Review: Maximum Likelihood Estimation

We have \( n \) i.i.d observations drawn from an unknown distribution \( Y_i \sim p_{\theta^*}, \quad i = \{1, \ldots, n\} \) where \( \theta^* \in \Theta \). We can view \( p_{\theta^*} \) 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 \( p_{\theta^*} \) in some sense.

We use the negative log-likelihood loss function, defined as

\[
\ell(\theta, Y_i) = -\log p_{\theta}(Y_i).
\]

The empirical risk is

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

We select the distribution that minimizes the empirical risk

\[
\min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^{n} \log p_{\theta}(Y_i).
\]

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

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

As shown above, \( \theta^* \) minimizes \( R(\theta) \) over \( \Theta \).

\[
\theta^* = \arg \min_{\theta \in \Theta} -E[\log p_{\theta}(Y)]
\]

\[
= \arg \min_{\theta \in \Theta} - \int \log p_{\theta}(y) \cdot p_{\theta^*}(y) \, dy.
\]

Finally, the excess risk of \( \theta \) 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 \( p_{\theta_1} = p_{\theta_2} \). KL divergence measures how different two probability distributions are and therefore is natural to measure convergence of the maximum likelihood procedures. However, \( K(p_{\theta_1}, p_{\theta_2}) \) is not a distance metric because it is not symmetric and does not satisfy the triangle inequality. For this reason, two other quantities play a key role in maximum likelihood estimation, namely Hellinger Distance and Affinity.
The **Hellinger distance** is defined as
\[
H(p_{\theta_1}, p_{\theta_2}) = \left( \int \left( \sqrt{p_{\theta_1}(y)} - \sqrt{p_{\theta_2}(y)} \right)^2 dy \right)^{\frac{1}{2}}.
\]

We proved that the squared Hellinger distance lower bounds the KL divergence:
\[
H^2(p_{\theta_1}, p_{\theta_2}) \leq K(p_{\theta_1}, p_{\theta_2})
\]
\[
H^2(p_{\theta_2}, p_{\theta_1}) \leq K(p_{\theta_2}, p_{\theta_1})
\]

The **affinity** is defined as
\[
A(p_{\theta_1}, p_{\theta_2}) = \int \sqrt{p_{\theta_1}(y)p_{\theta_2}(y)} \, dy.
\]
we also proved that
\[
H^2(p_{\theta_1}, p_{\theta_2}) \leq -2 \log \left( A(p_{\theta_1}, p_{\theta_2}) \right).
\]

**Example 1 (Gaussian Distribution)** \(Y\) is Gaussian with mean \(\theta\) and variance \(\sigma^2\).
\[
p_{\theta}(y) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}
\]
First, look at
\[
\log \frac{p_{\theta_2}}{p_{\theta_1}} = \frac{1}{2\sigma^2} [(\theta_1^2 - \theta_2^2) - 2(\theta_1 - \theta_2)y]
\]
Then,
\[
K(p_{\theta_1}, p_{\theta_2}) = E_{\theta_2} \left[ \log \frac{p_{\theta_2}}{p_{\theta_1}} \right]
= \frac{\theta_1^2 - \theta_2^2}{2\sigma^2} - \left( \frac{\theta_1^2}{4\sigma^2} - \frac{\theta_2^2}{4\sigma^2} \right) \int_{E[Y]=\theta_2} y \cdot p_{\theta_2}(y) \, dy
= \frac{1}{2\sigma^2} (\theta_1^2 + \theta_2^2 - 2\theta_1 \theta_2) = \frac{(\theta_1 - \theta_2)^2}{2\sigma^2},
\]
\[
-2 \log A(p_{\theta_1}, p_{\theta_2}) = -2 \log \left( \int \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\theta_1)^2}{2\sigma^2}} \cdot \left( \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\theta_2)^2}{2\sigma^2}} \right)^{1/2} \, dy \right)
= -2 \log \left( \int \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\theta_1)^2}{4\sigma^2}} \cdot \left( \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\theta_2)^2}{4\sigma^2}} \right)^{1/2} \, dy \right)
= -2 \log \left( \int \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{4\pi\sigma^2} [(y-\theta_1+\theta_2)^2 + (\theta_1 - \theta_2)^2]} \, dy \right)
= -2 \log e^{-\frac{(\theta_1 - \theta_2)^2}{4\sigma^2}}
= \frac{(\theta_1 - \theta_2)^2}{4\sigma^2} = \frac{1}{2} K(p_{\theta_1}, p_{\theta_2}) \geq H^2(p_{\theta_1}, p_{\theta_2}).
1 Maximum likelihood estimation and Complexity regularization

Suppose that we have \( n \) i.i.d training samples, \( \{X_i, Y_i\}_{i=1}^n \sim p_{XY} \).

Using conditional probability, \( p_{XY} \) can be written as
\[
p_{XY}(x, y) = p_X(x) \cdot p_{Y|X=x}(y).
\]

Let’s assume for the moment that \( p_X \) is completely unknown, but \( p_{Y|X=x}(y) \) has a special form:
\[
p_{Y|X=x}(y) = p_{f^*(x)}(y)
\]

where \( p_{Y|X=x}(y) \) is a known parametric density function with parameter \( f^*(x) \).

Example 2 (Signal-plus-noise observation model)

\[
y_i = f^*(x_i) + W_i, \quad i = 1, \ldots, n
\]

where \( W_i \sim N(0, \sigma^2) \) and \( X_i \sim p_X \).

\[
p_{f^*(x)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f^*(x))^2}{2\sigma^2}}
\]

\[
Y|X = x \sim \text{Poisson}(f^*(x))
\]

\[
p_{f^*(x)}(y) = e^{-f^*(x)} \left[ \frac{[f^*(x)]^y}{y!} \right].
\]

The likelihood loss function is
\[
l(f(x), y) = -\log p_{XY}(X, Y)
\]
\[
= -\log p_X(X) - \log p_{Y|X}(Y|X)
\]
\[
= -\log p_X(X) - \log p_{f(X)}(Y).
\]

The expected loss is
\[
E[l(f(X), Y)] = EX \left[ E_{Y|X}[l(f(X), Y)|X=x] \right]
\]
\[
= EX \left[ E_{Y|X}[-\log p_X(X) - \log p_{f(x)}(Y)|X=x] \right]
\]
\[
= -EX \left[ \log p_X(X) \right] - \log p_{f(X)}(Y).
\]

Notice that the first term is a constant with respect to \( f \).

Hence, we define our risk to be
\[
R(f) = -E[\log p_{f(X)}(Y)]
\]
\[
= -EX \left[ E_{Y|X}[\log p_{f(x)}(Y)|X=x] \right]
\]
\[
= -\int \left( \int \log p_{f(x)}(y) \cdot p_{f^*(x)}(y) \, dy \right) p_X(x) \, dx.
\]

The function \( f^* \) minimizes this risk since \( f^*(x) = f^*(x) \) minimizes the integrand.

Our empirical risk is the negative log-likelihood of the training samples:
\[
\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n -\log p_{f(X_i)}(Y_i)
\]
The value $\frac{1}{n}$ is the empirical probability of observing $X = X_i$.

Often in function estimation, we have control over where we sample $X$. Let’s assume that $X = [0, 1]^d$ and $Y = \mathbb{R}$. Suppose we sample $X$ uniformly with $n = m^d$ samples for some positive integer $m$ (i.e., take $m$ evenly spaced samples in each coordinate).

Let $x_i, i = 1, \ldots, n$ denote these sample points, and assume that $Y_i \sim p_{f^*}(x_i)(y)$. Then, our empirical risk is

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

Note that $x_i$ is now a deterministic quantity. Our risk is

$$R(f) = -\frac{1}{n} \sum_{i=1}^n E[\log p_{f(x_i)}(Y_i)]$$

$$= -\frac{1}{n} \sum_{i=1}^n \int \log p_{f(x_i)}(y_i) \cdot p_{f^*}(x_i)(y_i) dy_i.$$

The risk is minimized by $f^*$. However, $f^*$ is not a unique minimizer. Any $f$ that agrees with $f^*$ at the point $x_i$ also minimizes this risk.

Now, we will make use of the following vector and shorthand notation. The uppercase $Y$ denotes a random variable, while the lowercase $y$ and $x$ denote deterministic quantities.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Then,

$$p_f(Y) = \prod_{i=1}^n p(Y_i|f(x_i)) \quad \text{(random)}$$

$$p_f(y) = \prod_{i=1}^n p(y_i|f(x_i)) \quad \text{(deterministic)}.$$

With this notation, the empirical risk and the true risk can be written as

$$\hat{R}_n(f) = -\frac{1}{n} \log p_f(Y).$$

$$R(f) = -\frac{1}{n} E[\log p_f(Y)]$$

$$= -\frac{1}{n} \int \log p_f(y) \cdot p_{f^*}(y) dy.$$

2 Error Bound

Suppose that we have a pool of candidate functions $\mathcal{F}$, and we want to select a function $f$ from $\mathcal{F}$ using the training data. Our usual approach is to show that the distribution of $\hat{R}_n(f)$ concentrates about its mean as $n$ grows. First, we assign a complexity $c(f) > 0$ to each $f \in \mathcal{F}$ so that $\sum 2^{-c(f)} \leq 1$. Then, apply the union bound to get a uniform concentration inequality holding for all models in $\mathcal{F}$. Finally, we use this concentration inequality to bound the expected risk of our selected model.
We will essentially accomplish the same result here, but avoid the need for explicit concentration inequalities and instead make use of the information-theoretic bounds.

We would like to select an \( f \in F \) so that the excess risk is small.

\[
0 \leq R(f) - R(f^*) = \frac{1}{n}E[\log p_{f^*}(Y) - \log p_f(Y)] \\
= \frac{1}{n}E \left[ \log \frac{p_{f^*}(Y)}{p_f(Y)} \right] \\
\equiv \frac{1}{n}K(p_f, p_{f^*})
\]

where

\[
K(p_f, p_{f^*}) = \sum_{i=1}^{n} \left( \int \log \frac{p_{f^*}(x_i)}{p_f(x_i)} \cdot \frac{p_{f^*}(y_i)}{p_f(y_i)} \cdot p_f(x_i) \cdot p_{f^*}(y_i) \, dy_i \right)
\]

\( K(p_f, p_{f^*}) \) is again the KL divergence.

Unfortunately, as mentioned before, \( K(p_f, p_{f^*}) \) is not a true distance. So instead we will focus on the expected squared Hellinger distance as our measure of performance:

\[
H^2(p_f, p_{f^*}) = \sum_{i=1}^{n} \int \left( \sqrt{p_f(x_i)(y_i)} - \sqrt{p_{f^*}(x_i)(y_i)} \right)^2 \, dy_i
\]

### 3 Maximum Complexity-Regularized Likelihood Estimation

**Theorem 1 (Li-Barron 2000, Kolaczyk-Nowak 2002)** Let \( \{x_i, Y_i\}_{i=1}^{n} \) be a random sample of training data with \( \{Y_i\} \) independent,

\[
Y_i \sim p_{f^*(x_i)}(y_i) , i = 1, \ldots, n
\]

for some unknown function \( f^* \).

Suppose we have a collection of candidate functions \( F \), and complexities \( c(f) > 0, f \in F \), satisfying

\[
\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1.
\]

Define the complexity-regularized estimator

\[
\hat{f}_n \equiv \arg \min_{f \in \mathcal{F}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log p_f(Y_i) + \frac{2c(f) \log 2}{n} \right\}.
\]

Then,

\[
\frac{1}{n}E \left[ H^2(p_{\hat{f}_n}, p_{f^*}) \right] \leq \frac{2}{n} \left[ \log \left( A(p_{\hat{f}_n}, p_{f^*}) \right) \right] \\
\leq \min_{f \in \mathcal{F}} \left\{ \frac{1}{n}K(p_f, p_{f^*}) + \frac{2c(f) \log 2}{n} \right\}.
\]

Before proving the theorem, let’s look at a special case.
Example 3 (Gaussian noise) Suppose $Y_i = f(x_i) + W_i$, $W_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

$p_{f(x_i)}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - f(x_i))^2}{2\sigma^2}}$

Using results from example 1, we have

$$-2 \log A\left(p_{f_n}(Y), p_{f^*}(Y)\right) = \sum_{i=1}^{n} -2 \log A\left(p_{f_n(x_i)}(Y_i), p_{f^*(x_i)}(Y_i)\right)$$

$$= \sum_{i=1}^{n} -2 \log \int p_{f_n(x_i)}(y_i) \cdot p_{f^*(x_i)}(y_i) \, dy_i$$

$$= \frac{1}{4\sigma^2} \sum_{i=1}^{n} \left(\hat{f}_n(x_i) - f^*(x_i)\right)^2.$$

Then,

$$-\frac{2}{n} E\left[\log A(p_{\hat{f}_n}, p_{f^*})\right] = \frac{1}{4\sigma^2 n} \sum_{i=1}^{n} E\left[\left(\hat{f}_n(x_i) - f^*(x_i)\right)^2\right].$$

We also have,

$$\frac{1}{n} K(p_f, p_{f^*}) = \frac{1}{n} \sum_{i=1}^{n} \frac{(f(x_i) - f^*(x_i))^2}{2\sigma^2}$$

$$-\log p_f(Y) = \sum_{i=1}^{n} \frac{(Y_i - f(x_i))^2}{2\sigma^2}.$$

Combine everything together to get

$$\hat{f}_n = \arg \min_{f \in F} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - f(x_i))^2}{2\sigma^2} + \frac{2c(f) \log 2}{n} \right\}.$$

The theorem tells us that

$$\frac{1}{4n} \sum_{i=1}^{n} E\left[\frac{(\hat{f}_n(x_i) - f^*(x_i))^2}{\sigma^2}\right] \leq \min_{f \in F} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(f(x_i) - f^*(x_i))^2}{2\sigma^2} + \frac{2c(f) \log 2}{n} \right\}$$

or

$$\frac{1}{n} \sum_{i=1}^{n} E\left[\left(\hat{f}_n(x_i) - f^*(x_i)\right)^2\right] \leq \min_{f \in F} \left\{ \frac{2}{n} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2 + \frac{8\sigma^2 c(f) \log 2}{n} \right\}.$$

Now let’s come back to the proof.

Proof:

$$H^2(p_{\hat{f}_n}, p_{f^*}) = \int \left(\sqrt{p_{\hat{f}_n}(y)} - \sqrt{p_{f^*}(y)}\right)^2 \, dy$$

$$\leq -2 \log \left(\int \sqrt{p_{\hat{f}_n}(y)} \cdot p_{f^*}(y) \, dy\right)_{\text{affinity}}$$
Now, define the theoretical analog of $\hat{f}_n$:

$$f_n = \arg\min_{f \in F} \left\{ -\frac{1}{n} \log p_f(Y) + \frac{2c(f) \log 2}{n} \right\}.$$

Since

$$\hat{f}_n = \arg\min_{f \in F} \left\{ -\frac{1}{n} \log p_f(Y) + \frac{2c(f) \log 2}{n} \right\}$$

$$= \arg\max_{f \in F} \left\{ \frac{1}{n} \log p_f(Y) - 2c(f) \log 2 \right\}$$

$$= \arg\max_{f \in F} \left\{ \frac{1}{2} \log p_f(Y) - 2c(f) \log 2 \right\}$$

$$= \arg\max_{f \in F} \left\{ \log \left( \frac{p_f(Y) \cdot e^{-c(f) \log 2}}{\sqrt{p_{f,n}(Y)}} \right) \right\}$$

we can see that

$$\frac{\sqrt{p_{f,n}(Y)} e^{-c(f_n) \log 2}}{\sqrt{p_f(Y)} e^{-c(f_n) \log 2}} \geq 1.$$

Then can write

$$E \left[ H^2 \left( p_{f,n}, p_{f^*} \right) \right] \leq 2E \left[ \log \left( \frac{1}{\int \sqrt{p_{f,n}(y)} \cdot p_{f^*}(y) \, dy} \right) \right]$$

$$\leq 2E \left[ \log \left( \frac{\sqrt{p_{f,n}(Y)} e^{-c(f_n) \log 2}}{\sqrt{p_f(Y)} e^{-c(f_n) \log 2} \cdot \int \sqrt{p_{f,n}(y)} \cdot p_{f^*}(y) \, dy} \right) \right].$$

Now, simply multiply the argument inside the log by $\sqrt{p_{f^*}(Y)} / p_{f,n}(Y)$ to get

$$E \left[ H^2 \left( p_{f,n}, p_{f^*} \right) \right] \leq 2E \left[ \log \left( \frac{\sqrt{p_{f^*}(Y)} \cdot \sqrt{p_{f,n}(Y)}}{\sqrt{p_f(Y)} \cdot \sqrt{p_{f,n}(Y)} e^{-c(f_n) \log 2} \cdot \int \sqrt{p_{f,n}(y)} \cdot p_{f^*}(y) \, dy} \right) \right]$$

$$= E \left[ \log \left( \frac{p_{f^*}(Y)}{p_{f,n}(Y)} \right) \right] + 2c(f_n) \log 2$$

$$\quad + 2E \left[ \log \left( \frac{\sqrt{p_{f,n}(Y)}}{\sqrt{p_{f^*}(Y)}} \cdot \frac{e^{-c(f_n) \log 2}}{\int \sqrt{p_{f,n}(y)} \cdot p_{f^*}(y) \, dy} \right) \right]$$

$$= K(p_{f,n}, p_{f^*}) + 2c(f_n) \log 2$$

$$\quad + 2E \left[ \log \left( \frac{\sqrt{p_{f,n}(Y)}}{\sqrt{p_{f^*}(Y)}} \cdot \frac{e^{-c(f_n) \log 2}}{\int \sqrt{p_{f,n}(y)} \cdot p_{f^*}(y) \, dy} \right) \right].$$
The terms $K(p_{f_n}, p_{f^*}) + 2c(f_n) \log 2$ are precisely what we wanted for the upper bound of the theorem. So, to finish the proof we only need to show that the last term is non-positive. Applying Jensen's inequality, we get

$$2E \left[ \log \left( \frac{p_{f_n}(Y)}{\sqrt{p_{f^*}(Y)}} \cdot \frac{e^{-c(f_n) \log 2}}{\int \sqrt{p_{f_n}(y) \cdot p_{f^*}(y)} \, dy} \right) \right] \leq 2 \log \left( E \left[ e^{-c(f_n) \log 2} \cdot \frac{\sqrt{p_{f_n}(Y)}}{\sqrt{p_{f^*}(Y)}} \right] \right).$$

Both $Y$ and $\hat{f}_n$ are random, which makes the expectation difficult to compute. However, we can simplify the problem using the union bound, which eliminates the dependence on $\hat{f}_n$:

$$2E \left[ \log \left( \frac{p_{f_n}(Y)}{\sqrt{p_{f^*}(Y)}} \cdot \frac{e^{-c(f_n) \log 2}}{\int \sqrt{p_{f_n}(y) \cdot p_{f^*}(y)} \, dy} \right) \right] \leq 2 \log \left( \sum_{f \in F} e^{-c(f) \log 2} \cdot \frac{\sqrt{p_f(Y)}}{\sqrt{p_{f^*}(Y)}} \right).$$

where the last two lines come from

$$E \left[ \frac{p_f(Y)}{\sqrt{p_{f^*}(Y)}} \right] = \int p_f(y) \cdot p_{f^*}(y) \, dy = \int \sqrt{p_f(y) \cdot p_{f^*}(y)} \, dy$$

and

$$\sum_{f \in F} 2^{-c(f)} \leq 1.$$