ECE 901
Lecture 14: Maximum Likelihood Estimation and Complexity Regularization

R. Nowak

5/17/2009

1 Review : Maximum Likelihood Estimation

We have n i.i.d observations drawn from an unknown distribution

$$Y_i \overset{i.i.d.}{\sim} p_{\theta^*} \quad , \; i = \{1, \ldots, n\}$$

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

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

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

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

The **risk** is defined as

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

As shown before, θ^* minimizes $R(\theta)$ over Θ.

$$\begin{aligned} \theta^* &= \arg\min_{\theta \in \Theta} \quad -E[\log p_\theta(Y)] \\ &= \arg\min_{\theta \in \Theta} \quad -\int \log p_\theta(y) \cdot p_{\theta^*}(y)\, \mathrm{d}y. \end{aligned}$$

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

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

We seen 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 the analysis of maximum likelihood estimation, namely *Hellinger Distance* and *Affinity*.

The **Hellinger distance** is defined as

$$H(p_{\theta_1}, p_{\theta_2}) = \left(\int \left(\sqrt{p_{\theta_1}(y)} - \sqrt{p_{\theta_2}(y)}\right)^2 \mathrm{d}y\right)^{\frac{1}{2}} .$$

We proved that the squared Hellinger distance lower bounds the KL divergence:

$$\begin{aligned} H^2(p_{\theta_1}, p_{\theta_2}) &\leq K(p_{\theta_1}, p_{\theta_2}) \\ H^2(p_{\theta_1}, p_{\theta_2}) &\leq K(p_{\theta_2}, p_{\theta_1}) . \end{aligned}$$

The **affinity** is defined as

$$A(p_{\theta_1}, p_{\theta_2}) = \int \sqrt{p_{\theta_1}(y) p_{\theta_2}(y)} \, \mathrm{d}y .$$

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 θ and variance σ^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,

$$\begin{aligned} K(p_{\theta_1}, p_{\theta_2}) &= E_{\theta_2}\left[\log \frac{p_{\theta_2}}{p_{\theta_1}}\right] \\ &= \frac{\theta_1^2 - \theta_2^2}{2\sigma^2} - \frac{2(\theta_1 - \theta_2)}{2\sigma^2} \underbrace{\int y \cdot p_{\theta_2}(y) \, \mathrm{d}y}_{E[Y] = \theta_2} \\ &= \frac{1}{2\sigma^2}(\theta_1^2 + \theta_2^2 - 2\theta_1\theta_2) \quad = \frac{(\theta_1^2 - \theta_2)^2}{2\sigma^2} . \\ -2 \log A(p_{\theta_1}, p_{\theta_2}) &= -2 \log \left(\int \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta_1)^2}{2\sigma^2}} \right)^{1/2} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta_2)^2}{2\sigma^2}} \right)^{1/2} \mathrm{d}y \right) \\ &= -2 \log \left(\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta_1)^2}{4\sigma^2} - \frac{(y-\theta_2)^2}{4\sigma^2}} \, \mathrm{d}y \right) \\ &= -2 \log \left(\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left[\left(y - \frac{\theta_1+\theta_2}{2}\right)^2 + \left(\frac{\theta_1-\theta_2}{2}\right)^2\right]} \, \mathrm{d}y \right) \\ &= -2 \log e^{-\frac{\left(\frac{\theta_1-\theta_2}{2}\right)^2}{2\sigma^2}} \\ &= \frac{(\theta_1 - \theta_2)^2}{4\sigma^2} \quad = \frac{1}{2} K(p_{\theta_1}, p_{\theta_2}) \quad \geq H^2(p_{\theta_1}, p_{\theta_2}) . \end{aligned}$$

2 Maximum likelihood estimation and Complexity regularization

Suppose that we have n i.i.d training samples, $\{X_i, Y_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} p_{XY}$.
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 $W_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ *and* $X_i \overset{i.i.d.}{\sim} p_X$.

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

$Y|X = x \sim \mathit{Poisson}(f^*(x))$

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

The **likelihood loss function** is

$$\begin{aligned} l(f(x), y) &= -\log\, p_{XY}(X, Y) \\ &= -\log\, p_X(X) - \log\, p_{Y|X}(Y|X) \\ &= -\log\, p_X(X) - \log\, p_{f(X)}(Y) \,. \end{aligned}$$

The *expected loss* is

$$\begin{aligned} E[l(f(X), Y)] &= E\left[E[l(f(X), Y)|X]\right] \\ &= E[\,E[-\log\, p_X(X) - \log\, p_{f(X)}(Y)|X]\,] \\ &= -E[\,\log\, p_X(X)\,] - E[\,E[\log\, p_{f(X)}(Y)|X\,]\,] \\ &= -E[\,\log\, p_X(X)\,] - E[\,\log\, p_{f(X)}(Y)\,] \,. \end{aligned}$$

Notice that the first term is a constant with respect to f. With that in mind we define our **risk** to be

$$\begin{aligned} R(f) &= -E[\,\log\, p_{f(X)}(Y)\,] \\ &= -E[\,E[\log\, p_{f(X)}(Y)|X]\,] \\ &= -\int \left(\int \log\, p_{f(x)}(y) \cdot p_{f^*(x)}(y)\, \mathrm{d}y \right) p_X(x)\, \mathrm{d}x \,. \end{aligned}$$

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:

$$\hat{R_n}(f) = \frac{1}{n} \sum_{i=1}^{n} -\log\, p_{f(X_i)}(Y_i)$$

We can regard the value $\frac{1}{n}$ as the *empirical* probability of observing $X = X_i$ (since we are making no assumptions on P_X).

3 Deterministic Designs

Often in function estimation, we have control over where we sample. For illustration let's assume that $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \mathbf{R}$. Suppose we have the samples deterministically distributed somewhat uniformly over the domain $\mathcal{X}$. 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 (hence the name *deterministic design*). Our **risk** is

$$\begin{aligned} R(f) &= -\frac{1}{n} \sum_{i=1}^{n} E\left[\log\, p_{f(x_i)}(Y_i)\right] \\ &= -\frac{1}{n} \sum_{i=1}^{n} \left[\int \log\, p_{f(x_i)}(y_i) \cdot p_{f^*(x_i)}(y_i)\, \mathrm{d}y_i\right]\ . \end{aligned}$$

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,

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

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

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

$$\begin{aligned} \hat{R_n}(f) &= -\frac{1}{n}\, \log p_f(Y)\, . \\ R(f) &= -\frac{1}{n} E[\log p_f(Y)] \\ &= -\frac{1}{n} \int \log p_f(y) \cdot p_{f^*}(y)\, \mathrm{d}y\, . \end{aligned}$$

4 Constructing Error Bounds

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 $f \in \mathcal{F}$ so that the excess risk is small.

$$\begin{aligned} 0 &\le R(f) - R(f^*) \\ &= \frac{1}{n} E[\log p_{f^*}(Y) - \log p_f(Y)] \\ &= \frac{1}{n} E\left[\log \frac{p_{f^*}(Y)}{p_f(Y)}\right] \\ &\equiv \frac{1}{n} K(p_f, p_{f^*}) \end{aligned}$$

where

$$K(p_f, p_{f^*}) = \sum_{i=1}^{n} \underbrace{\left(\int \log \frac{p_{f^*(x_i)}(y_i)}{p_{f(x_i)}(y_i)} \cdot p_{f^*(x_i)}(y_i)\, \mathrm{d}y_i \right)}_{K(p_{f(x_i)}, p_{f^*(x_i)})}$$

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:

$$H^2(p_f, p_{f^*}) \;=\; \sum_{i=1}^{n} \int \left(\sqrt{p_{f(x_i)}(y_i)} - \sqrt{p_{f^*(x_i)}(y_i)} \right)^2 \mathrm{d}y_i$$

5 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, where* $\{x_i\}$ *are deterministic and* $\{Y_i\}$ *are independent random variables, distributed as*

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

for some unknown function f^**. Suppose we have a collection of candidate functions*

$$\mathcal{F} \subseteq \{\textit{measurable functions } f : \mathcal{X} \to \mathcal{Y}\} \;\;,$$

and complexities $c(f) > 0, f \in \mathcal{F}$*, satisfying*

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

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,

$$\begin{aligned} \frac{1}{n} E\left[H^2(p_{\hat{f}_n}, p_{f^*})\right] &\le -\frac{2}{n} E\left[\log\left(A(p_{\hat{f}_n}, p_{f^*})\right)\right] \\ &\le \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} K(p_f, p_{f^*}) + \frac{2c(f)\log 2}{n} \right\} \;. \end{aligned}$$

Before proving the theorem, let's look at a very special and important case. We will use this results quite a lot in the following lectures.

Example 3 (Gaussian noise)**.** *Suppose* $Y_i = f(x_i) + W_i \quad, W_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

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

Using results from example 1, we have

$$\begin{aligned}
-2\log A\left(p_{\hat{f}_n}(Y), p_{f^*}(Y)\right) &= \sum_{i=1}^{n} -2\log A\left(p_{\hat{f}_n(x_i)}(Y_i), p_{f^*(x_i)}(Y_i)\right) \\
&= \sum_{i=1}^{n} -2\log \int \sqrt{p_{\hat{f}_n(x_i)}(y_i) \cdot p_{f^*(x_i)}(y_i)}\, \mathrm{d}y_i \\
&= \frac{1}{4\sigma^2} \sum_{i=1}^{n} \left(\hat{f}_n(x_i) - f^*(x_i)\right)^2 .
\end{aligned}$$

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,

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

and $-\log\, p_f(Y) = \sum_{i=1}^{n} \frac{(Y_i - f(x_i))^2}{2\sigma^2}$. *Combine everything together to get*

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \frac{8\sigma^2 c(f) \log 2}{n} \right\}.$$

The theorem tells us that

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

or

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

We will now prove Theorem 1.

Proof:

$$\begin{aligned}
H^2\left(p_{\hat{f}_n}, p_{f^*}\right) &= \int \left(\sqrt{p_{\hat{f}_n}(y)} - \sqrt{p_{f^*}(y)}\right)^2 \mathrm{d}y \\
&\le -2\log \underbrace{\left(\int \sqrt{p_{\hat{f}_n}(y) \cdot p_{f^*}(y)}\, \mathrm{d}y\right)}_{\text{affinity}=A(p_{\hat{f}_n}, p_{f^*})} .
\end{aligned}$$

Notice that $A(p_{\hat{f}_n}, p_{f^*})$ is a random quantity (it dependents of the training set through $\hat{f}_n$. Keeping that in mind the above result tells us that

$$E\left[H^2\left(p_{\hat{f}_n}, p_{f^*}\right)\right] \le 2\, E\left[\log\left(\frac{1}{A(p_{\hat{f}_n}, p_{f^*})}\right)\right].$$

Now, define the theoretical analog of $\hat{f}_n$:

$$f_n = \arg\min_{f\in\mathcal{F}} \left\{ \frac{1}{n} K\left(p_f, p_{f^*}\right) + \frac{2c(f)\log 2}{n} \right\} \; .$$

Now rewrite the definition of $\hat{f}_n$.

$$\begin{aligned}
\hat{f}_n &= \arg\min_{f\in\mathcal{F}} \left\{ -\frac{1}{n}\log p_f(Y) + \frac{2c(f)\log 2}{n} \right\} \\
&= \arg\max_{f\in\mathcal{F}} \left\{ \frac{1}{n}\left(\log p_f(Y) - 2c(f)\log 2\right) \right\} \\
&= \arg\max_{f\in\mathcal{F}} \left\{ \frac{1}{2}\left(\log p_f(Y) - 2c(f)\log 2\right) \right\} \\
&= \arg\max_{f\in\mathcal{F}} \left\{ \log\left(\sqrt{p_f(Y)} \cdot 2^{-c(f)} \right) \right\} \\
&= \arg\max_{f\in\mathcal{F}} \left\{ \sqrt{p_f(Y)} \cdot 2^{-c(f)} \right\} \; .
\end{aligned}$$

Since $\hat{f}_n$ is the function $f \in \mathcal{F}$ that maximizes $\sqrt{p_f(Y)} \cdot 2^{-c(f)}$ we conclude that

$$\frac{\sqrt{p_{\hat{f}_n}(Y)}2^{-c(\hat{f}_n)}}{\sqrt{p_{f_n}(Y)}2^{-c(f_n)}} \geq 1 \; .$$

Then can write

$$\begin{aligned}
E\left[H^2\left(p_{\hat{f}_n}, p_{f^*}\right)\right] &\leq 2\,E\left[\log\left(\frac{1}{A(p_{\hat{f}_n}, p_{f^*})}\right)\right] \\
&\leq 2\,E\left[\log\left(\frac{\sqrt{p_{\hat{f}_n}(Y)}2^{-c(\hat{f}_n)}}{\sqrt{p_{f_n}(Y)}2^{-c(f_n)}} \cdot \frac{1}{A(p_{\hat{f}_n}, p_{f^*})}\right)\right] \; .
\end{aligned}$$

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

$$\begin{aligned}
E\left[H^2\left(p_{\hat{f}_n}, p_{f^*}\right)\right] &\leq 2\,E\left[\log\left(\frac{\sqrt{p_{f^*}(Y)}}{\sqrt{p_{f_n}(Y)}}\frac{\sqrt{p_{\hat{f}_n}(Y)}}{\sqrt{p_{f^*}(Y)}}\frac{2^{-c(\hat{f}_n)}}{2^{-c(f_n)}} \cdot \frac{1}{A(p_{\hat{f}_n}, p_{f^*})}\right)\right] \\
&= E\left[\log\left(\frac{p_{f^*}(Y)}{p_{f_n}(Y)}\right)\right] + 2c(f_n)\log 2 \\
&\quad +2E\left[\log\left(\frac{\sqrt{p_{\hat{f}_n}(Y)}}{\sqrt{p_{f^*}(Y)}} \cdot \frac{2^{-c(\hat{f}_n)}}{A(p_{\hat{f}_n}, p_{f^*})}\right)\right] \\
&= K\left(p_{f_n}, p_{f^*}\right) + 2c(f_n)\log 2 \\
&\quad +2E\left[\log\left(\frac{\sqrt{p_{\hat{f}_n}(Y)}}{\sqrt{p_{f^*}(Y)}} \cdot \frac{2^{-c(\hat{f}_n)}}{A(p_{\hat{f}_n}, p_{f^*})}\right)\right] \; .
\end{aligned}$$

The terms $K\left(p_{f_n}, p_{f^*}\right) + 2c(f_n)\log 2$ 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 not-positive.

Applying Jensen's inequality, we get

$$2E\left[\log\left(\frac{\sqrt{p_{\hat{f}_n}(Y)}}{\sqrt{p_{f^*}(Y)}}\cdot\frac{2^{-c(\hat{f}_n)\log 2}}{A(p_{\hat{f}_n},p_{f^*})}\right)\right]\leq 2\log E\left[2^{-c(\hat{f}_n)}\cdot\frac{\sqrt{\frac{p_{\hat{f}_n}(Y)}{p_{f^*}(Y)}}}{A(p_{\hat{f}_n},p_{f^*})}\right]\ . \tag{1}$$

. Note that in the above expression the random quantities are Y and $\hat{f}_n$. Because $\hat{f}_n \in \mathcal{F}$ and we know something about this space we can use a union bound to get the randomness of $\hat{f}_n$ out of the way. In this case the union bound is simply saying that any individual term in a summation of positive terms is smaller than the summation[1]

$$\begin{aligned}
2E\left[\log\left(\frac{\sqrt{p_{\hat{f}_n}(Y)}}{\sqrt{p_{f^*}(Y)}}\cdot\frac{e^{-c(\hat{f}_n)\log 2}}{A(p_{\hat{f}_n},p_{f^*}}\right)\right] &\leq 2\log\left(E\left[\sum_{f\in\mathcal{F}} e^{-c(f)\log 2}\cdot\frac{\sqrt{\frac{p_f(Y)}{p_{f^*}(Y)}}}{A(p_f,p_{f^*}}\right]\right)\\
&= 2\log\left(\sum_{f\in\mathcal{F}} 2^{-c(f)}\frac{E\left[\sqrt{\frac{p_f(Y)}{p_{f^*}(Y)}}\right]}{A(p_f,p_{f^*}}\right)\\
&= 2\log\left(\sum_{f\in\mathcal{F}} 2^{-c(f)}\right)\\
&\leq 0\ ,
\end{aligned}$$

where the last steps of the proof follow from the fact that

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

and

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

■

[1]Let $z_1, z_2, \ldots$ be non-negative. Then for all $i \in \mathbb{N}$ $x_i \leq \sum_{j=1}^{\infty} x_j$