General Discounting versus Average Reward

Marcus Hutter

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

January 2006

Abstract

Consider an agent interacting with an environment in cycles. In every interaction cycle the agent is rewarded for its performance. We compare the average reward U from cycle 1 to m (average value) with the future discounted reward V from cycle k to ∞ (discounted value). We consider essentially arbitrary (non-geometric) discount sequences and arbitrary reward sequences (non-MDP environments). We show that asymptotically U for $m \to \infty$ and V for $k \to \infty$ are equal, provided both limits exist. Further, if the effective horizon grows linearly with k or faster, then the existence of the limit of U implies that the limit of V exists. Conversely, if the effective horizon grows linearly with k or slower, then existence of the limit of V implies that the limit of U exists.

Contents

Keywords

reinforcement learning; average value; discounted value; arbitrary environment; arbitrary discount sequence; effective horizon; increasing farsightedness; consistent behavior.

1 Introduction

We consider the reinforcement learning setup [RN03, Hut05], where an agent interacts with an environment in cycles. In cycle k, the agent outputs (acts) a_k, then it makes observation o_k and receives reward r_k, both provided by the environment. Then the next cycle $k+1$ starts. For simplicity we assume that agent and environment are deterministic.

Typically one is interested in action sequences, called plans or policies, for agents that result in high reward. The simplest reasonable measure of performance is the total reward sum or equivalently the average reward, called average value $U_{1m} := \frac{1}{m}[r_1+...+r_m]$, where m should be the lifespan of the agent. One problem is that the lifetime is often not known in advance, e.g. often the time one is willing to let a system run depends on its displayed performance. More serious is that the measure is indifferent to whether an agent receives high rewards early or late if the values are the same.

A natural (non-arbitrary) choice for m is to consider the limit $m \to \infty$. While the indifference may be acceptable for finite m, it can be catastrophic for $m = \infty$. Consider an agent that receives no reward until its first action is $b_k = b$, and then once receives reward $\frac{k-1}{k}$. For finite m, the optimal k to switch from action a to b is $k_{opt} = m$. Hence $k_{opt} \to \infty$ for $m \to \infty$, so the reward maximizing agent for $m \to \infty$ actually always acts with a, and hence has zero reward, although a value arbitrarily close to 1 would be achievable. (Immortal agents are lazy [Hut05, Sec.5.7]). More serious, in general the limit $U_{1\infty}$ may not even exist.

Another approach is to consider a moving horizon. In cycle k, the agent tries to maximize $U_{km} := \frac{1}{m-k+1}[r_k+...+r_m]$, where m increases with k, e.g. $m = k+h-1$ with h being the horizon. This naive truncation is often used in games like chess (plus a heuristic reward in cycle m) to get a reasonably small search tree. While this can work in practice, it can lead to inconsistent optimal strategies, i.e. to agents that change their mind. Consider the example above with $h = 2$. In every cycle k it is better first to act a and then b ($U_{km} = r_k + r_{k+1} = 0 + \frac{k}{k+1}$), rather than immediately b ($U_{km} = r_k + r_{k+1} = \frac{k-1}{k} + 0$), or a,a ($U_{km} = 0+0$). But entering the next cycle $k+1$, the agent throws its original plan overboard, to now choose a in favor of b, followed by b. This pattern repeats, resulting in no reward at all.

The standard solution to the above problems is to consider geometrically=exponentially discounted reward [Sam37, BT96, SB98]. One discounts the reward for every cycle of delay by a factor $\gamma < 1$, i.e. considers $V_{k\gamma} := (1-\gamma)\sum_{i=k}^{\infty} \gamma^{i-k} r_i$. The $V_{1\gamma}$ maximizing policy is consistent in the sense that its actions a_k,a_{k+1},... coincide with the optimal policy based on $V_{k\gamma}$. At first glance, there seems to be no arbitrary lifetime m or horizon h, but this is an illusion. $V_{k\gamma}$ is dominated by contributions from rewards $r_k...r_{k+O(\ln\gamma^{-1})}$, so has an effective horizon $h^{eff} \approx \ln\gamma^{-1}$. While such a sliding effective horizon does not cause inconsistent policies, it can nevertheless lead to suboptimal behavior. For every (effective) horizon, there is a task that needs a larger horizon to be solved. For instance, while $h^{eff} = 5$ is sufficient

for tic-tac-toe, it is definitely insufficient for chess. There are elegant closed form solutions for Bandit problems, which show that for any $\gamma<1$, the Bayes-optimal policy can get stuck with a suboptimal arm (is not self-optimizing) [BF85, KV86].

For $\gamma\to 1$, $h^{eff}\to\infty$, and the defect decreases. There are various deep papers considering the limit $\gamma\to 1$ [Kel81], and comparing it to the limit $m\to\infty$ [Kak01]. The analysis is typically restricted to ergodic MDPs for which the limits $\lim_{\gamma\to 1}V_{1\gamma}$ and $\lim_{m\to\infty}U_{1m}$ exist. But like the limit policy for $m\to\infty$, the limit policy for $\gamma\to 1$ can display very poor performance, i.e. we need to choose $\gamma<1$ fixed in advance (but how?), or consider higher order terms [Mah96, AA99]. We also cannot consistently adapt γ with k. Finally, the value limits may not exist beyond ergodic MDPs.

There is little work on other than geometric discounts. In the psychology and economics literature it has been argued that people discount a one day=cycle delay in reward more if it concerns rewards now rather than later, e.g. in a year (plus one day) [FLO02]. So there is some work on "sliding" discount sequences $W_{k\gamma}\propto\gamma_0 r_k+\gamma_1 r_{k+1}+...$. One can show that this also leads to inconsistent policies if $\boldsymbol{\gamma}$ is non-geometric [Str56, VW04].

Is there any non-geometric discount leading to consistent policies? In [Hut02] the generally discounted value $V_{k\gamma}:=\frac{1}{\Gamma_k}\sum_{i=k}^{\infty}\gamma_i r_i$ with $\Gamma_k:=\sum_{i=k}^{\infty}\gamma_i<\infty$ has been introduced. It is well-defined for arbitrary environments, leads to consistent policies, and e.g. for quadratic discount $\gamma_k=1/k^2$ to an increasing effective horizon (proportionally to k), i.e. the optimal agent becomes increasingly farsighted in a consistent way, leads to self-optimizing policies in ergodic (kth-order) MDPs in general, Bandits in particular, and even beyond MDPs. See [Hut02] for these and [Hut05] for more results. The only other serious analysis of general discounts we are aware of is in [BF85], but their analysis is limited to Bandits and so-called regular discount. This discount has bounded effective horizon, so also does not lead to self-optimizing policies.

The *asymptotic* total average performance $U_{1\infty}$ and future discounted performance $V_{\infty\gamma}$ are of key interest. For instance, often we do not know the exact environment in advance but have to *learn* it from past experience, which is the domain of reinforcement learning [SB98] and adaptive control theory [KV86]. Ideally we would like a learning agent that performs *asymptotically* as well as the optimal agent that knows the environment in advance.

Contents and main results. The subject of study of this paper is the relation between $U_{1\infty}$ and $V_{\infty\gamma}$ for *general discount* $\boldsymbol{\gamma}$ and *arbitrary environment*. The importance of the performance measures U and V, and general discount $\boldsymbol{\gamma}$ has been discussed above. There is also a clear need to study general environments beyond ergodic MDPs, since the real world is neither ergodic (e.g. losing an arm is irreversible) nor completely observable.

The only restriction we impose on the discount sequence $\boldsymbol{\gamma}$ is summability ($\Gamma_1<\infty$) so that $V_{k\gamma}$ exists, and monotonicity ($\gamma_k\geq\gamma_{k+1}$). Our main result is that if both limits $U_{1\infty}$ and $V_{\infty\gamma}$ exist, then they are necessarily equal (Section 7, Theorem 19). Somewhat surprisingly this holds for *any* discount sequence $\boldsymbol{\gamma}$ and *any* environment

(reward sequence $\boldsymbol{r}$), whatsoever.

Note that limit $U_{1\infty}$ may exist or not, independent of whether $V_{\infty\gamma}$ exists or not. We present examples of the four possibilities in Section 2. Under certain conditions on $\boldsymbol{\gamma}$, existence of $U_{1\infty}$ implies existence of $V_{\infty\gamma}$, or vice versa. We show that if (a quantity closely related to) the effective horizon grows linearly with k or faster, then existence of $U_{1\infty}$ implies existence of $V_{\infty\gamma}$ and their equality (Section 5, Theorem 15). Conversely, if the effective horizon grows linearly with k or slower, then existence of $V_{\infty\gamma}$ implies existence of $U_{1\infty}$ and their equality (Section 6, Theorem 17). Note that apart from discounts with oscillating effective horizons, this implies (and this is actually the path used to prove) the first mentioned main result. In Sections 3 and 4 we define and provide some basic properties of average and discounted value, respectively.

2 Example Discount and Reward Sequences

In order to get a better feeling for general discount sequences, effective horizons, average and discounted value, and their relation and existence, we first consider various examples.

Notation. In the following we assume that $i,k,m,n \in \mathbb{N}$ are natural numbers, $\underline{F} := \underline{\lim}_n F_n = \lim_{k\to\infty}\inf_{n>k} F_n$ denotes the limit inferior and $\overline{F} := \overline{\lim}_n F_n = \lim_{k\to\infty}\sup_{n>k} F_n$ the limit superior of F_n, $\forall' n$ means for all but finitely many n, $\boldsymbol{\gamma} = (\gamma_1,\gamma_2,...)$ denotes a summable discount sequence in the sense that $\Gamma_k := \sum_{i=k}^{\infty}\gamma_i < \infty$ and $\gamma_k \in \mathbb{R}^+$ $\forall k$, $\boldsymbol{r} = (r_1,r_2,...)$ is a bounded reward sequence w.l.g. $r_k \in [0,1]$ $\forall k$, constants $\alpha,\beta \in [0,1]$, boundaries $0 \le k_1 < m_1 < k_2 < m_2 < k_3 < ...$, total average value $U_{1m} := \frac{1}{m}\sum_{i=1}^{m} r_i$ (see Definition 10) and future discounted value $V_{k\gamma} = \frac{1}{\Gamma_k}\sum_{i=k}^{\infty}\gamma_i r_i$ (see Definition 12). The derived theorems also apply to general bounded rewards $r_i \in [a,b]$ by linearly rescaling $r_i \rightsquigarrow \frac{r_i-a}{b-a} \in [0,1]$ and $U \rightsquigarrow \frac{U-a}{b-a}$ and $V \rightsquigarrow \frac{V-a}{b-a}$.

Discount sequences and effective horizons. Rewards r_{k+h} give only a small contribution to $V_{k\gamma}$ for large h, since $\gamma_{k+h} \overset{h\to\infty}{\longrightarrow} 0$. More important, the whole reward tail from $k+h$ to ∞ in $V_{k\gamma}$ is bounded by $\frac{1}{\Gamma_k}[\gamma_{k+h}+\gamma_{k+h+1}+...]$, which tends to zero for $h \to \infty$. So effectively $V_{k\gamma}$ has a horizon h for which the cumulative tail weight Γ_{k+h}/Γ_k is, say, about $\frac{1}{2}$, or more formally $h_k^{eff} := \min\{h \ge 0 : \Gamma_{k+h} \le \frac{1}{2}\Gamma_k\}$. The closely related quantity $h_k^{quasi} := \Gamma_k/\gamma_k$, which we call the quasi-horizon, will play an important role in this work. The following table summarizes various discounts with their properties.

Discounts	γ_k	Γ_k	h_k^{eff}	h_k^{quasi}	$k\gamma_k/\Gamma_k$	$\to$?
finite	$\frac{1 \text{ for } k\le m}{0 \text{ for } k>m}$	$m-k+1$	$\frac{1}{2}(m-k+1)$	$m-k+1$	$\frac{k}{m-k+1}$	
geometric	$\gamma^k,\ 0 \le \gamma < 1$	$\frac{\gamma^k}{1-\gamma}$	$\frac{\ln 2}{\ln \gamma^{-1}}$	$\frac{1}{1-\gamma}$	$(1-\gamma)k$	$\to \infty$
quadratic	$\frac{1}{k(k+1)}$	$\frac{1}{k}$	k	$k+1$	$\frac{k}{k+1}$	$\to 1$
power	$k^{-1-\varepsilon},\ \varepsilon > 0$	$\sim \frac{1}{\varepsilon}k^{-\varepsilon}$	$\sim (2^{1/\varepsilon}-1)k$	$\sim \frac{k}{\varepsilon}$	$\sim \varepsilon$	$\to \varepsilon$
harmonic$_\approx$	$\frac{1}{k\ln^2 k}$	$\sim \frac{1}{\ln k}$	$\sim k^2$	$\sim k\ln k$	$\sim \frac{1}{\ln k}$	$\to 0$

For instance, the standard discount is geometric $\gamma_k=\gamma^k$ for some $0\leq\gamma<1$, with constant effective horizon $\frac{\ln(1/2)}{\ln\gamma}$. (An agent with $\gamma=0.95$ can/will not plan farther than about 10-20 cycles ahead). Since in this work we allow for general discount, we can even recover the average value U_{1m} by choosing $\gamma_k=\{{}^{1 \text{ for } k\leq m}_{0 \text{ for } k>m}\}$. A power discount $\gamma_k=k^{-\alpha}$ $(\alpha>1)$ is very interesting, since it leads to a linearly increasing effective horizon $h_k^{eff}\propto k$, i.e. to an agent whose farsightedness increases proportionally with age. This choice has some appeal, as it avoids preselection of a global time-scale like m or $\frac{1}{1-\gamma}$, and it seems that humans of age k years usually do not plan their lives for more than, perhaps, the next k years. It is also the boundary case for which $U_{1\infty}$ exists if and only if $V_{\infty\gamma}$ exists.

Example reward sequences. Most of our (counter)examples will be for binary reward $\boldsymbol{r}\in\{0,1\}^\infty$. We call a maximal consecutive subsequence of ones a 1-run. We denote start, end, and length of the nth run by k_n, m_n-1, and $A_n=m_n-k_n$, respectively. The following 0-run starts at m_n, ends at $k_{n+1}-1$, and has length $B_n=k_{n+1}-m_n$. The (non-normalized) discount sum in 1/0-run n is denoted by a_n / b_n, respectively. The following definition and two lemmas facilitate the discussion of our examples. The proofs contain further useful relations.

Definition 1 (Value for binary rewards) *Every binary reward sequence* $\boldsymbol{r}\in\{0,1\}^\infty$ *can be defined by the sequence of change points* $0\leq k_1<m_1<k_2<m_2<...$ *with*

$$r_k=1 \quad\Longleftrightarrow\quad k\in\bigcup_n\mathcal{S}_n, \quad \textit{where} \quad \mathcal{S}_n:=\{k\in\mathbb{N}:k_n\leq k<m_n\}.$$

The intuition behind the following lemma is that the relative length A_n of a 1-run and the following 0-run B_n (previous 0-run B_{n-1}) asymptotically provides a lower (upper) limit of the average value U_{1m}.

Lemma 2 (Average value for binary rewards) *For binary* $\boldsymbol{r}$ *of Definition 1, let* $A_n:=m_n-k_n$ *and* $B_n:=k_{n+1}-m_n$ *be the lengths of the* n*th 1/0-run. Then*

$$\begin{aligned}&\textit{If}\quad \tfrac{A_n}{A_n+B_n}\to\alpha \quad \textit{then}\quad \underline{U}_{1\infty}=\lim_n U_{1,k_n-1}=\alpha\\ &\textit{If}\quad \tfrac{A_n}{B_{n-1}+A_n}\to\beta \quad \textit{then}\quad \overline{U}_{1\infty}=\lim_n U_{1,m_n-1}=\beta\end{aligned}$$

In particular, if $\alpha=\beta$*, then* $U_{1\infty}=\alpha=\beta$ *exists.*

Proof. The elementary identity $U_{1m}=U_{1,m-1}+\frac{1}{m}(r_m-U_{1,m-1})\gtrless U_{1,m-1}$ if $r_m=\{{}^1_0\}$ implies

$$\begin{aligned}U_{1k_n}\leq U_{1m}\leq U_{1,m_n-1} \quad &\text{for} \quad k_n\leq m<m_n\\ U_{1,k_{n+1}-1}\leq U_{1m}\leq U_{1,m_n} \quad &\text{for} \quad m_n\leq m<k_{n+1}\end{aligned}$$

$$\Rightarrow \quad \inf_{n\geq n_0} U_{1k_n} \;\leq\; U_{1m} \;\leq\; \sup_{m\geq n_0} U_{1,m_n-1} \quad \forall m\geq k_{n_0}$$
$$\Rightarrow \quad \underline{\lim_n} U_{1k_n} \;=\; \underline{U}_{1\infty} \;\leq\; \overline{U}_{1\infty} \;=\; \overline{\lim_n} U_{1,m_n-1} \qquad (1)$$

Note the equalities in the last line. The $\geq$ holds, since (U_{1k_n}) and (U_{1,m_n-1}) are subsequences of (U_{1m}). Now

$$\text{If} \quad \tfrac{A_n}{A_n+B_n}\geq\alpha\;\forall n \quad \text{then} \quad U_{1,k_n-1}=\tfrac{A_1+\ldots+A_{n-1}}{A_1+B_1+\ldots+A_{n-1}+B_{n-1}}\geq\alpha\;\forall n \qquad (2)$$

This implies $\inf_n \frac{A_n}{A_n+B_n} \leq \inf_n U_{1,k_n-1}$. If the condition in (2) is initially (for a finite number of n) violated, the conclusion in (2) still holds asymptotically. A standard argument along these lines shows that we can replace the inf by a $\underline{\lim}$, i.e.

$$\underline{\lim_n} \tfrac{A_n}{A_n+B_n} \leq \underline{\lim_n} U_{1,k_n-1} \quad \text{and similarly} \quad \overline{\lim_n} \tfrac{A_n}{A_n+B_n} \geq \overline{\lim_n} U_{1,k_n-1}$$

Together this shows that $\lim_n U_{1,k_n-1}=\alpha$ exists, if $\lim_n \frac{A_n}{A_n+B_n}=\alpha$ exists. Similarly

$$\text{If} \quad \tfrac{A_n}{B_{n-1}+A_n}\geq\beta\;\forall n \quad \text{then} \quad U_{1,m_n-1}=\tfrac{A_1+\ldots+A_n}{B_0+A_1+\ldots+B_{n-1}+A_n}\geq\beta\;\forall n \qquad (3)$$

where $B_0:=0$. This implies $\inf_n \frac{A_n}{B_{n-1}+A_n} \leq \inf_n U_{1,m_n-1}$, and by an asymptotic refinement of (3)

$$\underline{\lim_n} \tfrac{A_n}{B_{n-1}+A_n} \leq \underline{\lim_n} U_{1,m_n-1} \quad \text{and similarly} \quad \overline{\lim_n} \tfrac{A_n}{B_{n-1}+A_n} \geq \overline{\lim_n} U_{1,m_n-1}$$

Together this shows that $\lim_n U_{1,m_n-1}=\beta$ exists, if $\lim_n \frac{A_n}{B_{n-1}+A_n}=\beta$ exists. ■

Similarly to Lemma 2, the asymptotic ratio of the discounted value a_n of a 1-run and the discount sum b_n of the following (b_{n-1} of the previous) 0-run determines the upper (lower) limits of the discounted value $V_{k\gamma}$.

Lemma 3 (Discounted value for binary rewards) *For binary $\boldsymbol{r}$ of Definition 1, let $a_n:=\sum_{i=k_n}^{m_n-1}\gamma_i=\Gamma_{k_n}-\Gamma_{m_n}$ and $b_n:=\sum_{i=m_n}^{k_{n+1}-1}\gamma_i=\Gamma_{m_n}-\Gamma_{k_{n+1}}$ be the discount sums of the nth 1/0-run. Then*

$$\begin{aligned} &\textit{If} \quad \tfrac{a_{n+1}}{b_n+a_{n+1}}\to\alpha \quad \textit{then} \quad \underline{V}_{\infty\gamma}=\lim_n V_{m_n\gamma}=\alpha \\ &\textit{If} \quad \tfrac{a_n}{a_n+b_n}\to\beta \quad \textit{then} \quad \overline{V}_{\infty\gamma}=\lim_n V_{k_n\gamma}=\beta \end{aligned}$$

In particular, if $\alpha=\beta$, then $V_{\infty\gamma}=\alpha=\beta$ exists.

Proof. The proof is very similar to the proof of Lemma 2. The elementary identity $V_{k\gamma}=V_{k+1,\gamma}+\frac{\gamma_k}{\Gamma_k}(r_k-V_{k+1,\gamma})\gtrless V_{k+1,\gamma}$ if $r_k=\{{}^1_0\}$ implies

$$\begin{aligned} V_{m_n\gamma}\leq V_{k\gamma}\leq V_{k_n\gamma} \quad &\text{for} \quad k_n\leq k\leq m_n \\ V_{m_n\gamma}\leq V_{k\gamma}\leq V_{k_{n+1}\gamma} \quad &\text{for} \quad m_n\leq k\leq k_{n+1} \end{aligned}$$

$$\Rightarrow \quad \inf_{n\geq n_0} V_{m_n\gamma} \;\leq\; V_{k\gamma} \;\leq\; \sup_{m\geq n_0} V_{k_n\gamma} \quad \forall k\geq k_{n_0}$$
$$\Rightarrow \quad \underline{\lim_n} V_{m_n\gamma} \;=\; \underline{V}_{\infty\gamma} \;\leq\; \overline{V}_{\infty\gamma} \;=\; \overline{\lim_n} V_{k_n\gamma} \tag{4}$$

Note the equalities in the last line. The $\geq$ holds, since $(V_{k_n\gamma})$ and $(V_{m_n\gamma})$ are subsequences of $(V_{k\gamma})$. Now if $\frac{a_n}{a_n+b_n}\geq\beta\ \forall n\geq n_0$ then $V_{k_n\gamma}=\frac{a_n+a_{n+1}+...}{a_n+b_n+a_{n+1}+b_{n+1}+...}\geq\beta$ $\forall n\geq n_0$. This implies

$$\underline{\lim_n}\tfrac{a_n}{a_n+b_n}\leq\underline{\lim_n} V_{k_n\gamma} \quad \text{and similarly} \quad \overline{\lim_n}\tfrac{a_n}{a_n+b_n}\geq\overline{\lim_n} V_{k_n\gamma}$$

Together this shows that $\lim_n V_{k_n\gamma}=\beta$ exists, if $\lim_n\frac{a_n}{a_n+b_n}=\beta$ exists. Similarly if $\frac{a_{n+1}}{b_n+a_{n+1}}\geq\alpha\ \forall n\geq n_0$ then $V_{m_n\gamma}=\frac{a_{n+1}+a_{n+2}+...}{b_n+a_{n+1}+b_{n+1}+a_{n+2}+...}\geq\alpha\ \forall n\geq n_0$. This implies

$$\underline{\lim_n}\tfrac{a_{n+1}}{b_n+a_{n+1}}\leq\underline{\lim_n} V_{m_n\gamma} \quad \text{and similarly} \quad \overline{\lim_n}\tfrac{a_{n+1}}{b_n+a_{n+1}}\geq\overline{\lim_n} V_{m_n\gamma}$$

Together this shows that $\lim_n V_{m_n\gamma}=\alpha$ exists, if $\lim_n\frac{a_{n+1}}{b_n+a_{n+1}}=\alpha$ exists. ■

Example 4 ($U_{1\infty}=V_{\infty\gamma}$) Constant rewards $r_k\equiv\alpha$ is a trivial example for which $U_{1\infty}=V_{\infty\gamma}=\alpha$ exist and are equal.

A more interesting example is $\boldsymbol{r}=1^1 0^2 1^3 0^4...$ of linearly increasing 0/1-run-length with $A_n=2n-1$ and $B_n=2n$, for which $U_{1\infty}=\frac{1}{2}$ exists. For quadratic discount $\gamma_k=\frac{1}{k(k+1)}$, using $\Gamma_k=\frac{1}{k}$, $h_k^{quasi}=k+1=\Theta(k)$, $k_n=(2n-1)(n-1)+1$, $m_n=(2n-1)n+1$, $a_n=\Gamma_{k_n}-\Gamma_{m_n}=\frac{A_n}{k_n m_n}\sim\frac{1}{2n^3}$, and $b_n=\Gamma_{m_n}-\Gamma_{k_{n+1}}=\frac{B_n}{m_n k_{n+1}}\sim\frac{1}{2n^3}$, we also get $V_{\infty\gamma}=\frac{1}{2}$. The values converge, since they average over increasingly many 1/0-runs, each of decreasing weight.

Example 5 (simple $U_{1\infty}\not\Rightarrow V_{\infty\gamma}$) Let us consider a very simple example with alternating rewards $\boldsymbol{r}=101010...$ and geometric discount $\gamma_k=\gamma^k$. It is immediate that $U_{1\infty}=\frac{1}{2}$ exists, but $\underline{V}_{\infty\gamma}=V_{2k,\gamma}=\frac{\gamma}{1+\gamma}<\frac{1}{1+\gamma}=V_{2k-1,\gamma}=\overline{V}_{\infty\gamma}$.

Example 6 ($U_{1\infty}\not\Rightarrow V_{\infty\gamma}$) Let us reconsider the more interesting example $\boldsymbol{r}=1^1 0^2 1^3 0^4...$ of linearly increasing 0/1-run-length with $A_n=2n-1$ and $B_n=2n$ for which $U_{1\infty}=\frac{1}{2}$ exists, as expected. On the other hand, for geometric discount $\gamma_k=\gamma^k$, using $\Gamma_k=\frac{\gamma^k}{1-\gamma}$ and $a_n=\Gamma_{k_n}-\Gamma_{m_n}=\frac{\gamma^{k_n}}{1-\gamma}[1-\gamma^{A_n}]$ and $b_n=\Gamma_{m_n}-\Gamma_{k_{n+1}}=\frac{\gamma^{m_n}}{1-\gamma}[1-\gamma^{B_n}]$, i.e. $\frac{b_n}{a_n}\sim\gamma^{A_n}\to 0$ and $\frac{a_{n+1}}{b_n}\sim\gamma^{B_n}\to 0$, we get $\underline{V}_{\infty\gamma}=\alpha=0<1=\beta=\overline{V}_{\infty\gamma}$. Again, this is plausible since for k at the beginning of a long run, $V_{k\gamma}$ is dominated by the reward 0/1 in this run, due to the bounded effective horizon of geometric $\boldsymbol{\gamma}$.

Example 7 ($V_{\infty\gamma}\not\Rightarrow U_{1\infty}$) Discounted may not imply average value on sequences of exponentially increasing run-length like $\boldsymbol{r}=1^1 0^2 1^4 0^8 1^{16}...$ with $A_n=2^{2n-2}=k_n$ and $B_n=2^{2n-1}=m_n$ for which $\underline{U}_{1\infty}=\frac{A_n}{A_n+B_n}=\frac{1}{3}<\frac{2}{3}=\frac{A_n}{B_{n-1}+A_n}=\overline{U}_{1\infty}$, i.e. $U_{1\infty}$ does not exist. On the other hand, $V_{\infty\gamma}$ exists for a discount with super-linear horizon like $\gamma_k=[k\ln^2 k]^{-1}$, since an increasing number of runs contribute to $V_{k\gamma}$: $\Gamma_k\sim\frac{1}{\ln k}$, hence $\Gamma_{k_n}\sim\frac{1}{(2n-2)\ln 2}$ and $\Gamma_{m_n}\sim\frac{1}{(2n-1)\ln 2}$, which implies $a_n=\Gamma_{k_n}-\Gamma_{m_n}\sim[4n^2\ln 2]^{-1}\sim\Gamma_{m_n}-\Gamma_{k_{n+1}}=b_n$, i.e. $V_{\infty\gamma}=\frac{1}{2}$ exists.

Example 8 (Non-monotone discount γ, $U_{1\infty}\neq V_{\infty\gamma}$) Monotonicity of γ in Theorems 15, 17, and 19 is necessary. As a simple counter-example consider alternating rewards $r_{2k}=0$ with arbitrary γ_{2k} and $r_{2k-1}=1$ with $\gamma_{2k-1}=0$, which implies $V_{k\gamma}\equiv 0$, but $U_{1\infty}=\frac{1}{2}$.

The above counter-example is rather simplistic. One may hope equivalence to hold on smoother γ like $\frac{\gamma_{k+1}}{\gamma_k}\to 1$. The following example shows that this condition alone is not sufficient. For a counter-example one needs an oscillating γ of constant relative amplitude, but increasing wavelength, e.g. $\gamma_k=[2+\cos(\pi\sqrt{2k})]/k^2$. For the sequence $\boldsymbol{r}=1^1 0^2 1^3 0^4...$ of Example 6 we had $U_{1\infty}=\frac{1}{2}$. Using $m_n=\frac{1}{2}(2n-\frac{1}{2})^2+\frac{7}{8}$ and $k_{n+1}=\frac{1}{2}(2n+\frac{1}{2})^2+\frac{7}{8}$, and replacing the sums in the definitions of a_n and b_n by integrals, we get $a_n\sim\frac{1}{n^3}[\frac{1}{2}-\frac{1}{\pi}]$ and $b_n\sim\frac{1}{n^3}[\frac{1}{2}+\frac{1}{\pi}]$, which implies that $V_{\infty\gamma}=\frac{1}{2}-\frac{1}{\pi}$ exists, but differs from $U_{1\infty}=\frac{1}{2}$.

Example 9 (Oscillating horizon) It is easy to construct a discount γ for which $\sup_k\frac{\Gamma_k}{k\gamma_k}=\infty$ *and* $\sup_k\frac{k\gamma_k}{\Gamma_k}=\infty$ by alternatingly patching together discounts with super- and sub-linear quasi-horizon h_k^{quasi}. For instance choose $\gamma_k\propto\gamma^k$ geometric until $\frac{\Gamma_k}{k\gamma_k}<\frac{1}{n}$, then $\gamma_k\propto\frac{1}{k\ln^2 k}$ harmonic until $\frac{\Gamma_k}{k\gamma_k}>n$, then repeat with $n\leadsto n+1$. The proportionality constants can be chosen to insure monotonicity of γ. For such γ neither Theorem 15 nor Theorem 17 is applicable, only Theorem 19.

3 Average Value

We now take a closer look at the (total) average value U_{1m} and relate it to the future average value U_{km}, an intermediate quantity we need later. We recall the definition of the average value:

Definition 10 (Average value, U_{1m}) *Let $r_i\in[0,1]$ be the reward at time $i\in\mathbb{N}$. Then*

$$U_{1m}\;:=\;\frac{1}{m}\sum_{i=1}^{m}r_i\;\in[0,1]$$

is the average value from time 1 to m, and $U_{1\infty}:=\lim_{m\to\infty}U_{1m}$ the average value if it exists.

We also need the average value $U_{km}:=\frac{1}{m-k+1}\sum_{i=k}^{m}r_i$ from k to m and the following Lemma.

Lemma 11 (Convergence of future average value, $U_{k\infty}$) *For $k_m\leq m\to\infty$ and every k we have*

$$U_{1m}\to\alpha\quad\Leftrightarrow\quad U_{km}\to\alpha\qquad\begin{array}{ll}\Rightarrow & U_{k_m m}\to\alpha\quad \textit{if}\quad \sup_m\frac{k_m-1}{m}<1\\ \Leftarrow & U_{k_m m}\to\alpha\end{array}$$

The first equivalence states the obvious fact (and problem) that any finite initial part has no influence on the average value $U_{1\infty}$. Chunking together many $U_{k_m m}$ implies the last $\Leftarrow$. The $\Rightarrow$ only works if we average in $U_{k_m m}$ over sufficiently many rewards, which the stated condition ensures ($\boldsymbol{r}=101010...$ and $k_m=m$ is a simple counter-example). Note that $U_{km_k}\to\alpha$ for $m_k\geq k\to\infty$ implies $U_{1m_k}\to\alpha$, but not necessarily $U_{1m}\to\alpha$ (e.g. in Example 7, $U_{1m_k}=\frac{1}{3}$ and $\frac{k-1}{m_k}\to 0$ imply $U_{km_k}\to\frac{1}{3}$ by (5), but $U_{1\infty}$ does not exist).

Proof. The trivial identity $mU_{1m}=(k-1)U_{1,k-1}+(m-k+1)U_{km}$ implies $U_{km}-U_{1m}=\frac{k-1}{m-k+1}(U_{1m}-U_{1,k-1})$ implies

$$|U_{km}-U_{1m}| \;\leq\; \frac{|U_{1m}-U_{1,k-1}|}{\frac{m}{k-1}-1} \tag{5}$$

$\Leftrightarrow$) The numerator is bounded by 1, and for fixed k and $m\to\infty$ the denominator tends to ∞, which proves $\Leftrightarrow$.

$\Rightarrow$) We choose (small) $\varepsilon>0$, m_ε large enough so that $|U_{1m}-\alpha|<\varepsilon\ \forall m\geq m_\varepsilon$, and $m\geq\frac{m_\varepsilon}{\varepsilon}$. If $k:=k_m\leq m_\varepsilon$, then (5) is bounded by $\frac{1}{1/\varepsilon-1}$. If $k:=k_m>m_\varepsilon$, then (5) is bounded by $\frac{2\varepsilon}{1/c-1}$, where $c:=\sup_k\frac{k_m-1}{m}<1$. This shows that $|U_{k_m m}-U_{1m}|=O(\varepsilon)$ for large m, which implies $U_{k_m m}\to\alpha$.

$\Leftarrow$) We partition the time-range $\{1...m\}=\bigcup_{n=1}^{L}\{k_{m_n}...m_n\}$, where $m_1:=m$ and $m_{n+1}:=k_{m_n}-1$. We choose (small) $\varepsilon>0$, m_ε large enough so that $|U_{k_m m}-\alpha|<\varepsilon$ $\forall m\geq m_\varepsilon$, $m\geq\frac{m_\varepsilon}{\varepsilon}$, and l so that $k_{m_l}\leq m_\varepsilon\leq m_l$. Then

$$\begin{aligned}
U_{1m} &= \frac{1}{m}\left[\sum_{n=1}^{l}+\sum_{n=l+1}^{L}\right](m_n-k_{m_n}+1)U_{k_{m_n}m_n} \\
&\leq \frac{1}{m}\sum_{n=1}^{l}(m_n-k_{m_n}+1)(\alpha+\varepsilon)+\frac{m_{l+1}-k_{m_L}+1}{m} \\
&\leq \frac{m_1-k_{m_l}+1}{m}(\alpha+\varepsilon)+\frac{k_{m_l}}{m} \;\leq\; (\alpha+\varepsilon)+\varepsilon \\
\text{Similarly}\quad U_{1m} &\geq \frac{m_1-k_{m_l}+1}{m}(\alpha-\varepsilon)\geq\frac{m-m_\varepsilon}{m}(\alpha-\varepsilon)\geq(1-\varepsilon)(\alpha-\varepsilon)
\end{aligned}$$

This shows that $|U_{1m}-\alpha|\leq 2\varepsilon$ for sufficiently large m, hence $U_{1m}\to\alpha$. ■

4 Discounted Value

We now take a closer look at the (future) discounted value $V_{k\gamma}$ for general discounts $\boldsymbol{\gamma}$, and prove some useful elementary asymptotic properties of discount γ_k and normalizer Γ_k. We recall the definition of the discounted value:

Definition 12 (Discounted value, $V_{k\boldsymbol{\gamma}}$) *Let $r_i \in [0,1]$ be the reward and $\gamma_i \geq 0$ a discount at time $i \in \mathbb{N}$, where $\boldsymbol{\gamma}$ is assumed to be summable in the sense that $0 < \Gamma_k := \sum_{i=k}^{\infty} \gamma_i < \infty$. Then*

$$V_{k\gamma} \;:=\; \frac{1}{\Gamma_k} \sum_{i=k}^{\infty} \gamma_i r_i \;\in [0,1]$$

is the $\boldsymbol{\gamma}$-discounted future value and $V_{\infty\gamma} := \lim_{k\to\infty} V_{k\gamma}$ its limit if it exists.

We say that $\boldsymbol{\gamma}$ is *monotone* if $\gamma_{k+1} \leq \gamma_k \forall k$. Note that monotonicity and $\Gamma_k > 0$ $\forall k$ implies $\gamma_k > 0$ $\forall k$ and convexity of Γ_k.

Lemma 13 (Discount properties, $\boldsymbol{\gamma}/\Gamma$)

$$\begin{aligned} &i) \quad \frac{\gamma_{k+1}}{\gamma_k} \to 1 \quad \Leftrightarrow \quad \frac{\gamma_{k+\Delta}}{\gamma_k} \to 1 \quad \forall \Delta \in \mathbb{N} \\ &ii) \quad \frac{\gamma_k}{\Gamma_k} \to 0 \quad \Leftrightarrow \quad \frac{\Gamma_{k+1}}{\Gamma_k} \to 1 \quad \Leftrightarrow \quad \frac{\Gamma_{k+\Delta}}{\Gamma_k} \to 1 \quad \forall \Delta \in \mathbb{N} \end{aligned}$$

Furthermore, (i) implies (ii), but not necessarily the other way around (even not if $\boldsymbol{\gamma}$ is monotone).

Proof. $(i) \Rightarrow \frac{\gamma_{k+\Delta}}{\gamma_k} = \prod_{i=k}^{\Delta-1} \frac{\gamma_{i+1}}{\gamma_i} \stackrel{k\to\infty}{\longrightarrow} 1$, since Δ is finite.
$(i) \Leftarrow$ Set $\Delta = 1$.
(ii) The first equivalence follows from $\Gamma_k = \gamma_k + \Gamma_{k+1}$. The proof for the second equivalence is the same as for (i) with γ replaced by Γ.
$(i) \Rightarrow (ii)$ Choose $\varepsilon > 0$. (i) implies $\frac{\gamma_{k+1}}{\gamma_k} \geq 1-\varepsilon$ $\forall' k$ implies

$$\Gamma_k \;=\; \sum_{i=k}^{\infty} \gamma_i \;=\; \gamma_k \sum_{i=k}^{\infty} \prod_{j=k}^{i-1} \frac{\gamma_{i+1}}{\gamma_i} \;\geq\; \gamma_k \sum_{i=k}^{\infty} (1-\varepsilon)^{i-k} \;=\; \gamma_k/\varepsilon$$

hence $\frac{\gamma_k}{\Gamma_k} \leq \varepsilon$ $\forall' k$, which implies $\frac{\gamma_k}{\Gamma_k} \to 0$.
$(i) \not\Leftarrow (ii)$ Consider counter-example $\gamma_k = 4^{-\lceil \log_2 k \rceil}$, i.e. $\gamma_k = 4^{-n}$ for $2^{n-1} < k \leq 2^n$. Since $\Gamma_k \geq \sum_{i=2^n}^{\infty} \gamma_i = 2^{-n-1}$ we have $0 \leq \frac{\gamma_k}{\Gamma_k} \leq 2^{1-n} \to 0$, but $\frac{\gamma_{k+1}}{\gamma_k} = \frac{1}{4} \not\to 1$ for $k = 2^n$. ■

5 Average Implies Discounted Value

We now show that existence of $\lim_m U_{1m}$ can imply existence of $\lim_k V_{k\gamma}$ and their equality. The necessary and sufficient condition for this implication to hold is roughly that the effective horizon grows linearly with k or faster. The auxiliary quantity U_{km} is in a sense closer to $V_{k\gamma}$ than U_{1m} is, since the former two both average from k (approximately) to some (effective) horizon. If $\boldsymbol{\gamma}$ is sufficiently smooth, we can chop the area under the graph of $V_{k\gamma}$ (as a function of k) "vertically" approximately into a sum of average values, which implies

Proposition 14 (Future average implies discounted value, $U_\infty \Rightarrow V_{\infty\gamma}$)
Assume $k \leq m_k \to \infty$ and monotone $\boldsymbol{\gamma}$ with $\frac{\gamma_{m_k}}{\gamma_k} \to 1$. If $U_{km_k} \to \alpha$, then $V_{k\gamma} \to \alpha$.

The proof idea is as follows: Let $k_1 = k$ and $k_{n+1} = m_{k_n} + 1$. Then for large k we get

$$\begin{aligned} V_{k\gamma} &= \frac{1}{\Gamma_k} \sum_{n=1}^{\infty} \sum_{i=k_n}^{m_{k_n}} \gamma_i r_i \approx \frac{1}{\Gamma_k} \sum_{n=1}^{\infty} \gamma_{k_n} (k_{n+1} - k_n) U_{k_n m_{k_n}} \\ &\approx \frac{\alpha}{\Gamma_k} \sum_{n=1}^{\infty} \gamma_{k_n} (k_{n+1} - k_n) \approx \frac{\alpha}{\Gamma_k} \sum_{n=1}^{\infty} \sum_{i=k_n}^{m_{k_n}} \gamma_i = \alpha \end{aligned}$$

The (omitted) formal proof specifies the approximation error, which vanishes for $k \to \infty$.

Actually we are more interested in relating the (total) average value $U_{1\infty}$ to the (future) discounted value $V_{k\gamma}$. The following (first main) Theorem shows that for linearly or faster increasing quasi-horizon, we have $V_{\infty\gamma} = U_{1\infty}$, provided the latter exists.

Theorem 15 (Average implies discounted value, $U_{1\infty} \Rightarrow V_{\infty\gamma}$)
Assume $\sup_k \frac{k\gamma_k}{\Gamma_k} < \infty$ and monotone $\boldsymbol{\gamma}$. If $U_{1m} \to \alpha$, then $V_{k\gamma} \to \alpha$.

For instance, quadratic, power and harmonic discounts satisfy the condition, but faster-than-power discount like geometric do not. Note that Theorem 15 does not imply Proposition 14.

The intuition of Theorem 15 for binary reward is as follows: For U_{1m} being able to converge, the length of a run must be small compared to the total length m up to this run, i.e. $o(m)$. The condition in Theorem 15 ensures that the quasi-horizon $h_k^{quasi} = \Omega(k)$ increases faster than the run-lengths $o(k)$, hence $V_{k\gamma} \approx U_{k\Omega(k)} \approx U_{1m}$ (Lemma 11) asymptotically averages over many runs, hence should also exist. The formal proof "horizontally" slices $V_{k\gamma}$ into a weighted sum of average rewards U_{1m}. Then $U_{1m} \to \alpha$ implies $V_{k\gamma} \to \alpha$.

Proof. We represent $V_{k\gamma}$ as a δ_j-weighted mixture of U_{1j}'s for $j \geq k$, where $\delta_j := \gamma_j - \gamma_{j+1} \geq 0$. The condition $\infty > c \geq \frac{k\gamma_k}{\Gamma_k} =: c_k$ ensures that the excessive initial part $\propto U_{1,k-1}$ is "negligible". It is easy to show that

$$\sum_{j=i}^{\infty} \delta_j = \gamma_i \quad \text{and} \quad \sum_{j=k}^{\infty} j\delta_j = (k-1)\gamma_k + \Gamma_k$$

We choose some (small) $\varepsilon>0$, and m_ε large enough so that $|U_{1m}-\alpha|<\varepsilon\ \forall m\geq m_\varepsilon$. Then, for $k>m_\varepsilon$ we get

$$\begin{aligned}
V_{k\gamma} &= \frac{1}{\Gamma_k}\sum_{i=k}^{\infty}\gamma_i r_i \;=\; \frac{1}{\Gamma_k}\sum_{i=k}^{\infty}\sum_{j=i}^{\infty}\delta_j r_i \;=\; \frac{1}{\Gamma_k}\sum_{j=k}^{\infty}\sum_{i=k}^{j}\delta_j r_i \\
&= \frac{1}{\Gamma_k}\sum_{j=k}^{\infty}\delta_j[jU_{1j}-(k-1)U_{1,k-1}] \\
&\leqslant \frac{1}{\Gamma_k}\sum_{j=k}^{\infty}\delta_j[j(\alpha\pm\varepsilon)-(k-1)(\alpha\mp\varepsilon)] \\
&= \frac{1}{\Gamma_k}[(k-1)\gamma_k+\Gamma_k](\alpha\pm\varepsilon)-\frac{1}{\Gamma_k}\gamma_k(k-1)(\alpha\mp\varepsilon) \\
&= \alpha\pm\Big(1+\frac{2(k-1)\gamma_k}{\Gamma_k}\Big)\varepsilon \;\leqslant\; \alpha\pm(1+2c_k)\varepsilon
\end{aligned}$$

i.e. $|V_{k\gamma}-\alpha|<(1+2c_k)\varepsilon\leq(1+2c)\varepsilon\ \forall k>m_\varepsilon$, which implies $V_{k\gamma}\to\alpha$. ■

Theorem 15 can, for instance, be applied to Example 4. Examples 5, 6, and 8 demonstrate that the conditions in Theorem 15 cannot be dropped. The following proposition shows more strongly, that the sufficient condition is actually necessary (modulo monotonicity of $\boldsymbol{\gamma}$), i.e. cannot be weakened.

Proposition 16 ($\boldsymbol{U_{1\infty}\not\Rightarrow V_{\infty\gamma}}$)
For every monotone $\boldsymbol{\gamma}$ with $\sup_k\frac{k\gamma_k}{\Gamma_k}=\infty$, there are $\boldsymbol{r}$ for which $U_{1\infty}$ exists, but not $V_{\infty\gamma}$.

The proof idea is to construct a binary $\boldsymbol{r}$ such that all change points k_n and m_n satisfy $\Gamma_{k_n}\approx 2\Gamma_{m_n}$. This ensures that $V_{k_n\gamma}$ receives a significant contribution from 1-run n, i.e. is large. Choosing $k_{n+1}\gg m_n$ ensures that $V_{m_n\gamma}$ is small, hence $V_{k\gamma}$ oscillates. Since the quasi-horizon $h_k^{quasi}\neq\Omega(k)$ is small, the 1-runs are short enough to keep U_{1m} small so that $U_{1\infty}=0$.

Proof. The assumption ensures that there exists a sequence m_1, m_2, m_3, ... for which

$$\frac{m_n\gamma_{m_n}}{\Gamma_{m_n}}\geq n^2 \quad \text{We further (can) require } \Gamma_{m_n}<\tfrac{1}{2}\Gamma_{m_{n-1}+1} \qquad (m_0:=0)$$

For each m_n we choose k_n such that $\Gamma_{k_n}\approx 2\Gamma_{m_n}$. More precisely, since Γ is monotone decreasing and $\Gamma_{m_n}<2\Gamma_{m_n}\leq\Gamma_{m_{n-1}+1}$, there exists (a unique) k_n in the range $m_{n-1}<k_n<m_n$ such that $\Gamma_{k_n+1}<2\Gamma_{m_n}\leq\Gamma_{k_n}$. We choose a binary reward sequence with

$r_k=1$ iff $k_n\leq k<m_n$ for some n. This implies

$$\begin{aligned}
n^2 &\leq \frac{m_n\gamma_{m_n}}{\Gamma_{m_n}} = \frac{m_n}{m_n-k_n-1}\frac{(m_n-k_n-1)\gamma_{m_n}}{\Gamma_{m_n}} \\
&\leq \frac{m_n}{m_n-k_n-1}\frac{\Gamma_{k_n+1}-\Gamma_{m_n}}{\Gamma_{m_n}} \leq \frac{m_n}{m_n-k_n-1} \\
\Longrightarrow \quad \frac{m_n-k_n}{m_n} &= \frac{m_n-k_n-1}{m_n}+\frac{1}{m_n} \leq \frac{1}{n^2}+\frac{\gamma_{m_n}}{\Gamma_{m_n}}\frac{1}{n^2} \leq \frac{2}{n^2} \\
\Longrightarrow \quad U_{1m_n} &\leq \frac{1}{m_n}[k_l-1]+\frac{1}{m_n}\sum_{n'=l}^{n}[m_{n'}-k_{n'}] \leq \frac{k_l}{m_n}+\sum_{n'=l}^{n}\frac{m_{n'}-k_{n'}}{m_{n'}} \\
&\leq \frac{k_l}{m_n}+\sum_{n'=l}^{n}\frac{2}{n'^2} \leq \frac{k_l}{m_n}+\frac{2}{l-1}
\end{aligned}$$

hence by (1) we have $\overline{U}_{1\infty}=\overline{\lim}_n U_{1,m_n-1}\leq\frac{2}{l-1}\ \forall l$, hence $U_{1\infty}=0$. On the other hand

$$\Gamma_{k_n}V_{k_n\gamma} = [\Gamma_{k_n}-\Gamma_{m_n}]+\Gamma_{m_n}V_{m_n\gamma} \quad\Rightarrow\quad \frac{1-V_{k_n\gamma}}{1-V_{m_n\gamma}} = \frac{\Gamma_{m_n}}{\Gamma_{k_n}} \leq \tfrac{1}{2}$$

This shows that $V_{k\gamma}$ cannot converge to an $\alpha<1$. Theorem 19 and $U_{1\infty}=0$ implies that $V_{k\gamma}$ can also not converge to 1, hence $V_{\infty\gamma}$ does not exist. ■

6 Discounted Implies Average Value

We now turn to the converse direction that existence of $V_{\infty\gamma}$ can imply existence of $U_{1\infty}$ and their equality, which holds under a nearly converse condition on the discount: Roughly, the effective horizon has to grow linearly with k or slower.

Theorem 17 (Discounted implies average value, $V_{\infty\boldsymbol{\gamma}}\Rightarrow U_{1\infty}$)
Assume $\sup_k\frac{\Gamma_k}{k\gamma_k}<\infty$ *and monotone* $\boldsymbol{\gamma}$*. If* $V_{k\gamma}\to\alpha$*, then* $U_{1m}\to\alpha$*.*

For instance, power or faster and geometric discounts satisfy the condition, but harmonic does not. Note that power discounts satisfy the conditions of Theorems 15 *and* 17, i.e. $U_{1\infty}$ exists iff $V_{\infty\gamma}$ in this case.

The intuition behind Theorem 17 for binary reward is as follows: The run-length needs to be small compared to the quasi-horizon, i.e. $o(h_k^{quasi})$, to ensure convergence of $V_{k\gamma}$. The condition in Theorem 17 ensures that the quasi-horizon $h_k^{quasi}=O(k)$ grows at most linearly, hence the run-length $o(m)$ is a small fraction of the sequence up to m. This ensures that U_{1m} ceases to oscillate. The formal proof slices U_{1m} in "curves" to a weighted mixture of discounted values $V_{k\gamma}$. Then $V_{k\gamma}\to\alpha$ implies $U_{1m}\to\alpha$.

Proof. We represent U_{km} as a ($0 \le b_j$-weighted) mixture of $V_{j\gamma}$ for $k \le j \le m$. The condition $c := \sup_k \frac{\Gamma_k}{k\gamma_k} < \infty$ ensures that the redundant tail $\propto V_{m+1,\gamma}$ is "negligible". Fix k large enough so that $|V_{j\gamma} - \alpha| < \varepsilon \ \forall j \ge k$. Then

$$\begin{aligned} \sum_{j=k}^{m} b_j(\alpha \mp \varepsilon) \ &\lessgtr \ \sum_{j=k}^{m} b_j U_{1j} \ = \ \sum_{j=k}^{m} \frac{b_j}{\Gamma_j} \sum_{i=j}^{m} \gamma_i r_i \ + \ \sum_{j=k}^{m} \frac{b_j}{\Gamma_j} \sum_{i=m+1}^{\infty} \gamma_i r_i \qquad (6) \\ &= \ \sum_{i=k}^{m} \left(\sum_{j=k}^{i} \frac{b_j}{\Gamma_j} \right) \gamma_i r_i \ + \ \left(\sum_{j=k}^{m} \frac{b_j}{\Gamma_j} \right) \Gamma_{m+1} V_{m+1,\gamma} \end{aligned}$$

In order for the first term on the r.h.s. to be a uniform mixture, we need

$$\sum_{j=k}^{i} \frac{b_j}{\Gamma_j} \ = \ \frac{1}{\gamma_i} \frac{1}{m-k+1} \quad (k \le i \le m) \qquad (7)$$

Setting $i = k$ and, respectively, subtracting an $i \rightsquigarrow i-1$ term we get

$$\frac{b_k}{\Gamma_k} = \frac{1}{\gamma_k} \frac{1}{m-k+1} \quad \text{and} \quad \frac{b_i}{\Gamma_i} = \left(\frac{1}{\gamma_i} - \frac{1}{\gamma_{i-1}} \right) \frac{1}{m-k+1} \ge 0 \quad \text{for} \quad k < i \le m$$

So we can evaluate the b-sum in the l.h.s. of (6) to

$$\begin{aligned} \sum_{j=k}^{m} b_j \ &= \ \frac{1}{m-k+1} \left[\sum_{j=k+1}^{m} \left(\frac{\Gamma_j}{\gamma_j} - \frac{\Gamma_j}{\gamma_{j-1}} \right) + \frac{\Gamma_k}{\gamma_k} \right] \\ &= \ \frac{1}{m-k+1} \left[\sum_{j=k}^{m} \left(\frac{\Gamma_j}{\gamma_j} - \frac{\Gamma_{j+1}}{\gamma_j} \right) + \frac{\Gamma_{m+1}}{\gamma_m} \right] \\ &= \ 1 + \frac{\Gamma_{m+1}}{\gamma_m (m-k+1)} \ =: \ 1 + c_m \qquad (8) \end{aligned}$$

where we shifted the sum index in the second equality, and used $\Gamma_j - \Gamma_{j+1} = \gamma_j$ in the third equality. Inserting (7) and (8) into (6) we get

$$(1+c_m)(\alpha \mp \varepsilon) \ \lessgtr \ \sum_{i=k}^{m} \frac{1}{m-k+1} r_i + \frac{\Gamma_{m+1}}{\gamma_m (m-k+1)} V_{m+1,\gamma} \ \lessgtr \ U_{km} + c_m(\alpha \pm \varepsilon)$$

Note that the excess c_m over unity in (8) equals the coefficient of the tail contribution $V_{m+1,\gamma}$. The above bound shows that

$$|U_{km} - \alpha| \ \le \ (1 + 2c_m)\varepsilon \ \le (1 + 4c)\varepsilon \quad \text{for} \quad m \ge 2k$$

Hence $U_{m/2,m} \to \alpha$, which implies $U_{1m} \to \alpha$ by Lemma 11. ■

Theorem 17 can, for instance, be applied to Example 4. Examples 7 and 8 demonstrate that the conditions in Theorem 17 cannot be dropped. The following proposition shows more strongly, that the sufficient condition is actually necessary, i.e. cannot be weakened.

Proposition 18 ($V_{\infty\gamma} \not\Rightarrow U_{1\infty}$)
For every monotone $\boldsymbol{\gamma}$ with $\sup_k \frac{\Gamma_k}{k\gamma_k} = \infty$, there are $\boldsymbol{r}$ for which $V_{\infty\gamma}$ exists, but not $U_{1\infty}$.

Proof. The assumption ensures that there exists a sequence k_1, k_2, k_3, ... for which

$$\frac{k_n\gamma_{k_n}}{\Gamma_{k_n}} \leq \frac{1}{n^2} \quad \text{We further choose} \quad k_{n+1} > 8k_n$$

We choose a binary reward sequence with $r_k = 1$ iff $k_n \leq k < m_n := 2k_n$.

$$\begin{aligned} V_{k_n\gamma} &= \frac{1}{\Gamma_{k_n}} \sum_{l=n}^{\infty} \gamma_{k_l} + ... + \gamma_{2k_l-1} \leq \frac{1}{\Gamma_{k_n}} \sum_{l=n}^{\infty} k_l\gamma_{k_l} \\ &\leq \sum_{l=n}^{\infty} \frac{k_l\gamma_{k_l}}{\Gamma_{k_l}} \leq \sum_{l=n}^{\infty} \frac{1}{l^2} \leq \frac{1}{n-1} \to 0 \end{aligned}$$

which implies $V_{\infty\gamma} = 0$ by (4). In a sense the 1-runs become asymptotically very sparse. On the other hand,

$$\begin{aligned} U_{1,m_n-1} &\geq \tfrac{1}{m_n}[r_{k_n} + ... + r_{m_n-1}] = \tfrac{1}{m_n}[m_n - k_n] = \tfrac{1}{2} \quad \text{but} \\ U_{1,k_{n+1}-1} &\leq \tfrac{1}{k_{n+1}-1}[r_1 + ... + r_{m_{n-1}}] \leq \tfrac{1}{8k_n}[m_n - 1] \leq \tfrac{1}{4}, \end{aligned}$$

hence $U_{1\infty}$ does not exist. ■

7 Average Equals Discounted Value

Theorem 15 and 17 together imply for nearly all discount types (all in our table) that $U_{1\infty} = V_{\infty\gamma}$ if $U_{1\infty}$ and $V_{\infty\gamma}$ both exist. But Example 9 shows that there are $\boldsymbol{\gamma}$ for which simultaneously $\sup_k \frac{\Gamma_k}{k\gamma_k} = \infty$ *and* $\sup_k \frac{k\gamma_k}{\Gamma_k} = \infty$, i.e. neither Theorem 15, nor Theorem 17 applies. This happens for quasi-horizons that grow alternatingly super- and sub-linear. Luckily, it is easy to also cover this missing case, and we get the remarkable result that $U_{1\infty}$ equals $V_{\infty\gamma}$ if both exist, for *any* monotone discount sequence $\boldsymbol{\gamma}$ and *any* reward sequence $\boldsymbol{r}$, whatsoever.

Theorem 19 (Average equals discounted value, $U_{1\infty} = V_{\infty\gamma}$)
Assume monotone $\boldsymbol{\gamma}$ and that $U_{1\infty}$ and $V_{\infty\gamma}$ exist. Then $U_{1\infty} = V_{\infty\gamma}$.

Proof. Case 1, $\sup_k \frac{\Gamma_k}{k\gamma_k} < \infty$: By assumption, there exists an α such that $V_{k\gamma} \to \alpha$. Theorem 17 now implies $U_{1m} \to \alpha$, hence $U_{1\infty} = V_{\infty\gamma} = \alpha$.

Case 2, $\sup_k \frac{\Gamma_k}{k\gamma_k} = \infty$: This implies that there is an infinite subsequence $k_1 < k_2 < k_3$,... for which $\Gamma_{k_i}/k_i\gamma_{k_i} \to \infty$, i.e. $c_{k_i} := k_i\gamma_{k_i}/\Gamma_{k_i} \leq c < \infty$. By assumption, there exists an α such that $U_{1m} \to \alpha$. If we look at the proof of Theorem 15, we

see that it still implies $|V_{k_i\gamma}-\alpha|<(1+c_{k_i})\varepsilon\leq(1+2c)\varepsilon$ on this subsequence. Hence $V_{k_i\gamma}\to\alpha$. Since we assumed existence of the limit $V_{k\gamma}$ this shows that the limit necessarily equals α, i.e. again $U_{1\infty}=V_{\infty\gamma}=\alpha$. ■

Considering the simplicity of the statement in Theorem 19, the proof based on the proofs of Theorems 15 and 17 is remarkably complex. A simpler proof, if it exists, probably avoids the separation of the two (discount) cases.

Example 8 shows that the monotonicity condition in Theorem 19 cannot be dropped.

8 Discussion

We showed that asymptotically, discounted and average value are the same, provided both exist. This holds for essentially arbitrary discount sequences (interesting since geometric discount leads to agents with bounded horizon) and arbitrary reward sequences (important since reality is neither ergodic nor MDP). Further, we exhibited the key role of power discounting with linearly increasing effective horizon. First, it separates the cases where existence of $U_{1\infty}$ implies/is-implied-by existence of $V_{\infty\gamma}$. Second, it neither requires nor introduces any artificial time-scale; it results in an increasingly farsighted agent with horizon proportional to its own age. In particular, we advocate the use of quadratic discounting $\gamma_k=1/k^2$. All our proofs provide convergence rates, which could be extracted from them. For simplicity we only stated the asymptotic results. The main theorems can also be generalized to probabilistic environments. Monotonicity of $\boldsymbol{\gamma}$ and boundedness of rewards can possibly be somewhat relaxed. A formal relation between effective horizon and the introduced quasi-horizon may be interesting.

References

[AA99] K. E. Avrachenkov and E. Altman. Sensitive discount optimality via nested linear programs for ergodic Markov decision processes. In *Proceedings of Information Decision and Control 99*, pages 53–58, Adelaide, Australia, 1999. IEEE.

[BF85] D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, 1985.

[BT96] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

[FLO02] S. Frederick, G. Loewenstein, and T. O'Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40:351–401, 2002.

[Hut02] M. Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proc. 15th Annual Conf. on Computational Learning Theory (COLT'02)*, volume 2375 of *LNAI*, pages 364–379, Sydney, 2002. Springer, Berlin.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

[Kak01] S. Kakade. Optimizing average reward using discounted rewards. In *Proc. 14th Conf. on Computational Learning Theory (COLT'01)*, volume 2111 of *LNCS*, pages 605–615, Amsterdam, 2001. Springer.

[Kel81] F. P. Kelly. Multi-armed bandits with discount factor near one: The Bernoulli case. *Annals of Statistics*, 9:987–1001, 1981.

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

[Mah96] S. Mahadevan. Sensitive discount optimality: Unifying discounted and average reward reinforcement learning. In *Proc. 13th International Conference on Machine Learning*, pages 328–336. Morgan Kaufmann, 1996.

[RN03] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 2003.

[Sam37] P. Samuelson. A note on measurement of utility. *Review of Economic Studies*, 4:155–161, 1937.

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[Str56] R. H. Strotz. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, 23:165–180, 1955–1956.

[VW04] N. Vieille and J. W. Weibull. Dynamic optimization with non-exponential discounting: On the uniqueness of solutions. Technical Report WP No. 577, Department of Economics, Boston Univeristy, Boston, MA, 2004.