respect to f is equivalent to maximizing

$$\frac{1}{n} \sum_{i=1}^n \log p_{f_i}(Y_i)$$

or

$$\prod_{i=1}^n p_{f_i}(Y_i).$$

Thus, using the negative log-likelihood as a loss function leads to maximum likelihood estimation. If the W_i are iid zero-mean Gaussian r.v.s then this is just the squared error loss we considered last time. If the W_i are Laplacian distributed e.g. $p(w) \propto e^{-|w|}$, then we obtain the absolute error, or L_1 , loss function. We can also handle non-additive models such as the Poisson model

$$Y_i \sim P(y|f(i/n)) = \frac{e^{-f(i/n)} [f(i/n)]^y}{y!}$$

In this case

$$-\log P(Y_i|f(i/n)) = f(i/n) - Y_i \log(f(i/n)) + \text{constant}$$

which is a very different loss function, but quite appropriate for many imaging problems.

Before we investigate maximum likelihood estimation for model selection, let's review some of the basis concepts. Let Θ denote a parameter space (e.g., $\Theta = R$), and assume we have observations

$$Y_i \stackrel{iid}{\sim} p_{\theta^*}(y), \quad i = 1, \dots, n$$

where $\theta^* \in \Theta$ is a parameter determining the density of the $\{Y_i\}$. The ML estimator of θ^* is

$$\begin{aligned} \hat{\theta}_n &= \arg \max_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(Y_i) \\ &= \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log p_{\theta}(Y_i) \\ &= \arg \min_{\theta \in \Theta} \sum_{i=1}^n -\log p_{\theta}(Y_i). \end{aligned}$$

$\hat{\theta}$ maximizes the expected log-likelihood. To see this, let's compare the expected log-likelihood of θ^* with any other $\theta \in \Theta$.

$$\begin{aligned} E[\log p_{\theta^*}(Y) - \log p_{\theta}(Y)] &= E \left[\log \frac{p_{\theta^*}(Y)}{p_{\theta}(Y)} \right] \\ &= \int \log \frac{p_{\theta^*}(y)}{p_{\theta}(y)} p_{\theta^*}(y) dy \\ &= K(p_{\theta}, p_{\theta^*}) \quad \text{the KL divergence} \\ &\geq 0 \quad \text{with equality iff } p_{\theta^*} = p_{\theta}. \end{aligned}$$

Why?

$$\begin{aligned} -E \left[\log \frac{p_{\theta^*}(y)}{p_{\theta}(y)} \right] &= E \left[\log \frac{p_{\theta}(y)}{p_{\theta^*}(y)} \right] \\ &\leq \log E \left[\frac{p_{\theta}(y)}{p_{\theta^*}(y)} \right] \\ &= \log \int p_{\theta}(y) dy = 0 \\ &\Rightarrow K(p_{\theta}, p_{\theta^*}) \geq 0 \end{aligned}$$

2.1 Likelihood as a Loss Function

We can restate the maximum likelihood estimator in the general terms we are using in this course. We have n i.i.d observations drawn from an unknown distribution

$$Y_i \stackrel{i.i.d.}{\sim} p_{\theta^*}, \quad i = \{1, \dots, n\}$$

where $\theta^* \in \Theta$. We can view p_{θ^*} as a member of a parametric class of distributions, $\mathcal{P} = \{p_{\theta}\}_{\theta \in \Theta}$. Our goal is to use the observations $\{Y_i\}$ to *select* an appropriate distribution (e.g., model) from \mathcal{P} . We would like the selected distribution to be close to p_{θ^*} in some sense.

We use the negative log-likelihood *loss function*, defined as $l(\theta, Y_i) = -\log p_{\theta}(Y_i)$. The **empirical risk** is

$$\hat{R}_n = -\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(Y_i).$$

We select the distribution that minimizes the empirical risk

$$\min_{p \in \mathcal{P}} -\sum_{i=1}^n \log p(Y_i) = \min_{\theta \in \Theta} -\sum_{i=1}^n \log p_{\theta}(Y_i)$$

In other words, the distribution we select is $\hat{p} := p_{\hat{\theta}_n}$, where

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} -\sum_{i=1}^n \log p_{\theta}(Y_i)$$

The **risk** is defined as

$$R(\theta) = E[l(\theta, Y)] = -E[\log p_{\theta}(Y)].$$

And, the **excess risk** of θ is defined as

$$R(\theta) - R(\theta^*) = \int \log \frac{p_{\theta^*}(y)}{p_{\theta}(y)} p_{\theta^*}(y) dy \equiv K(p_{\theta}, p_{\theta^*}).$$

We recognized that the excess risk corresponding to this loss function is simply the *Kullback-Leibler (KL) Divergence* or *Relative Entropy*, denoted by $K(p_{\theta_1}, p_{\theta_2})$. It is easy to see that $K(p_{\theta_1}, p_{\theta_2})$ is always non-negative and is zero if and only if $p_{\theta_1} = p_{\theta_2}$. This shows that θ^* minimizes the risk. The KL divergence measures how different two probability distributions are and therefore is natural to measure convergence of the maximum likelihood procedures.

2.2 Convergence of Log-Likelihood to KL Divergence

Since $\hat{\theta}_n$ maximizes the likelihood over $\theta \in \Theta$, we have

$$\sum_{i=1}^n \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} = \sum_{i=1}^n \log p_{\theta^*}(Y_i) - \log p_{\hat{\theta}_n}(Y_i) \leq 0$$

Therefore,

$$\frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} - K(p_{\hat{\theta}_n}, p_{\theta^*}) + K(p_{\hat{\theta}_n}, p_{\theta^*}) \leq 0$$

or re-arranging

$$K(p_{\hat{\theta}_n}, p_{\theta^*}) \leq \left| \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} - K(p_{\hat{\theta}_n}, p_{\theta^*}) \right|$$

Notice that the quantity

$$\frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta^*}(Y_i)}{p_{\theta}(Y_i)}$$

is an empirical average whose mean is $K(p_{\theta}, p_{\theta^*})$. By the law of large numbers, for each $\theta \in \Theta$,

$$\left| \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta^*}(Y_i)}{p_{\theta}(Y_i)} - K(p_{\theta}, p_{\theta^*}) \right| \xrightarrow{a.s.} 0$$

If this also holds for the sequence $\{\hat{\theta}_n\}$, then we have

$$K(p_{\hat{\theta}_n}, p_{\theta^*}) \leq \left| \frac{1}{n} \sum \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} - K(p_{\hat{\theta}_n}, p_{\theta^*}) \right| \rightarrow 0 \text{ as } n \rightarrow \infty$$

which implies that

$$p_{\hat{\theta}_n} \rightarrow p_{\theta^*}$$

which often implies that

$$\hat{\theta}_n \rightarrow \theta^*$$

in some appropriate sense (e.g., point-wise or in norm).

Example 1 *Gaussian Distributions*

$$p_{\theta^*}(y) = \frac{1}{\sqrt{\pi}} e^{-(y-\theta^*)^2}$$

$$\Theta = \mathbb{R}, \quad \{Y_i\}_{i=1}^n \stackrel{iid}{\sim} p_{\theta^*}(y)$$

$$\begin{aligned} K(p_{\theta}, p_{\theta^*}) &= \int \log \frac{p_{\theta^*}(y)}{p_{\theta}(y)} p_{\theta^*}(y) dy \\ &= \int [(y-\theta)^2 - (y-\theta^*)^2] p_{\theta^*}(y) dy \\ &= E_{\theta^*}[(y-\theta)^2] - E_{\theta^*}[(y-\theta^*)^2] \\ &= E_{\theta^*}[Y^2 - 2Y\theta + \theta^2] - 1/2 \\ &= (\theta^*)^2 + 1/2 - 2\theta^*\theta + \theta^2 - 1/2 \\ &= (\theta^* - \theta)^2 \end{aligned}$$

$\Rightarrow \theta^*$ maximizes $E[\log p_{\theta}(Y)]$ wrt $\theta \in \Theta$

$$\begin{aligned} \hat{\theta}_n &= \arg \max_{\theta} \left\{ - \sum (Y_i - \theta)^2 \right\} \\ &= \arg \min_{\theta} \left\{ \sum (Y_i - \theta)^2 \right\} \\ &= \frac{1}{n} \sum_{i=1}^n Y_i \end{aligned}$$

2.3 Hellinger Distance

The KL divergence is not a distance function.

$$K(p_{\theta_1}, p_{\theta_2}) \neq K(p_{\theta_2}, p_{\theta_1})$$

Therefore, it is often more convenient to work with the Hellinger metric,

$$H(p_{\theta_1}, p_{\theta_2}) = \left(\int \left(p_{\theta_1}^{\frac{1}{2}} - p_{\theta_2}^{\frac{1}{2}} \right)^2 dy \right)^{\frac{1}{2}}$$

The Hellinger metric is symmetric, non-negative and

$$H(p_{\theta_1}, p_{\theta_2}) = H(p_{\theta_2}, p_{\theta_1})$$

and therefore it is a distance measure. Furthermore, the squared Hellinger distance lower bounds the KL divergence, so convergence in KL divergence implies convergence of the Hellinger distance.

Proposition 1

$$H^2(p_{\theta_1}, p_{\theta_2}) \leq K(p_{\theta_1}, p_{\theta_2})$$

Proof:

$$\begin{aligned} H^2(p_{\theta_1}, p_{\theta_2}) &= \int \left(\sqrt{p_{\theta_1}(y)} - \sqrt{p_{\theta_2}(y)} \right)^2 dy \\ &= \int p_{\theta_1}(y) dy + \int p_{\theta_2}(y) dy - 2 \int \sqrt{p_{\theta_1}(y)} \sqrt{p_{\theta_2}(y)} dy \\ &= 2 - 2 \int \sqrt{p_{\theta_1}(y)} \sqrt{p_{\theta_2}(y)} dy, \quad \text{since } \int p_{\theta}(y) dy = 1 \forall \theta \\ &= 2 \left(1 - E_{\theta_2} \left[\sqrt{p_{\theta_1}(Y)/p_{\theta_2}(Y)} \right] \right) \\ &\leq 2 \log \left(E_{\theta_2} \left[\sqrt{p_{\theta_2}(Y)/p_{\theta_1}(Y)} \right] \right), \quad \text{since } 1 - x \leq -\log x \\ &\leq 2 E_{\theta_2} \left[\log \sqrt{p_{\theta_2}(Y)/p_{\theta_1}(Y)} \right], \quad \text{by Jensen's inequality} \\ &= E_{\theta_2} [\log(p_{\theta_2}(Y)/p_{\theta_1}(Y))] \equiv K(p_{\theta_1}, p_{\theta_2}) \end{aligned}$$

■

Note that in the proof we also showed that

$$H(p_{\theta_1}, p_{\theta_2}) = 2 \left(1 - \int \sqrt{p_{\theta_1}(y)} \sqrt{p_{\theta_2}(y)} dy \right)$$

and using the fact $\log x \leq x - 1$ again, we have

$$H(p_{\theta_1}, p_{\theta_2}) \leq -2 \log \left(\int \sqrt{p_{\theta_1}(y)} \sqrt{p_{\theta_2}(y)} dy \right)$$

The quantity inside the log is called the *affinity* between p_{θ_1} and p_{θ_2} :

$$A(p_{\theta_1}, p_{\theta_2}) = \int \sqrt{p_{\theta_1}(y)} \sqrt{p_{\theta_2}(y)} dy$$

This is another measure of closeness between p_{θ_1} and p_{θ_2} .

Example 2 *Gaussian Distributions*

$$p_\theta(y) = \frac{1}{\sqrt{\pi}} e^{-(y-\theta)^2}$$

$$\begin{aligned} & -2 \log \int \sqrt{p_{\theta_1}(y)} \sqrt{p_{\theta_2}(y)} dy \\ &= -2 \log \int \left(\frac{1}{\sqrt{\pi}} e^{-(y-\theta_1)^2} \right)^{\frac{1}{2}} \left(\frac{1}{\sqrt{\pi}} e^{-(y-\theta_2)^2} \right)^{\frac{1}{2}} dy \\ &= -2 \log \left(\int \frac{1}{\sqrt{\pi}} e^{-\left[\frac{(y-\theta_1)^2}{2} + \frac{(y-\theta_2)^2}{2}\right]} dy \right) \\ &= -2 \log \left(\int \frac{1}{\sqrt{\pi}} e^{-\left[(y-\frac{\theta_1+\theta_2}{2})^2 + (\frac{\theta_1-\theta_2}{2})^2\right]} dy \right) \\ &= -2 \log e^{-\left(\frac{\theta_1-\theta_2}{2}\right)^2} \\ &= \frac{1}{2}(\theta_1 - \theta_2)^2 \end{aligned}$$

$$\Rightarrow -2 \log A(p_{\theta_1}, p_{\theta_2}) = \frac{1}{2}(\theta_1 - \theta_2)^2 \quad \text{for Gaussian distributions}$$

$$\Rightarrow H^2(p_{\theta_1}, p_{\theta_2}) \leq \frac{1}{2}(\theta_1 - \theta_2)^2 \quad \text{for Gaussian.}$$

Summary

$$Y_i \stackrel{iid}{\sim} p_{\theta^*}$$

1. Maximum likelihood estimator maximizes the empirical average

$$\frac{1}{n} \sum_{i=1}^n \log p_\theta(Y_i)$$

(our empirical risk is negative log-likelihood)

2. θ^* maximizes the expectation

$$E \left[\frac{1}{n} \sum_{i=1}^n \log p_\theta(Y_i) \right]$$

(the risk is the expected negative log-likelihood)

- 3.

$$\frac{1}{n} \sum_{i=1}^n \log p_\theta(Y_i) \xrightarrow{a.s.} E \left[\frac{1}{n} \sum_{i=1}^n \log p_\theta(Y_i) \right]$$

so we expect some sort of concentration of measure.

4. In particular, since

$$\frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta^*}(Y_i)}{p_\theta(Y_i)} \xrightarrow{a.s.} K(p_\theta, p_{\theta^*})$$

we might expect that $K(p_{\hat{\theta}_n}, p_{\theta^*}) \rightarrow 0$ for the sequence of estimates $\{p_{\hat{\theta}_n}\}_{n=1}^\infty$.

So, the point is that maximum likelihood estimator is just a special case of a loss function in learning. Due to its special structure, we are naturally led to consider KL divergences, Hellinger distances, and Affinities.