Universal Sequential Decisions in Unknown Environments

Marcus Hutter MARCUS@IDSIA.CH
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Galleria 2, CH-6928 Manno-Lugano, Switzerland

We give a brief introduction to the AIXI model, which unifies and overcomes the limitations of sequential decision theory and universal Solomonoff induction. While the former theory is suited for active agents in known environments, the latter is suited for passive prediction of unknown environments.

Introduction: Every inductive inference problem can be brought into the following form: Given a string $x_1x_2...x_{t-1} \equiv x_{1:t-1} \equiv x_{<t}$, take a guess at its continuation x_t. We will assume that the strings which have to be continued are drawn from a probability distribution μ. The maximal prior information a prediction algorithm can possess is the exact knowledge of μ, but often the true distribution is unknown. Instead, prediction is based on a guess ρ of μ. We expect that a predictor based on ρ performs well, if ρ is close to μ or converges to μ.

Universal probability distribution: Let $\mathcal{M} := \{\mu_1, \mu_2, ...\}$ be a finite or countable set of candidate probability distributions on strings. We define a weighted average on $\mathcal{M}$,

$$\xi(x_{1:n}) := \sum_{\mu_i \in \mathcal{M}} w_{\mu_i} \cdot \mu_i(x_{1:n}), \qquad \sum_{\mu_i \in \mathcal{M}} w_{\mu_i} = 1, \quad w_{\mu_i} > 0.$$

We call ξ universal relative to $\mathcal{M}$, as it multiplicatively dominates all distributions in $\mathcal{M}$, i.e. $\xi(x_{1:n}) \geq w_{\mu_i} \cdot \mu_i(x_{1:n})$ for all $\mu_i \in \mathcal{M}$. In the following, we assume that $\mathcal{M}$ is known and contains the true distribution from which $x_1x_2...$ is sampled, i.e. $\mu \in \mathcal{M}$. The condition $\mu \in \mathcal{M}$ is not a serious constraint if we include *all* computable probability distributions in $\mathcal{M}$ with high weights assigned to simple μ_i. Solomonoff-Levin's universal semi-measure is obtained if we include all enumerable semi-measures in $\mathcal{M}$ with weights $w_{\mu_i} \sim 2^{-K(\mu_i)}$, where $K(\mu_i)$ is the length of the shortest program for μ_i [1, 3]. One can show that the conditional ξ and μ probabilities rapidly converge to each other:

$$\xi(x_t|x_{<t}) \to \mu(x_t|x_{<t}) \quad \text{with } \mu \text{ probability 1.} \qquad (1)$$

Since the conditional probabilities are the basis of the decision algorithms considered in this work, we expect a good prediction performance if we use ξ as a guess of μ.

Bayesian decisions: Let $\ell_{x_t y_t} \in [0,1]$ be the received loss when predicting $y_t \in \mathcal{Y}$, but $x_t \in \mathcal{X}$ turns out to be the true t^{th} symbol of the sequence. Let $L_{n\Lambda_\rho}$ be the total expected loss for the first n symbols of the Bayes predictor Λ_ρ which minimizes the ρ expected loss. For instance for $\mathcal{X} = \mathcal{Y} = \{0,1\}$, Λ_ρ is a threshold strategy with $y_t^{\Lambda_\rho} = 0/1$ for $\rho(1|x_{<t}) \gtrless \gamma$, where $\gamma := \frac{\ell_{01} - \ell_{00}}{\ell_{01} - \ell_{00} + \ell_{10} - \ell_{11}}$. Let Λ be *any* prediction scheme (deterministic or probabilistic) with no constraint at all, taking *any* action $y_t^\Lambda \in \mathcal{Y}$ with total expected loss $L_{n\Lambda}$. If μ is known, Λ_μ is obviously the best prediction scheme in the sense of achieving minimal expected loss $L_{n\Lambda_\mu} \leq L_{n\Lambda}$ for any Λ. For the predictor Λ_ξ based on the universal distribution ξ, on can show $L_{n\Lambda_\xi}/L_{n\Lambda_\mu} = 1 + O(\sqrt{K(\mu)/L_{n\Lambda_\mu}})$, i.e. Λ_ξ has optimal asymptotics for $L_{n\Lambda_\mu} \to \infty$ with rapid convergence of the quotient to 1. If $L_{\infty\Lambda_\mu}$ is finite, then also $L_{\infty\Lambda_\xi}$ [1, 3].

More active systems: Prediction means guessing the future, but not influencing it. One step in the direction to more active systems was to allow the Λ system to act and to receive a loss $\ell_{x_t y_t}$ depending on the action y_t and the outcome x_t. The probability μ is still independent of the action, and the loss function ℓ^t has to be known in advance. This ensures that the greedy Λ_μ strategy is still optimal. The loss function can also be generalized to depend on the history $x_{<t}$ and on t.

Agents in known probabilistic environments: The full model of an acting agent influencing the environment has been developed in [2, 3]. The probability of the next symbol (input, perception) x_t depends in this case not only on the past sequence $x_{<t}$ but also on the past actions (outputs) $y_{1:t}$, i.e. $\mu = \mu(x_t|x_{<t}y_{1:t})$. We call probability distributions of this form *chronological*. The total μ expected loss is $\sum_{x_{1:n}} (\ell^1 + ... + \ell^n)\mu(x_{1:n}|y_{1:n})$, where we assumed a total number of n interaction cycles. Action $y_t(x_{<t}y_{<t})$ and loss function $\ell^t(x_{1:t}y_{1:t})$ may depend on the complete history, which allows planning and delayed loss assignment.

Sequential decision theory: The goal is to perform the actions which minimize the total μ expected loss:

$$y_t := \arg\min_{y_t} \sum_{x_t} ... \min_{y_n} \sum_{x_n} (\ell^1 + ... + \ell^n)\mu(x_{1:n}|y_{1:n}), \qquad (2)$$

$$L_{n\Lambda_\mu} = \min_{y_1} \sum_{x_1} ... \min_{y_n} \sum_{x_n} (\ell^1 + ... + \ell^n)\mu(x_{1:n}|y_{1:n}). \qquad (3)$$

The minimization over y_t is in chronological order to correctly incorporate the dependency of x_t and y_t on the history. Note that y_t only depends on the known history $x_{<t}y_{<t}$, whereas minima and expectations are taken over the unknown $x_{t:n}y_{t:n}$ variables. The policy (2) (called AIμ model) is optimal in the sense that no other policy leads to lower μ-expected loss.

Bellman equations: In the case that ℓ^t is independent of $y_{<t}$ and μ is independent of $y_{1:n}$, policy (2) reduces to the greedy Bayes Λ_μ strategy. For (completely observable) Markov Decision Processes $\mu = \mu(x_t|x_{t-1}y_t)$ (2) and (3) can be written as recursive Bellman equations of sequential decision theory with state space $\mathcal{X}$, action space $\mathcal{Y}$, state transition matrix $\mu(x_t|x_{t-1}y_t)$, rewards $-\ell^t$, *etc.* The general (non-MDP) case may also be (artificially) reduced to Bellman equations by identifying complete histories $x_{<t}y_{<t}$ with states and $\mu(x_t|x_{<t}y_{1:t})$ with the state transition matrix. Due to the use of complete histories as state space, the AIμ model neither assumes stationarity, nor the Markov property, nor complete accessibility of the environment. But since every state occurs at most once in the lifetime of the system the explicit formulation (2) is more useful than a pseudo-recursive Bellman equation form. There is no principle problem in determining y_k as long as μ is known and computable and $\mathcal{X}$, $\mathcal{Y}$ and n are finite.

Reinforcement learning for unknown environment: Things dramatically change if μ is unknown. Reinforcement learning algorithms are commonly used in this case to learn the unknown μ (or directly a value function). They succeed if the state space is either small or has effectively been made small by generalization or function approximation techniques. In almost all approaches, the solutions are either *ad hoc*, or work in restricted domains only, or have serious problems with state space exploration versus exploitation, or have non-optimal learning rate. Below we propose the AIξ model as a universal and optimal solution to these problems.

Unknown loss function: Furthermore, the loss function $\ell^t(x_{1:t}y_{1:t})$ may also be unknown, but there is an easy "solution" to this problem. The specification of the loss function can be absorbed in the probability distribution μ by increasing the input space $\mathcal{X}$. Let $x_t \equiv x'_t l_t$, where x'_t is the regular input, l_t is interpreted as the loss, $\ell^t(x_{1:t}y_{1:t})$ is replaced by l_t in (2) and (3), and μ is only non-zero if l_t is consistent with the loss, i.e. $l_t = \ell^t(x_{1:t}y_{1:t})$. In this way all possible unknowns are absorbed in μ.

The universal AIξ model: Encouraged by the good performance of the universal sequence predictor Λ_ξ, we propose a new model, where the probability distribution μ is learned indirectly by replacing it with a universal prior ξ. We define $\xi(x_{1:n}|y_{1:n}) := \sum_{\mu_i \in \mathcal{M}} w_{\mu_i} \cdot \mu_i(x_{1:n}|y_{1:n})$ as a weighted sum over chronological probability distributions in $\mathcal{M}$. Convergence $\xi(x_n|x_{<n}y_{1:n}) \to \mu(x_n|x_{<n}y_{1:n})$ can be proven analogously to (1). Replacing μ by ξ in (2) the *AIξ system* outputs

$$y_t := \arg\min_{y_t} \sum_{x_t} ... \min_{y_n} \sum_{x_n} (l_t + ... + l_n)\xi(x_{1:n}|y_{1:n}) \quad (4)$$

in cycle t given the history $x_{<t}y_{<t}$, where $x_t \equiv x'_t l_t$. The largest class $\mathcal{M}$ which is necessary from a computational point of view is the set of all enumerable chronological semi-measures with weights $w_{\mu_i} \sim 2^{-K(\mu_i)}$, where $K(\mu_i)$ is the Kolmogorov complexity of μ_i. Apart from the dependence on the horizon n and unimportant details, the AIξ system is uniquely defined by (4) without adjustable parameters. It does not depend on any assumption about the environment apart from being generated by some computable (but unknown!) probability distribution in $\mathcal{M}$.

Universally optimal AI systems: We want to call an AI model *universal*, if it is μ-independent (unbiased, model-free) and is able to solve any solvable problem and learn any learnable task. Further, we call a universal model, *universally optimal*, if there is no program which can solve or learn significantly faster (in terms of interaction cycles). As the AIξ model is parameterless, ξ rapidly converges to μ in the sense of (1), the AIμ model is itself optimal, and we expect no other model to converge faster to AIμ (in some sense) by analogy to the sequence prediction case, we risk the conjecture that AIξ is such a universally optimal system. Further support is given in [2, 3] by a detailed analysis of the behaviour of AIξ for various problem classes, including prediction, optimization, games, and supervised learning. We discuss in which sense AIξ overcomes some fundamental problems in reinforcement learning, like generalization, optimal learning rates, exploration versus exploitation, *etc.* Computational issues are also addressed.

References

[1] M. Hutter. General loss bounds for universal sequence prediction. In *Proc. 18th International Conf. on Machine Learning (ICML-2001)*, pages 210–217, Williamstown, MA, 2001. Morgan Kaufmann.

[2] M. Hutter. Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions. In *Proc. 12th European Conf. on Machine Learning (ECML-2001)*, volume 2167 of *LNAI*, pages 226–238, Freiburg, 2001. Springer, Berlin.

[3] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2004. 300 pages, http://www.idsia.ch/ai/~marcus/uaibook/.