Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

25 Asymptotics of Inference and Universality

In these notes, we develop the asymptotics of inference. In particular, we examine the behavior of model capacity in the regime of a large number of samples, and identify the asymptotic least-informative prior.

Finally, we apply these results to the problem of universal inference, and sharply characterize when such inference is and is not possible.

25.1 Asymptotics of Model Capacity

To characterize the asymptotic behavior of the model capacity, we use the Bayesian terminology and analyze the mutual information between the hidden random variable x and the observed variable $\mathbf{y}$, which consists of N i.i.d. random variables distributed according to $p_{y|x}(\cdot; x)$. We assume that x is distributed according to the prior $p_{\mathsf{x}}(\cdot)$.

As the number of samples changes, the likelihood model of the full sequence $\mathbf{y}$ also changes. Naturally, we would expect the capacity of the model and the least informative prior to depend on the number of samples. However, this creates an inconvenience of having to change the least informative prior as the amount of available data changes. In practice, for large N, we could use as a prior a probability distribution that is equal to the limit of the least informative prior when $N \to \infty$. Such priors are called Jeffreys' priors. Our goal in this section is to characterize the dependency of the model capacity on the number of samples N and to determine the Jeffreys' prior for a broad class of problems.

We first represent the mutual information as the mean divergence between the likelihood model and the marginal distribution of $\mathbf{y}$,

$$I(\mathsf{x}; \mathbf{y}) = \mathbb{E}_{p_{\mathsf{x}}(\cdot)} \left[\mathbb{E}_{p_{\mathbf{y}|\mathsf{x}}(\cdot|\mathsf{x})} \left[\log \frac{p_{\mathbf{y},x}(\mathbf{y}, \mathsf{x})}{p_{\mathbf{y}}(\mathbf{y}) p_{\mathsf{x}}(\mathsf{x})} \middle| \mathsf{x} = x \right] \right] \tag{1}$$

$$= \mathbb{E}_{p_{\mathsf{x}}(\cdot)} \left[\mathbb{E}_{p_{\mathbf{y}|\mathsf{x}}(\cdot|\mathsf{x})} \left[\log \frac{p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|\mathsf{x})}{p_{\mathbf{y}}(\mathbf{y})} \middle| \mathsf{x} = x \right] \right] \tag{2}$$

$$= \mathbb{E}_{p_{\mathsf{x}}(\cdot)} \left[D(p_{\mathbf{y}|\mathsf{x}}(\cdot|\mathsf{x}) \| p_{\mathbf{y}}(\cdot)) \right] \tag{3}$$

and analyze the behavior of $D(p_{\mathbf{y}|\mathsf{x}}(\cdot|x) \| p_{\mathbf{y}}(\cdot))$ as $N \to \infty$.

25.1.1 Model Capacity for Finite Model Classes

If x is a random variable defined over a finite alphabet $\mathcal{X}$,

$$D(p_{\mathbf{y}|\mathsf{x}}(\cdot|x)\|p_{\mathbf{y}}(\cdot)) = \mathbb{E}_{p_{\mathbf{y}|\mathsf{x}}(\cdot|x)}\left[\log \frac{p_{\mathbf{y}|x}(\mathbf{y}|x)}{p_{\mathbf{y}}(\mathbf{y})}\right] \tag{4}$$

$$= \mathbb{E}_{p_{\mathbf{y}|\mathsf{x}}(\cdot|x)}\left[\log \frac{p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x)}{\sum_{x'} p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x')p_{\mathsf{x}}(x')}\right] \tag{5}$$

$$= \mathbb{E}_{p_{\mathbf{y}|\mathsf{x}}(\cdot|x)}\left[\log \frac{1}{p_{\mathsf{x}}(x) + \sum_{x'\neq x} p_{\mathsf{x}}(x')\frac{p_{\mathbf{y}|x}(\mathbf{y}|x')}{p_{\mathbf{y}|x}(\mathbf{y}|x)}}\right]. \tag{6}$$

Since $\mathcal{X}$ is finite, the sum in the denominator contains a finite number of terms. For any sequence $\mathbf{y}$ generated i.i.d. from the distribution $p_{y|x}(\cdot|x)$, each term in the sum decays exponentially with the number of samples N:

$$\frac{p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x')}{p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x)} \doteq e^{-ND(p_{y|\mathsf{x}}(\cdot|x')\|p_{y|x}(\cdot|x))}. \tag{7}$$

Therefore, the whole sum decays exponentially. For large N,

$$D(p_{\mathbf{y}|x}(\cdot|x)\|p_{\mathbf{y}}) \cong \mathbb{E}_{p_{\mathbf{y}|\mathsf{x}}(\cdot|x)}\left[\log \frac{1}{p_{\mathsf{x}}(x)}\right] = -\log p_{\mathsf{x}}(x) \tag{8}$$

and

$$I(\mathsf{x};\mathbf{y}) = \mathbb{E}_{p_{\mathsf{x}}(\cdot)}\left[D(p_{\mathbf{y}|\mathsf{x}}(\cdot|\mathsf{x})\|p_{\mathbf{y}}(\cdot))\right] \cong \mathbb{E}_{p_{\mathsf{x}}(\cdot)}\left[-\log p_{\mathsf{x}}(\mathsf{x})\right] = H(p_{\mathsf{x}}). \tag{9}$$

The model capacity is equal to

$$C_N = \max_{p_{\mathsf{x}}} I(\mathsf{x};\mathbf{y}) \cong \max_{p_{\mathsf{x}}} H(p_{\mathsf{x}}) = \log|\mathcal{X}| \tag{10}$$

and is achieved for $p_{\mathsf{x}}(x) = |\mathcal{X}|^{-1}$. Since neither the model capacity nor the prior that achieves capacity depends on N, this is also the Jeffreys' prior:

$$p_{\mathsf{x}}^{\infty}(x) = \frac{1}{|\mathcal{X}|}. \tag{11}$$

25.1.2 Model Capacity for Continuous Model Classes

The analysis for a continuous parameter x is a bit more involved. We present it here for completeness, but note that the approximations taken in the derivations require particular care in specifying the conditions on the likelihood model to ensure that such approximations are valid.

We start by manipulating the log ratio of the likelihood and the marginal distribution of $\mathbf{y}$:

$$D(p_{\mathbf{y}|x}(\cdot|x)\|p_{\mathbf{y}}(\cdot)) = \mathbb{E}_{p_{\mathbf{y}|x}(\cdot|x)}\left[\log\frac{p_{\mathbf{y}|x}(\mathbf{y}|x)}{p_{\mathbf{y}}(\mathbf{y})}\right] \tag{12}$$

$$= \mathbb{E}_{p_{\mathbf{y}|x}(\cdot|x)}\left[\log\frac{p_{\mathbf{y}|x}(\mathbf{y}|x)}{p_{\mathbf{y}|x}(\mathbf{y}|\hat{x}_N)}\right] + \mathbb{E}_{p_{\mathbf{y}|x}(\cdot|x)}\left[\log\frac{p_{\mathbf{y}|x}(\mathbf{y}|\hat{x}_N)}{p_{\mathbf{y}}(\mathbf{y})}\right] \tag{13}$$

where $\hat{x}_N$ is the ML estimate of the parameter x based on the observations $\mathbf{y}$.

Using the Taylor series expansion of the log-likelihood function around the ML estimate $\hat{x}_N$, we obtain

$$\frac{1}{N}\log p_{\mathbf{y}|x}(\mathbf{y}|x) = \frac{1}{N}\log p_{\mathbf{y}|x}(\mathbf{y}|\hat{x}_N) + \frac{1}{N}(x-\hat{x}_N)\left.\frac{\partial \log p_{\mathbf{y}|x}(\mathbf{y}|x)}{\partial x}\right|_{\hat{x}_N} \tag{14}$$

$$+ \frac{1}{2N}(x-\hat{x}_N)^2\left.\frac{\partial^2 \log p_{\mathbf{y}|x}(\mathbf{y}|x)}{\partial x^2}\right|_{\hat{x}_N} + o((x-\hat{x}_N)^2) \tag{15}$$

$$= \frac{1}{N}\log p_{\mathbf{y}|x}(\mathbf{y}|\hat{x}_N) \tag{16}$$

$$+ \frac{1}{2N}(x-\hat{x}_N)^2\sum_{n=1}^{N}\left.\frac{\partial^2 \log p_{y|x}(y_n|x)}{\partial x^2}\right|_{\hat{x}_N} + o((x-\hat{x}_N)^2), \tag{17}$$

which implies

$$\frac{1}{N}\log\frac{p_{\mathbf{y}|x}(\mathbf{y}|x)}{p_{\mathbf{y}|x}(\mathbf{y}|\hat{x}_N)} = \frac{1}{2}(x-\hat{x}_N)^2\frac{1}{N}\sum_{n=1}^{N}\left.\frac{\partial^2 \log p_{y|x}(y_n|x)}{\partial x^2}\right|_{\hat{x}_N} + o((x-\hat{x}_N)^2). \tag{18}$$

Since

$$\hat{x}_N \xrightarrow{\text{w.p.1}} x \tag{19}$$

$$\frac{1}{N}\sum_{n=1}^{N}\left.\frac{\partial^2 \log p_{y|x}(y_n|x)}{\partial x^2}\right|_{\hat{x}_N} \xrightarrow{\text{w.p.1}} \mathbb{E}_{p_{y|x}(\cdot|x)}\left[\left.\frac{\partial^2 \log p_{y|x}(y|x)}{\partial x^2}\right|_{\hat{x}_N}\right] = -J_y(\hat{x}_N), \tag{20}$$

for sufficiently large N

$$\frac{1}{N}\log\frac{p_{\mathbf{y}|x}(\mathbf{y}|x)}{p_{\mathbf{y}|x}(\mathbf{y}|\hat{x}_N)} \cong -\frac{1}{2}J_y(\hat{x}_N)(x-\hat{x}_N)^2. \tag{21}$$

This implies that the first term in the right-hand side of (13) is equal to

$$\mathbb{E}_{p_{\mathbf{y}|x}(\cdot|x)}\left[\log\frac{p_{\mathbf{y}|x}(\mathbf{y}|x)}{p_{\mathbf{y}|x}(\mathbf{y}|\hat{x}_N)}\right] \cong \mathbb{E}_{p_{\mathbf{y}|x}(\cdot|x)}\left[-\frac{1}{2}(NJ_y(\hat{x}_N))(x-\hat{x}_N)^2\right] \tag{22}$$

$$\cong \mathbb{E}_{p_{\mathbf{y}|x}(\cdot|x)}\left[-\frac{1}{2}(NJ_y(x))(x-\hat{x}_N)^2\right] \tag{23}$$

$$\cong -\frac{1}{2}(NJ_y(x))\operatorname{var}(\hat{x}_N) \cong -\frac{1}{2}. \tag{24}$$

We can also use (21) to show that

$$\frac{1}{N}\log\frac{p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})}{p_{\mathsf{x}|\mathbf{y}}(\hat{x}_N|\mathbf{y})} = \frac{1}{N}\log\frac{p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x)}{p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|\hat{x}_N)} + \frac{1}{N}\log\frac{p_{\mathsf{x}}(x)}{p_{\mathsf{x}}(\hat{x}_N)} \tag{25}$$

$$\cong -\frac{1}{2}J_y(\hat{x}_N)(x-\hat{x}_N)^2 + \frac{1}{N}\log\frac{p_{\mathsf{x}}(x)}{p_{\mathsf{x}}(\hat{x}_N)}. \tag{26}$$

As $N \to \infty$, the effect of the second term is clearly removed and

$$\frac{1}{N}\log\frac{p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})}{p_{\mathsf{x}|\mathbf{y}}(\hat{x}_N|\mathbf{y})} \cong -\frac{1}{2}J_y(\hat{x}_N)(x-\hat{x}_N)^2. \tag{27}$$

By exponentiating both sides of the equation, we realize that

$$p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y}) \propto e^{-\frac{1}{2}NJ_y(\hat{x}_N)(x-\hat{x}_N)^2}, \tag{28}$$

i.e., as $N \to \infty$, the posterior probability distribution of the hidden parameter x can be approximated by a Gaussian distribution with mean $\hat{x}_N$ and variance $(NJ_y(\hat{x}_N))^{-1}$:

$$p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y}) = \frac{(NJ_y(\hat{x}_N))^{1/2}}{(2\pi)^{1/2}}e^{-\frac{1}{2}NJ_y(\hat{x}_N)(x-\hat{x}_N)^2}. \tag{29}$$

Now we analyze the asymptotic behavior of the second term in (13):

$$\log\frac{p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|\hat{x}_N)}{p_{\mathbf{y}}(\mathbf{y})} = \log\frac{p_{\mathsf{x}|\mathbf{y}}(\hat{x}_N|\mathbf{y})}{p_{\mathsf{x}}(\hat{x}_N)} \tag{30}$$

$$\cong \frac{1}{2}\log\frac{N}{2\pi} + \log\sqrt{J_y(\hat{x}_N)} - \log p_{\mathsf{x}}(\hat{x}_N) \tag{31}$$

$$\cong \frac{1}{2}\log\frac{N}{2\pi} + \log\frac{\sqrt{J_y(x)}}{p_{\mathsf{x}}(x)}. \tag{32}$$

We made use of $\hat{\mathsf{x}}_N \xrightarrow{\text{w.p.1}} x$ to arrive at the last approximation. Substituting into (13), we obtain

$$D(p_{\mathbf{y}|\mathsf{x}}(\cdot|x)\|p_{\mathbf{y}}(\cdot)) \cong -\frac{1}{2} + \frac{1}{2}\log\frac{N}{2\pi} + \log\frac{\sqrt{J_y(x)}}{p_{\mathsf{x}}(x)} \tag{33}$$

$$= \frac{1}{2}\log\frac{N}{2\pi e} + \log\frac{\sqrt{J_y(x)}}{p_{\mathsf{x}}(x)}. \tag{34}$$

Here we assume natural logarithms.

To compute the model capacity, we evaluate

$$I(\mathsf{x};\mathbf{y}) = \mathbb{E}_{p_{\mathsf{x}}(\cdot)}\left[D(p_{\mathbf{y}|\mathsf{x}}(\cdot|\mathsf{x})\|p_{\mathbf{y}}(\cdot))\right] \cong \frac{1}{2}\log\frac{N}{2\pi e} + \mathbb{E}_{p_{\mathsf{x}}(\cdot)}\left[\log\frac{\sqrt{J_y(x)}}{p_{\mathsf{x}}(x)}\right] \tag{35}$$

$$= \frac{1}{2}\log\frac{N}{2\pi e} - D(p_{\mathsf{x}}(\cdot)\|q(\cdot)), \tag{36}$$

where

$$q(x) = \frac{\sqrt{J_y(x)}}{\int_x \sqrt{J_y(x)}dx}. \tag{37}$$

The model capacity is equal to

$$C_N = \max_{p_x} I(x; \mathbf{y}) \cong \frac{1}{2} \log \frac{N}{2\pi e}. \tag{38}$$

The Jeffreys' prior is equal to $q(\cdot)$:

$$p_x^{\infty}(x) = \frac{\sqrt{J_y(x)}}{\int_x \sqrt{J_y(x)}\, dx}. \tag{39}$$

We note that $p_x^{\infty}(x)$ might be an improper prior if the integral of $\sqrt{J_y(x)}$ does not exist.

25.1.3 Vector Parameters

We can also extend the analysis to the case of a vector parameter $\mathbf{x}$ whose dimensionality is k. The derivation uses the Fisher information matrix in place of the Fisher information function. It is easy to show that (21) becomes

$$\frac{1}{N} \log \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\hat{\mathbf{x}}_N)} \cong -\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}}_N)^{\mathrm{T}} J_y(\hat{\mathbf{x}}_N)(\mathbf{x} - \hat{\mathbf{x}}_N). \tag{40}$$

Since the covariance of the ML estimator is equal to $(NJ_y(\mathbf{x}))^{-1}$, the first term in the right-hand side of (13) becomes

$$\mathbb{E}_{p_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x})} \left[\log \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\hat{\mathbf{x}}_N)} \right] \cong \mathbb{E}_{p_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x})} \left[-\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}}_N)^{\mathrm{T}} (NJ_y(\mathbf{x}))(\mathbf{x} - \hat{\mathbf{x}}_N) \right] = -\frac{k}{2}. \tag{41}$$

Similarly,

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \cong \frac{(N)^{k/2}}{(2\pi)^{k/2}} (\det(J_y(\hat{\mathbf{x}}_N)))^{1/2} e^{-\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}}_N)^{\mathrm{T}} (NJ_y(\hat{\mathbf{x}}_N))(\mathbf{x} - \hat{\mathbf{x}}_N)} \tag{42}$$

is a multivariate Gaussian distribution. The second term in the right-hand side of (13) therefore becomes

$$\mathbb{E}_{p_{\mathbf{x}}(\cdot)} \left[\log \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\hat{\mathbf{x}}_N)}{p_{\mathbf{y}}(\mathbf{y})} \right] \cong \frac{k}{2} \log \frac{N}{2\pi} + \log \frac{\sqrt{\det(J_y(\mathbf{x}))}}{p_{\mathbf{x}}(\mathbf{x})}. \tag{43}$$

Combining the two terms and maximizing, we arrive at the model capacity:

$$C_N = \max_{p_{\mathbf{x}}} I(\mathbf{x}; \mathbf{y}) \cong \frac{k}{2} \log \frac{N}{2\pi e}, \tag{44}$$

and the Jeffreys' prior:

$$p_{\mathbf{x}}^{\infty}(x) = \frac{\sqrt{\det(J_y(\mathbf{x}))}}{\int_{\mathbf{x}} \sqrt{\det(J_y(\mathbf{x}))}\, d\mathbf{x}}. \tag{45}$$

25.2 Asymptotics of Inference: Universal Inference

The concept of *universal inference* reflects a notion that after we observe a sufficient amount of data, we could perform inference and make predictions about future observations as well as if we knew the true probability distribution that generates the data.

A formal definition is as follows.

Definition 1 (Universal Predictor)**.** *Let $p_{\mathsf{y}}^N(\cdot; x)$ be an observation model, with $x \in \mathcal{X}$ denoting an unknown parameter. Let $q_1, q_2, \ldots, q_N$ be a sequence of predictors, where $q_n(\cdot; y^{n-1})$ is an estimate of the distribution for y_n based on $\mathsf{y}^{n-1} = y^{n-1} = (y_1, \ldots, y_{n-1})$. Then this sequence of predictors is* universal *if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{p_{\mathsf{y}}^N(\cdot;x)} \left[\log \frac{p_{\mathsf{y}_n|\mathsf{y}^{n-1}}(y_n|y^{n-1}; x)}{q_n(y_n; y^{n-1})} \right] = 0, \quad \textit{for all } x \in \mathcal{X}, \tag{46}$$

where

$$p_{\mathsf{y}_n|\mathsf{y}^{n-1}}(y_n|y^{n-1}; x) = \frac{p_{\mathsf{y}^n}(y^n; x)}{p_{\mathsf{y}^{n-1}}(y^{n-1}; x)}. \tag{47}$$

Notice that the expectation in (46) corresponds to the approximation loss for the nth predictor, and that this loss is averaged over the N predictors, so any initial transient behavior has negligible effect with such a measure. Note too that when the model is i.i.d., i.e.,

$$p_{\mathsf{y}}^N(\mathbf{y}; x) = p_{\mathsf{y}^N}(y^N; x) = \prod_{n=1}^{N} p_{\mathsf{y}}(y_n; x), \tag{48}$$

then the numerator in the logarithm in (46) is simply $p_{\mathsf{y}}(y_n; x)$.

Given this definition of universality, the following theorem characterizes when universal inference is possible. We state the theorem for the case of i.i.d. models, but it can be shown to hold more generally.

Theorem 1 (Universal Prediction Theorem)**.** *For an i.i.d. source $\mathbf{y} = \mathsf{y}^N = (\mathsf{y}_1, \ldots, \mathsf{y}_N)$ whose distribution depends on an unknown parameter x, universal inference is possible if and only if*

$$\lim_{N\to\infty} \frac{C_N}{N} = 0, \tag{49}$$

where C_N is the capacity of the parameterized model.

Proof. We begin by showing the "if" part of the theorem. In particular, we show by construction that when $C_N/N \to 0$ for all $x \in \mathcal{X}$, there exists a universal predictor.

First, let $p_{\mathsf{y}^n}(\cdot)$ denote the mixture distribution for the (partial) observations y^n that is induced by the least informative prior $p_*(\cdot)$ over $\mathcal{X}$, i.e.,

$$p_{\mathsf{y}^n}(y^n) = \sum_{x\in\mathcal{X}} p_*(x) \prod_{i=1}^{n} p_{\mathsf{y}}(y_i; x), \tag{50}$$

and, in turn, let

$$\delta(x) = \frac{1}{N} D(p_{\mathbf{y}}^N(\cdot; x) \| p_{\mathbf{y}}^N(\cdot)), \tag{51}$$

where $p_{\mathbf{y}}^N = p_{y^N}$.

Now

$$\begin{aligned}
\frac{C_N}{N} &= \frac{1}{N} \min_{q^N(\mathbf{y})} \max_{x \in \mathcal{X}} D(p_{\mathbf{y}}^N(\cdot; x) \| q^N(\cdot)) && (52)\\
&= \frac{1}{N} \max_{x \in \mathcal{X}} D(p_{\mathbf{y}}^N(\cdot; x) \| p_{\mathbf{y}}^N(\cdot)) && (53)\\
&= \frac{1}{N} D(p_{\mathbf{y}}^N(\cdot; x) \| p_{\mathbf{y}}^N(\cdot)) && (54)\\
&= \max_{x \in \mathcal{X}} \delta(x), && (55)
\end{aligned}$$

where to obtain (52) we have used the definition of C_N as the minimax redundancy, where to obtain (53) we have used the redundancy-capacity theorem to allow interchanging the order of minimization and maximization, and the fact that the minimization is achieved by the mixture model derived from the least informative prior, and where in (55) we have used (51).

Next, note that the mixture-predictor

$$q_n(y_n; y^{n-1}) = p_{y_n|y^{n-1}}(y_n|y^{n-1}) = \frac{p_{y^n}(y^n)}{p_{y^{n-1}}(y^{n-1})}, \tag{56}$$

derived from (50) has average prediction loss given by

$$\begin{aligned}
\delta_-(x) &= \frac{1}{N} \mathbb{E}_{p_{\mathbf{y}}^N(\cdot;x)} \left[\log \frac{p_{\mathbf{y}}^N(\mathbf{y}; x)}{p_{\mathbf{y}}^N(\mathbf{y})} \right] && (57)\\
&= \frac{1}{N} D(p_{\mathbf{y}}^N(\cdot; x) \| p_{\mathbf{y}}^N(\cdot)). && (58)
\end{aligned}$$

Thus, since for all $x \in \mathcal{X}$,

$$\frac{C_N}{N} = \max_{x \in \mathcal{X}} \delta(x) \geq \delta_-(x),$$

it follows that

$$\lim_{N \to \infty} \frac{C_N}{N} = 0 \quad \Rightarrow \quad \lim_{N \to \infty} \delta_-(x) = 0 \text{ for all } x \in \mathcal{X},$$

and thus the mixture-predictor derived from the least-informative prior is universal.

We now show the "only if" part of the theorem. In particular, we show that if a universal predictor exists, then $C_N/N \to 0$ as $N \to \infty$.

First, note that

$$\frac{C_N}{N} = \frac{1}{N} \min_{q^N} \max_{x \in \mathcal{X}} D(p_{\mathbf{y}}^N(\cdot; x) \| q^N(\cdot)) \tag{59}$$

$$\leq \max_{x \in \mathcal{X}} \frac{1}{N} D(p_{\mathbf{y}}^N(\cdot; x) \| q^N(\cdot)) \tag{60}$$

$$= \frac{1}{N} D(p_{\mathbf{y}}^N(\cdot; x_*) \| q^N(\cdot)) \tag{61}$$

where (60) holds for any particular choice of q^N, and where in (61) x_* is the maximizing x in (60) for the particular choice of q^N.

Now let $q_n(y_n; y^{n-1})$ for $n = 1, \ldots, N$ denote our universal predictor, and note that its average loss $\delta_+(x)$ satisfies

$$\delta_+(x) \triangleq \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{p_{\mathbf{y}}^N(\cdot;x)} \left[\log \frac{p_y(y_n; x)}{q_n(y_n; y^{n-1})} \right] \tag{62}$$

$$= \frac{1}{N} \sum_{n=1}^{N} D(p_{\mathbf{y}}^N(\cdot; x) \| q_{\mathbf{y}}^N(\mathbf{y})), \tag{63}$$

where in (63) we have

$$q_{\mathbf{y}}^N(\mathbf{y}) \triangleq \prod_{n=1}^{N} q_n(y_n; y^{n-1}). \tag{64}$$

Note that if each of the predictors are valid distributions (nonnegative and integrate to unity), (64) is also a valid distribution (nonnegative and integrates to unity), which can be verified by successive marginalization.

Finally, using (61) with (63) we see that

$$\max_{x \in \mathcal{X}} \delta_+(x) = \delta_+(x_*) \geq \frac{C_N}{N},$$

and thus

$$\lim_{N \to \infty} \delta_+(x) = 0 \text{ for all } x \in \mathcal{X} \quad \Rightarrow \quad \lim_{N \to \infty} \delta_+(x_*) = 0 \quad \Rightarrow \quad \lim_{N \to \infty} \frac{C_N}{N} = 0. \tag{65}$$

□

Some final comments. From our analysis of the model capacity, the condition of the theorem implies a condition of the distances from the likelihood models in the class and the marginal (mixture) probability distribution. We associate universal inference with those distances growing slower than the number of observations.

In both cases we analyzed in the previous section, the condition is satisfied, therefore we expect universal inference in the limit. But what is the data is generated by a distribution $q_y(\cdot)$ that is not in the likelihood family? In this case, we can show that $D(q_{\mathbf{y}}(\cdot) \| p_{\mathbf{y}}(\cdot))$ grows linearly with N, implying that universal inference cannot be achieved. This again highlights the importance of choosing the model class that is large enough to include the true distribution.

25.3 Further Reading

Clarke & Barron. *Information-Theoretic Asymptotics of Bayes Methods*. In IEEE Transactions on Information Theory. This paper reviews and proves the relevant results on asymptotic behavior of model capacity.

Merhav & Feder. Universal Prediction. In IEEE Transactions on Information Theory. This paper discusses universal prediction.