HAL
open science

Asymptotic Properties of Optimal Trajectories in Dynamic Programming

Sylvain Sorin, Xavier Venel, Guillaume Vigeral

To cite this version:

Sylvain Sorin, Xavier Venel, Guillaume Vigeral. Asymptotic Properties of Optimal Trajectories in Dynamic Programming. Sankhya: The Indian Journal of Statistics, 2010, 72-A, pp.237-245. hal-00549788

HAL Id: hal-00549788

https://hal.science/hal-00549788v1

Submitted on 22 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ASYMPTOTIC PROPERTIES OF OPTIMAL TRAJECTORIES IN DYNAMIC PROGRAMMING

SYLVAIN SORIN, XAVIER VENEL, GUILLAUME VIGERAL

Abstract. We show in a dynamic programming framework that uniform convergence of the finite horizon values implies that asymptotically the average accumulated payoff is constant on optimal trajectories. We analyze and discuss several possible extensions to two-person games.

1. Presentation

Consider a dynamic programming problem as described in Lehrer and Sorin [1]. Given a set of states S, a correspondence Φ from S to itself with non empty values and a payoff function f from S to $[0,1]$, a feasible play at $s \in S$ is a sequence $\{s_m\}$ of states with $s_1 = s$ and $s_{m+1} \in \Phi(s_m)$. It induces a sequence of payoffs $\{f_m = f(s_m)\}, m = 1, ..., n, ...$ Recall that starting from a standard problem with random transitions and/or signals on the state, this presentation amounts to work on the set of probabilities on S and to consider expected payoffs.

Let $v_n(s)$ (resp. $v_\lambda(s)$) be the value of the n stage program $G_n(s)$ (resp. λ discounted program $G_\lambda(s)$) starting from state s. The **asymptotic approach** deals with asymptotic properties of the values v_n and v_λ as n goes to ∞ or λ goes to 0.

The **uniform approach** focuses on properties of the strategies that hold uniformly in long horizons. v_∞ is the uniform value if for each $\varepsilon > 0$ there exists N such that for each $s \in S$:
1) there is a feasible play $\{s_m\}$ at s with

$$\frac{1}{n}\sum_{m=1}^{n} f(s_m) \geq v_\infty(s) - \varepsilon, \qquad \forall n \geq N$$

2) for any feasible play $\{s'_m\}$ at s and any $n \geq N$

$$\frac{1}{n}\sum_{m=1}^{n} f(s'_m) \leq v_\infty(s) + \varepsilon.$$

Obviously the second approach is more powerful than the second (existence of a uniform value implies existence of an asymptotic value : the limit of v_n exists) but it is also more demanding: there are problems without uniform value where the asymptotic value exists (see Section 2).
Note that the condition for the existence of a uniform value implies that the average accumulated payoff on optimal trajectories remains close to the value.

We will prove that a similar phenomenon holds true under conditions that are stronger than the existence of an asymptotic value but weaker than the existence of a uniform value.

Say that the dynamic programming problem is **regular** if :
i) $\lim v_n(s) = v(s)$ exists for each $s \in S$.
ii) the convergence is uniform.
This condition was already introduced and studied in Lehrer and Sorin [1] (see Section 2).

We consider the following property **P**:

Date: October 2009.

For any $\varepsilon > 0$, there exists n_0, such that for all $n \geq n_0$, for any state s and any feasible play $\{s_m\}$ ε-optimal for $G_n(s)$ and for any $t \in [0,1]$:

$$3\varepsilon \geq \frac{1}{n}(\sum_{m=1}^{[tn]} f_m) - tv(s) \geq -3\varepsilon. \tag{1}$$

where $[tn]$ stands for the integer part of tn.

This condition says that the average payoff remains close to the value on every almost-optimal trajectory with long duration (but the trajectory may depend on this duration).
It also implies a similar property on every time interval.

2. Examples and comments

1) The existence of the asymptotic value is not enough to control the payoff as required in property **P**.
An example is given in Lehrer and Sorin [1] (Section 2), where both $\lim v_n$ and $\lim v_\lambda$ exist on S but where the asymptotic average payoff is not constant on the unique optimal trajectory, nor on ε-optimal trajectories: in G_{2n}, an optimal play will induce n times 0 then n times 1 while $v = 1/2$.
Note that this example is not regular: the convergence of v_n to v is not uniform.

2) Recall that in the framework of dynamic programming, regularity is also equivalent to uniform convergence of v_λ (and with the same limit), see Lehrer and Sorin [1] (Section 3).
Note also that this regularity condition is not sufficient to obtain the existence of a uniform value, see Monderer and Sorin [2] (Section 2).

3) General conditions for regularity can be found in Renault [5].

3. Main result

Theorem 3.1. *Assume that the program is regular, then* **P** *holds.*

Proof

Let us start with the upper bound inequality in (1).
The result is clear for $t \leq \varepsilon$ (recall that that the payoff is in $[0,1]$). Otherwise let n_1 large enough so that $n \geq n_1$ implies $||v_n - v|| \leq \varepsilon$ by uniform convergence. Then the required inequality holds for $n \geq n_2$ with $[\varepsilon n_2] \geq n_1$.

Consider now the lower bound inequality in (1).
The result holds for $t \geq 1 - \varepsilon$ by the ε-optimal property of the play, for $n \geq n_1$. Otherwise we use the following lemma from Lehrer and Sorin [1] (Proposition 1).

Lemma 3.1. *Both* $\limsup v_n$ *and* $\limsup v_\lambda$ *decresase on feasible histories.*

In particular, starting from $s_{[tn]}$ the value of the program for the last $n - [tn]$ stages is at most $v(s_{[tn]}) + \varepsilon$ for $n \geq n_2$, by uniform convergence, hence less than the initial $v(s) + \varepsilon$, using the previous Lemma. Since the play is ε-optimal in $G_n(s)$, this implies that

$$\sum_{m=1}^{[tn]} f_m + (n - [tn])(v(s) + \varepsilon) \geq n(v_n(s) - \varepsilon) \geq n(v(s) - 2\varepsilon) \tag{2}$$

hence the required inequality. ∎

4. Extensions

4.1. Discounted case.

A similar result holds for the program G_λ corresponding to the evaluation $\sum_{m=1}^{\infty} \lambda(1-\lambda)^{m-1} f_m$.
Explicitly, one introduces the property $\mathbf{P}'$:

For any $\varepsilon > 0$, there exists λ_0, such that for all $\lambda \leq \lambda_0$, for any state s and any feasible play $\{s_m\}$ ε-optimal for $G_\lambda(s)$ and for any $t \in [0,1]$:

$$3\varepsilon \geq \sum_{m=1}^{n(t;\lambda)} \lambda(1-\lambda)^{m-1} f_m) - tv(s) \geq -3\varepsilon. \tag{3}$$

where $n(t;\lambda) = \inf\{p \in I\!N; \sum_{m=1}^{p} \lambda(1-\lambda)^{m-1} \geq t\}$. Stage $n(t;\lambda)$ corresponds to the fraction t of the total duration of the program.

Theorem 4.1. *Assume that the program is regular, then* **P'** *holds.*

Proof

The proof follows the same lines than the proof of Theorem 3.1.
Recall that by regularity both v_n and v_λ converge uniformly to v. Moreover the discounted sums $(1-\lambda)^{-N} \sum_{m=1}^{N} \lambda(1-\lambda)^{m-1} f_m$ belong to the convex hull of the averages $\frac{1}{n}\sum_{m=1}^{n} f_m; 1 \leq n \leq N$. The counterpart of equation (2) is now

$$\sum_{m=1}^{n(t;\lambda)} \lambda(1-\lambda)^{m-1} f_m + (1-t)(v(s)+\varepsilon) \geq (v_\lambda(s) - \varepsilon) \geq v(s) - 2\varepsilon \tag{4}$$

■

4.2. **Continuous time.**
Similar results holds in the following set-up: $v_T(x)$ is the value of the control problem Γ_T with control set U where the state variable in X is governed by a differential equation (or more generally a differential inclusion)

$$\dot{x}_t = f(x_t, u_t)$$

starting from x at time 0. The real payoff function is $g(x,u)$ and the evaluation is given by:

$$\frac{1}{T}\int_0^T g(x_t, u_t)dt.$$

Regularity in this framework amounts to uniform convergence (on X) of V_T to some V. (Sufficient conditions for regularity can be found in Quincampoix and Renault [4]). The corresponding property is now **P"**:
For any $\varepsilon > 0$, there exists T_0, such that for all $T \geq T_0$, for any state x and any feasible trajectory ε-optimal for $\Gamma_T(x)$ and for any $\theta \in [0,1]$:

$$3\varepsilon \geq \frac{1}{T}\int_0^{\theta T} g(x_t, u_t)dt - \theta V(x) \geq -3\varepsilon. \tag{5}$$

Theorem 4.2. *Assume that the optimal control problem is regular, then* **P"** *holds.*

Proof

Follow exactly the same lines than the proof of Theorem (2). ■

Finally the same tools can be used for an evaluation of the form $\lambda \int_0^{+\infty} e^{-\lambda t} g(x_t, u_t)dt$.

5. Two-player zero-sum games

In trying to extend this result to a two-person zero-sum framework, several problems occurs.

5.1. **Optimal strategies on both sides.**
First it is necessary, to obtain good properties on the trajectory, to ask for optimality on both sides.
For example in the Big Match with no signals,

	α	β
a	1*	0*
b	0	1

where a $*$ denotes an absorbing payoff, the optimal strategy of player 1 in the "asymptotic game" on $[0,1]$ is to play "a before time t" with probability t, see Sorin [6] Section 5.3.2. Obviously, if there is no restrictions on player 2's moves the average payoff will not be constant. However, the optimal strategy of player 2 is "always $(1/2, 1/2)$" hence time independent on $[0,1]$. It thus induces a constant payoff and it is easy to see that the property is robust to small perturbations in the evaluation of the payoff.

5.2. **Player 1 controls the transition.**
Consider a repeated game with finite characteristics (states, moves, signals, ...) and use the recursive formula corresponding to the canonical representation with entrance laws being consistent probabilities on the universal belief space, see Mertens, Sorin and Zamir [3], Chapters III.1, IV.3. This representation preserves the values but in the auxiliary game, if player 1 controls the transition an optimal strategy of player 2 is to play a stage by stage best reply. Hence the model reduces to the dynamic programming framework and the results of the previous sections apply.
A simple example corresponds to a game with incomplete information on one side where asymptotically an optimal strategy of the uniform player 1 is a splitting at time 0, while player 2 can obain $u(p_t)$ at time t where u is the value of the non-revealing game and p_t the martingale of posteriors at time t, see Sorin [6], 3.7.2.

5.3. **Example.**
Back to the general framework of two person zero-sum repeated games, the following example shows that in addition one has to strengthen the conditions on the pair of ε-optimal strategies. We exhibit a game having a uniform value v but for some state s with $v(s) = 0$ one can construct, for each n, optimal strategies in $\Gamma_n(s)$ inducing essentially a constant payoff 1 during the first half of the game.

Starting from the initial state s, the tree representing the game Γ has countably many subgames $\tilde{\Gamma}_{2n}$, the transition being controlled by player 1 (with payoff 0). In $\tilde{\Gamma}_{2n}$ there are at most n stages before reaching an absorbing state. At each of these stages of the form $(2n, m), m = 1, ...n$, the players plays a "jointly controlled" process leading either to a payoff 1 and the next stage $(2n, m+1)$ (if they agree) or an absorbing payoff $x_{2n,m}$ with $(m-1) + (2n - (m-1))x_{2n,m} = 0$, otherwise. Hence every feasible path of length $2n$ in $\tilde{\Gamma}_{2n}$ gives a total payoff 0. Obviously the uniform value exists since each player can stop the game at each node, inducing the same absorbing payoff. The representation is as follows:

Notice that in the $2n+1$ stage game, after a move of player 1 to $\tilde{\Gamma}_{2n}$, any play is compatible with optimal strategies, in particular those leading to the sequence of payoffs $2n$ times 0 or n times 1 then n times -1.

5.4. **Conjectures.**
A natural conjecture is that in any regular game (i.e. where v_n converges uniformly to v):
for any $\varepsilon > 0$, there exists n_0, such that for all $n \geq n_0$, for any initial state s, there exists a couple (σ_n, τ_n) of ε-optimal strategies in $G_n(s)$ such that for any $t \in [0,1]$:

$$3\varepsilon \geq \frac{1}{n}\mathbf{E}^s_{\sigma_n,\tau_n}(\sum_{m=1}^{[tn]} f_m) - tv(s) \geq -3\varepsilon. \tag{6}$$

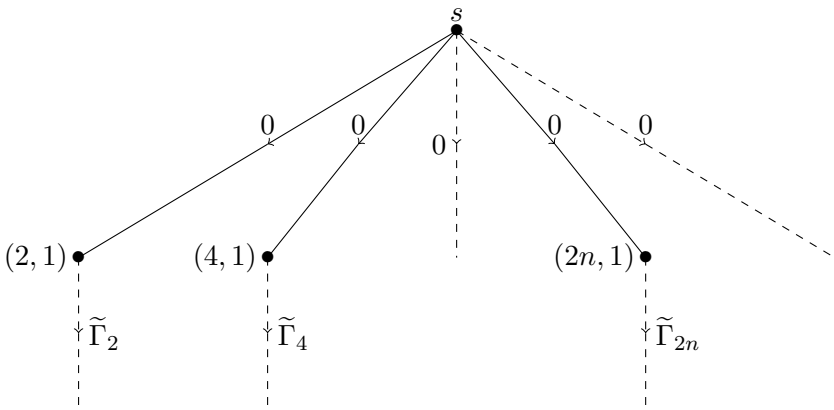

FIGURE 1. The game Γ starting from state s

	C	A
C	$\underset{\rightarrow}{1}$	0*
A	0*	0*

$(2n,1)$

	C	A
C	$\underset{\rightarrow}{1}$	$x_{2n,2}$*
A	$x_{2n,2}$*	$x_{2n,2}$*

$(2n,2)$ $\cdots$

	C	A
C	$\underset{\rightarrow}{1}$	$x_{2n,m}$*
A	$x_{2n,m}$*	$x_{2n,m}$*

$(2n,m)$ $\cdots$

	C	A
C	$\underset{\rightarrow}{1}$	$x_{2n,n}$*
A	$x_{2n,n}$*	$x_{2n,n}$*

$(2n,n)$

	C	A
C	-1*	-1*
A	-1*	-1*

-1^*

FIGURE 2. The subgame $\widetilde{\Gamma}_{2n}$ starting from state $(2n,1)$

where $[tn]$ stands for the integer part of tn and f_m is the payoff at stage m.
A more elaborate conjecture would rely on the existence of an asymptotic game Γ^* played in continuous time on $[0,1]$ with value v (as in Section 5.1). We use the representation of the repeated game as a stochastic game trough the recursive structure as above, see Mertens, Sorin, Zamir [3], Chapter IV. The condition is now the existence of a couple of strategies (σ,τ) in the asymptotic game that would depend only on the time $t \in [0,1]$ and on the current state s such that for any $\varepsilon > 0$, there exists η with the following property: in any repeated game where the (relative) weight of stage m is α_m, with $\{\alpha_m\}$ decreasing and less than η, thus defining a partition Π of $[0,1]$, the strategies (σ_Π, τ_Π) induced in the repeated game by (σ,τ) satisfies (6).

Acknowledgment:
This work was done while the three authors were members of the Equipe Combinatoire et Optimisation.
Sorin's research was supported by grant ANR-08-BLAN-0294-01 (France).

REFERENCES

[1] Lehrer E. and S. Sorin (1992) A Uniform tauberian theorem in dynamic programming, *Mathematics of Operations Research*, **17**, 303-307.
[2] Monderer D. and S. Sorin (1993) Asymptotic properties in dynamic programming, *International Journal of Game Theory*, **22**, 1-11.
[3] Mertens J.-F. , S. Sorin and S. Zamir (1994) Repeated Games, CORE Discussion Papers 9420, 9421, 9422.
[4] Quincampoix M. and J. Renault (2009) On the existence of a limit value in some non expansive optimal control problems, preprint.
[5] Renault J. (2007) Uniform value in dynamic programming, Cahier du CEREMADE, 2007-1.
[6] Sorin S. (2002) A first course on zero-sum repeated games, Mathmatiques et Applications, **37**, Springer.
[7] Sorin S. (2005) New approaches and recent advances in two-person zero-sum repeated games, *Advances in Dynamic Games*, A. Nowak and K. Szajowski (eds.), Birkhauser, 67-93.

Equipe Combinatoire et Optimisation, CNRS FRE 3232, Faculté de Mathématiques, UPMC-Paris 6, 175 Rue du Chevaleret, 75013 Paris, France
GREMAQ Universit de Toulouse 1 Manufacture des Tabacs, Aile J.J. Laffont 21 alle de Brienne 31000 Toulouse, France
INRIA Saclay - Ile-de-France and CMAP, Ecole Polytechnique, route de Saclay, 91128 Palaiseau cedex, France
E-mail address: sorin@math.jussieu.fr, xavier.venel@sip.univ-tlse1.fr, guillaumevigeral@gmail.com