Deterministic Setting of Universal Prediction

Last time, we assumed a probabilistic setting where data is generated according to some underlying statistical mechanism, e.g. Markov source or stationary, ergodic source, i.e. [estimate entire either explicitly (plug-in approach) or implicitly (mixture approach) the conditional probability of next outcome given past, and then use.

Deterministic setting of individual sequences

→ no ensemble of sequences, rather just one arbitrary, deterministic individual sequence.

→ formally, for any sequence, there is a perfect predictor that has zero loss for the particular sequence, but might be overly bad for any other sequence

→ to avoid such overfitting, constrain the choice of the predictor.

Limited class of allowable predictors, the comparison class, denoted $\mathcal{B}$.

Want a single universal predictor b_t^* that competes with best predictor in $\mathcal{B}$, simultaneously for every x^n, i.e. want

$$\frac{1}{n}\sum_{t=1}^{n} l(b_t^*, x_t) \quad \text{to be asymptotically same as} \quad \min_{\mathcal{B}} \frac{1}{n}\sum_{t=1}^{n} l(b_t, x_t)$$

The universal predictor need not be in $\mathcal{B}$ but must be the same predictor for every x^n. The choice of reference predictor in $\mathcal{B}$ depends on entire sequence x^n.

The difference between sequential universal predictor and best predictor in $\mathcal{B}$ is the regret.

To do so, minimize Q over

$$\max_{x^n} \frac{1}{n}\left[-\log Q(x^n) - \left(-\log \max_{\theta} \prod_{t=1}^{n} b^{\theta}(x_t|x^{t-1})\right)\right]$$

Specifically, minimax-optimal probability assignment is achieved by normalized maximum-likelihood function

$$Q_n^*(x^n) = \frac{1}{K_n} \max_{\theta} \prod_{t=1}^{n} b^{\theta}(x_t|x^{t-1})$$

where $K_n = \sum_{x^n} \max_{\theta} \prod_{t=1}^{n} b^{\theta}(x_t|x^{t-1})$.

Notice by definition of Q_n^* that

$$-\frac{1}{n}\log Q_n^*(x^n) = -\frac{1}{n}\log \max_{\theta} \prod_{t=1}^{n} b^{\theta}(x_t|x^{t-1}) + \frac{1}{n}\log K_n.$$

So Q_n^* essentially assigns uniformly as high probabilities as those assigned by best member in comparison class, if K_n doesn't grow exponentially with n.

$\Gamma_n = \frac{1}{n}\log K_n$ is interpreted as deterministic analog to minimax redundancy-capacity C_n from probabilistic setting. where maximization of redundancy over θ is replaced by maximization over all possible sequences x^n.

Γ_n is never smaller than C_n for some class of sources or probability assignments indexed by n.

Necessary condition for existence of minimax universality in deterministic setting is existence of parallel property in dual probabilistic setting.

Very Large Comparison Class

One approach: universal probability measure proportional to $2^{-LZ(x^n)}$ where $LZ(x^n)$ is LZ codeword length for x^n.

extension to ML probability assignment since $2^{-LZ(x^n)}$ upper bounds $\max_P P(x^n)$ over all finite-state sources with fixed number of states