1 Summary of Lecture 12

In the last lecture we derived a risk (MSE) bound for regression problems; i.e., select an \( f \in F \) 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 \( X = \mathbb{R}^d \), \( Y = [-b/2, b/2] \), \( \{X_i, Y_i\}_{i=1}^n \) iid, \( P_{XY} \) unknown, \( F = \{ \text{collection of candidate functions} \} \), \( f : \mathbb{R}^d \rightarrow Y \), \( R(f) = E[(f(X) - Y)^2] \).

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

\[
\hat{f}_n = \arg \min_{f \in F} \left\{ \hat{R}_n(f) + \frac{1}{\epsilon} \frac{c(f) \log 2}{n} \right\},
\]

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

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

where \( \alpha = \frac{\epsilon b^2}{1 - 2 \epsilon b^2 / 3} \).

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:

\[
Y_i = f \left( \frac{i}{n} \right) + W_i, i = 1, \cdots, n
\]

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

\[
\frac{1}{n} \sum_{i=1}^n \log P_{f_i}(Y_i)
\]
Maximum Likelihood Estimation 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
\[ Y_i \sim P(y|f(i/n)) = e^{-f(i/n)} \left[ \frac{f(i/n)^y}{y!} \right] \]
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 \( \Theta \) denote a parameter space (e.g., \( \Theta = R \)), and assume we have observations
\[ Y_i \overset{iid}{\sim} P_{\theta^*}(y), \quad i = 1, \ldots, n \]
where \( \theta^* \in \Theta \) is a parameter determining the density of the \( \{Y_i\} \). The ML estimator of \( \theta^* \) is
\[ \hat{\theta}_n = \arg \max_{\theta \in \Theta} \prod_{i=1}^{n} P_{\theta}(Y_i) = \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log P_{\theta}(Y_i) = \arg \min_{\theta \in \Theta} \sum_{i=1}^{n} -\log P_{\theta}(Y_i). \]
\( \hat{\theta} \) maximizes the expected log-likelihood. To see this, let’s compare the expected log-likelihood of \( \theta^* \) with any other \( \theta \in \Theta \).
\[ E[\log P_{\theta^*}(Y) - \log P_{\theta}(Y)] = E \left[ \log \frac{P_{\theta^*}(Y)}{P_{\theta}(Y)} \right] \]
\[ = \int \log \frac{P_{\theta^*}(y)}{P_{\theta}(y)} P_{\theta^*}(y) dy \]
\[ = K(P_{\theta}, P_{\theta^*}) \quad \text{the KL divergence} \]
\[ \geq 0 \quad \text{with equality iff } P_{\theta^*} = P_{\theta}. \]
Why?
\[ -E \left[ \log \frac{P_{\theta^*}(y)}{P_{\theta}(y)} \right] = E \left[ \log \frac{P_{\theta}(y)}{P_{\theta^*}(y)} \right] \leq \log E \left[ \frac{P_{\theta}(y)}{P_{\theta^*}(y)} \right] \]
\[ = \log \int P_{\theta}(y) dy = 0 \]
\[ \Rightarrow K(P_{\theta}, P_{\theta^*}) \geq 0 \]
Maximum Likelihood Estimation

On the other hand, 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,

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

or re-arranging

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

Notice that the quantity

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

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

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

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

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

which implies that

$$P_{\theta_n} \rightarrow P_{\theta^*}$$

which often implies that

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

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

**Example 1 Gaussian Distributions**

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

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

$$K(P_{\theta}, P_{\theta^*}) = \int \log \frac{P_{\theta^*}(y)}{P_{\theta}(y)} P_{\theta^*}(y)dy$$

$$= \int [(y - \theta)^2 - (y - \theta^*)^2] P_{\theta^*}(y)dy$$

$$= E_{\theta}[(y - \theta)^2] - E_{\theta^*}[(y - \theta^*)^2]$$

$$= E_{\theta}[y^2 - 2Y\theta + \theta^2] - 1/2$$

$$= (\theta^*)^2 + 1/2 - 2\theta^* \theta + \theta^2 - 1/2$$

$$= (\theta^* - \theta)^2$$

$$\Rightarrow \theta^* \text{ maximizes } E[\log P_{\theta}(Y)] \text{ wrt } \theta \in \Theta$$
\[ \hat{\theta}_n = \arg \max_\theta \left\{ - \sum (Y_i - \theta)^2 \right\} \]
\[ = \arg \min_\theta \left\{ \sum (Y_i - \theta)^2 \right\} \]
\[ = \frac{1}{n} \sum_{i=1}^n Y_i \]

2.1 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,

\[ H(P_{\theta_1}, P_{\theta_2}) = \left( \int \left( P_{\theta_1}^y - P_{\theta_2}^y \right)^2 dy \right)^{\frac{1}{2}} \]

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:

\[
H(P_{\theta_1}, P_{\theta_2}) = \int \left( \sqrt{P_{\theta_1}(y)} - \sqrt{P_{\theta_2}(y)} \right)^2 dy \\
= \int P_{\theta_1}(y)dy + \int P_{\theta_2}(y)dy - 2 \int \sqrt{P_{\theta_1}(y)} \sqrt{P_{\theta_2}(y)}dy \\
= 2 - 2 \int \sqrt{P_{\theta_1}(y)} \sqrt{P_{\theta_2}(y)}dy, \text{ since } \int P_\theta(y)dy = 1 \forall \theta \\
= 2 \left( 1 - E_{\theta_2} \left[ \sqrt{P_{\theta_1}(Y)}/P_{\theta_2}(Y) \right] \right) \\
\leq 2 \log \left( E_{\theta_2} \left[ \sqrt{P_{\theta_2}(Y)/P_{\theta_1}(Y)} \right] \right), \text{ since } 1 - x \leq -\log x \\
\leq 2E_{\theta_2} \left[ \log \sqrt{P_{\theta_2}(Y)/P_{\theta_1}(Y)} \right], \text{ by Jensen’s inequality} \\
= E_{\theta_2} \left[ \log(P_{\theta_2}(Y)/P_{\theta_1}(Y)) \right] \equiv K(P_{\theta_1}, P_{\theta_2})
\]

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 $P_{\theta_1}$ and $P_{\theta_2}$:

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

This is another measure of closeness between $P_{\theta_1}$ and $P_{\theta_2}$.

**Example 2 Gaussian Distributions**

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

$$-2 \log \int \sqrt{P_{\theta_1}(y)} \sqrt{P_{\theta_2}(y)} dy$$

$$= -2 \log \int \left( \frac{1}{\sqrt{\pi}} e^{-(y-\theta_1)^2} \right)^\frac{1}{2} \left( \frac{1}{\sqrt{\pi}} e^{-(y-\theta_2)^2} \right)^\frac{1}{2} dy$$

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

$$= -2 \log e^{-\left(\frac{1}{2}(\theta_1 - \theta_2)^2\right)}$$

$$= \frac{1}{2}(\theta_1 - \theta_2)^2$$

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

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

**Example 3 Poisson Distributions**

If $P_{\theta}(y) = e^{-\theta y} \frac{\theta^y}{y!}, \theta \geq 0$, then

$$-2 \log A(P_{\theta_1}, P_{\theta_2}) = \left( \sqrt{\theta_1} - \sqrt{\theta_2} \right)^2$$

**Summary**

$Y_i \overset{iid}{\sim} P_{\theta^*}$

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. $\theta^*$ 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. 

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

so we expect some sort of concentration of measure.

4. In particular, since

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

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.