$P_{\mathcal{M}_e}$ can be proven to be invariant, again up to a small multiplicative constant. For a proof of this, as well as further powerful properties of the universal prior distribution, see the paper that this section is based on (Hutter, 2007a).

2.7 Solomonoff induction

Given a prior distribution ξ over $\mathbb{B}^\infty$, it is straightforward to predict the continuation of a binary sequence using the same approach as we used in Section 2.3. Given prior distribution ξ and the observed string $\omega_{1:t} \in \mathbb{B}^\infty$ from a sequence $\omega \in \mathbb{B}^\infty$ that has been sampled from an unknown computable distribution $\mu \in \mathcal{M}_c$, our estimate of the probability that the next bit will be 0 is,

$$\xi(\omega_{1:t}\underline{0}) = \frac{\xi(\underline{\omega_{1:t}0})}{\xi(\underline{\omega}_{1:t})}.$$

Is this predictor based on ξ any good? By definition, the best possible predictor would be based on the unknown true distribution μ that ω has been sampled from. That is, the true probability that the next bit is a 0 given an observed initial string $\omega_{1:t}$ is,

$$\mu(\omega_{1:t}\underline{0}) = \frac{\mu(\underline{\omega_{1:t}0})}{\mu(\underline{\omega}_{1:t})}.$$

As this predictor is optimal by construction, it can be used to quantify the relative performance of the predictor based on ξ. For example, consider the expected squared error in the estimated probability that the t^{th} bit will be a 0:

$$S_t = \sum_{x \in \mathbb{B}^{t-1}} \mu(\underline{x})\big(\xi(x\underline{0}) - \mu(x\underline{0})\big)^2.$$

If ξ is a good predictor, then its predictions should be close to those made by the optimal predictor μ, and thus S_t will be small.

Solomonoff (1978) was able to prove the following remarkable convergence theorem:

2.7.1 Theorem. *For any computable probability measure $\mu \in \mathcal{M}_c$,*

$$\sum_{t=1}^{\infty} S_t \;\leq\; \frac{\ln 2}{2}\, K(\mu).$$

That is, the total of *all* the prediction errors over the length of the infinite sequence ω is bounded by a constant. This implies rapid convergence for *any* unknown hypothesis that can be described by a computable distribution (for a precise analysis see Hutter, 2007a). This set includes all computable hypotheses over binary strings, which is essentially the set of all well defined hypotheses. If it were not for the fact that the universal prior ξ is not computable, Solomonoff induction would be the ultimate all purpose universal predictor.

Although we will not present Solomonoff's proof, the following highlights the key step required to obtaining the convergence result. For any probability measure μ the following relation can be proven,

$$\sum_{t=1}^{n} S_t \;\leq\; \frac{1}{2} \sum_{x \in \mathbb{B}^n} \mu(\underline{x}) \ln \frac{\mu(\underline{x})}{\xi(\underline{x})}.$$

This in fact holds for any semi-measure ξ, thus no special properties of the universal distribution have been used up to this point in the proof. Now, by the universal dominance property of ξ, we know that $\forall x \in \mathbb{B}^* : \xi(\underline{x}) \geq 2^{-K(\mu)} \mu(\underline{x})$. Substituting this into the above equation,

$$\sum_{t=1}^{n} S_t \;\leq\; \frac{1}{2} \sum_{x \in \mathbb{B}^n} \mu(\underline{x}) \ln \frac{\mu(\underline{x})}{2^{-K(\mu)} \mu(\underline{x})} \;=\; \frac{\ln 2}{2} K(\mu) \sum_{x \in \mathbb{B}^n} \mu(\underline{x}) \;=\; \frac{\ln 2}{2} K(\mu).$$

As this holds for all $n \in \mathbb{N}$, the result follows. It is this application of dominance to obtain powerful convergence results that lies at the heart of Solomonoff induction, and indeed universal artificial intelligence in general.

Although Solomonoff induction is not computable and is thus impractical, it nevertheless has many connections to practical principles and methods that are used for inductive inference. Clearly, if we define a computable prior rather than ξ, we recover normal Bayesian inference. If we define our prior to be uniform, for example by assuming that all models have the same complexity, then the result is maximum a posteriori (MAP) estimation, which in turn is related to maximum likelihood (ML) estimation. Relations can also be established to Minimum Message Length (MML), Minimum Description Length (MDL), and Maximum entropy (ME) based prediction (see Chapter 5 of Li and Vitányi, 1997). Thus, although Solomonoff induction does not yield a prediction algorithm itself, it does provide a theoretical framework that can be used to understand various practical inductive inference methods. It is a kind of ideal, but unattainable, model of optimal inductive inference.

2.8 Agent-environment model

Up to this point we have only considered the inductive inference problem, either in terms of inferring hypotheses, or predicting the continuation of a sequence. In both cases the agents were *passive* in the sense that they were unable to take actions that affect the future. Obviously this greatly limits them. More powerful is the class of *active* agents which not only observe their environment, they are also able to take actions that may affect the environment. Such agents are able to explore and achieve goals in their environment.