Self-information (log-loss)

$\ell(b,x) = -\log b(x)$ is the ideal codeword length of x with probability function $b(\cdot)$ in lossless source coding

Also in gambling, $b_t(\cdot|x^{t-1})$ represents distribution of money invested in each of the possible values of next outcome. The log-loss then dictates exponential growth rate of money with time.

Observe that any sequential probability assignment mechanism yields probability assignment for entire observation vector x^n as

$$Q(x^n) = \prod_{t=1}^{n} b_t(x_t|x^{t-1})$$

and any consistent $Q(x^n)$ provides valid sequential probability assignment as:

$$b_t(x_t|x^{t-1}) = \frac{Q(x^t)}{Q(x^{t-1})}$$

So log-loss is completely equivalent to choice of Q that assigns maximum probability to x^n, i.e. maximum likelihood estimation

Focus on log-loss

Probabilistic setting for universal ~~compressi~~ prediction under log-loss is equivalent to finding probability assignment Q for entire sequence.

If P is known, optimal Q is just $Q=P$, so induced prediction is

$$b_t(\cdot|x^{t-1}) = Q(\cdot|x^{t-1}) = P(\cdot|x^{t-1})$$

Average cumulative loss in then $H_n(P) = -E[\log P(X^n)]$, the entropy where

$$w(\theta|x^{t-1}) = \frac{w(\theta) P_\theta(x^{t-1})}{\int_\Lambda dw(\theta') P_{\theta'}(x^{t-1})} = \frac{w(\theta)\, 2^{-\log 1/P_\theta(x^{t-1})}}{\int_\Lambda dw(\theta') P_{\theta'}(x^{t-1})}$$

exponential weighting view.

How to choose the weight function $w(\cdot)$ of the mixture Q_w?

minimax problem:

$$\inf_Q \sup_{\theta\in\Lambda} D_n(P_\theta \| Q) = \inf_Q \sup_w \int_\Lambda dw(\theta)\, D_n(P_\theta\|Q).$$

n-normalized quantity is minimax redundancy.

Alternate is ~~average~~ maximin redundancy rather than minimax; here judge by weighted average

$$R_n(Q,w) = \frac{1}{n}\int_\Lambda dw(\theta)\, D_n(P_\theta\|Q).$$

The Q that minimizes this for a given w is the Q we had before

$Q_w = \int_\Lambda dw(\theta) P_\theta(x^n)$ and the resultant average redundancy $R_n(Q_w, w)$

is the mutual information $I_w(\Theta; X^n)$ between random variables Θ and X^n

with joint density $\mu(\theta, x^n) = w(\theta) P_\theta(x^n)$.

But what is good choice of w?

$$\sup_w \inf_Q R_n(Q,w).$$

whose n-normalized value is maximin redundancy

One can think of $\sup_w I_w(\Theta; X^n)$ as capacity of channel from Θ to X^n.

In fact the minimax- and maximin solutions are equivalent

The mixture Q_{w^*} where w^* is capacity-achieving prior is what to use.

This is redundancy-capacity theorem of universal coding (and of prediction)

What about very large classes of sources?

For example Stationary and ergodic sources.

These classes are so rich that under log-loss, for every finite n, and every predictive probability assignment Q, there exists a source in the class such that

$$E[-\log Q(x^n)] \geq n \log A - o(n)$$ when A is alphabet size.

Aim for weakly universal schemes instead, not requiring uniform redundancy rates

For example based on predictive probability assignments induced by Lempel-Ziv algorithm.