Sankhyā : The Indian Journal of Statistics
2010, Volume 72-A, Part special, pp. 237-245
© 2010, Indian Statistical Institute

Asymptotic Properties of Optimal Trajectories in Dynamic Programming

Sylvain Sorin
UPMC-Paris 6, Paris, France
Xavier Venel
Université de Toulouse 1, Toulouse, France
Guillaume Vigeral
Ecole Polytechnique, Palaiseau, France

Abstract

We prove 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.

AMS (2000) *subject classification.* Primary 49L20, 90C39, 91A15.
Keywords and phrases. Dynamic programming, asymptotic value.

1 Presentation

Consider a dynamic programming problem as described in Lehrer and Sorin (1992). 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 $G_\lambda(s)$): maximize $\frac{1}{n}\sum_{m=1}^{n} f_m$ (resp. $\sum_{m=1}^{+\infty} \lambda(1-\lambda)^{m-1} f_m$) over the set of feasible plays at 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 $\varepsilon > 0$ and each $s \in S$, there exists N such that:

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 first (existence of a uniform value $v_\infty(s)$ implies existence of an asymptotic value :$v(s)$ which is the limit of $v_n(s)$ and $v_\infty(s) = v(s)$) but it is also more demanding: there are problems without uniform value where the asymptotic value exists (see Section 2).

Note that the existence of a uniform value says that the average accumulated payoff on optimal trajectories remains close to the value.

We will study a related phenomenon in the asymptotic framework and consider the following property **P**:

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

$$-3\varepsilon \leq \frac{1}{n}(\sum_{m=1}^{[tn]} f_m) - tw(s) \leq 3\varepsilon. \tag{1.1}$$

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.

Say that the dynamic programming problem is **regular** if :

1. $\lim v_n(s) = v(s)$ exists for each $s \in S$.

2. the convergence is uniform on S.

This property was already introduced and studied in Lehrer and Sorin (1992, Section 2). Note that **P** implies regularity.

Our main result is:

Theorem 1.1. *Assume that the program is regular, then* **P** *holds (with* $w = v$*).*

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 (1992, Section 2), where $\lim v_n(s) = v(s)$ exists on S but where the asymptotic average payoff is not constant on the unique optimal trajectory, nor on ε-optimal trajectories at some state s_0: in $G_{2n}(s_0)$, an optimal play will induce n times 0 then n times 1 while $v(s_0) = 1/2$.

Note that this example is not regular: the convergence of v_n to v is not uniform.

2) In the framework of dynamic programming, regularity is also equivalent to uniform convergence of v_λ (and with the same limit v), see Lehrer and Sorin (1992, Section 3).

Note also that this regularity condition is not sufficient to obtain the existence of a uniform value, see Monderer and Sorin (1993, Section 2).

3) General sufficient conditions for regularity can be found in Renault (2007).

3 Proof of the main result

Take $w = v$ and let us start with the upper bound inequality in (1.1.).

The result is clear for $t \leq \varepsilon$ (recall that that the payoff is in $[0, 1]$). Otherwise let n_1 be 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 (1.1). 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 (1992, 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

$$\sum_{m=1}^{[tn]} f_m + (n - [tn])(v(s) + \varepsilon) \geq n(v_n(s) - \varepsilon) \geq n(v(s) - 2\varepsilon) \qquad (3.1)$$

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 $\mathbf{P}'$:

There exists $w : S \to \mathbb{R}$ satisfying: 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 $t \in [0,1]$:

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

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

Theorem 4.1. *Assume that the program is regular, then* $\mathbf{P}'$ *holds (with* $w = v$*).*

Proof. The proof follows the same lines than the proof of Theorem 1.1.

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 $\frac{1}{n} \sum_{m=1}^{n} f_m; 1 \leq n \leq N$, see Lehrer and Sorin (1992). The counterpart of equation (3.1) is now

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

and the result follows. □

4.2 Continuous time. Similar results holds in the following set-up: Let $v_T(x)$ be the value of the control problem $\Gamma_T(x)$ 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, 2009). The corresponding property is now $\mathbf{P}''$:

There exists $W : X \to \mathbb{R}$ satisfying: 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 $\theta \in [0, 1]$:

$$-3\varepsilon \leq \frac{1}{T}\int_0^{\theta T} g(x_t, u_t)\ dt - \theta W(x) \leq 3\varepsilon.$$

THEOREM 4.2. *Assume that the optimal control problem is regular, then* $\mathbf{P}''$ *holds (with $W = V$).*

PROOF. Follow exactly the same lines than the proof of Theorem 1.1. □

Finally similar tools can be used for an evaluation of the form

$$\lambda \int_0^{+\infty} e^{-\lambda t} g(x_t, u_t)\ dt,$$

see Oliu-Barton and Vigeral (2009).

5 Two-player zero-sum games

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

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 consider the Big Match with no signals, which is a stochastic game described by the matrix

	α	β
a	1^*	0^*
b	0	1

where a $*$ denotes an absorbing payoff. Assume that the players receive no information during the play. Then the asymptotic properties of the repeated game can be analyzed trough an "asymptotic game" played on $[0, 1]$, see Sorin (2002, Section 5.3.2) and Sorin (2005, Section 4) and the optimal strategy of player 1 is to play "a before time t" with probability t. Obviously, without restriction on player 2's moves, the average payoff along the play will not be closed to the asymptotic value $v = \frac{1}{2}$. (For example if Player 2 plays α during the first half of the game the corresponding average payoff at time $t = \frac{1}{2}$ is $\frac{1}{4}$). However, the optimal strategy of player 2 is "always $(1/2, 1/2)$" hence time independent on $[0, 1]$and thus induces a constant payoff.

5.2 Player 1 controls the transition. Consider a repeated game with finite characteristics (states, moves, signals, ...) and use the recursive formula for the values corresponding to the canonical representation with entrance laws being consistent probabilities on the universal belief space, see Mertens, Sorin and Zamir (1994, 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 (2002, section 3.7.2).

Another class corresponds to the Markov games with incomplete information introduced by Renault (2006).

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 regular game where for some state s with $v(s) = 0$ one can construct, for each n, optimal strategies in $\Gamma_n(s)$ inducing roughly 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 $\tilde{\Gamma}_{2n}, n = 1, ...$, 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 play 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. At stage $(2n, n + 1)$ the payoff is -1 and absorbing. Hence every feasible path of length $2n$ in $\tilde{\Gamma}_{2n}$ gives a total payoff 0. Obviously the game is regular since each player can stop the game at each node $(2n, m)$, inducing the same absorbing payoff $x_{2n,m}$. The representation is as follows:

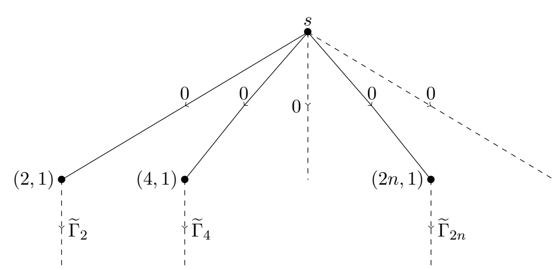

FIGURE 1. The game Γ starting from state s

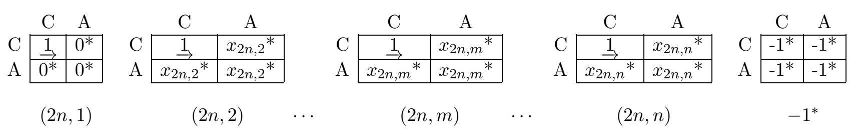

FIGURE 2. The subgame $\tilde{\Gamma}_{2n}$ starting from state $(2n, 1)$

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 $t \in [0,1]$:

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

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), see Sorin(2005, Section 4). We use the representation of the repeated game as a stochastic game trough the recursive structure as above, see Mertens, Sorin, Zamir (1993, 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 and which would generate a constant payoff along the play. Then the couple (σ_n, τ_n) would correspond to the strategies induced by (σ, τ) for a discretization of $[0,1]$ of width $\frac{1}{n}$.

Acknowledgement. This work was done while the three authors were members of the Equipe Combinatoire et Optimisation.

The research of the first author was supported by grant ANR-08-BLAN-0294-01 (France).

References

Lehrer E. and Sorin, S. (1992). A Uniform tauberian theorem in dynamic programming. *Math. Oper. Res.*, **17**, 303–307.

Monderer D. and Sorin, S. (1993). Asymptotic properties in dynamic programming. *Internat. J. Game Theory*, **22**, 1–11.

Mertens J.-F. , Sorin, S. and Zamir, S. (1994). Repeated games. CORE Discussion Papers 9420, 9421, 9422.

Oliu-Barton, M. and Vigeral, G. (2009). A uniform Tauberian theorem in optimal control. Preprint.

Quincampoix, M. and Renault, J. (2009). On the existence of a limit value in some non expansive optimal control problems. Preprint.

RENAULT, J. (2006). The value of Markov chain repeated games with lack of information on one side. *Math. Oper. Res.*, **31**, 490–512.

RENAULT, J. (2007). Uniform value in dynamic programming. Cahier du CEREMADE, 2007-1.

SORIN, S. (2002). *A first course on zero-sum repeated games.* Mathématiques et Applications, **37**, Springer.

SORIN, S. (2005). New approaches and recent advances in two-person zero-sum repeated games. In *Advances in Dynamic Games*, (A. Nowak and K. Szajowski, eds.). Birkhauser, 67–93.

SYLVAIN SORIN
EQUIPE COMBINATOIRE ET OPTIMISATION
CNRS FRE 3232,
FACULTÉ DE MATHÉMATIQUES,
UPMC-PARIS 6,
175 RUE DU CHEVALERET,
75013 PARIS, FRANCE
E-mail: sorin@math.jussieu.fr

XAVIER VENEL
GREMAQ
UNIVERSITÉ DE TOULOUSE 1
MANUFACTURE DES TABACS,
AILE J.J. LAFFONT
21 ALLÉE DE BRIENNE
31000 TOULOUSE, FRANCE
E-mail: xavier.venel@sip.univ-tlse1.fr

GUILLAUME VIGERAL
INRIA SACLAY - ILE-DE-FRANCE
AND CMAP, ECOLE POLYTECHNIQUE
ROUTE DE SACLAY
91128 PALAISEAU CEDEX, FRANCE
E-mail: guillaumevigeral@gmail.com

Paper received October 2009; revised January 2010.