Technical Report IDSIA-04-02 – 17 April 2002 – Proceedings of COLT-2002

# Self-Optimizing and Pareto-Optimal Policies in General Environments based on Bayes-Mixtures

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland*
marcus@idsia.ch – http://www.idsia.ch/~marcus

**Keywords**

Rational agents, sequential decision theory, reinforcement learning, value function, Bayes mixtures, self-optimizing policies, Pareto-optimality, unbounded effective horizon, (non) Markov decision processes.

**Abstract**

The problem of making sequential decisions in unknown probabilistic environments is studied. In cycle $t$ action $y_t$ results in perception $x_t$ and reward $r_t$, where all quantities in general may depend on the complete history. The perception $x_t$ and reward $r_t$ are sampled from the (reactive) environmental probability distribution $\mu$. This very general setting includes, but is not limited to, (partial observable, k-th order) Markov decision processes. Sequential decision theory tells us how to act in order to maximize the total expected reward, called value, if $\mu$ is known. Reinforcement learning is usually used if $\mu$ is unknown. In the Bayesian approach one defines a mixture distribution $\xi$ as a weighted sum of distributions $\nu \in \mathcal{M}$, where $\mathcal{M}$ is any class of distributions including the true environment $\mu$. We show that the Bayes-optimal policy $p^\xi$ based on the mixture $\xi$ is self-optimizing in the sense that the average value converges asymptotically for all $\mu \in \mathcal{M}$ to the optimal value achieved by the (infeasible) Bayes-optimal policy $p^\mu$ which knows $\mu$ in advance. We show that the necessary condition that $\mathcal{M}$ admits self-optimizing policies at all, is also sufficient. No other structural assumptions are made on $\mathcal{M}$. As an example application, we discuss ergodic Markov decision processes, which allow for self-optimizing policies. Furthermore, we show that $p^\xi$ is Pareto-optimal in the sense that there is no other policy yielding higher or equal value in *all* environments $\nu \in \mathcal{M}$ and a strictly higher value in at least one.

*This work was supported by SNF grant 2000-61847.00 to Jürgen Schmidhuber.

# 1 Introduction

**Reinforcement learning:** There exists a well developed theory for reinforcement learning agents in known probabilistic environments (like Blackjack) called sequential decision theory [Bel57, Ber95]. The optimal agent is the one which maximizes the future expected reward sum. This setup also includes deterministic environments (like static mazes). Even adversarial environments (like Chess or Backgammon) may be seen as special cases in some sense [Hut00, ch.6] (the reverse is also true [BT00]). Sequential decision theory deals with a wide range of problems, and provides a general formal solution in the sense that it is mathematically rigorous and (uniquely) specifies the optimal solution (leaving aside computational issues). The theory breaks down when the environment is unknown (like when driving a car in the real world). Reinforcement learning algorithms exist for unknown Markov decision processes (MDP*s*) with small state space, and for other restricted classes [KLM96, SB98, Ber95, KV86], but even in these cases their learning rate is usually far from optimum.

**Performance measures:** In this work we are interested in general (probabilistic) environmental classes $\mathcal{M}$. We assume $\mathcal{M}$ is given, and that the true environment $\mu$ is in $\mathcal{M}$, but is otherwise unknown. The expected reward sum (value) $V_\mu^p$ when following policy $p$ is of central interest. We are interested in policies $\tilde{p}$ which perform well (have high value) independent of what the true environment $\mu\in\mathcal{M}$ is. A natural demand from an optimal policy is that there is no other policy yielding higher or equal value in *all* environments $\nu\in\mathcal{M}$ and a strictly higher value in one $\nu\in\mathcal{M}$. We call such a property *Pareto-optimality*. The other quantity of interest is how close $V_\mu^{\tilde{p}}$ is to the value $V_\mu^*$ of the optimal (but infeasible) policy $p^\mu$ which knows $\mu$ in advance. We call a policy whose average value converges asymptotically for all $\mu\in\mathcal{M}$ to the optimal value $V_\mu^*$ if $\mu$ is the true environment, *self-optimizing*.

**Main new results for Bayes-mixtures:** We define the Bayes-mixture $\xi$ as a weighted average of the environments $\nu\in\mathcal{M}$ and analyze the properties of the Bayes-optimal policy $p^\xi$ which maximizes the mixture value $V_\xi$. One can show that not all environmental classes $\mathcal{M}$ admit self-optimizing policies. One way to proceed is to search for and prove weaker properties than self-optimizingness [Hut00]. Here we follow a different approach: Obviously, the least we must demand from $\mathcal{M}$ to have a chance of finding a self-optimizing policy is that there exists some self-optimizing policy $\tilde{p}$ at all. The main new result of this work is that this necessary condition is also sufficient for $p^\xi$ to be self-optimizing. No other properties need to be imposed on $\mathcal{M}$. The other new result is that $p^\xi$ is always Pareto-optimal, with no conditions at all imposed on $\mathcal{M}$.

**Contents:** Section 2 defines the model of agents acting in general probabilistic environments and defines the finite horizon value of a policy and the optimal value-maximizing policy. Furthermore, the mixture-distribution is introduced and the fundamental linearity and convexity properties of the mixture-values is stated. Section 3 defines and proves Pareto-optimality of $p^\xi$. The concept is refined to balanced Pareto-optimality, showing

that a small increase of the value for some environments only leaves room for a small decrease in others. Section 4 shows that $p^\xi$ is self-optimizing if $\mathcal{M}$ admits self-optimizing policies, and also gives the speed of convergence in the case of finite $\mathcal{M}$. The finite horizon model has several disadvantages. For this reason Section 5 defines the discounted (infinite horizon) future value function, and the corresponding optimal value-maximizing policy. Pareto-optimality and self-optimizingness of $p^\xi$ are shown shown for this model. As an application we show in Section 6 that the class of ergodic MDPs admits self-optimizing policies w.r.t. the undiscounted model and w.r.t. the discounted model if the effective horizon tends to infinity. Together with the results from the previous sections this shows that $p^\xi$ is self-optimizing for erdodic MDPs. Conclusions and outlook can be found in Section 7.

# 2 Rational Agents in Probabilistic Environments

**The agent model:** A very general framework for intelligent systems is that of rational agents [RN95]. In cycle $k$, an agent performs *action* $y_k \in \mathcal{Y}$ (output) which results in a *perception* or *observation* $x_k \in \mathcal{X}$ (input), followed by cycle $k+1$ and so on. We assume that the action and perception spaces $\mathcal{X}$ and $\mathcal{Y}$ are finite. We write $p(x_{<k}) = y_{1:k}$ to denote the output $y_{1:k} \equiv y_1 ... y_k$ of the agents policy $p$ on input $x_{<k} \equiv x_1 ... x_{k-1}$ and similarly $q(y_{1:k}) = x_{1:k}$ for the environment $q$ in the case of deterministic environments. We call policy $p$ and environment $q$ behaving in this way *chronological*. Note that policy and environment are allowed to depend on the complete history. We do not make any MDP or POMDP assumption here, and we don't talk about states of the environment, only about observations. In the more general case of a *probabilistic environment*, given the history $yx_{<k}y_k \equiv yx_1 ... yx_{k-1}y_k \equiv y_1x_1 ... y_{k-1}x_{k-1}y_k$, the probability that the environment leads to perception $x_k$ in cycle $k$ is (by definition) $\rho(yx_{<k}y\underline{x}_k)$. The underlined argument $\underline{x}_k$ in $\rho$ is a random variable and the other non-underlined arguments $yx_{<k}y_k$ represent conditions.[1] We call probability distributions like $\rho$ *chronological*. Since value optimizing policies can always be chosen deterministic, there is no real need to generalize the setting to probabilistic policies. Arbitrarily we formalize Sections 3 and 4 in terms of deterministic policies and Section 5 in terms of probabilistic policies.

**Value functions and optimal policies:** The goal of the agent is to maximize future *rewards*, which are provided by the environment through the inputs $x_k$. The inputs $x_k \equiv x'_k r_k$ are divided into a regular part $x'_k$ and some (possibly empty or delayed) reward $r_k \in [0\,, r_{max}]$.[2] We use the abbreviation

$$\rho(yx_{<k}y\underline{x}_{k:m}) \;=\; \rho(yx_{<k}y\underline{x}_k)\cdot\rho(yx_{1:k}y\underline{x}_{k+1})\cdot\,...\,\cdot\rho(yx_{<m}y\underline{x}_m), \tag{1}$$

[1] The standard notation $\rho(x_k|yx_{<k}y_k)$ for conditional probabilities destroys the chronological order and would become quite confusing in later expressions.

[2] In the reinforcement learning literature when dealing with (PO)MDPs the reward is usually considered to be a function of the environmental state. The zero-assumption analogue here is that the reward $r_k$ is some probabilistic function $\rho'$ depending on the complete history. It is very convenient to integrate $r_k$ into $x_k$ and $\rho'$ into $\rho$.

which is essentially Bayes rules, and $\varepsilon=yx_{<1}$ for the empty string. The $\rho$-expected reward sum (value) of future cycles $k$ to $m$ with outputs $y_{k:m}$ generated by the agent's policy $p$, the optimal policy $p^\rho$ which maximizes the value, its action $y_k$ and the corresponding value can formally be defined as follows.

**Definition 1 (Value function and optimal policy)** *We define the* value *of policy $p$ in environment $\rho$ given history $yx_{<k}$, or shorter, the $\rho$-value of $p$ given $yx_{<k}$, as*

$$V_{km}^{p\rho}(yx_{<k}) \;:=\; \sum_{x_{k:m}}(r_k+...+r_m)\rho(yx_{<k}\underline{yx}_{k:m})_{|y_{1:m}=p(x_{<m})}. \tag{2}$$

*$m$ is the* lifespan *or initial* horizon *of the agent. The $\rho$-optimal policy $p^\rho$ which maximizes the (total) value $V_\rho^p:=V_{1m}^{p\rho}(\varepsilon)$ is*

$$p^\rho := \arg\max_p V_\rho^p, \quad V_{km}^{*\rho}(yx_{<k}) := V_{km}^{p^\rho\rho}(yx_{<k}). \tag{3}$$

*Explicit expressions for the action $y_k$ in cycle $k$ of the $\rho$-optimal policy $p^\rho$ and their value $V_{km}^{*\rho}(yx_{<k})$ are*

$$y_k \;=\; \arg\max_{y_k}\sum_{x_k}\max_{y_{k+1}}\sum_{x_{k+1}}\;...\;\max_{y_m}\sum_{x_m}(r_k+...+r_m)\cdot\rho(yx_{<k}\underline{yx}_{k:m}), \tag{4}$$

$$V_{km}^{*\rho}(yx_{<k}) \;=\; \max_{y_k}\sum_{x_k}\max_{y_{k+1}}\sum_{x_{k+1}}\;...\;\max_{y_m}\sum_{x_m}(r_k+...+r_m)\cdot\rho(yx_{<k}\underline{yx}_{k:m}). \tag{5}$$

*where $yx_{<k}$ is the actual history.*

One can show [Hut00] that these definitions are consistent and correctly capture our intention. For instance, consider the expectimax expression (5): The best expected reward is obtained by averaging over possible perceptions $x_i$ and by maximizing over the possible actions $y_i$. This has to be done in chronological order $y_kx_k...y_mx_m$ to correctly incorporate the dependency of $x_i$ and $y_i$ on the history. Obviously

$$V_{km}^{*\rho}(yx_{<k}) \;\geq\; V_{km}^{p\rho}(yx_{<k})\;\forall p, \quad \text{especially} \quad V_\rho^* \;\geq\; V_\rho^p\;\forall p. \tag{6}$$

**Known environment $\mu$:** Let us now make a change in conventions and assume that $\mu$ is the true environment in which the agent operates and that we know $\mu$ (like in Blackjack).[3] Then, policy $p^\mu$ is optimal in the sense that no other policy for an agent leads to higher $\mu$-expected reward. This setting includes as special cases deterministic environments, Markov decision processes (MDP*s*), and even adversarial environments for special choices of $\mu$ [Hut00]. There is no principle problem in determining the optimal action $y_k$ as long as $\mu$ is known and computable and $\mathcal{X}$, $\mathcal{Y}$ and $m$ are finite.

[3]If the existence of true objective probabilities violates the philosophical attitude of the reader he may assume a deterministic environment $\mu$.

**The mixture distribution $\xi$:** Things drastically change if $\mu$ is unknown. For (parameterized) MDPs with small state (parameter) space, suboptimal reinforcement learning algorithms may be used to learn the unknown $\mu$ [KLM96, SB98, Ber95, KV86]. In the Bayesian approach the true probability distribution $\mu$ is not learned directly, but is replaced by a Bayes-mixture $\xi$. Let us assume that we know that the true environment $\mu$ is contained in some known set $\mathcal{M}$ of environments. For convenience we assume that $\mathcal{M}$ is finite or countable. The Bayes-mixture $\xi$ is defined as

$$\xi(\underline{yx}_{1:m}) = \sum_{\nu\in\mathcal{M}} w_\nu \nu(\underline{yx}_{1:m}) \quad \text{with} \quad \sum_{\nu\in\mathcal{M}} w_\nu = 1, \quad w_\nu > 0 \quad \forall \nu \in \mathcal{M} \tag{7}$$

The weights $w_\nu$ may be interpreted as the prior degree of belief that the true environment is $\nu$. Then $\xi(\underline{yx}_{1:m})$ could be interpreted as the prior subjective belief probability in observing $x_{1:m}$, given actions $y_{1:m}$. It is, hence, natural to follow the policy $p^\xi$ which maximizes $V_\xi^p$. If $\mu$ is the true environment the expected reward when following policy $p^\xi$ will be $V_\mu^{p^\xi}$. The optimal (but infeasible) policy $p^\mu$ yields reward $V_\mu^{p^\mu} \equiv V_\mu^*$. It is now of interest ($a$) whether there are policies with uniformly larger value than $V_\mu^{p^\xi}$ and ($b$) how close $V_\mu^{p^\xi}$ is to $V_\mu^*$. These are the main issues of the remainder of this work.

**A universal choice of $\xi$ and $\mathcal{M}$:** One may also ask what the most general class $\mathcal{M}$ and weights $w_\nu$ could be. Without any prior knowledge we should include *all* environments in $\mathcal{M}$. In this generality this approach leads at best to negative results. More useful is the assumption that the environment possesses some structure, we just don't know which. From a computational point of view we can only unravel effective structures which are describable by (semi)computable probability distributions. So we may include *all* (semi)computable (semi)distributions in $\mathcal{M}$. Occam's razor tells us to assign high prior belief to simple environments. Using Kolmogorov's universal complexity measure $K(\nu)$ for environments $\nu$ one should set $w_\nu \sim 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program on a universal Turing machine computing $\nu$. The resulting policy $p^\xi$ has been developed and intensively discussed in [Hut00]. It is a unification of sequential decision theory [Bel57, Ber95] and Solomonoff's celebrated universal induction scheme [Sol78, LV97]. In the following we consider generic $\mathcal{M}$ and $w_\nu$. The following property of $V_\rho$ is crucial.

**Theorem 1 (Linearity and convexity of $V_\rho$ in $\rho$)** *$V_\rho^p$ is a linear function in $\rho$ and $V_\rho^*$ is a convex function in $\rho$ in the sense that*

$$V_\xi^p = \sum_{\nu\in\mathcal{M}} w_\nu V_\nu^p \quad \textit{and} \quad V_\xi^* \le \sum_{\nu\in\mathcal{M}} w_\nu V_\nu^* \quad \textit{where} \quad \xi(\underline{yx}_{1:m}) = \sum_{\nu\in\mathcal{M}} w_\nu \nu(\underline{yx}_{1:m})$$

**Proof:** Linearity is obvious from the definition of $V_\rho^p$. Convexity follows from $V_\xi^* \equiv V_\xi^{p^\xi} = \sum_\nu w_\nu V_\nu^{p^\xi} \le \sum_\nu w_\nu V_\nu^*$, where the identity is definition (3), the equality uses linearity of $V_\rho^{p^\xi}$ just proven, and the last inequality follows from the dominance (6) and non-negativity of the weights $w_\nu$. □

One loose interpretation of the convexity is that a mixture can never increase performance. In the remainder of this work $\mu$ denotes the true environment, $\rho$ any distribution, and $\xi$ the Bayes-mixture of distributions $\nu \in \mathcal{M}$.

# 3 Pareto Optimality of policy $p^\xi$

The total $\mu$-expected reward $V_\mu^{p^\xi}$ of policy $p^\xi$ is of central interest in judging the performance of policy $p^\xi$. We know that there *are* policies (e.g. $p^\mu$) with higher $\mu$-value ($V_\mu^* \geq V_\mu^{p^\xi}$). In general, every policy based on an estimate $\rho$ of $\mu$ which is closer to $\mu$ than $\xi$ is, outperforms $p^\xi$ in environment $\mu$, simply because it is more taylored toward $\mu$. On the other hand, such a system probably performs worse than $p^\xi$ in other environments. Since we do not know $\mu$ in advance we may ask whether there exists a policy $p$ with better or equal performance than $p^\xi$ in *all* environments $\nu \in \mathcal{M}$ and a strictly better performance for one $\nu \in \mathcal{M}$. This would clearly render $p^\xi$ suboptimal. We show that there is no such $p$.

**Theorem 2 (Pareto optimality)** *Policy $p^\xi$ is Pareto-optimal in the sense that there is no other policy $p$ with $V_\nu^p \geq V_\nu^{p^\xi}$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one $\nu$.*

**Proof:** We want to arrive at a contradiction by assuming that $p^\xi$ is not Pareto-optimal, i.e. by assuming the existence of a policy $p$ with $V_\nu^p \geq V_\nu^{p^\xi}$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one $\nu$:

$$V_\xi^p \;=\; \sum_\nu w_\nu V_\nu^p \;>\; \sum_\nu w_\nu V_\nu^{p^\xi} \;=\; V_\xi^{p^\xi} \;\equiv\; V_\xi^* \;\geq\; V_\xi^p$$

The two equalities follow from linearity of $V_\rho$ (Theorem 1). The strict inequality follows from the assumption and from $w_\nu > 0$. The identity is just Definition 1(3). The last inequality follows from the fact that $p^\xi$ maximizes by definition the universal value (6). The contradiction $V_\xi^p > V_\xi^p$ proves Pareto-optimality of policy $p^\xi$.□

Pareto-optimality should be regarded as a necessary condition for an agent aiming to be optimal. From a practical point of view a significant increase of $V$ for many environments $\nu$ may be desirable even if this causes a small decrease of $V$ for a few other $\nu$. The impossibility of such a "balanced" improvement is a more demanding condition on $p^\xi$ than pure Pareto-optimality. The next theorem shows that $p^\xi$ is also balanced-Pareto-optimal in the following sense:

**Theorem 3 (Balanced Pareto optimality)**

$$\Delta_\nu := V_\nu^{p^\xi} - V_\nu^{\tilde{p}}, \quad \Delta := \sum_{\nu \in \mathcal{M}} w_\nu \Delta_\nu \quad \Rightarrow \quad \Delta \geq 0.$$

*This implies the following: Assume $\tilde{p}$ has lower value than $p^\xi$ on environments $\mathcal{L}$ by a total weighted amount of $\Delta_\mathcal{L} := \sum_{\lambda \in \mathcal{L}} w_\lambda \Delta_\lambda$. Then $\tilde{p}$ can have higher value on $\eta \in \mathcal{H} := \mathcal{M} \backslash \mathcal{L}$, but the improvement is bounded by $\Delta_\mathcal{H} := |\sum_{\eta \in \mathcal{H}} w_\eta \Delta_\eta| \leq \Delta_\mathcal{L}$. Especially $|\Delta_\eta| \leq w_\eta^{-1} \max_{\lambda \in \mathcal{L}} \Delta_\lambda$.*

This means that a weighted value increase $\Delta_\mathcal{H}$ by using $\tilde{p}$ instead of $p^\xi$ is compensated by an at least as large weighted decrease $\Delta_\mathcal{L}$ on other environments. If the decrease is small, the increase can also only be small. In the special case of only a single environment

with decreased value $\Delta_\lambda$, the increase is bound by $\Delta_\eta \leq \frac{w_\lambda}{w_\eta}|\Delta_\lambda|$, i.e. a decrease by an amount $\Delta_\lambda$ can only cause an increase by at most the same amount times a factor $\frac{w_\lambda}{w_\eta}$. For the choice of the weights $w_\nu \sim 2^{-K(\nu)}$, a decrease can only cause a smaller increase in simpler environments, but a scaled increase in more complex environments. Finally note that pure Pareto-optimality (Theorem 2) follows from balanced Pareto-optimality in the special case of no decrease $\Delta_{\mathcal{L}} \equiv 0$.

**Proof:** $\Delta \geq 0$ follows from $\Delta = \sum_\nu w_\nu [V_\nu^{p^\xi} - V_\nu^{\tilde{p}}] = V_\xi^{p^\xi} - V_\xi^{\tilde{p}} \geq 0$, where we have used linearity of $V_\rho$ (Theorem 1) and dominance $V_\xi^{p^\xi} \geq V_\xi^p$ (6). The remainder of Theorem 3 is obvious from $0 \leq \Delta = \Delta_{\mathcal{L}} - \Delta_{\mathcal{H}}$ and by bounding the weighted average $\Delta_\eta$ by its maximum.□

# 4 Self-optimizing Policy $p^\xi$ w.r.t. Average Value

In the following we study under which circumstances[4]

$$\tfrac{1}{m} V_{1m}^{p^\xi \nu} \to \tfrac{1}{m} V_{1m}^{*\nu} \quad \text{for} \quad m \to \infty \quad \text{for } \textit{all} \quad \nu \in \mathcal{M}. \tag{8}$$

The least we must demand from $\mathcal{M}$ to have a chance that (8) is true is that there exists some policy $\tilde{p}$ at all with this property, i.e.

$$\exists \tilde{p} : \ \tfrac{1}{m} V_{1m}^{\tilde{p}\nu} \to \tfrac{1}{m} V_{1m}^{*\nu} \quad \text{for} \quad m \to \infty \quad \text{for } \textit{all} \quad \nu \in \mathcal{M}. \tag{9}$$

Luckily, this necessary condition will also be sufficient. This is another (asymptotic) optimality property of policy $p^\xi$. *If* universal convergence in the sense of (9) is possible at all in a class of environments $\mathcal{M}$, then policy $p^\xi$ converges in the sense of (8). We will call policies $\tilde{p}$ with a property like (9) *self-optimizing* [KV86]. The following two Lemmas pave the way for proving the convergence Theorem.

**Lemma 1 (Value difference relation)**

$$0 \ \leq \ V_\nu^* - V_\nu^{\tilde{p}} \ =: \ \Delta_\nu \quad \Rightarrow \quad 0 \ \leq \ V_\nu^* - V_\nu^{p^\xi} \ \leq \ \tfrac{1}{w_\nu}\Delta \quad \textit{with} \quad \Delta := \sum_{\nu \in \mathcal{M}} w_\nu \Delta_\nu$$

**Proof:** The following sequence of inequalities proves the lemma:

$$0 \ \leq \ w_\nu [V_\nu^* - V_\nu^{p^\xi}] \ \leq \ \textstyle\sum_\nu w_\nu [V_\nu^* - V_\nu^{p^\xi}] \ \leq \ \sum_\nu w_\nu [V_\nu^* - V_\nu^{\tilde{p}}] \ = \ \sum_\nu w_\nu \Delta_\nu \ \equiv \ \Delta$$

In the first and second inequality we used $w_\nu \geq 0$ and $V_\nu^* - V_\nu^{p^\xi} \geq 0$. The last inequality follows from $\sum_\nu w_\nu V_\nu^{p^\xi} = V_\xi^{p^\xi} \equiv V_\xi^* \geq V_\xi^{\tilde{p}} = \sum_\nu w_\nu V_\nu^{\tilde{p}}$. □

We also need some results for averages of functions $\delta_\nu(m) \geq 0$ converging to zero.

[4]Here and elsewhere we interpret $a_m \to b_m$ as an abbreviation for $a_m - b_m \to 0$. $\lim_{m\to\infty} b_m$ may not exist.

**Lemma 2 (Convergence of averages)** *For $\delta(m):=\sum_{\nu\in\mathcal{M}}w_\nu\delta_\nu(m)$ the following holds (we only need $\sum_\nu w_\nu\leq 1$):*

$$\begin{aligned} &i) \quad \delta_\nu(m)\leq f(m) \quad \forall\nu \quad \textit{implies} \quad \delta(m)\leq f(m).\\ &ii) \quad \delta_\nu(m)\stackrel{m\to\infty}{\longrightarrow}0 \quad \forall\nu \quad \textit{implies} \quad \delta(m)\stackrel{m\to\infty}{\longrightarrow}0 \quad \textit{if} \quad 0\leq\delta_\nu(m)\leq c. \end{aligned}$$

**Proof:** $(i)$ immediately follows from $\delta(m)=\sum_\nu w_\nu\delta_\nu(m)\leq\sum_\nu w_\nu f(m)\leq f(m)$. For $(ii)$ we choose some order on $\mathcal{M}$ and some $\nu_0\in\mathcal{M}$ large enough such that $\sum_{\nu\geq\nu_0}w_\nu\leq\frac{\varepsilon}{c}$. Using $\delta_\nu(m)\leq c$ this implies

$$\sum_{\nu\geq\nu_0}w_\nu\delta_\nu(m) \;\leq\; \sum_{\nu\geq\nu_0}w_\nu c \;\leq\; \varepsilon.$$

Furthermore, the assumption $\delta_\nu(m)\to 0$ means that there is an $m_{\nu\varepsilon}$ depending on $\nu$ and $\varepsilon$ such that $\delta_\nu(m)\leq\varepsilon$ for all $m\geq m_{\nu\varepsilon}$. This implies

$$\sum_{\nu\leq\nu_0}w_\nu\delta_\nu(m) \;\leq\; \sum_{\nu\leq\nu_0}w_\nu\varepsilon \;\leq\; \varepsilon \quad \text{for all} \quad m \;\geq\; \max_{\nu\leq\nu_0}\{m_{\nu\varepsilon}\} \;=:\; m_\varepsilon.$$

$m_\varepsilon<\infty$, since the maximum is over a finite set. Together we have

$$\delta(m) \;\equiv\; \sum_{\nu\in\mathcal{M}}w_\nu\delta_\nu(m) \;\leq\; 2\varepsilon \quad \text{for} \quad m \;\geq\; m_\varepsilon \quad \Rightarrow \quad \delta(m)\to 0 \quad \text{for} \quad m\to\infty$$

since $\varepsilon$ was arbitrary and $\delta(m)\geq 0$. □

**Theorem 4 (Self-optimizing policy $p^\xi$ w.r.t. average value)** *There exists a sequence of policies $\tilde{p}_m$, $m$=1,2,3,... with value within $\Delta(m)$ to optimum for all environments $\nu\in\mathcal{M}$, then, save for a constant factor, this also holds for the sequence of universal policies $p^\xi_m$, i.e.*

$$i) \quad \textit{If} \quad \exists\tilde{p}_m\forall\nu : V^{*\nu}_{1m}-V^{\tilde{p}_m\nu}_{1m}\leq\Delta(m) \quad\Longrightarrow\quad V^{*\mu}_{1m}-V^{p^\xi_m\mu}_{1m}\leq\tfrac{1}{w_\mu}\Delta(m).$$

*If there exists a sequence of self-optimizing policies $\tilde{p}_m$ in the sense that their expected average reward $\frac{1}{m}V^{\tilde{p}_m\nu}_{1m}$ converges to the optimal average $\frac{1}{m}V^{*\nu}_{1m}$ for all environments $\nu\in\mathcal{M}$, then this also holds for the sequence of universal policies $p^\xi_m$, i.e.*

$$ii) \quad \textit{If} \quad \exists\tilde{p}_m\forall\nu : \tfrac{1}{m}V^{\tilde{p}_m\nu}_{1m}\stackrel{m\to\infty}{\longrightarrow}\tfrac{1}{m}V^{*\nu}_{1m} \quad\Longrightarrow\quad \tfrac{1}{m}V^{p^\xi_m\mu}_{1m}\stackrel{m\to\infty}{\longrightarrow}\tfrac{1}{m}V^{*\mu}_{1m}.$$

The beauty of this theorem is that if universal convergence in the sense of (9) is possible at all in a class of environments $\mathcal{M}$, then policy $p^\xi$ converges (in the sense of (8)). The necessary condition of convergence is also sufficient. The unattractive point is that this is not an asymptotic convergence statement for $V^{p^\xi\mu}_{km}$ of a single policy $p^\xi$ for $k\to\infty$ for some fixed $m$, and in fact no such theorem could be true, since always $k\leq m$. The theorem merely says that under the stated conditions the average value of $p^\xi_m$ can be arbitrarily close to optimum for sufficiently large (pre-chosen) horizon $m$. This weakness will be resolved in the next subsection.

**Proof:** $(i)$ $\Delta_\nu(m)=f(m)$ implies $\Delta(m)=f(m)$ by Lemma 2$(i)$. Inserting this in Lemma 1 proves Theorem 4$(i)$ (recovering the $m$ dependence and finally renaming $f\rightsquigarrow\Delta$).

$(ii)$ We define $\delta_\nu(m) := \frac{1}{m}\Delta_\nu(m) = \frac{1}{m}[V^*_\nu - V^{\tilde{p}}_\nu]$. Since we assumed bounded rewards $0 \leq r \leq r_{max}$ we have

$$V^*_\nu \leq m r_{max} \quad \text{and} \quad V^{\tilde{p}}_\nu \geq 0 \quad \Rightarrow \quad \Delta_\nu \leq m r_{max} \quad \Rightarrow \quad 0 \leq \delta_\nu(m) \leq c := r_{max}.$$

The premise in Theorem 4$(ii)$ is that $\delta_\nu(m) = \frac{1}{m}[V^{*\nu}_{1m} - V^{\tilde{p}\nu}_{1m}] \to 0$ which implies

$$0 \;\leq\; \tfrac{1}{m}[V^{*\nu}_{1m} - V^{p^\xi \nu}_{1m}] \;\leq\; \tfrac{1}{w_\nu}\tfrac{\Delta(m)}{m} \;=\; \tfrac{1}{w_\nu}\delta(m) \;\to\; 0.$$

The inequalities follow from Lemma 1 and convergence to zero from Lemma 2$(ii)$. This proves Theorem 4$(ii)$. $\Box$.

In Section 6 we show that a converging $\tilde{p}$ exists for ergodic MDP$s$, and hence $p^\xi$ converges in this environmental class too (in the sense of Theorem 4).

# 5 Discounted Future Value Function

We now shift our focus from the total value $V_{1m}$, $m \to \infty$ to the future value (value-to-go) $V_{k?}$, $k \to \infty$. The main reason is that we want to get rid of the horizon parameter $m$. In the last subsection we have shown a convergence theorem for $m \to \infty$, but a specific policy $p^\xi$ is defined for all times relative to a fixed horizon $m$. Current time $k$ is moving, but $m$ is fixed[5]. Actually, to use $k \to \infty$ arguments we *have* to get rid of $m$, since $k \leq m$. This is the reason for the question mark in $V_{k?}$ above.

We eliminate the horizon by discounting the rewards $r_k \rightsquigarrow \gamma_k r_k$ with $\sum_{i=1}^\infty \gamma_i < \infty$ and letting $m \to \infty$. The analogue of $m$ is now an effective horizon $h_k^{eff}$ which may be defined by $\sum_{i=k}^{k+h_k^{eff}} \gamma_k \sim \sum_{i=k+h_k^{eff}}^\infty \gamma_k$. See [Hut00, Ch.4] for a detailed discussion of the horizon problem. Furthermore, we renormalize $V_{k\infty}$ by $\sum_{i=k}^\infty \gamma_i$ and denote it by $V_{k\gamma}$. It can be interpreted as a future expected weighted-average reward. Furthermore we extend the definition to probabilistic policies $\pi$.

**Definition 2 (Discounted value function and optimal policy)** *We define the $\gamma$ discounted weighted-average future* value *of (probabilistic) policy $\pi$ in environment $\rho$ given history $yx_{<k}$, or shorter, the $\rho$-value of $\pi$ given $yx_{<k}$, as*

$$V^{\pi\rho}_{k\gamma}(yx_{<k}) \;:=\; \frac{1}{\Gamma_k} \lim_{m\to\infty} \sum_{yx_{k:m}} (\gamma_k r_k + ... + \gamma_m r_m)\rho(yx_{<k}yx_{k:m})\pi(yx_{<k}yx_{k:m})$$

*with $\Gamma_k := \sum_{i=k}^\infty \gamma_i$. The policy $p^\rho$ is defined as to maximize the future value $V^{\pi\rho}_{k\gamma}$:*

$$p^\rho := \arg\max_\pi V^{\pi\rho}_{k\gamma}, \qquad V^{*\rho}_{k\gamma} := V^{p^\rho\rho}_{k\gamma} = \max_\pi V^{\pi\rho}_{k\gamma} \geq V^{\pi\rho}_{k\gamma} \;\forall \pi.$$

[5] A dynamic horizon like $m \rightsquigarrow m_k = k^2$ can lead to policies with very poor performance [Hut00, Ch.4].

**Remarks:**

- $\pi(\underline{yx}_{<k}\underline{yx}_{k:m})$ is actually independent of $x_m$, since $\pi$ is chronological.
- Normalization of $V_{k\gamma}$ by $\Gamma_k$ does not affect the policy $p^\rho$.
- The definition of $p^\rho$ is independent of $k$.
- Without normalization by $\Gamma_k$ the future values would converge to zero for $k\to\infty$ in every environment for every policy.
- For an MDP environment, a stationary policy, and geometric discounting $\gamma_k\sim\gamma^k$, the future value is independent of $k$ and reduces to the well-known MDP value function.
- There is always a deterministic optimizing policy $p^\rho$ (which we use).
- For a deterministic policy there is exactly one $y_{k:m}$ for each $x_{k:m}$ with $\pi\neq 0$. The sum over $y_{k:m}$ drops in this case.
- An iterative representation as in Definition 1 is possible.
- Setting $\gamma_k=1$ for $k\leq m$ and $\gamma_k=0$ for $k>m$ gives back the undiscounted model (1) with $V_{1\gamma}^{p\rho}=\frac{1}{m}V_{1m}^{p\rho}$.
- $V_{k\gamma}$ (and $w_k^\nu$ defined below) depend on the realized history $\underline{yx}_{<k}$.

Similarly to the previous sections one can prove the following properties:

**Theorem 5 (Linearity and convexity of $V_\rho$ in $\rho$)** *$V_{k\gamma}^{\pi\rho}$ is a linear function in $\rho$ and $V_{k\gamma}^{*\rho}$ is a convex function in $\rho$ in the sense that*

$$V_{k\gamma}^{\pi\xi} = \sum_{\nu\in\mathcal{M}} w_k^\nu V_{k\gamma}^{\pi\nu} \quad \textit{and} \quad V_{k\gamma}^{*\xi} \leq \sum_{\nu\in\mathcal{M}} w_k^\nu V_{k\gamma}^{*\nu}$$

$$\textit{where} \quad \xi(\underline{yx}_{<k}\underline{yx}_{k:m}) = \sum_{\nu\in\mathcal{M}} w_k^\nu \nu(\underline{yx}_{<k}\underline{yx}_{k:m}) \quad \textit{with} \quad w_k^\nu := w_\nu \frac{\nu(\underline{yx}_{<k})}{\xi(\underline{yx}_{<k})}$$

The conditional representation of $\xi$ can be proven by dividing the definition (7) of $\xi(\underline{yx}_{1:m})$ by $\xi(\underline{yx}_{<k})$ and by using Bayes rules (1). The posterior weight $w_k^\nu$ may be interpreted as the posterior belief in $\nu$ and is related to learning aspects of policy $p^\xi$.

**Theorem 6 (Pareto optimality)** *For every $k$ and history $\underline{yx}_{<k}$ the following holds: $p^\xi$ is Pareto-optimal in the sense that there is no other policy $\pi$ with $V_{k\gamma}^{\pi\nu}\geq V_{k\gamma}^{p^\xi\nu}$ for all $\nu\in\mathcal{M}$ and strict inequality for at least one $\nu$.*

**Lemma 3 (Value difference relation)**

$$0\leq V_{k\gamma}^{*\nu}-V_{k\gamma}^{\tilde{\pi}_k\nu}=:\Delta_k^\nu \quad \Rightarrow \quad 0\leq V_{k\gamma}^{*\nu}-V_{k\gamma}^{p^\xi\nu}\leq\tfrac{1}{w_k^\nu}\Delta_k \quad \textit{with} \quad \Delta_k:=\sum_{\nu\in\mathcal{M}} w_k^\nu\Delta_k^\nu$$

The proof of Theorem 6 and Lemma 3 follows the same steps as for Theorem 2 and Lemma 1 with appropriate replacements. The proof of the analogue of the convergence Theorem 4 involves one additional step. We abbreviate "with $\mu$ probability 1" by w.$\mu$.p.1.

**Theorem 7 (Self-optimizing policy $p^\xi$ w.r.t. discounted value)** *For any $\mathcal{M}$, if there exists a sequence of self-optimizing policies $\tilde\pi_k$ $k=1,2,3,...$ in the sense that their expected weighted-average reward $V_{k\gamma}^{\tilde\pi_k\nu}$ converges for $k\to\infty$ with $\nu$-probability one to the optimal value $V_{k\gamma}^{*\nu}$ for all environments $\nu\in\mathcal{M}$, then this also holds for the universal policy $p^\xi$ in the true $\mu$-environment, i.e.*

$$\text{If } \exists\tilde\pi_k\forall\nu : V_{k\gamma}^{\tilde\pi_k\nu} \stackrel{k\to\infty}{\longrightarrow} V_{k\gamma}^{*\nu} \quad w.\nu.p.1 \quad\Longrightarrow\quad V_{k\gamma}^{p^\xi\mu} \stackrel{k\to\infty}{\longrightarrow} V_{k\gamma}^{*\mu} \quad w.\mu.p.1.$$

*The probability qualifier refers to the historic perceptions $x_{<k}$. The historic actions $y_{<k}$ are arbitrary.*

The conclusion is valid for action histories $y_{<k}$ if the condition is satisfied for this action history. Since we usually need the conclusion for the $p^\xi$-action history, which is hard to characterize, we usually need to prove the condition for *all* action histories. Theorem 7 is a powerful result: An (inconsistent) sequence of probabilistic policies $\tilde\pi_k$ suffices to prove the existence of a (consistent) deterministic policy $p^\xi$. A result similar to Theorem 4($i$) also holds for the discounted case, roughly saying that $V^{\tilde\pi}-V^*=O(\Delta(k))$ implies $V^{p^\xi}-V^*=\frac{1}{\varepsilon}O(\Delta(k))$ with $\mu$ probability $1-\varepsilon$ for finite $\mathcal{M}$.

**Proof:** We define $\delta_\nu(k):=\Delta_k^\nu=V_{k\gamma}^{*\nu}-V_{k\gamma}^{\tilde\pi\nu}$. Since we assumed bounded rewards $0\le r\le r_{max}$ and $V_{k\gamma}^{*\nu}$ is a weighted average of rewards we have

$$V_{k\gamma}^{*\mu}\le r_{max} \quad\text{and}\quad V_{k\gamma}^{\tilde\pi\mu}\ge 0 \quad\Rightarrow\quad 0\le\delta_\nu(k)=\Delta_k^\nu\le c:=r_{max}.$$

The following inequalities follow from Lemma 3:

$$0 \;\le\; V_{k\gamma}^{*\mu}-V_{k\gamma}^{p^\xi\mu} \;\le\; \tfrac{1}{w_k^\mu}\Delta_k \;=\; \tfrac{1}{w_k^\mu}\delta(k) \;\stackrel{?}{\to}\; 0 \tag{10}$$

The premise in Theorem 7 is that $\delta_\nu(k)=V_{k\gamma}^{*\nu}-V_{k\gamma}^{\tilde\pi\nu}\to 0$ for $k\to\infty$ which implies $\delta(k)\to 0$ (w.$\mu$.p.1) by Lemma 2($ii$). What is new and what remains to be shown is that $w_k^\mu$ is bounded from below in order to have convergence of (10) to zero. We show that $z_{k-1}:=\frac{w_\mu}{w_k^\mu}=\frac{\xi(\underline{yx}_{<k})}{\mu(\underline{yx}_{<k})}\ge 0$ converges to a finite value, which completes the proof. Let $\mathbf{E}$ denote the $\mu$ expectation. Then

$$\mathbf{E}[z_k|x_{<k}]=\sum_{x_k}{}'\mu(yx_{<k}\underline{yx}_k)\frac{\xi(\underline{yx}_{1:k})}{\mu(\underline{yx}_{1:k})}=\frac{\sum'_{x_k}\xi(yx_{<k}\underline{yx}_k)\xi(\underline{yx}_{<k})}{\mu(\underline{yx}_{<k})}\le\frac{\xi(\underline{yx}_{<k})}{\mu(\underline{yx}_{<k})}=z_{k-1}$$

$\sum'_{x_k}$ runs over all $x_k$ with $\mu(\underline{yx}_{1:k})\neq 0$. The first equality holds w.$\mu$.p.1. In the second equality we have used Bayes rule twice. $\mathbf{E}[z_k|x_{<k}]\le z_{k-1}$ shows that $-z_k$ is a semi-martingale. Since $-z_k$ is non-positive, [Doo53, Th.4.1$s(i)$,p324] implies that $-z_k$ converges for $k\to\infty$ to a finite value w.$\mu$.p.1. □

# 6 Markov Decision Processes

From all possible environments, Markov (decision) processes are probably the most intensively studied ones. To give an example, we apply Theorems 4 and 7 to ergodic Markov decision processes, but we will be very brief.

**Definition 3 (Ergodic Markov Decision Processes)** *We call $\mu$ a (stationary)* Markov Decision Process (MDP) *if the probability of observing $x_k \in \mathcal{X}$, given history $yx_{<k}y_k$ does only depend on the last action $y_k \in \mathcal{Y}$ and the last observation $x_{k-1}$, i.e. if $\mu(yx_{<k}y_k\underline{x}_k) = \mu(yx_{k-1}\underline{x}_k)$. In this case $x_k$ is called a* state*, $\mathcal{X}$ the* state space*, and $\mu(yx_{k-1}\underline{x}_k)$ the* transition matrix*. An* MDP *$\mu$ is called* ergodic *if there exists a policy under which every state is visited infinitely often with probability 1. Let $\mathcal{M}_{MDP}$ be the set of* MDP*s and $\mathcal{M}_{MDP1}$ be the set of ergodic* MDP*s. If an* MDP *$\mu(yx_{k-1}\underline{x}_k)$ is independent of the action $y_{k-1}$ it is a* Markov process*, if it is independent of the last observation $x_{k-1}$ it is an* i.i.d. *process.*

Stationary MDP*s* $\mu$ have stationary optimal policies $p^\mu$ mapping the same state / observation $x_t$ always to the same action $y_t$. On the other hand a mixture $\xi$ of MDP*s* is itself not an MDP, i.e. $\xi \notin \mathcal{M}_{MDP}$, which implies that $p^\xi$ is, in general, not a stationary policy. The definition of ergodicity given here is least demanding, since it only demands on the existence of a single policy under which the Markov process is ergodic. Often, stronger assumptions, e.g. that every policy is ergodic or that a stationary distribution exists, are made. We now show that there are self-optimizing policies for the class of ergodic MDP*s* in the following sense.

**Theorem 8 (Self-optimizing policies for ergodic** MDP*s*) *There exist self-optimizing policies $\tilde{p}_m$ for the class of ergodic* MDP*s in the sense that*

$$i) \quad \exists \tilde{p}_m \forall \nu \in \mathcal{M}_{MDP1} : \tfrac{1}{m}V^{*\nu}_{1m} - \tfrac{1}{m}V^{\tilde{p}_m\nu}_{1m} \;\leq\; c_\nu m^{-1/3} \;\stackrel{m\to\infty}{\longrightarrow}\; 0,$$

*where $c_\nu$ are some constants. In the discounted case, if the discount sequence $\gamma_k$ has unbounded effective horizon $h^{eff}_k \stackrel{k\to\infty}{\longrightarrow} \infty$, then there exist self-optimizing policies $\tilde{\pi}_k$ for the class of ergodic* MDP*s in the sense that*

$$ii) \quad \exists \tilde{\pi}_k \forall \nu \in \mathcal{M}_{MDP1} : V^{\tilde{\pi}_k\nu}_{k\gamma} \;\stackrel{k\to\infty}{\longrightarrow}\; V^{*\nu}_{k\gamma} \quad \text{if} \;\; \tfrac{\gamma_{k+1}}{\gamma_k} \to 1.$$

There is much literature on constructing and analyzing self-optimizing learning algorithms in MDP environments. The assumptions on the structure of the MDP*s* vary, all include some form of ergodicity, often stronger than Definition 3, demanding that the Markov process is ergodic under *every* policy. See, for instance, [KV86, Ber95]. We will only briefly outline one algorithm satisfying Theorem 8 without trying to optimize performance.

**Proof idea:** For $(i)$ one can choose a policy $\tilde{p}_m$ which performs (uniformly) random actions in cycles $1...k_0-1$ with $1 \ll k_0 \ll m$ and which follows thereafter the optimal policy based on an estimate of the transition matrix $T^a_{ss'} \equiv \nu(as\underline{s}')$ from the initial $k_0-1$ cycles. The existence of an ergodic policy implies that for every pair of states $s_{start}, s \in \mathcal{X}$ there is a sequence of actions and transitions of length at most $|\mathcal{X}|-1$ such that state $s$ is reached from state $s_{start}$. The probability that the "right" transition occurs is at least $T_{min}$ with $T_{min}$ being the smallest non-zero transition probability in $T$. The probability that a random action is the "right" action is at least $|\mathcal{Y}|^{-1}$. So the probability of reaching a state $s$ in $|\mathcal{X}|-1$ cycles via a random policy is at least $(T_{min}/|\mathcal{Y}|)^{|\mathcal{X}|-1}$. In state $s$ action $a$ is taken with probability $|\mathcal{Y}|^{-1}$ and leads to state $s'$ with probability $T^a_{ss'} \geq$

$T_{min}$. Hence, the expected number of transitions $s \xrightarrow{a} s'$ to occur in the first $k_0$ cycles is $\geq \frac{k_0}{|\mathcal{X}|}(T_{min}/|\mathcal{Y}|)^{|\mathcal{X}|} \sim k_0$.[6] The accuracy of the frequency estimate $\hat{T}^a_{ss'}$ of $T^a_{ss'}$ hence is $\sim k_0^{-1/2}$. Similar MDP*s* lead to "similar" optimal policies, which lead to similar values. More precisely, one can show that $\hat{T}-T \sim k_0^{-1/2}$ implies the same accuracy in the average value, i.e. $|\frac{1}{m}V^{\tilde{p}_m\nu}_{k_0 m} - \frac{1}{m}V^{*\nu}_{k_0 m}| \sim k_0^{-1/2}$, where $\tilde{p}_m$ is the optimal policy based on $\hat{T}$ and $*$ is the optimal policy based on $T(=\nu)$. Since $\frac{1}{m}V_{1k_0} \sim \frac{k_0}{m}$, $(i)$ follows (with probability 1) by setting $k_0 \sim m^{2/3}$. The policy $\tilde{p}_m$ can be derandomized, showing $(i)$ for sure.

The discounted case $(ii)$ can be proven similarly. The history $yx_{<k}$ is simply ignored and the analogue to $m \to \infty$ is $h_k^{eff} \to \infty$ for $k \to \infty$, which is ensured by $\frac{\gamma_{k+1}}{\gamma_k} \to \infty$. Let $\tilde{\pi}_k$ be the policy which performs (uniformly) random actions in cycles $k...k_0-1$ with $k \ll k_0 \ll h_k^{eff}$ and which follows thereafter the optimal policy[7] based on an estimate $\hat{T}$ of the transition matrix $T$ from cycles $k...k_0-1$. The existence of an ergodic policy, again, ensures that the expected number of transitions $s \xrightarrow{a} s'$ occurring in cycles $k...k_0-1$ is proportional to $\Delta := k_0 - k$. The accuracy of the frequency estimate $\hat{T}$ of $T$ is $\sim \Delta^{-1/2}$ which implies

$$V^{\tilde{\pi}_k\nu}_{k_0\gamma} \to V^{*\nu}_{k_0\gamma} \quad \text{for} \quad \Delta = k_0 - k \to \infty, \tag{11}$$

where $\tilde{\pi}_k$ is the optimal policy based on $\hat{T}$ and $*$ is the optimal policy based on $T(=\nu)$. It remains to show that the achieved reward in the random phase $k...k_0-1$ gives a negligible contribution to $V_{k\gamma}$. The following implications for $k \to \infty$ are easy to show:

$$\frac{\gamma_{k+1}}{\gamma_k} \to 1 \;\Rightarrow\; \frac{\gamma_{k+\Delta}}{\gamma_k} \to 1 \;\Rightarrow\; \frac{\Gamma_{k+\Delta}}{\Gamma_k} \to 1 \;\Rightarrow\; \frac{1}{\Gamma_k}\sum_{i=k}^{k_0-1}\gamma_i r_i \leq \frac{r_{max}}{\Gamma_k}[\Gamma_{k+\Delta} - \Gamma_k] \to 0.$$

Since convergence to zero is true for all fixed finite $\Delta$ it is also true for sufficiently slowly increasing $\Delta(k) \to \infty$. This shows that the contribution of the first $\Delta$ rewards $r_k + ... + r_{k_0-1}$ to $V_{k\gamma}$ is negligible. Together with (11) this shows $V^{\tilde{\pi}_k\nu}_{k\gamma} \to V^{*\nu}_{k\gamma}$ for $k_0 := k + \Delta(k)$.□

The conditions $\Gamma_k < \infty$ and $\frac{\gamma_{k+1}}{\gamma_k} \to 1$ on the discount sequence are, for instance, satisfied for $\gamma_k = 1/k^2$, so the Theorem is not vacuous. The popular geometric discount $\gamma_k = \gamma^k$ fails the latter condition; it has finite effective horizon. [Hut00] gives a detailed account on discount and horizon issues, and motivates $h_k^{eff} \to \infty$ philosophically.

Together with Theorems 4 and 7, Theorem 8 immediately implies that policy $p^\xi$ is self-optimizing for the class of to ergodic MDP*s*.

**Corollary 1 (Policy $p^\xi$ is self-optimizing for ergodic MDP*s*)** *If $\mathcal{M}$ is a finite or countable class of ergodic* MDP*s, and $\xi() := \sum_{\nu\in\mathcal{M}} w_\nu \nu()$, then policies $p^\xi_m$ maximizing $V^{p\xi}_{1m}$ and $p^\xi$ maximizing $V^{\pi\xi}_{k\gamma}$ are self-optimizing in the sense that*

$$\forall \nu \in \mathcal{M} : \tfrac{1}{m}V^{p^\xi_m\nu}_{1m} \overset{m\to\infty}{\longrightarrow} \tfrac{1}{m}V^{*\nu}_{1m} \quad \textit{and} \quad V^{p^\xi\nu}_{k\gamma} \overset{k\to\infty}{\longrightarrow} V^{*\nu}_{k\gamma} \quad \textit{if} \;\; \tfrac{\gamma_{k+1}}{\gamma_k} \to 1.$$

*If $\mathcal{M}$ is finite, then the speed of the first convergence is at least $O(m^{-1/3})$.*

[6] For $T^a_{ss'} = 0$ the estimate $\hat{T}^a_{ss'} = 0$ is exact.

[7] For non-geometric discounts as here, optimal policies are, in general, *not* stationary.

# 7 Conclusions

**Summary:** We studied agents acting in general probabilistic environments with reinforcement feedback. We only assumed that the true environment $\mu$ belongs to a known class of environments $\mathcal{M}$, but is otherwise unknown. We showed that the Bayes-optimal policy $p^\xi$ based on the Bayes-mixture $\xi=\sum_{\nu\in\mathcal{M}}w_\nu\nu$ is Pareto-optimal and self-optimizing if $\mathcal{M}$ admits self-optimizing policies. The class of ergodic MDPs admitted self-optimizing policies w.r.t. the average value and w.r.t. the discounted value if the effective horizon grew indefinitely.

**Continuous classes $\mathcal{M}$:** There are uncountably many (ergodic) MDPs. Since we have restricted our development to countable classes $\mathcal{M}$ we had to give the Corollary for a countable subset of $\mathcal{M}_{MDP1}$. We may choose $\mathcal{M}$ as the set of all ergodic MDPs with rational (or computable) transition probabilities. In this case $\mathcal{M}$ is a dense subset of $\mathcal{M}_{MDP1}$ which is, from a practical point of view, sufficiently rich. On the other hand, it is possible to extend the theory to continuously parameterized families of environments $\mu_\theta$ and $\xi=\int w_\theta\mu_\theta\,d\theta$. Under some mild (differentiability and existence) conditions, most results of this work remain valid in some form, especially Corollary 1 for *all* ergodic MDPs.

**Bayesian self-optimizing policy:** Policy $p^\xi$ with unbounded effective horizon for ergodic MDPs is the first purely Bayesian self-optimizing consistent policy for ergodic MDPs. The policies of all previous approaches were either hand crafted, like the ones in the proof of Theorem 8, or were Bayesian with a pre-chosen horizon $m$, or with geometric discounting $\gamma$ with finite effective horizon (which does not allow self-optimizing policies) [KV86, Ber95]. The combined conditions $\Gamma_k<\infty$ and $\frac{\gamma_{k+1}}{\gamma_k}\to 1$ allow a consistent self-optimizing Bayes-optimal policy based on mixtures.

**Bandits:** Bandits are a special subclass of ergodic MDPs. In a two-armed bandit problem you pull repeatedly one of two levers resulting in a gain of A\$1 with probability $p_i$ for arm number $i$. The game can be described as an MDP with parameters $p_i$. If the $p_i$ are unknown, Corollary 1 shows that policy $p^\xi$ yields asymptotically optimal payoff. The discounted unbounded horizon approach and result is, to the best of our knowledge, even new when restricted to Bandits.

**Other environmental classes:** Bandits, i.i.d. processes, classification tasks, and many more are all special (degenerate) cases of ergodic MDPs, for which Corollary 1 shows that $p^\xi$ is self-optimizing. But the existence of self-optimizing policies is not limited to (subclasses of ergodic) MDPs. Certain classes of POMDPs, $k^{th}$ order ergodic MDPs, factorizable environments, repeated games, and prediction problems are not MDPs, but nevertheless admit self-optimizing policies (to be shown elsewhere), and hence the corresponding Bayes-optimal mixture policy $p^\xi$ is self-optimizing by Theorems 4 and 7.

**Outlook:** Future research could be the derivation of non-asymptotic bounds, possibly along the lines of [Hut01]. To get good bounds one may have to exploit extra properties of the environments, like the mixing rate of MDPs [KS98]. Another possibility is to search for other performance criteria along the lines of [Hut00, Ch.6], especially for the universal prior [Sol78] and for the Speed prior [Sch02]. Finally, instead of convergence of the expected reward sum, studying convergence with high probability of the actual reward sum would be interesting.

# References


[Bel57] R. Bellman. *Dynamic Programming*. Princeton University Press, New Jersey, 1957.

[Ber95] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. (I) and (II)*. Athena Scientific, Belmont, Massachusetts, 1995. Volumes 1 and 2.

[BT00] R. I. Brafman and M. Tennenholtz. A near-optimal polynomial time algorithm for learning in certain classes of stochastic games. *Artificial Intelligence*, 121(1–2):31–47, 2000.

[Doo53] J. L. Doob. *Stochastic Processes*. John Wiley & Sons, New York, 1953.

[Hut00] M. Hutter. A theory of universal artificial intelligence based on algorithmic complexity. Technical Report cs.AI/0004001, 62 pages, 2000. http://arxiv.org/abs/cs.AI/0004001.

[Hut01] M. Hutter. General loss bounds for universal sequence prediction. *Proceedings of the $18^{th}$ International Conference on Machine Learning (ICML-2001)*, pages 210–217, 2001.

[KLM96] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: a survey. *Journal of AI research*, 4:237–285, 1996.

[KS98] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. In *Proc. 15th International Conf. on Machine Learning*, pages 260–268. Morgan Kaufmann, San Francisco, CA, 1998.

[KV86] P. R. Kumar and P. P. Varaiya. *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice Hall, Englewood Cliffs, NJ, 1986.

[LV97] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2nd edition, 1997.

[RN95] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, 1995.

[SB98] R. Sutton and A. Barto. *Reinforcement learning: An introduction*. Cambridge, MA, MIT Press, 1998.

[Sch02] J. Schmidhuber. The Speed Prior: a new simplicity measure yielding near-optimal computable predictions. *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002)*, 2002.

[Sol78] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Inform. Theory*, IT-24:422–432, 1978.