Limits of Multi–Discounted Markov Decision Processes

Hugo Gimbert
LIX, École Polytechnique
Palaiseau, France
gimbert@lix.polytechnique.fr

Wiesław Zielonka
LIAFA, Université Denis Diderot
Paris, France
zielonka@liafa.jussieu.fr

Abstract

Markov decision processes (MDPs) are controllable discrete event systems with stochastic transitions. The payoff received by the controller can be evaluated in different ways, depending on the payoff function the MDP is equipped with. For example a mean–payoff *function evaluates average performance, whereas a* discounted *payoff function gives more weights to earlier performance by means of a discount factor. Another well–known example is the* parity *payoff function which is used to encode logical specifications [14].*

Surprisingly, parity and mean–payoff MDPs share two non–trivial properties: they both have pure stationary optimal strategies [4, 15] and they both are approximable by discounted MDPs with multiple discount factors (multi–discounted MDPs) [5, 15].

In this paper we unify and generalize these results. We introduce a new class of payoff functions called the priority weighted payoff functions, which are generalization of both parity and mean–payoff functions. We prove that priority weighted MDPs admit optimal strategies that are pure and stationary, and that the priority weighted value of an MDP is the limit of the multi–discounted value when discount factors tend to 0 *simultaneously at various speeds.*

1. Introduction

Markov decision processes (MDPs) are controllable discrete event systems with stochastic transitions. An MDP evolves through an infinite sequence of stages. At each stage, the system is in some state and the controller chooses an action among several available. Together with the current state, this action determines transition probabilities to the new state of the system. A payoff function associates with each infinite sequence of states a real number called the payoff, and the controller seeks to maximize his expected payoff.

Various payoff functions define various kind of MDPs, such as mean–payoff, discounted or parity MDPs. In a mean–payoff MDP a real number called the reward is associated with each state and the controller seeks to maximize the average value of the infinite stream of rewards. In a discounted MDP, this stream of rewards is evaluated according to the principle "the sooner, the better" by means of a discount factor $0 < \mu \leq 1$. Another well–known example is the parity payoff function which is used to encode logical specifications [14]. Different attempts to combine both *quantitative* aspects of the discounted and mean–payoff functions and *qualitative* aspects of the parity payoff function led to the definition of the discounted μ–calculus [6], the mean–payoff parity function [3] and the priority mean–payoff function [12].

Surprisingly, parity and mean–payoff MDPs share two common properties: they both have pure stationary optimal strategies and they both are approximable by multi–discounted MDPs.

Existence of pure stationary strategies in parity and mean–payoff MDPs is a non–trivial property, for two reasons. First in general only the existence of ϵ–optimal strategies is guaranteed and second in numerous examples the controller needs memory to play optimally.

A second property shared by the mean–payoff and the parity MDPs is their approximability by discounted MDPs. In the case of the mean–payoff MDPs a well–known results (see [15] for example) establishes that when the discount factor of a discounted MDP tends to 0, the value of the discounted MDP converges to the mean–payoff value. For approximating parity MDPs, it is necessary to consider multi–discounted MDPs with multiple discount factors (multi–discounted MDPs). In that case discount factors may converge to 0 in various ways. A first type of convergence is convergence of the various discount factors to 0 *one after another*. In that case, results of [5] imply that when the rewards are either 0 or 1 then the value of the multi–discounted MDP converges to the parity value. A second type of convergence is *simultaneous* convergence of the discount factors to 0, at various speeds. A special case of such simultaneous convergence is geometric conver-

gence where each discount factor converges to 0 at geometric speed $\left(\frac{1}{n^{k+1}}\right)_{n\in\mathbb{N}}$ for some $k \in \mathbb{N}$. In the non–stochastic case, [13] establishes that the multi–discounted value converges to the parity value.

In fact, [13] goes further and unify these results about parity and mean–payoff MDPs, in the non–stochastic case. That unification is realized by means of the class of priority mean–payoff functions, which are generalizations of both parity and mean–payoff functions. The two following results hold. First, one–player priority mean–payoff games on graphs admit pure stationary optimal strategies. Second the priority mean–payoff value is the limit of the multi–discounted value, When discount factors converge to 0 one after another or also simultaneously at various geometric speeds $\left(\frac{1}{n^{k+1}}\right)_{n\in\mathbb{N}}$ for some $k \in \mathbb{N}$.

These last results suffer restrictions that leave some questions open. First restriction is technical: only the non–stochastic case is considered, hence it is natural to wonder whether part of results of [13] about (one–player) games on graphs are extendable to MDPs? Second restriction is about the convergence of discount factors: only the case of "one after another" or geometric convergence to 0 is considered. It is natural to wonder what happens when discount factors tend to 0 at various speeds, with no restriction on these speeds: does the multi–discounted value converges? Towards which value?

In this paper, we give answers to these three questions. For that purpose we introduce the class of *priority weighted* MDPs which generalize priority and mean–payoff functions.

1. We prove that priority weighted MDPs admit pure stationary optimal strategies (Theorem 2).

2. We prove that the priority weighted value of an MDP is the limit of the multi–discounted value when discount factors converge simultaneously to 0, at comparable speeds (Theorem 6).

3. Moreover, we prove that in some sense, the class of priority weighted MDPs is the most general class of MDPs whose values are approximable by the multi–discounted value when discount factors tend to 0 simultaneously (Theorem 6).

These results have several interests. First one is algorithmic: priority weighted mean–payoff MDPs provide a new example of MDPs whose values are computable (cf. Section 7). Second is theoretic: we extend several results about existence of pure stationary optimal strategies [4, 15] or limits of multi–discounted MDPs [15, 5, 13].

2. Markov Decision Processes

Notations. By $\mathbb{R}_{>0}$ and $\mathbb{R}_{\geq 0}$ we denote respectively the sets of strictly positive and non–negative real numbers. If $\mathbf{S}$ is a finite set then by $\mathbf{S}^*$ and $\mathbf{S}^\omega$ we denote respectively the sets of finite and infinite words on $\mathbf{S}$, and we denote $\mathcal{D}(\mathbf{S})$ the set of probability distributions on $\mathbf{S}$, i.e. $\mathcal{D}(\mathbf{S}) = \{\delta : \mathbf{S} \to \mathbb{R} \mid \forall s \in \mathbf{S},\ 0 \leq \delta(s) \leq 1 \text{ and } \sum_{s\in\mathbf{S}} \delta(s) = 1\}$.

We fix a finite set of states $\mathbf{S}$ for the rest of this paper, and call it the set of *states*.

Definition 1 (Markov chains and controllable Markov chains). *A* controllable Markov chain $\mathcal{A} = (\mathbf{S}, \mathbf{A}, (\mathbf{A}(s))_{s\in\mathbf{S}}, p)$ *consists of a finite set of states* $\mathbf{S}$ *and a finite set of actions* $\mathbf{A}$. *For each state* $s \in \mathbf{S}$, $\mathbf{A}(s) \subseteq \mathbf{A}$ *is the set of actions available at* s. *For each* $s, t \in \mathbf{S}$ *and* $a \in \mathbf{A}(s)$, $p(t|s,a)$ *is the conditional probability to go to state* t *from state* s *upon the execution of action* a. *In the special case where, for each* $s \in \mathbf{S}$ $\mathbf{A}(s)$ *is a singleton,* $\mathcal{A}$ *is called a* Markov chain.

Intuitively, a controllable Markov chain evolves in discrete steps. At each step, the chain is in some state s and the controller chooses an available action $a \in \mathbf{A}(s)$. Then the state changes to state t with probability $p(t|s,a)$. For choosing his actions the controller uses a strategy:

Definition 2 (Finite histories and strategies). *A finite history is a finite sequence* $h = s_0 a_1 s_1 \cdots s_n \in \mathbf{S}(\mathbf{AS})^*$ *such that, for each* $0 \leq i < n$ $a_{i+1} \in \mathbf{A}(s_i)$. *States* s_0 *and* s_n *are respectively the source and the target of* h. *The set of finite histories is denoted* $\mathcal{H}^*_{\mathcal{A}}$ *or* $\mathcal{H}^*_{\mathcal{A},s}$ *if we fix the source* s. *A* strategy *is a mapping* $\sigma : \mathcal{H}^*_{\mathcal{A}} \to \mathcal{D}(\mathbf{A})$ *such that, for any finite history* $h \in \mathcal{H}^*_{\mathcal{A}}$ *with target* t *the distribution* $\sigma(h)$ *puts non–zero probabilities only on actions available in* t, *i.e.* $\in \mathbf{S}$, $\sigma(h) \in \mathcal{D}(\mathbf{A}(t))$.

Thus, in general, the actions prescribed by a strategy depend on the entire history and the strategy uses randomization. On the other hand, a strategy is said to be *pure* if actions are chosen deterministically, i.e. for each finite history h and each action a either $\sigma(h)(a) = 0$ or $\sigma(h)(a) = 1$. A strategy σ is *stationary* if for any two finite histories $h, h' \in \mathcal{H}^*_{\mathcal{A}}$ with same target state t, $\sigma(h) = \sigma(h')(= \sigma(t))$. We can identify pure stationary strategies with mappings $\sigma : \mathbf{S} \to \mathbf{A}$.

Let us fix a strategy σ. Then, intuitively, the probability of the finite history $s_0 a_1 \cdots a_n s_n$ is $\sigma(s_0)(a_1) \cdot p(s_1|s_0,a_1) \cdots \sigma(s_0 \cdots s_{n-1})(a_n) \cdot p(s_n|s_{n-1},a_n)$.

Definition 3 (Probability measure induced by a strategy). *An infinite history is an infinite sequence* $h = s_0 a_1 s_1 \cdots \in \mathbf{S}(\mathbf{AS})^\omega$ *such that, for each* n, $a_{n+1} \in \mathbf{A}(s_n)$. *The set of infinite histories with source* s *is denoted* $\mathcal{H}^\omega_{\mathcal{A},s}$. *It is equipped with the* σ*-field generated by the random variables* $S_n, A_n, n \in \mathbb{N}$, *where* $S_n(s_0 a_1 s_1 \cdots) = s_n$

and $A_n(s_0a_1s_1\cdots) = a_n$. *In the sequel, a measurable set of infinite paths will be called an* event. *According to a theorem of Ionescu Tulcea [1], for each strategy* σ, *there exists a unique probability measure* $\mathbb{P}_s^\sigma$ *on* $\mathcal{H}_{\mathcal{A},s}^\omega$ *such that, for each finite history* $s_0a_1s_1\cdots s_na_{n+1}s_{n+1} \in \mathcal{H}_{\mathcal{A},s}^*$, $\mathbb{P}_s^\sigma(s_0) = 1$, $\mathbb{P}_s^\sigma(A_{n+1} = a_{n+1} \mid S_0A_1\cdots S_n = s_0a_1\cdots s_n) = \sigma(s_0a_1\cdots s_n)(a)$ *and* $\mathbb{P}_s^\sigma(S_{n+1} = s_{n+1} \mid S_0A_1\cdots S_nA_{n+1} = s_0a_1\cdots s_na_{n+1}) = p(s_{n+1}|s_n, a_{n+1})$. *The expectation of a real–valued random variable* X *under the probability measure* $\mathbb{P}_s^\sigma$ *is denoted* $\mathbb{E}_s^\sigma[X]$.

After an infinite history the controller gets some payoff, which is computed by a payoff function ϕ. Once the controllable Markov chain $\mathcal{A}$ and the payoff function ϕ are chosen, we obtain a *Markov decision process* $(\mathcal{A}, \phi)$ in which the controller seeks to use strategies which maximize his expected payoff:

Definition 4 (Payoff functions, Markov decision processes, expected payoff, value of a state and optimal strategies). *A payoff function is a bounded measurable function* $\phi : \mathbf{S}^\omega \to \mathbb{R}$. *A* Markov decision process *is a couple* $\mathcal{M} = (\mathcal{A}, \phi)$ *where* $\mathcal{A}$ *is a controllable Markov chain and* ϕ *is a payoff function. Let* $s \in \mathbf{S}$ *be a state and* σ *a strategy. The* expected payoff *under probability* $\mathbb{P}_s^\sigma$ *is* $\mathbb{E}_s^\sigma[\phi(S_0S_1\cdots)]$ *and will often be denoted* $\mathbb{E}_s^\sigma[\phi]$. *The* value of a state $s \in \mathbf{S}$ *in the MDP* $\mathcal{M}$ *is the supremum of expected payoffs over all strategies:* $\mathrm{val}(\mathcal{M})(s) = \sup_\sigma \mathbb{E}_s^\sigma[\phi]$. *A strategy* σ *is said to be* optimal *if* $\forall s \in \mathbf{S},\ \mathbb{E}_s^\sigma[\phi] = \mathrm{val}(\mathcal{M})(s)$.

In this paper, we are interested in properties of MDPs $(\mathcal{A}, \phi)$ that hold for a fixed payoff function ϕ, *independently of the controllable Markov chain* $\mathcal{A}$. In particular, we are interested in the convergence of the values of MDPs. For that purpose, it will be convenient to use the following notions of convergence:

Definition 5 (MC–convergence and MDP–convergence). *Let* ϕ_∞ *be a payoff function and* $(\phi_n)_{n\in\mathbb{N}}$ *be a sequence of payoff functions. We say that* $(\phi_n)_{n\in\mathbb{N}}$ MDP–converges *(resp. MC–converges) to* ϕ_∞ *if for any Markov decision process (resp. any Markov chain)* $\mathcal{A}$ *and any state* s *of* $\mathcal{A}$, $\mathrm{val}(\mathcal{A}, \phi_n)(s) \longrightarrow_n \mathrm{val}(\mathcal{A}, \phi_\infty)(s)$.

We will use later the fact that for the special class of payoff functions that ensure existence of pure positional optimal strategies, notions of MC–convergence and MDP–convergence coincide:

Proposition 1 (Equivalence of MC–convergence and MDP–convergence). *Let* ϕ_∞ *be a payoff function and* $(\phi_n)_{n\in\mathbb{N}}$ *be a sequence of payoff functions. Suppose that for each* $n \in \mathbb{N} \cup \{\infty\}$ *and each controllable Markov chain* $\mathcal{A}$ *there exists pure stationary optimal strategies in the MDP* $(\mathcal{A}, \phi_n)$. *Then* $(\phi_n)_{n\in\mathbb{N}}$ *MDP–converges to* ϕ_∞ *if and only if* $(\phi_n)_{n\in\mathbb{N}}$ *MC–converges to* ϕ_∞.

Proof. Since Markov chains are special cases of MDPs, MDP–convergence implies MC–convergence. Conversely, suppose that $(\phi_n)_{n\in\mathbb{N}}$ MC–converges to ϕ_∞. We prove that $(\phi_n)_{n\in\mathbb{N}}$ MDP–converges to ϕ_∞. Let $\mathcal{A} = (\mathbf{S}, \mathbf{A}, (\mathbf{A}(s))_{s\in\mathbf{S}}, p)$ be a controllable Markov chain. Let Σ_{ps} be the set of pure stationary strategies for $\mathcal{A}$. If we fix a pure stationary strategy $\sigma : \mathbf{S} \to \mathbf{A}$ for $\mathcal{A}$ we obtain a Markov chain that we denote $\mathcal{A}[\sigma]$. Since $(\phi_n)_{n\in\mathbb{N}}$ MC–converges to ϕ_∞, $\mathbb{E}_s^\sigma[\phi_n] = \mathrm{val}(\mathcal{A}[\sigma], \phi_n)(s) \longrightarrow_n \mathrm{val}(\mathcal{A}[\sigma], \phi_\infty)(s) = \mathbb{E}_s^\sigma[\phi_\infty]$. Since Σ_{ps} is finite and by existence of pure stationary optimal strategies this implies $\mathrm{val}(\mathcal{A}, \phi_n) = \max_{\sigma\in\Sigma_{ps}} \mathbb{E}_s^\sigma[\phi_n] \longrightarrow_n \max_{\sigma\in\Sigma_{ps}} \mathbb{E}_s^\sigma[\phi_\infty] = \mathrm{val}(\mathcal{A}, \phi_\infty)$. □

3. Discounted and mean–payoff MDPs

From this moment onward with each state s is associated a reward $r(s) \in \mathbb{R}$. Thus for an infinite history $s_0a_1s_1\cdots \in \mathcal{H}_\mathcal{A}^\omega$ we obtain an infinite sequence $r(s_0), r(s_1), \ldots$ of rewards. The objective of the controller is to maximize a specific evaluation of this sequence.

Discounted MDPs. In a discounted MDP, a sequence $r(s_0), r(s_1), \ldots$ of rewards is evaluated according to the principle "the sooner the better". Formally, we fix a discount factor $\mu \in]0, 1]$ and for $s_0s_1\cdots \in \mathbf{S}^\omega$ the value of the discounted payoff function is:

$$\mathrm{disc}_{r,\mu}(s_0s_1\cdots) = \sum_{n=0}^{\infty} \mu(1-\mu)^n r(s_n) \ . \qquad (1)$$

Mean–payoff MDPs. In mean–payoff MDPs we seek to maximize the *average value* of the sequence of rewards. This is done using the *mean–payoff* function:

$$\mathrm{mean}_r(s_0s_1\cdots) = \limsup_{n\in\mathbb{N}} \frac{1}{n+1} \sum_{i=0}^{n} r(s_i) \ . \qquad (2)$$

4. Priority weighted MDPs

4.1. Definition

Weighted MDPs. In a weighted MDP, with each state $s \in \mathbf{S}$ is associated not only a reward $r(s) \in \mathbb{R}$ but also a *weight* $w(s) \in \mathbb{R}_{>0}$. The weighted mean–payoff function $\mathrm{mean}_{w,r}$ evaluates an infinite sequence $s_0s_1\cdots \in \mathbf{S}^\omega$ as

follows:

$$\mathrm{mean}_{w,r}(s_0 s_1 \cdots) = \limsup_n \frac{1}{\sum_{i=0}^{n} w(s_i)} \sum_{i=0}^{n} w(s_i) r(s_i) \ . \tag{3}$$

Thus, the usual mean–payoff function of (2) is just a special case of (3) with each weight $w(s)$ equal to 1.

Priority weighted MDPs. In a priority weighted MDP, with each state s is associated not only a reward $r(s) \in \mathbb{R}$ and a weight $w(s) \in \mathbb{R}_{>0}$ but also a priority $c(s) \in \mathbb{N}$. The priority weighted payoff $\mathrm{mean}_{c,w,r}$ function computes payoff in several steps. Let $u = s_0 s_1 s_2 \cdots$. First we consider the highest priority occurring infinitely often in u:

$$\mathbf{c}(u) = \limsup_n c(s_n) \ .$$

Then we compute the set of indices of states whose priority is $\mathbf{c}(u)$:

$$N(u) = \{n \in \mathbb{N} \mid c(s_n) = \mathbf{c}(u)\} \ ,$$

and extract the corresponding sub–sequence of states:

$$\pi(u) = (s_n)_{n \in N(u)} \ . \tag{4}$$

Finally we apply the weighted mean–payoff function to this sub–sequence:

$$\mathrm{mean}_{c,w,r}(u) = \mathrm{mean}_{w,r}(\pi(u)) \ . \tag{5}$$

The weighted payoff function (3) is just a special case of (5) with each priority $c(s)$ equal to 0. Intuitively, weights and priorities are used to give respectively finitely or infinitely more weight to a state than to another. For example, if s and t have the same priority and $w(s) = 1$ whereas $w(t) = 2$ then occurrences of t will influence twice more the payoff (3) than occurrences of s. If s and t have different priorities, for example $c(s) = 0$ and $c(t) = 1$ and t occurs infinitely often then occurrences of s will not influence at all the payoff (5).

Parity MDPs are also a special case of priority weighted MDPs. In a parity MDP the payoff has value 0 if the highest priority seen infinitely often is even, and 1 if it is odd. Thus, parity MDPs are priority weighted MDPs where weights are constant equal to 1 and reward $r(s)$ is equal to 1 if $c(s)$ is odd and equal to 0 if $c(s)$ is even. The parity payoff function is a central tool in Model–checking used to encode logical specifications [14].

The priority weighted payoff function is not the only payoff function that generalizes both parity and mean–payoff function: recently also the mean–payoff parity [3] and the priority mean–payoff [12] functions have been considered. Since framework of [3, 12] is *two–player zero–sum games played on finite graphs*, the suitable framework to compare the present paper to [3, 12] is one–player games on finite graphs. Both the mean–payoff parity function and the priority mean–payoff functions are defined by mean of a reward mapping $r \in \mathbb{R}^{\mathbf{S}}$ and a priority mapping $c \in \mathbb{N}^{\mathbf{S}}$. The priority mean–payoff function is just a special case of the priority weighted payoff function with each weight $w(s)$ equal to 1. The mean–payoff parity payoff function is rather different: the controller seeks to satisfy the parity condition associated with c. If the controller succeeds then his payoff is the mean–payoff whereas if he fails he is heavily punished: his payoff is $-\infty$.

Although the mean–payoff parity and the priority weighted function are both generalizations of parity and mean–payoff functions they have radically different properties. The main difference is that using the mean–payoff parity function does not guarantee the existence of pure stationary optimal strategies, and the controller may even need an infinite amount of memory to play optimally [3]. On the other hand, the use of a priority mean–payoff function guarantees the existence of optimal strategies that are pure and stationary (cf. Theorem 2 in this paper or also [12]). Another difference between the mean–payoff parity and the priority mean–payoff function arises when we consider the stochastic framework. The mean–payoff parity function may have value $-\infty$ and as soon as this occurs with positive probability the expected payoff is $-\infty$. Hence, if there is a non–zero probability that the parity condition is violated, the controller of a mean–payoff parity MDP becomes totally indifferent to the mean–payoff evaluation of rewards. Such a phenomenon does not occur in a priority mean–payoff MDP. Thus when MDPs are used to model stochastic systems with both fairness assumption and quantitative constraints, using a priority mean–payoff function guarantees that the expected payoff always depends on both qualitative (parity) and quantitative (mean–payoff) aspects of the specification.

4.2. Pure stationary optimal strategies in priority mean–payoff MDPs

Although priority weighted payoff MDPs are a radical generalization of both mean–payoff and parity MDPs, we do not leave the comfortable framework of MDPs with pure and stationary optimal strategies:

Theorem 2. *In any priority weighted MDP there exists an optimal strategy which is pure and stationary.*

This result is well–known in the special cases of parity MDPs [4] or mean–payoff MDPs [15, 16, 2]. However, we could not adapt existing proofs to the case of priority weighted MDPs. Instead, we make use of a criterion for the existence of pure stationary optimal strategies in MDPs, established recently by the first author of this paper :

Theorem 3 ([11] A criterion for the existence of pure stationary optimal strategies in MDPs). *Let ϕ be a payoff function. Suppose that ϕ is* prefix–independent *i.e. for $u \in \mathbf{S}^*$ and $v \in \mathbf{S}^\omega$, $\phi(uv) = \phi(v)$ and that ϕ is* sub–mixing *i.e. for any sequence $u_1, v_1, u_2, v_2, \ldots \in \mathbf{S}^*$ of non–empty finite words, $\phi(u_0v_0u_1v_1\cdots) \leq \max\{\phi(u_0u_1\cdots), \phi(v_0v_1\cdots)\}$. Then in any MDP $(\mathcal{A}, \phi)$ there exists pure stationary optimal strategies.*

We can now prove Theorem 2:

Proof of Theorem 2. According to Theorem 3, it is enough to prove that $\mathrm{mean}_{c,w,r}$ is *prefix–independent* and *sub–mixing*.

Prefix–independency of $\mathrm{mean}_{c,w,r}$ is easy to establish. Let $u \in \mathbf{S}^*$ and $v \in \mathbf{S}^\omega$. Then (4) implies that $\pi(uv)$ and $\pi(v)$ differ only by a finite prefix. According to (3) $\mathrm{mean}_{w,r}$ is prefix–independent hence $\mathrm{mean}_{w,r}(\pi(uv)) = \mathrm{mean}_{w,r}(\pi(v))$ and finally $\mathrm{mean}_{c,w,r}(uv) = \mathrm{mean}_{c,w,r}(v)$.

Before proving that $\mathrm{mean}_{c,w,r}$ is sub–mixing, we first show that the weighted payoff function $\mathrm{mean}_{w,r}$ defined by (3) is sub–mixing. Let us extend the definition domain of $\mathrm{mean}_{w,r} : \mathbf{S}^\omega \to \mathbb{R}$ to finite words:

$$\mathrm{mean}_{w,r}(s_0s_1\cdots s_n) = \frac{1}{\sum_{i=0}^n w(s_i)} \sum_{i=0}^n w(s_i)r(s_i) \ .$$

Then for $s_0s_1\cdots \in \mathbf{S}^\omega$,

$$\mathrm{mean}_{w,r}(s_0s_1\cdots) = \limsup_n(\mathrm{mean}_{w,r}(s_0s_1\cdots s_n)). \tag{6}$$

Let $u_0, v_0, u_1, v_1, \ldots \in \mathbf{S}^\omega$ be a sequence of finite non–empty words over $\mathbf{S}$. Let $u, v, w \in \mathbf{S}^\omega$ defined by $u = u_0u_1\cdots$, $v = v_0v_1\cdots$ and $w = u_0v_0u_1v_1\cdots$. Let $(s_i)_{i\in\mathbb{N}} \in \mathbf{S}^\mathbb{N}$ be the sequence of states such that $w = s_0s_1\cdots$. Since word w is a shuffle of words u and v, there exists a partition (I_0, I_1) of $\mathbb{N}$ such that $u = (s_i)_{i\in I_0}$ and $v = (s_i)_{i\in I_1}$. Let $n \in \mathbb{N}$ and let $I_0^n = I_0 \cap \{0, \ldots, n\}$ and $I_1^n = I_1 \cap \{0, \ldots, n\}$. Then:

$$\begin{aligned}
&\mathrm{mean}_{w,r}(s_0s_1\cdots s_n) \\
&\quad= \frac{\sum_{i\in I_0^n} w(s_i)}{\sum_{i=0}^n w(s_i)} \mathrm{mean}_{w,r}((s_i)_{i\in I_0^n}) \\
&\quad\quad+ \frac{\sum_{i\in I_1^n} w(s_i)}{\sum_{i=0}^n w(s_i)} \mathrm{mean}_{w,r}((s_i)_{i\in I_1^n}) \\
&\quad\leq \max\left\{\mathrm{mean}_{w,r}((s_i)_{i\in I_0^n}), \mathrm{mean}_{w,r}((s_i)_{i\in I_1^n})\right\}.
\end{aligned}$$

The inequality holds since (I_0^n, I_1^n) is a partition of $\{0, \ldots, n\}$. Taking the superior limit of this inequality, we obtain $\mathrm{mean}_{w,r}(w) \leq \max\{\mathrm{mean}_{w,r}(u), \mathrm{mean}_{w,r}(v)\}$, which proves that $\mathrm{mean}_{w,r}$ is sub–mixing.

Now let us prove that $\mathrm{mean}_{c,w,r}$ is sub–mixing. Results of [11] could be directly used since $\mathrm{mean}_{c,w,r}$ is the priority product of d payoff functions of type $\mathrm{mean}_{w,r}$. Alternatively we give a direct proof. Let $u, v, w \in \mathbf{S}^\omega$, $(s_i)_{i\in\mathbb{N}}$ and (I_0, I_1) be as above. Let $c(w) = \limsup_{i\in\mathbb{N}} c(s_i)$, $c(u) = \limsup_{i\in I_0} c(s_i)$ and $c(v) = \limsup_{i\in I_1} c(s_i)$. Then $c(w) = \max\{c(u), c(v)\}$. If $c(w) = c(u) > c(v)$ then by (4), we have $\pi(w) = \pi(u)$. Then by definition of $\mathrm{mean}_{c,w,r}$, we deduce $\mathrm{mean}_{c,w,r}(w) = \mathrm{mean}_{c,w,r}(u)$. Symmetrically, in the case where $c(w) = c(v) > c(u)$ we get $\mathrm{mean}_{c,w,r}(w) = \mathrm{mean}_{c,w,r}(v)$. In the remaining case, $c(w) = c(u) = c(v)$ hence $\pi(w)$ is a shuffle of $\pi(u)$ and $\pi(v)$. Since $\mathrm{mean}_{w,r}$ is sub–mixing, $\mathrm{mean}_{c,w,r}(w) = \mathrm{mean}_{w,r}(\pi(w)) \leq \max\{\mathrm{mean}_{w,r}(\pi(u)), \mathrm{mean}_{w,r}(\pi(v))\} = \max\{\mathrm{mean}_{c,w,r}(u), \mathrm{mean}_{c,w,r}(v)\}$. Hence, in every three cases $\mathrm{mean}_{c,w,r}(w) = \max\{\mathrm{mean}_{c,w,r}(u), \mathrm{mean}_{c,w,r}(v)\}$, which achieves to prove that $\mathrm{mean}_{c,w,r}$ is sub–mixing. □

5. Multi–discounted MDPs.

In a multi–discounted MDP, each state $s \in \mathbf{S}$ is labelled with a discount factor $\mu(s) \in]0, 1]$. Let $s_0s_1\cdots \in \mathbf{S}^\omega$ and for $i \in \mathbb{N}$ let $\mu_i = \mu(s_i)$ and $r_i = r(s_i)$. Then the value of the multi–discounted payoff function is:

$$\begin{aligned}
&\mathrm{disc}_{r,\mu}(s_0s_1\cdots) \\
&\qquad= \sum_{n\in\mathbb{N}} (1-\mu_0)\cdots(1-\mu_{n-1})\mu_n r_n \ . \qquad (7)
\end{aligned}$$

Difference with the discounted payoff defined by (1) is that the discount factor is not fix but depends on the current state.

The use of multiple discount factors appeared with the seminal paper of Shapley [17] and was also considered in [5, 13]. Framework of [17] is different from this paper since Shapley considered two–player zero–sum stochastic games with stopping probabilities. However there exists a natural correspondence between multi–discounted MDPs and one–player Shapley game and this correspondence preserves expected payoff of strategies. Both the MDP and the corresponding one–player Shapley game have the same states, actions and transition probabilities. In the one–player Shapley game, the stopping probability in state s is $\mu(s)$ and the daily reward is $\mu(s)r(s)$. Using this correspondence, one of the results of [17] rephrases as:

Theorem 4 ([17]). *In any multi–discounted MDP, there exists an optimal strategy which is pure and stationary.*

6. Limits of multi–discounted MDPs

In this section, we establish that limits of multi–discounted MDPs are priority weighted MDPs (Theorem 6).

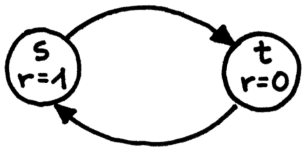

Figure 1. A simple MDP.

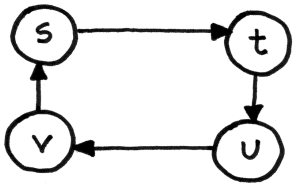

Figure 2. Another simple MDP.

Our starting point is the following well–known theorem:

Theorem 5 ([15]Limits of discounted MDPs are mean–payoff MDPs). *Let $r \in \mathbb{R}^{\mathbf{S}}$ a reward mapping and $(\mu_n)_{n\in\mathbb{N}}$ a sequence of discount factors that converges to 0. Then for any controllable Markov chain $\mathcal{A}$ and any state s:*

$$\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) \underset{n}{\longrightarrow} \mathrm{val}(\mathcal{A}, \mathrm{mean}_r)(s) \ .$$

We seek to generalize this result to the case where $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$ is a sequence of *multi–discounted* payoff functions, and $(\mu_n)_{n\in\mathbb{N}}$ is a sequence of discount mappings that converges to $0^{\mathbf{S}}$ (i.e for each state s, $(\mu_n(s))_{n\in\mathbb{N}}$ converges to 0).

The rest of this section is organized as follows. In sub–section 6.1 we analyze an example and obtain a necessary hypothesis to generalize Theorem 5: the convergence speeds to 0 of the various discount factors should be comparable, in the sense of (9). In sub–section 6.2, we analyze a second example and show how to compute priority weighted MDPs that are good candidates for being limits of multi–discounted MDPs. Finally in sub–section 6.3, we prove that condition (9) is not only necessary but also sufficient to obtain the generalization of Theorem 5 to the multi–discounted case.

6.1. Comparing convergence speeds of the discount factors

Do we need any extra hypothesis to generalize Theorem 5 to the multi–discounted case?

Consider the example $\mathcal{A}$ depicted on Fig. 1, and let us suppose that $(\mu_n)_{n\in\mathbb{N}}$ converges to $0^{\mathbf{S}}$ and $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$ MDP–converges in the sense of Def. 5. In $\mathcal{A}$ the controller has no choice and transitions are deterministic hence:

$$\begin{aligned}\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) &= \mathrm{disc}_{r,\mu}(ststs\cdots) \\ &= \frac{1}{1+\frac{\mu_n(s)}{\mu_n(t)}-\mu_n(t)} \ . \end{aligned} \quad (8)$$

Since we require the MDP–convergence of disc_{r,μ_n} then in particular (8) should converge when n tends to ∞. Moreover $\mu_n(t) \longrightarrow_n 0$, hence (8) converges if and only if:

$$\left(\frac{\mu_n(s)}{\mu_n(t)}\right)_{n\in\mathbb{N}} \text{ converges in } \mathbb{R}_{\geq 0} \cup \{+\infty\} \ . \quad (9)$$

In the above example, condition (9) is not only necessary but is also sufficient to obtain the MDP–convergence of $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$. It is not too hard to prove that more generally, for any controllable Markov chain $\mathcal{A}$ such that the controller has no choice and transitions are deterministic, $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$ MDP–converges if and only if (9) holds for any states $s, t \in \mathbf{S}$. In fact this equivalence remains true for any controllable Markov chain $\mathcal{A}$ as we will prove in Theorem 6.

6.2. Computing priorities and weights

Suppose now that the necessary hypothesis (9) is satisfied for each $s, t \in \mathbf{S}$ and that $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$ MDP–converges in the sense of Def. 5.

Then is there a good candidate for the MDP–limit of $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$ and can we compute it?

Let us analyse the example depicted in Fig.2, together with the sequence $(\mu_n)_{n\in\mathbb{N}}$ of discount mappings defined for each $n \in \mathbb{N}$ by $\mu_n(s) = \frac{2}{n}$, $\mu_n(t) = \frac{3}{n}$, $\mu_n(u) = \frac{4}{n^2}$ and $\mu_n(v) = \frac{5}{n^2}$. Let $v_n = \mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})$.

Again the controller has no choice and transitions are deterministic hence:

$$\begin{aligned} v_n &= \mathrm{disc}_{r,\mu_n}(stuvst\cdots) \\ &= \frac{\frac{2}{n}r(s)+\frac{3}{n}r(t)+\frac{4}{n^2}r(u)+\frac{5}{n^2}r(v)+o(\frac{1}{n^2})}{\frac{2}{n}+\frac{3}{n}+\frac{4}{n^2}+\frac{5}{n^2}+o(\frac{1}{n^2})} \quad (10) \\ &\underset{n}{\longrightarrow} \frac{2}{2+3}r(s)+\frac{3}{2+3}r(t) \ . \quad (11) \end{aligned}$$

Let us detail what happens in (10) when n tends to ∞. Discount factors $\mu_n(u) = \frac{4}{n^2}$ and $\mu_n(v) = \frac{5}{n^2}$ converge to 0 much faster than $\mu_n(s) = \frac{2}{n}$ and $\mu_n(t) = \frac{3}{n}$. Thus the weight of $r(u)$ and $r(v)$ in (10) vanishes when n tends to $+\infty$. On the other hand, convergence speeds of $\mu_n(s)$ and $\mu_n(t)$ are comparable hence neither $r(s)$ nor $r(t)$ vanishes. The respective weights of $r(s)$ and $r(t)$ in (10) even converge and $\lim_n v_n$, given by (11), is a convex combination of $r(s)$ and $r(t)$ whose weights are proportional to the speeds of convergence of $\mu_n(s)$ and $\mu_n(t)$ to 0.

Value (11) can be computed by means of a priority weighted payoff function $\mathrm{mean}_{c,w,r}$, using adequate priorities c and weights w. Since u and v are negligible compared

to s and t, we give to u and v low priorities $c(u) = c(v) = 0$ and to s and t high priorities $c(s) = c(t) = 1$. Then for each set of states with the same priority, we attribute weights that are proportional to the speeds of convergence to 0. For priority 1 for example, we set $w(s) = \frac{2}{2+3}$ and $w(t) = \frac{3}{2+3}$.

This way we have defined c and w such that:

$$\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) \underset{n}{\longrightarrow} \mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(s) \ . \quad (12)$$

Let us prove quickly that (12) holds. The value of state s in the MDP $(\mathcal{A}, \mathrm{mean}_{c,w,r})$ is $\mathrm{mean}_{c,w,r}(stuvst\cdots)$. Since u and v have priority strictly less than s and t, then according to (5), $\mathrm{mean}_{c,w,r}(stuvstu\cdots) = \mathrm{mean}_{w,r}(stst\cdots)$. Using the definition of $\mathrm{mean}_{w,r}$ given by (3), we obtain $\mathrm{mean}_{w,r}(stst\cdots) = \frac{2}{2+3}r(s) + \frac{3}{2+3}r(t)$. This last value is exactly the limit of $(v_n)_{n\in\mathbb{N}}$ given in (11) hence we obtain (12).

In fact this example and the subsequent analysis can be easily extended to the case where $\mathcal{A}$ is any controllable Markov chain where the controller has no choice and transitions are deterministic. Priorities and weights should be defined according to two constraints. First for each $s, t \in \mathbf{S}$, if $(\mu_n(s))_{n\in\mathbb{N}}$ converges faster than $(\mu_n(t))_{n\in\mathbb{N}}$ to 0 then the priority of t is strictly bigger than the priority of s. Second, if $(\mu_n(s))_{n\in\mathbb{N}}$ and $(\mu_n(t))_{n\in\mathbb{N}}$ converge to 0 at comparable speeds then their weights are proportional to those speeds. Formally:

$$\begin{cases} \text{if } \frac{\mu_n(s)}{\mu_n(t)} \longrightarrow_n 0 \text{ then } c(s) < c(t), \\ \text{if } \frac{\mu_n(t)}{\mu_n(s)} \longrightarrow_n 0 \text{ then } c(t) < c(s), \\ \text{otherwise } c(s) = c(t) \text{ and } \frac{w(s)}{w(t)} = \lim_n \frac{\mu_n(s)}{\mu_n(t)} \ . \end{cases} \quad (13)$$

The following definition gives a procedure to construct such priorities c and weights w.

Definition 6. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of discount mappings that converges to $0^{\mathbf{S}}$ and such that, for each $s, t \in \mathbf{S}$, $\left(\frac{\mu_n(s)}{\mu_n(t)}\right)_{n\in\mathbb{N}}$ converges in $\mathbb{R}_{\geq 0} \cup \{+\infty\}$. Then we define the priority and weight mappings $c \in \mathbb{N}^{\mathbf{S}}$ and $w \in \mathbb{R}^{\mathbf{S}}_{>0}$ associated with $(\mu_n)_{n\in\mathbb{N}}$ as follows. Let $\prec$ be the total pre–order on $\mathbf{S}$ defined by:*

$$(s \prec t) \iff \left(\lim_n \frac{\mu_n(s)}{\mu_n(t)} = 0\right) \ , \quad (14)$$

and let $\equiv$ be the associated equivalence relation:

$$(s \equiv t) \iff \left(\lim_n \frac{\mu_n(s)}{\mu_n(t)} \text{ is neither } 0 \text{ nor } +\infty\right) \ . \quad (15)$$

Let $(\mathbf{S}_0, \ldots, \mathbf{S}_k)$ be the collection of $\equiv$–equivalence classes and suppose that this collection is $\prec$–sorted i.e., for each $s \in \mathbf{S}_i$ and $t \in \mathbf{S}_j$, if $i < j$ then $s \prec t$. Then for $s \in \mathbf{S}$:

$$c(s) \text{ is the unique } i \in \{0, \ldots, k\} \text{ such that } s \in \mathbf{S}_i \ , \quad (16)$$

$$w(s) = \lim_n \frac{\mu_n(s)}{\sum_{t\equiv s} \mu_n(t)} \ . \quad (17)$$

This construction of c and w satisfies the constraints (13). If $\frac{\mu_n(s)}{\mu_n(t)} \longrightarrow_n 0$ then $s \prec t$ hence the equivalence class of s has index strictly smaller than the equivalence class of t and according to (16), $c(s) < c(t)$. The case where $\frac{\mu_n(t)}{\mu_n(s)} \longrightarrow_n 0$ is symmetric. In the remaining case, s and t are in the same equivalence class, hence $\sum_{t\equiv s} \mu_n(t) = \sum_{s\equiv t} \mu_n(s)$ and according to (17) we get (13).

6.3. Limits of multi–discounted MDPs

We are now ready to state the main result of this section: the class of priority weighted MDPs is exactly the class of MDPs whose values are limits of multi–discounted values when discount factors converge simultaneously to 0.

Theorem 6 (Priority weighted MDPs are the limits of multi–discounted MDPs). *Let $r \in \mathbb{R}^{\mathbf{S}}$ be a reward mapping and $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of discount mappings that converges to $0^{\mathbf{S}}$. Suppose that for each $s, t \in \mathbf{S}$,*

$$\left(\frac{\mu_n(s)}{\mu_n(t)}\right)_{n\in\mathbb{N}} \text{ converges in } \mathbb{R}_{\geq 0} \cup \{+\infty\} \ . \quad (18)$$

Let $c \in \mathbb{N}^{\mathbf{S}}$ and $w \in \mathbb{R}^{\mathbf{S}}_{>0}$ that satisfy (13)*. Then for each controllable Markov chain $\mathcal{A}$ and for each $s \in \mathbf{S}$:*

$$\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) \underset{n}{\longrightarrow} \mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(s) \ . \quad (19)$$

Conversely, suppose that (19) *holds for each controllable Markov chain $\mathcal{A}$ and state s. Then* (18) *holds for each states $s, t \in \mathbf{S}$.*

Theorem 6 unifies and generalizes several results.

First, this theorem extends the classical result stated in Theorem 5 to the multi–discounted case.

Second, Theorem 6 establishes a new way of obtaining parity MDPs as limits of multi–discounted MDPs. As a corollary of results of [5] interpreted in the framework of MDPs, we know that if a multi–discounted values has rewards equal to either 0 or 1 and its discount factors tend to 0 one after another, then the limit value is the value of the parity MDP (where lower priorities are given to states whose discount factors converge first to 0). Theorem 6 extends this result to the case where discount factors tend to 0 simultaneously, provided that for any $s, t \in \mathbf{S}$, $\left(\frac{\mu_n(s)}{\mu_n(t)}\right)_{n\in\mathbb{N}}$ converges either to 0 or to $+\infty$ or to 1. In particular this holds if we fix some priorities $c \in \mathbb{R}^{\mathbf{S}}_{>0}$ and define $\mu_n(s) = \frac{1}{n^{c(s)}}$. Moreover,Theorem 6 proves that the restriction to rewards 0 or 1 can be removed and in that case we obtain priority mean–payoff MDPs.

Third, some of the results of [13] about the *priority mean–payoff* function are extended in two directions. Priority mean–payoff MDPs are a special case of priority

weighted MDPs where each weight $w(s)$ is equal to 1. In [13] a priority mapping $c \in \mathbb{N}^{\mathbf{S}}$ was fixed and a very specific sequence of discount mapping is considered: for $n \in \mathbb{N}$, $\mu_n(s) = \frac{1}{n^{c(s)}}$. One of the results of [13] establishes that for this particular sequence of discount mappings the multi–discounted value converges to the priority mean–payoff value. This result is a simple corollary of Theorem 6, since hypothesis of Theorem 6 are satisfied: if $c(s) = c(t)$ then $\mu_n(s) = \mu_n(t)$ and $\left(\frac{\mu_n(s)}{\mu_n(t)}\right)_{n\in\mathbb{N}}$ converges to 1, otherwise if $c(s) < c(t)$ then $\left(\frac{\mu_n(s)}{\mu_n(t)}\right)_{n\in\mathbb{N}}$ converges to 0. Moreover, Theorem 6 generalizes that result of [13] not only to any type of simultaneous convergence of discount factors but also to the stochastic case.

6.4. A proof of Theorem 6

The main step for proving Theorem 6 has already been done with Theorem 2, which establishes the existence of pure stationary optimal strategies in priority weighted mean–payoff MDPs. Since this is also the case for multi–discounted MDPs (cf. Theorem 4). Thus, according to Proposition 1, $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$ MDP–converges to $\mathrm{mean}_{c,w,r}$ if and only if $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$ MC–converges to $\mathrm{mean}_{c,w,r}$. Hence it is sufficient to prove Theorem 6 in the much simpler framework of Markov chains, which is done in Theorem 7. The reciprocal implication of Theorem 6 has already been proven in sub–section 6.1.

Theorem 7. *Let $r \in \mathbb{R}^{\mathbf{S}}$, $(\mu_n)_{n\in\mathbb{N}}$, $c \in \mathbb{N}^{\mathbf{S}}$ and $w \in \mathbb{R}^{\mathbf{S}}_{>0}$ that satisfy hypothesis of Theorem 6. Then for each Markov chain $\mathcal{A}$ and each state s:*

$$\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) \underset{n}{\longrightarrow} \mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(s) \ .$$

Proof. We only give a sketch of proof of Theorem 7 in the special case where all the states of $\mathcal{A}$ are recurrent and form a recurrence class. Full proof can be found in the appendix.

Let $s \in \mathbf{S}$ be a state and $\tau_0, \tau_1, \ldots$ be the sequence of return time in the initial state, i.e. the sequence of random variables defined by:

$$\tau_0 = 0 \text{ and } \tau_{n+1} = \min\{i > \tau_n \mid S_i = S_0\} \ . \tag{20}$$

For any $n \in \mathbb{N}$, let

$$H_n = (S_{\tau_n}, S_{\tau_{n+1}}, \ldots, S_{\tau_{n+1}-1}) \ . \tag{21}$$

Then by properties of Markov chains with finitely many states (see [7] for example):

$$H_0, H_1, \ldots \text{ are independent and identically distributed.} \tag{22}$$

First, we establish an equality about expected value of a multi–discounted MDP. We extend the definition domain of $\mathrm{disc}_{r,\mu} : \mathbf{S}^\omega \to \mathbb{R}$ to finite sequences of states:

$$\mathrm{disc}_{r,\mu}(s_0 s_1 \cdots s_n) = \sum_{i=0}^{n} (1-\mu(s_0)) \cdots (1-\mu(s_{i-1}))\mu(s_i) r(s_i) \ .$$

Then using basic properties of stopping time in Markov chains, we prove that:

$$\mathbb{E}_s\left[\mathrm{disc}_{r,\mu}(S_0 S_1 \cdots)\right] = \frac{\mathbb{E}_s\left[\mathrm{disc}_{r,\mu}(S_0 \cdots S_{\tau_1-1})\right]}{1 - \mathbb{E}_s\left[(1-\mu(S_0)) \cdots (1-\mu(S_{\tau_1-1}))\right]} \ . \tag{23}$$

Using equation (23), we prove that $(\mathbb{E}_s\left[\mathrm{disc}_{r,\mu_n}\right])_{n\in\mathbb{N}}$ converges and we compute its limit. Let I be the random variable defined by $I = \{i \in \mathbb{N} \mid 0 \le i < \tau_1 \text{ and } c(S_i) = \max_{s\in\mathbf{S}} c(s)\}$. Then:

$$\mathbb{E}_s\left[\mathrm{disc}_{r,\mu_n}\right] \underset{n\to\infty}{\longrightarrow} \frac{1}{\mathbb{E}_s\left[\sum_{i\in I} w(S_i)\right]} \mathbb{E}_s\left[\sum_{i\in I} w(S_i) r(S_i)\right] \ . \tag{24}$$

Let $t \in \mathbf{S}$ such that $c(t) = d = \max_{s\in\mathbf{S}} c(s)$. By hypothesis, w and c satisfy (13) and since priority of t is maximal it implies that the two following limits hold $\mathbb{P}^\sigma_s$–a.s.:

$$\frac{1}{\mu_n(t)}\left(1 - (1-\mu_n(S_0)) \cdots (1-\mu_n(S_{\tau_1-1}))\right) \underset{n\to\infty}{\longrightarrow} \sum_{i\in I} \frac{w(S_i)}{w(t)} \ , \tag{25}$$

and

$$\frac{1}{\mu_n(t)} \mathrm{disc}_{r,\mu_n}(S_0 \cdots S_{\tau_1-1}) = \sum_{i=0}^{\tau_1-1} (1-\mu_n(S_0) \cdots (1-\mu_n(S_{i-1})) \frac{\mu_n(S_i)}{\mu_n(t)} r(S_i) \underset{n\to\infty}{\longrightarrow} \sum_{i\in I} \frac{w(S_i)}{w(t)} r(S_i) \ . \tag{26}$$

Putting (25) and (26) in (23), we get equality (24).

Last step of the proof consists in showing that:

$$\mathbb{E}_s\left[\mathrm{mean}_{c,w,r}\right] = \frac{1}{\mathbb{E}_s\left[\sum_{i\in I} w(S_i)\right]} \mathbb{E}_s\left[\sum_{i\in I} w(S_i) r(S_i)\right] \ , \tag{27}$$

which is based on the law of large numbers.

Equations (24) and (27) together prove that $(\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s))_{n\in\mathbb{N}}$ converges to $\mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(s)$. Since this holds for any Markov chain $\mathcal{A}$ and any state s, this achieves the proof of Theorem 7. □

We achieve this subsection with a comparison of the proof of Theorem 6 together with proofs of results similar to Theorem 5 that can be found for example in [15, 16, 2, 8].

Generally the central tool for proving convergence of values of discounted MDPs is the existence of pure stationary optimal strategies. In fact, although values of a discounted MDP may converge when the discount factor converges to 0, there exists in general strategies whose associated sequence of expected payoffs does not converge. For that reason extra knowledge about the structure of optimal strategies is needed.

Three different approaches are possible, all based on the fact that discounted MDPs admit pure stationary optimal strategies. In a classical approach (see for example [15]), existence of pure stationary optimal strategies in discounted MDPs is used to show that the function which maps the discount factor to the value of the discounted MDP is a rational function. This implies existence of pure stationary strategies that are optimal for every small values of discount factors, a phenomenon called Blackwell optimality. Once Blackwell optimality is proven, the convergence of values is straightforward, as well as the existence of a pure stationary strategy which is optimal in the limit MDP. This technique has been adapted to the non–stochastic multi–discounted case in [13] to prove that the priority mean–payoff value can be approximated by means of multi–discounted values.

Another approach, in two steps, consists in first considering a weak form of mean–payoff MDPs. In weak mean–payoff MDPs payoffs are computed taking the average value of expectations of rewards rather than the expectation of the average value of rewards (see [16] for example). Simple matrix calculation shows that weak mean–payoff MDPs are limits of discounted MDPs. Then one can conclude using a result of [2] that the same hold for (strong) mean–payoff MDPs. We did not succeed in adapting this last approach to the case of priority weighted MDPs.

Third approach consists in proving the existence of pure stationary optimal strategies not only in the discounted MDPs but also directly in the MDP which is candidate for the limit. Once this is done, it is enough to prove convergence of values for the easy case of MDPs with no choice, i.e. Markov chains. This second approach was used in [8] in the non–stochastic case to prove the convergence of the discounted value to the mean–payoff value. It was also used in [13], still in the non–stochastic case, to prove the convergence of the multi–discounted value to the priority mean–payoff value in the case where discount factors converge to 0 one after another (see [10] for details). In this paper, we use this third approach in the stochastic case. This is made possible by a recent result [11] that gives a criterion for the existence of pure stationary optimal strategies in MDPs.

7. Computing values of a priority weighted MDP

The central algorithmic problem about MDPs (and more generally about two–player zero–sum games) is the computation of values and optimal strategies.

Although there exists no general algorithm to achieve this computation, the class of MDPs with pure stationary optimal strategies has good algorithmic properties: under weak hypothesis, the values of these MDPs are computable. Let us fix a payoff function ϕ and take reasonable assumption about the computability of ϕ: we suppose that for any *Markov chain* $\mathcal{A}$, and any state s of $\mathcal{A}$, the expected value $\mathrm{val}(\phi, \mathcal{A})(s)$ is computable. Now let $(\mathcal{A}, \phi)$ be an MDP with pure stationary optimal strategies. Then each positional strategy σ induces naturally a Markov chain $\mathcal{A}[\sigma]$ obtained from $\mathcal{A}$ by restriction to actions allowed by σ, and moreover $\mathbb{E}_s^\sigma[\phi] = \mathrm{val}(\mathcal{A}[\sigma], \phi)(s)$. There exists a natural enumerative algorithm to compute values and some optimal strategies of $(\mathcal{A}, \phi)$: it consists in enumerating the *finite* collection of pure stationary strategies $\sigma : \mathbf{S} \to \mathbf{A}$ and selecting one who maximizes $\mathrm{val}(\mathcal{A}[\sigma], \phi)(s)$ for every $s \in \mathbf{S}$.

This naïve enumerative algorithm establishes computability of values and optimal strategies, but it is not optimal in general. For example, values of mean–payoff, discounted and parity MDPs are computable in polynomial time, via reductions to linear programming [16, 4], whereas the complexity of the enumerative algorithm is at least EXPTIME.

For computing values of priority weighted MDPs, the enumerative algorithm gives an EXPTIME upper bound. Indeed, values of Markov chains equipped with $\mathrm{mean}_{c,w,r}$ are computable in polynomial time: first compute recurrence classes, then for each recurrence class compute the stationary distribution, then apply (27) to obtain values of recurrent states.

Is there a polynomial time algorithm to compute values of priority weighted MDPs? This is an open problem.

8. Conclusion

We studied discounted Markov decision processes with multiple discount factors (multi–discounted MDPs) and priority weighted MDPs. Priority weighted MDPs are a generalization of both mean–payoff and parity MDPs. For any $\epsilon > 0$, the existence of ϵ–optimal strategies is guaranteed in these MDPs. We proved the existence of optimal (not only

ϵ–optimal) strategies for these various MDPs. Moreover we showed that there exists optimal strategies that are pure and stationary (Theorem 2). As a corollary, we proved that when the discount factors of a multi–discounted MDP converge simultaneously to 0 at various but comparable speeds, the value of the multi–discounted MDP converge to the priority weighted value (Theorem 6). Moreover, we proved that the only limits of multi–discounted MDPs are priority weighted MDPs (Theorem 6).

These results lead to several algorithmic and theoretic questions. Algorithmic aspects are discussed in Section 7. The most challenging theoretic question is the following: to what extent can we adapt results of this paper to the framework of two–player zero–sum stochastic games? The special case of perfect information games (turn–based games) is easy, this is ongoing work which will be published soon. In the general case of concurrent games the existence of pure stationary optimal strategies established by Theorem 2 is no longer guaranteed since it is not even guaranteed in matrix games. However, it seems plausible that the convergence of values established by Theorem 6 still holds. In the case of discounted games with a single discount factor this result is well–known but the complexity of existing proofs (see [9] by example) indicates that extension of Theorem 6 to concurrent games may be hard.

References

[1] D. Bertsekas and S. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, 1978.

[2] K.J. Bierth. An expected average reward criterion. *Stochastic Processes and Applications*, 26:133–140, 1987.

[3] K. Chatterjee, T.A. Henzinger, and M. Jurdzinski. Mean–payoff parity games. In *Proc. of LICS'05*, pages 178–187. IEEE, 2005.

[4] C. Courcoubetis and M. Yannakakis. Markov decision processes and regular events. In *ICALP'90*, volume 443 of *LNCS*, pages 336–349. Springer, 1990.

[5] L. de Alfaro. Quantitative verification and control via the μ–calculus. In *Proc. of CONCUR'03*, volume 2761 of *LNCS*, pages 102–126. springer, 2003.

[6] L. de Alfaro, T. A. Henzinger, and R. Majumdar. Discounting the future in systems theory. In *Proc. of ICALP'03*, volume 2719 of *LNCS*, pages 1022–1037. Springer, 2003.

[7] R. Durett. *Probability Theory and Examples*. Duxbury Press, 1996.

[8] A. Ehrenfeucht and J. Mycielski. Positional strategies for mean–payoff games. *International Journal of Game Theory*, 8:109–113, 1979.

[9] J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer, 1997.

[10] H. Gimbert. *Jeux Positionnels*. PhD thesis, 2006.

[11] H. Gimbert. Pure stationary optimal strategies in markov decision processes. In *Proc. of STACS'07*, 2007. Full proofs can be found in the technical report 2006–02 of LIAFA, Université Denis Diderot. `http://www.liafa.jussieu.fr/`.

[12] H. Gimbert and W. Zielonka. Games where you can play optimally without any memory. In *CONCUR 2005*, volume 3653 of *LNCS*, pages 428–442. Springer, 2005.

[13] H. Gimbert and W. Zielonka. Deterministic priority mean–payoff games as limits of discounted games. In *Proc. of ICALP 06*, volume 4052 of *LNCS*, pages 312–323. Springer, 2006.

[14] E. Grädel, W. Thomas, and T. Wilke. *Automata, Logics and Infinite Games*, volume 2500 of *LNCS*. Springer, 2002.

[15] A. Neyman. Chapter 2: Discounted stochastic games: The finite case. In S. Sorin A. Neyman, editor, *Stochastic Games and Applications*. Kluwer Academic Publishers, 2003.

[16] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[17] L. S. Shapley. Stochastic games. In *Proceedings of the National Academy of Science USA*, volume 39, pages 1095–1100, 1953.

Appendix

Theorem 8. *Let $(\mathrm{disc}_{r,\mu_n})_{n\in\mathbb{N}}$, prior and w satisfy hypothesis of Theorem 6. Let $\mathcal{A}$ be a Markov chain and $s \in \mathbf{S}$ a state of $\mathcal{A}$. Then*

$$\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) \underset{n}{\longrightarrow} \mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(s) \ . \quad (28)$$

Proof. We first prove (28) in the special case where all states of $\mathcal{A}$ are recurrent and $\mathcal{A}$ has a unique recurrence class. Let $s \in \mathbf{S}$ be a state and $\tau_0, \tau_1, \ldots$ be the sequence of return time in the initial state, i.e. the sequence of random variables defined by:

$$\tau_0 = 0 \ ,$$
$$\tau_{n+1} = \min\{i > \tau_n \mid S_i = S_0\} \ ,$$

and for any $n \in \mathbb{N}$, let

$$H_n = (S_{\tau_n}, S_{\tau_{n+1}}, \ldots, S_{\tau_{n+1}-1}) \ . \quad (29)$$

Then by properties of Markov chains with finitely many states (see [7] for example):

$$H_0, H_1, \ldots \text{ are independent and identically distributed.} \quad (30)$$

For shortening notations, we define for any $n \in \mathbb{N}$:

$$\nu_n = \mu(S_n) \ . \quad (31)$$

First step consists in establishing the following equality about expected value of a discounted MDP.

$$\mathbb{E}_s\left[\mathrm{disc}_{r,\mu}(S_0 S_1 \cdots)\right] = \frac{\mathbb{E}_s\left[\mathrm{disc}_{r,\mu}(S_0 \cdots S_{\tau_1-1})\right]}{1 - \mathbb{E}_s\left[(1-\nu_0)\cdots(1-\nu_{\tau_1-1})\right]} \ , \quad (32)$$

In fact, by definition of $\mathrm{disc}_{r,\mu}$,

$$\mathbb{E}_s\left[\mathrm{disc}_{r,\mu}(S_0 S_1 \cdots)\right] = \mathbb{E}_s\left[\mathrm{disc}_{r,\mu}(S_0 \cdots S_{\tau_1-1})\right] + \mathbb{E}_s\left[(1-\nu_0)\cdots(1-\nu_{\tau_1-1}) \cdot \mathrm{disc}_{r,\mu}(S_{\tau_1} S_{\tau_1+1} \cdots)\right] \ .$$

According to (30), we deduce:

$$\mathbb{E}_s\left[(1-\nu_0)\cdots(1-\nu_{\tau_1-1}) \cdot \mathrm{disc}_{r,\mu}(S_{\tau_1} S_{\tau_1+1} \cdots)\right] = \mathbb{E}_s\left[(1-\nu_0)\cdots(1-\nu_{\tau_1-1})\right] \cdot \mathbb{E}_s\left[\mathrm{disc}_{r,\mu}(S_0 S_1 \cdots)\right] \ .$$

This two last equations give (32).

Now, using equation (32)**, we are going to compute the limit of $\mathbb{E}_s\left[\mathrm{disc}_{r,\mu_n}\right]$ when n tends to ∞.** Let I be the random variable defined by $I = \{i \in \mathbb{N} \mid 0 \leq i < \tau_1 \text{ and } c(S_i) = d\}$. We are going to show that:

$$\mathbb{E}_s\left[\mathrm{disc}_{r,\mu_n}\right] \underset{n\to\infty}{\longrightarrow} \frac{1}{\mathbb{E}_s\left[\sum_{i\in I} w(S_i)\right]} \mathbb{E}_s\left[\sum_{i\in I} w(S_i) r(S_i)\right] \ . \quad (33)$$

Let $t \in \mathbf{S}$ such that $c(t) = d = \max_{s\in\mathbf{S}} c(s)$. Since by hypothesis, w and c satisfy (13), then for any state $q \in \mathbf{S}$, $\mu_n(q) \longrightarrow_n 0$ and:

$$\begin{cases} \frac{\mu_n(q)}{\mu_n(t)} \longrightarrow_n 0 & \text{if } c(q) < d \ , \\ \frac{\mu_n(q)}{\mu_n(t)} \longrightarrow_n \frac{w(q)}{w(t)} & \text{if } c(q) = d \ . \end{cases}$$

Since priority of t is maximal, it implies that the two following limits hold $\mathbb{P}_s^\sigma$–a.s.:

$$\frac{1}{\mu_n(t)}\left(1 - (1-\mu_n(S_0))\cdots(1-\mu_n(S_{\tau_1-1}))\right) \underset{n\to\infty}{\longrightarrow} \sum_{i\in I} \frac{w(S_i)}{w(t)} \ , \quad (34)$$

and

$$\frac{1}{\mu_n(t)}\mathrm{disc}_{r,\mu_n}(S_0 \cdots S_{\tau_1-1}) = \sum_{i=0}^{\tau_1-1} (1-\mu_n(S_0)\cdots(1-\mu_n(S_{i-1}))\frac{\mu_n(S_i)}{\mu_n(t)} r(S_i) \underset{n\to\infty}{\longrightarrow} \sum_{i\in I} \frac{w(S_i)}{w(t)} r(S_i) \ . \quad (35)$$

Putting (34) and (35) in (32), we finally get equality (33).

Last step of the proof consists in showing that:

$$\mathbb{E}_s\left[\mathrm{mean}_{c,w,r}\right] = \frac{1}{\mathbb{E}_s\left[\sum_{i\in I} w(S_i)\right]} \cdot \mathbb{E}_s\left[\sum_{i\in I} w(S_i) r(S_i)\right] \ . \quad (36)$$

Since all states of $\mathcal{A}$ are recurrent and form a recurrence class, $\mathbb{P}_s^\sigma(\limsup_n c(S_n) = d) = 1$. Hence, by definition of $\mathrm{mean}_{c,w,r}$:

$$\mathbb{E}_s\left[\mathrm{mean}_{c,w,r}\right] = \mathbb{E}_s\left[\mathrm{mean}_{w,r}(\pi(S_0 S_1 \cdots))\right] \quad (37)$$

$$= \mathbb{E}_s\left[\limsup_n \frac{1}{\sum_{\substack{0\leq i<n \\ c(S_i)=d}} w(S_i)} \sum_{\substack{0\leq i<n \\ c(S_i)=d}} w(S_i) r(S_i)\right] \ . \quad (38)$$

Let us define $w' \in [0, +\infty[^{\mathbf{S}}$ in the following way. For $t \in \mathbf{S}$, if $c(t) = d$ then $w'(t) = w(t)$ and if $c(t) < d$ we define $w'(t) = 0$. Then equation (38) can be rewritten as:

$$\mathbb{E}_s\left[\mathrm{mean}_{c,w,r}\right] = \mathbb{E}_s\left[\limsup_n \frac{1}{\sum_{0\leq i<n} w'(S_i)} \sum_{0\leq i<n} w'(S_i) r(S_i)\right] \ . \quad (39)$$

For any $l \in \mathbb{N}$, we define the random variable

$$\mathrm{ind}(l) = \mid \{0 < n \leq l \mid S_n = S_0\} \mid = \max\{n \in \mathbb{N} \mid \tau_n \leq l\} \ ,$$

i.e. $\mathrm{ind}(l)$ is the number of return to initial state before instant l. Then since all states of $\mathcal{A}$ are recurrent and form a recurrence class,

$$\mathrm{ind}(l) \underset{l\to\infty}{\longrightarrow} \infty \ . \tag{40}$$

Equation (39) can be rewritten:

$$\mathbb{E}_s\left[\mathrm{mean}_{c,w,r}\right] = \mathbb{E}_s\left[\limsup_n \frac{\mathrm{ind}(n)}{\sum_{i=0}^{n-1} w'(S_i)} \cdot \frac{1}{\mathrm{ind}(n)} \sum_{0\leq i<n} w'(S_i) r(S_i)\right] \tag{41}$$

We are going to prove that each factor of right part of equation (41) converges $\mathbb{P}_s^\sigma$–a.s. everywhere. Let $n \in \mathbb{N}$. Then by definition of return times τ_i in initial state, $0 = \tau_0 < \tau_1 < \ldots < \tau_{\mathrm{ind}(n)} \leq n$, hence:

$$\begin{aligned}\frac{1}{\mathrm{ind}(n)} \sum_{0\leq i<n} w'(S_i) &= \frac{1}{\mathrm{ind}(n)} \sum_{l=0}^{\mathrm{ind}(n)-1} \left(\sum_{k=\tau_l}^{\tau_{l+1}-1} w'(S_k)\right) \\ &\quad + \frac{1}{\mathrm{ind}(n)} \sum_{k=\tau_{\mathrm{ind}(n)}}^{n} w'(S_k) \ . \end{aligned} \tag{42}$$

For any $i \in \mathbb{N}$, let

$$X_i = \sum_{k=\tau_i}^{\tau_{i+1}-1} w'(S_k) \ .$$

According to (30), the random variables $X_0, X_1, \ldots$ are independent and identically distributed. According to (40), we can apply the strong law of large numbers to left summand of (42) and we get that:

$$\frac{1}{\mathrm{ind}(n)} \sum_{l=0}^{\mathrm{ind}(n)-1} \left(\sum_{k=\tau_l}^{\tau_{l+1}-1} w'(S_k)\right) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_s\left[X_0\right] = \mathbb{E}_s\left[\sum_{i\in I} w(S_i)\right] \ . \tag{43}$$

Let $Y_n = \frac{1}{\mathrm{ind}(n)} \sum_{k=\tau_{\mathrm{ind}(n)}}^{n} w'(S_k)$ be the right summand in (42). Then $0 \leq Y_n \leq \frac{1}{\mathrm{ind}(n)} X_n$. Since $X_n$is i.i.d with X_0 and X_0 has finite value $\mathbb{P}_s^\sigma$–a.s., Y_n converges $\mathbb{P}_s^\sigma$–a.s. to 0. Together with (42) and (43), we get that:

$$\frac{1}{\mathrm{ind}(n)} \sum_{0\leq i<n} w'(S_i) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_s\left[\sum_{i\in I} w(S_i)\right] \ \ \mathbb{P}_s^\sigma\text{–a.s.} \ . \tag{44}$$

Similarly, we get that $\mathbb{P}_s^\sigma$–a.s.:

$$\frac{1}{\mathrm{ind}(n)} \sum_{0\leq i<n} w'(S_i) r(S_i) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_s\left[\sum_{i\in I} w(S_i) r(S_i)\right] \ . \tag{45}$$

Putting (44) and (45) together in (41), we finally obtain (36).

Equations (36) and (33) together give (28), which achieves the proof of Theorem 6, in the special case where all states of $\mathcal{A}$ are recurrent and form a unique recurrence class.

Case of multiple recurrence classes In the general case, the set of states $\mathbf{S}$ is partitioned in $\mathbf{S} = (T, R_0, \ldots, R_k)$, where T is the set of transient states and $R_0, \ldots, R_k$ are the recurrence classes. Let $R = R_0 \cup \ldots \cup R_k$ be the set of recurrent states and let $T_R = \min\{n \in \mathbb{N} \mid S_n \in R\}$ be the first time of visit to R. By definition of recurrence classes, T_R takes finite value $\mathbb{P}_s^\sigma$-a.s.. Since $\mathrm{mean}_{c,w,r}$ is prefix–independent, we have for each $s \in \mathbf{S}$:

$$\mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(s) = \sum_{r\in R} \mathbb{P}_s(S_{T_R} = r) \cdot \mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(r) \ . \tag{46}$$

Moreover,

$$\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) = \sum_{r\in R} \mathbb{P}_s(S_{T_R} = r) \cdot \mathbb{E}_s\left[\mathrm{disc}_{r,\mu_n} \mid S_{T_R} = r\right] \ . \tag{47}$$

Let $\nu_n = \mu(S_n)$. Using an independency property similar to (40):

$$\mathbb{E}_s\left[\mathrm{disc}_{r,\mu} \mid S_{T_R} = r\right] = \mathbb{E}_s\left[(1-\nu_0)\cdots(1-\nu_{T_R-1})\right] \cdot \mathbb{E}_r\left[\mathrm{disc}_{r,\mu}\right] \ . \tag{48}$$

Then since r in the equation above is a recurrent state, we can apply (28) and since $\forall s \in \mathbf{S}$, $\mu_n(s) \longrightarrow_n 0$, (47) gives:

$$\mathrm{val}(\mathcal{A}, \mathrm{disc}_{r,\mu_n})(s) \underset{n}{\longrightarrow} \sum_{r\in R} \mathbb{P}_s(S_{T_R} = r) \cdot \mathrm{val}(\mathcal{A}, \mathrm{mean}_{c,w,r})(s) \ . \tag{49}$$

Together with (46), we obtain (28), which achieves the proof of Theorem 6. □