General Discounting versus Average Reward

Marcus Hutter

http://www.hutter1.de

IDSIA RSISE ANU NICTA

ALT, 7 - 10 October 2006

Contents

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.

Setup: Rewards, Values, Discounts

Bounded reward: $r_k \in [a,b]$ at time $k \in I\!N$

Total average value: $U_{1m} := \frac{1}{m}[r_1 + ... + r_m]$

Monotone discount sequence: $\gamma_1 \geq \gamma_2 \geq \gamma_3 ... > 0$

Summable normalizer: $\Gamma_k := \gamma_k + \gamma_{k+1} + ... < \infty$

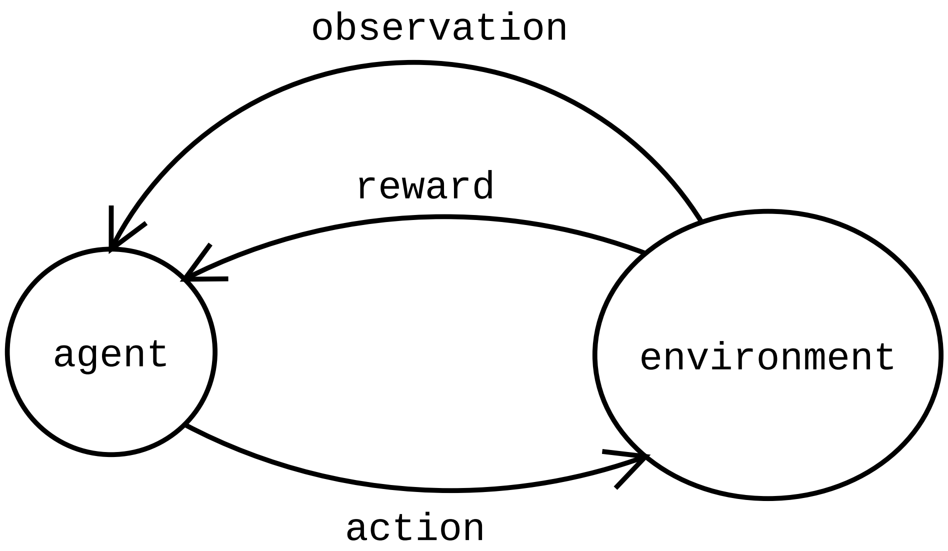
Future discounted value: $V_{k\gamma} := \frac{1}{\Gamma_k} \sum_{i=k}^{\infty} \gamma_i r_i$

Main Result

Theorem 1 (Average equals discounted value, $U_{1\infty} = V_{\infty\gamma}$)
Asymptotically, the average value coincides with the discounted value, i.e. $\lim_{m\to\infty} U_{1m} = \lim_{k\to\infty} V_{k\gamma}$, provided both limits exist.

Reinforcement Learning Setup

- An agent acts and gets rewarded for his actions in cycles.
  [Russell&Norvig 2003, Hutter 2005]



- Simplifying assumption: agent and environment are deterministic.

- Generic goal: find action sequence (policy) that maximizes reward.

Which reward $r_1, r_2, r_3, \ldots$?

Average Reward

Consider total reward sum or equivalently the average reward:

Definition 2 (Average value) $U_{1m} := \frac{1}{m}[r_1 + ... + r_m]$

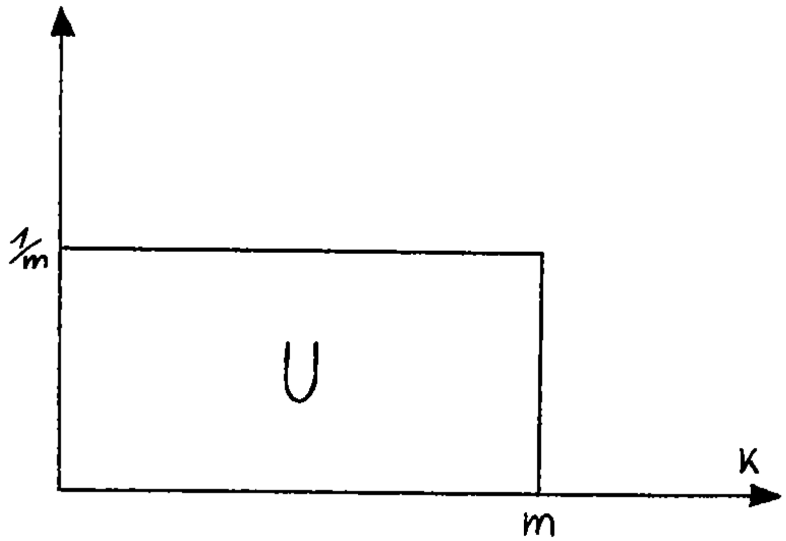
where m should be the lifespan of the agent.

Pro:

- Simplest reasonable measure of performance.

Problems:

- lifetime m is often not known in advance.
- no bias towards early rewards.



Idea: Infinite horizon $m \to \infty$: Problems:

- immortal agents are lazy. [Hutter 2005]
- limit $U_{1\infty}$ may not exist.

Geometric≡Exponential Discount

Geometrically discounted reward sum: $V_{k\gamma} := (1-\gamma)\sum_{i=k}^{\infty}\gamma^{i-k}r_i$ with $0 \leq \gamma < 1$. [Samuelson 1937, Bertsekas&Tsitsiklis 1996, Sutton&Barto 1998, ...]

Pro: Preference towards early rewards and leads to consistent policies in the sense that the $V_{k\gamma}$ maximizing policies are the same for all k (the agent does not change his mind).

Problems:

Effective finite moving horizon $h^{eff} \approx \ln\gamma^{-1}$

can lead to suboptimal behavior:

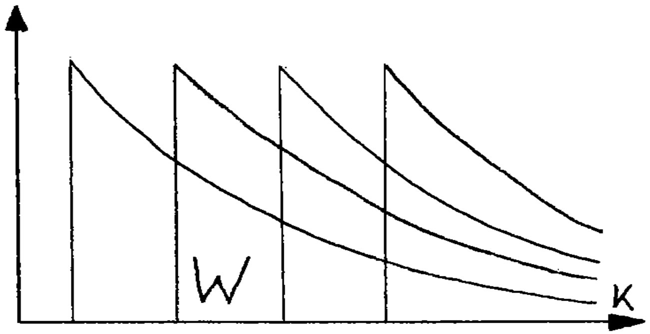
- not self-optimizing for Bandits [Berry&Fristedt 1985, Kumar&Varaiya 1986].
- for every h^{eff} there is a "game" needing larger h^{eff}.

Solution Attempts

Moving horizon: $U_{k,k+h-1} := \frac{1}{h}[r_k + ... + r_{k+h-1}]$
(popular for minimax tree truncation in zero sum games)
Problem: Can lead to inconsistent strategies (agent changes his mind)

Discount $\gamma \to 1$: $\Rightarrow h^{eff} \to \infty \Rightarrow$ defect decreases [Kelly 1981].
Similar and related to $m \to \infty$ [Kakade 2001].
Problems: - limits $\lim_{\gamma\to 1} V_{1\gamma}$ and $\lim_{m\to\infty} U_{1m}$ exist may not exist beyond ergodic MDPs.
[Mahadevan 1996 and Avrachenkov&Altman 1999 consider higher order terms]
- but real world is neither ergodic nor completely observable.
- Either fix $\gamma < 1$ (how?) or dynamically adapt $\gamma \overset{k\to\infty}{\longrightarrow} 1$ (inconsistent)

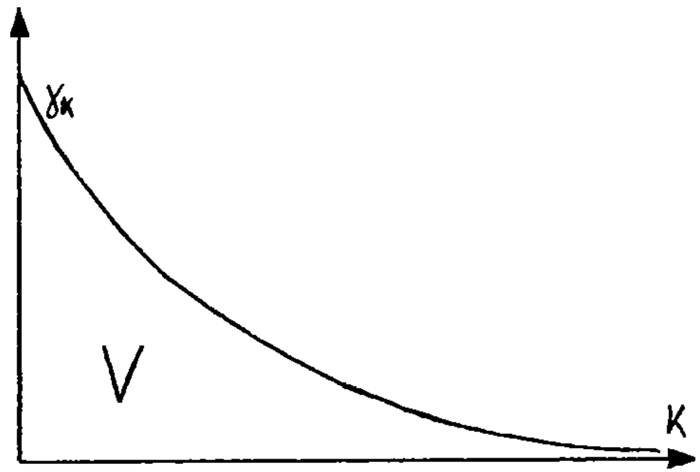
Sliding Discount: $W_{k\gamma} \propto \gamma_0 r_k + \gamma_1 r_{k+1} + ...$
(in psychology & economy)
Problem: also inconsistent for general $\boldsymbol{\gamma}$.
[Strotz 1955, Vieille&Weibull 2004]



Consistent General (Non-Geometric) Discount

Definition 3 (Discounted value)
$V_{k\gamma} := \frac{1}{\Gamma_k}\sum_{i=k}^{\infty}\gamma_i r_i$ with normalizer $\Gamma_k := \sum_{i=k}^{\infty}\gamma_i < \infty$

- is well-defined for arbitrary environments,
- leads to consistent policies,
- leads to an increasing effective horizon (proportionally to k) for e.g. quadratic discount $\gamma_k = 1/k^2$,
- 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.

[Hutter 2002 and 2005]

Asymptotics

If the exact environment is not known in advance it has to be learned by reinforcement [Sutton&Barto 1998] or adaptation [Kumar&Varaiya 1986].

In this case
the asymptotic total average performance $U_{1\infty} := \lim_{m\to\infty} U_{1m}$ and
the asymptotic future discounted performance $V_{\infty\gamma} := \lim_{k\to\infty} V_{k\gamma}$
are more relevant than finite values.

Subject of Study in this Talk

Relation between $U_{1\infty}$ and $V_{\infty\gamma}$
for general discount $\boldsymbol{\gamma}$ and arbitrary environment $\boldsymbol{r}$.

Effective and Quasi-Horizon

- Rewards $r_{k+h}, r_{k+h+1,...}$ give only a small contribution to $V_{k\gamma}$ for large h, since $\Gamma_{k+h} \equiv \gamma_{k+h} + \gamma_{k+h+1} + ... \to 0$ for $h \to \infty$

$\Rightarrow$ $V_{k\gamma}$ has effective horizon h^{eff} for which the cumulative tail weight $\Gamma_{k+h^{eff}}/\Gamma_k \approx \frac{1}{2}$

- Quasi-horizon $h_k^{quasi} := \Gamma_k/\gamma_k \approx h_k^{eff}$

- Super|sub|linear quasi-horizon: $h_k^{quasi}/k \to \infty|0|$finite

Example Discount Sequences & Quasi-Horizons

Discounts	γ_k	Γ_k	h_k^{quasi}	is	growth	h^{quasi}/k	
finite	$1_{k\leq m}$	$m-k+1$	$m-k+1$	is	decreasing	$\frac{m-k+1}{k}$	
geometric	γ^k	$\frac{\gamma^k}{1-\gamma}$	$\frac{1}{1-\gamma}$	is	constant= sublinear	$\frac{1}{(1-\gamma)k}$	$\to 0$
quadratic	$\frac{1}{k(k+1)}$	$\frac{1}{k}$	$k+1$	is	linear	$\frac{k+1}{k}$	$\to 1$
power	$k^{-1-\varepsilon}$	$\frac{1}{\varepsilon}k^{-\varepsilon}$	$\frac{k}{\varepsilon}$	is	linear	$\frac{1}{\varepsilon}$	$\to \frac{1}{\varepsilon}$
harmonic	$\frac{1}{k\ln^2 k}$	$\frac{1}{\ln k}$	$k\ln k$	is	superlinear	$\ln k$	$\to \infty$

Example Reward Sequences

- Limit $U_{1\infty}$ may exist or not, independent of whether $V_{\infty\gamma}$ exists.
- Examples for all four possibilities in the table below, with
- asymptotic value for the considered discount and reward sequences
- $\sim$ means oscillation.

Value$_\infty$	$\gamma \diagdown r$	1^∞	$101010...$	$1^1 0^2 1^3 0^4 ...$	$1^1 0^2 1^4 0^8 ...$
finite	$1_{k\leq m}$	1	${}^1/_2$	${}^1/_2$	$\frac{1}{3} \sim \frac{2}{3}$
geometric	γ^k	1	$\frac{\gamma}{1+\gamma} \sim \frac{1}{1+\gamma}$	$0 \sim 1$	$0 \sim 1$
quadratic	$\frac{1}{k(k+1)}$	1	${}^1/_2$	${}^1/_2$	$\frac{1}{3} \sim \frac{2}{3}$
power	$k^{-1-\varepsilon}$	1	${}^1/_2$	${}^1/_2$	$\frac{1}{1+2^\varepsilon} \sim \frac{1}{1+2^{-\varepsilon}}$
harmonic	$\frac{1}{k \ln^2 k}$	1	${}^1/_2$	${}^1/_2$	${}^1/_2$
oscillating	h^{quasi}	1	${}^1/_2$ or $\sim$	${}^1/_2$ or $\sim$	$\sim$

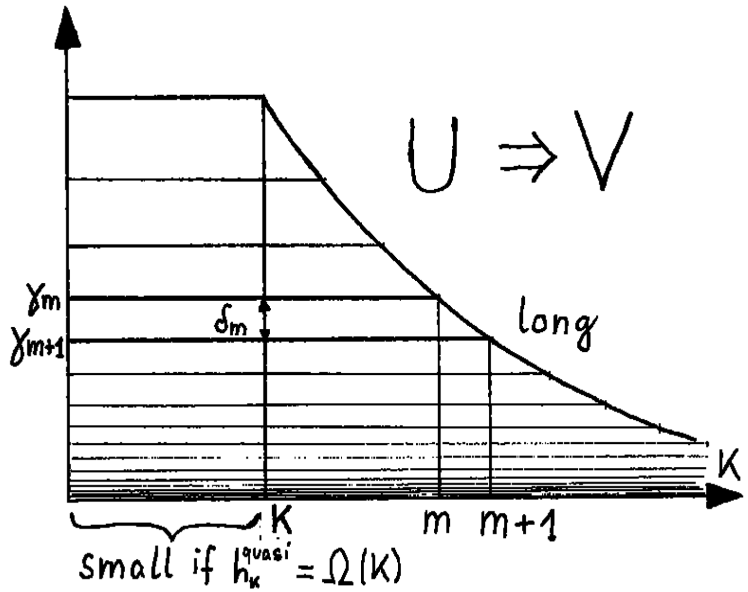
Average Implies Discounted Value

... if the quasi-horizon grows linearly with k or faster.

Theorem 4 ($\boldsymbol{U_{1\infty} \Rightarrow V_{\infty\gamma}}$) Assume $h_k^{quasi}=\Omega(k)=$(super)linear:

If $U_{1m} \to \alpha$ then $V_{k\gamma} \to \alpha$ $(\forall\boldsymbol{\gamma})$.

For instance, quadratic, power and harmonic discounts satisfy the condition, but faster-than-power discount like geometric do not.



Proof "horizontally" slices $V_{k\gamma}$ (as a function of k) into a weighted sum of average rewards U_{1m}.

The condition is actually necessary in the sense that

Proposition 5 ($\boldsymbol{U_{1\infty} \not\Rightarrow V_{\infty\gamma}}$) $\forall\boldsymbol{\gamma}$ with $h_k^{quasi}\neq\Omega(k)$

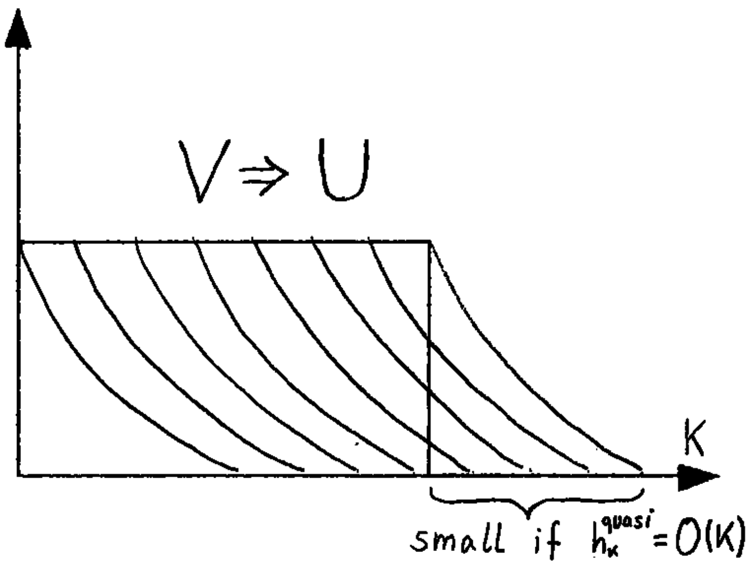
$\exists\boldsymbol{r}$ for which $U_{1\infty}$ exists, but not $V_{\infty\gamma}$.

Discounted Implies Average Value

... if the effective horizon grows linearly with k or slower.

Theorem 6 ($V_{\infty\gamma} \Rightarrow U_{1\infty}$) Assume $h_k^{quasi} = O(k) =$ (sub)linear:
If $V_{k\gamma} \to \alpha$ then $U_{1m} \to \alpha$ $(\forall\gamma)$.

For instance, power or faster and geometric discounts satisfy the condition, but harmonic does not.



Proof slices U_{1m} in "curves" to a weighted mixture of discounted values $V_{k\gamma}$.

The condition is necessary in the sense that

Proposition 7 ($V_{\infty\gamma} \not\Rightarrow U_{1\infty}$) $\forall\gamma$ with $h_k^{quasi} \neq O(k)$
$\exists \boldsymbol{r}$ for which $V_{\infty\gamma}$ exists, but not $U_{1\infty}$.

Average Equals Discounted Value

Theorem 4 and 6 nearly imply

Theorem 1 ($U_{1\infty} = V_{\infty\gamma}$)
Assume $U_{1\infty}$ and $V_{\infty\gamma}$ exist. Then $U_{1\infty} = V_{\infty\gamma}$.

Missing case to prove: Oscillating quasi-horizon $h_k^{quasi}/k \in [0,\infty]$:
$\underline{\lim}\, h_k^{quasi}/k = 0 < \infty = \overline{\lim}\, h_k^{quasi}/k$

Reminder: Theorem 1 holds for arbitrary monotone discount sequences (interesting since geometric discount leads to agents with bounded horizon) and arbitrary bounded reward sequences (important since reality is neither ergodic nor MDP).

Appeal and Key Role of Power Discounting

- separates the cases where existence of $U_{1\infty}$ implies/is-implied-by existence of $V_{\infty\gamma}$ ($U_{1\infty}$ exists iff $V_{\infty\gamma}$ exists),

- has linearly increasing effective/quasi horizon,

- neither requires nor introduces any artificial global time-scale,

- results in an increasingly farsighted agent with horizon proportional to its own age (realistic model for humans?)

- In particular I advocate using quadratic discounting $\gamma_k = 1/k^2$.

Outlook

- All proofs in the paper provide convergence rates.

- Generalization to probabilistic environments possible.

- Monotonicity of γ and boundedness of rewards can possibly be somewhat relaxed.

- Is there an easier direct way of proving Theorem 1 w/o separation of the two (discount) cases?

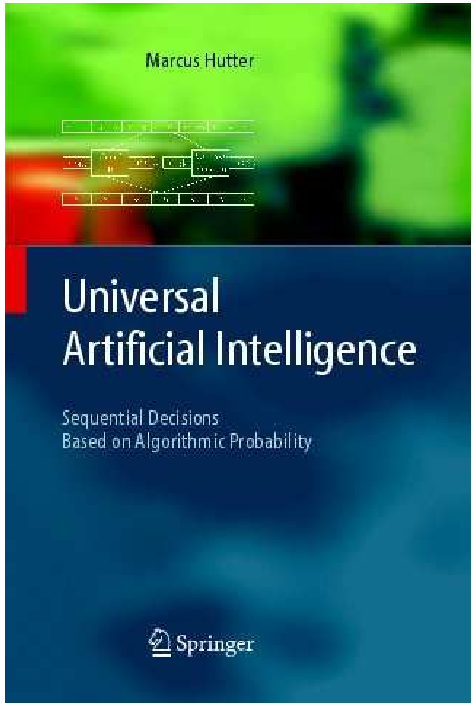
- A formal relation between effective horizon and the introduced quasi-horizon may be interesting.

Thanks! Questions? Details:

- M. Hutter, *General Discounting versus Average Reward.* Proc. 17th International Conf. on Algorithmic Learning Theory (ALT 2006) http://arxiv.org/abs/cs.LG/0605040

- M. Hutter, *Self-optimizing and Pareto-Optimal Policies in General Environments.* In Proc. 15th International Conf. on Computational Learning Theory (COLT 2002) 364–379, Springer. http://arxiv.org/abs/cs.AI/0204040

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



Decision Theory = Probability + Utility Theory

\+ +

Universal Induction = Ockham + Bayes + Turing

= =

A Unified View of Artificial Intelligence