Stochastic Games with Imperfect Monitoring

Dinah Rosenberg[*], Eilon Solan[†] and Nicolas Vieille[‡§]

March 10, 2002

Abstract

We study zero-sum stochastic games in which players do not observe the actions of the opponent. Rather, they observe a stochastic signal that may depend on the state, and on the pair of actions chosen by the players. We assume each player observes the state and his own action.

In a companion paper we proposed a candidate for the max-min value, we proved that player 2 can defend this value, and that player 1 can guarantee it in the class of absorbing games. In the present paper we prove that player 1 can guarantee this quantity in general stochastic games.

An analogous result holds for the min-max value.

[*] Laboratoire d'Analyse Géométrie et Applications, Institut Galilée, Université Paris Nord, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France. e-mail: dinah@zeus.math.univ-paris13.fr

[†] MEDS Department, Kellogg School of Management, Northwestern University, *and* the School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: eilons@post.tau.ac.il

[‡] Département Finance et Economie, HEC, 1, rue de la Libération, 78 Jouy-en-Josas, France. e-mail: vieille@hec.fr

[§] We acknowledge the financial support of the Arc-en-Ciel/Keshet program for 2001/2002. The research of the second author was supported by the Israel Science Foundation (grant No. 03620191).

1 Introduction

A celebrated result of Mertens and Neyman (1981) states that in every two-player zero sum stochastic game with finitely many states and actions, the uniform value exists, provided the players observe the stage payoff.

The requirement that players observe the stage payoff is crucial for their construction, since players determine how to play in each stage as a function of the stream of payoffs they have received so far. It is not difficult to construct examples where the value fails to exist if players do not observe the stage payoff (e.g. the "Big Match" with no signals).

Coulomb (1999, 2001) studied the class of absorbing games with imperfect monitoring. Those are stochastic games in which only one state is non absorbing, and players do not observe the action made by their opponent nor the stage payoff. Rather, they observe a signal that depends on the actions chosen by both players. Coulomb (1999, 2001) proved that in this class of games the uniform max-min value exists. Moreover, Coulomb provides an explicit formula for the uniform max-min value, which is independent of the signalling structure of player 2. Analogous results hold for the uniform min-max value.

In a companion paper (Rosenberg, Solan and Vieille, 2002a, henceforth RSVa) we studied general stochastic games with imperfect monitoring. We proposed a candidate for the uniform max-min value, and proved that player 2 can defend the proposed value. For the class of absorbing games, we also proved that player 1 can guarantee the proposed value. As in the analysis of Coulomb, our candidate is independent of the signalling structure of player 2.

In the present paper, we prove that player 1 can guarantee the proposed value in every stochastic game with imperfect monitoring, thereby completing the proof that our candidate is indeed the uniform max-min value of the game.

Along the paper we use the same notations and definition as in RSVa, and we use some of the results proven there. Though we provide all the necessary definitions, we urge the interested reader to read that paper first.

The paper is arranged as follows. The model and the main results are presented in Section 2. Few preliminary facts appear in Section 3. The proof of the main theorem appears in Sections 4 - 6.

2 The model and the main result

For every finite set K, $\Delta(K)$ is the space of probability distributions over K. We identify each element $k \in K$ with the probability distribution in $\Delta(K)$ that gives weight 1 to k.

2.1 The model

We consider the standard model of finite two-person zero-sum stochastic games with signals. Such a game is described by: (i) a finite set S of states, (ii) finite action sets A and B for the two players, (iii) a daily reward function $r : S \times A \times B \to \mathbf{R}$, (iv) finite sets of signals M^1 and M^2 of signals for the two players and (v) a transition function $\psi : S \times A \times B \to \Delta(M^1 \times M^2 \times S)$.

The game is played in stages. The initial state s_1 is known to both players. At each stage $n \in \mathbf{N}$, (a) the players choose independent of each other actions a_n and b_n; (b) player 1 gains $r(s_n, a_n, b_n)$, and player 2 looses the same amount; (c) a triple (m_n^1, m_n^2, s_{n+1}) is drawn according to $\psi(s_n, a_n, b_n)$; (d) players 1 and 2 are told respectively m^1_n and m^2_n, but they are *not* informed of a_n, b_n, or $r(s_n, a_n, b_n)$; and (e) the game proceeds to stage $n + 1$.

We assume throughout that each player always knows the current state, and the action he is playing.

We assume w.l.o.g. that payoffs are non-negative and bounded by 1. All norms are supremum norms.

We denote by $H_n = S \times (A \times B \times M^1 \times M^2 \times S)^{n-1}$ the set of histories up to stage n,[1] by $H^1_n = S \times (M^1)^{n-1}$ and $H^2_n = S \times (M^2)^{n-1}$ the set of histories to players 1 and 2 respectively. We equip these spaces with the discrete topology. We also let $H_\infty = (S \times A \times B \times M^1 \times M^2)^{\mathbf{N}}$ denote the set of infinite plays, $\mathcal{H}^i_n$ denote the cylinder algebra over H_∞ induced by H^i_n, and we set $\mathcal{H}_\infty = \sigma(\mathcal{H}^1_n \cup \mathcal{H}^2_n, n \geq 1)$, the σ-algebra generated by all those cylinder algebras. We let $\mathbf{s}_n, \mathbf{a}_n$ and $\mathbf{b}_n$ denote respectively the state at stage n, and the action played at stage n: these are random variables, and are respectively $\mathcal{H}^i_n$, $\mathcal{H}^1_{n+1}$ and $\mathcal{H}^2_{n+1}$ measurable.

Whenever convenient, we use the convention that boldfaced letters denote random variables, and non boldfaced letters denote the value of the random variable.

A (behavioral) strategy of player 1 (resp. player 2) is a sequence $\sigma = (\sigma_n)_{n\geq 1}$ (resp. $\tau = (\tau_n)_{n\geq 1}$) of functions $\sigma_n : H^1_n \to \Delta(A)$ (resp. $\tau_n : H^2_n \to \Delta(B)$). Such a strategy is *stationary* if the mixed move used at stage n depends only on $\mathbf{s}_n$ (which is known to both players). Every stationary strategy of player 1 (resp. player 2) can be identified with a vector $x \in (\Delta(A))^S$ (resp. $y \in (\Delta(B))^S$), with the interpretation that x^s is the lottery used by player 1 whenever the play visits state s.

Given a pair (σ, τ) of strategies and an initial state s, we denote by $\mathbf{P}_{s,\sigma,\tau}$ the probability distribution induced over $(H_\infty, \mathcal{H}_\infty)$ by (σ, τ) and s, and by $\mathbf{E}_{s,\sigma,\tau}$ the corresponding expectation operator. The expected average payoff up to stage n is

$$\gamma_n(s, \sigma, \tau) = \mathbf{E}_{s,\sigma,\tau}\left[\frac{1}{n}\sum_{k=1}^{n} r(\mathbf{s}_k, \mathbf{a}_k, \mathbf{b}_k)\right].$$

2.2 An equivalence relation

In RSVa we defined the following equivalence relation over mixed actions of player 2. The reader is referred to RSVa for various properties of this relation.

Given $\varepsilon, \lambda > 0$,[2] $s \in S$ and $x \in \Delta(A)$, we define a binary relation $\sim_{\lambda,\varepsilon,s,x}$ over $\Delta(B)$ as follows:

$$y \sim_{\lambda,\varepsilon,s,x} y' \text{ if and only if } \psi(s, a, y) = \psi(s, a, y') \text{ whenever } x[a] \geq \lambda/\varepsilon$$

and we set

$$\widetilde{r}^{\varepsilon}_{\lambda}(s, x, y) = \inf_{z \sim_{\lambda,\varepsilon,s,x} y} r(s, x, z). \qquad (1)$$

We define moreover $\widetilde{r}^{\varepsilon}_{\lambda}(s, x, y) = \inf_{z \sim_{\lambda,\varepsilon,s,x} y} r(s, x, z)$.

An important property of this relation is the following, proved in RSVa.

Lemma 1 *For every $\delta > 0$, there is $\eta > 0$ such that for every $s \in S$, every $x \in \Delta(A)$, and every $y, z \in \Delta(B)$, the following is satisfied: if $\|\psi(s, a, y) - \psi(s, a, z)\| < \eta$ for every $a \in A$ that satisfies $x[a] \geq \lambda/\varepsilon$, then $|\widetilde{r}^{\varepsilon}_{\lambda}(s, x, y) - \widetilde{r}^{\varepsilon}_{\lambda}(s, x, z)| < \delta$.*

[1] Since the signal of each player contains the current state and his action, some information in this representation is redundant.

[2] λ always stands for a discount factor. Here and in the sequel we omit the condition $\lambda \leq 1$.

We denote by $q : S \times A \times B \to \Delta(S)$ the transition function induced by ψ :

$$q(s' \mid s, a, b) = \psi(s, a, b)[\{s'\} \times M^1 \times M^2].$$

the multi-linear extension of q to $S \times \Delta(A) \times \Delta(B)$ is still denoted by q. For every $s \in S$, and every $(x, y) \in \Delta(A) \times \Delta(B)$, denote by $\mathbf{E}_{q(\cdot|s,x,y)}$ the expectation with respect to $q(\cdot \mid s, x, y)$.

For every $\lambda, \varepsilon > 0$ let $v^\varepsilon_\lambda : S \to [0, 1]$ be the unique function that solves the system

$$v^\varepsilon_\lambda(s) = \sup_{x\in\Delta(A)} \inf_{y\in\Delta(B)} \{\lambda \widetilde{r}^\varepsilon_\lambda(s, x, y) + (1 - \lambda)\mathbf{E}[v^\varepsilon_\lambda \mid s, x, y]\} \quad \forall s \in S. \tag{2}$$

The fact that the system (2) has a unique solution is proven in RSVa.

2.3 The main result

Definition 2 *$v(s)$ is the (uniform)* max-min value *of the game with initial state s if:*

- *Player 1 can* guarantee *$v(s)$: for every $\varepsilon > 0$, there exists a strategy σ of player 1 and $N \in \mathbf{N}$, such that:*
$$\forall \tau, \forall n \geq N, \ \gamma_n(s, \sigma, \tau) \geq v(s) - \varepsilon.$$
- *Player 2 can* defend *$v(s)$: for every $\varepsilon > 0$ and every strategy σ of player 1 there exists a strategy τ of player 2 and $N \in \mathbf{N}$, such that:*
$$\forall n \geq N, \gamma_n(s, \sigma, \tau) \leq v(s) + \varepsilon.$$

The definition of the (uniform) min-max value is obtained by exchanging the roles of the two players.

For every initial state s define

$$v(s) = \lim_{\varepsilon\to 0} \lim_{\lambda\to 0} v^\varepsilon_\lambda(s).$$

The existence of the limit is proven in RSVa.

Our main result is:

Theorem 3 *For every initial state s, $v(s)$ is the max-min value of the game.*

Exchanging the roles of the two players, one deduces that the min-max value exists as well.

In RSVa we proved that player 2 can defend $v(s)$, and that in the class of absorbing games, player 1 can guarantee it. Here we prove that player 1 can guarantee $v(s)$ in a general stochastic game:

Proposition 4 *For every initial state $s \in S$, player 1 can guarantee $v(s)$.*

2.4 A reduction: the signal coincides with the state

It is conceptually convenient to assume that the signal structure ψ^1 is such that player 1 gets no information apart from the current state.

This assumption entails no loss of generality. Indeed, let Γ be a two-person zero-sum stochastic game with imperfect monitoring. We let Γ_1 be the game deduced from Γ by changing the signal structure to player 2: in Γ_1, player 2 observes all past play, including signals to player 1. The strategy set of player 2 is larger in Γ_1, whereas the strategy set for player 1 in Γ_1 remains the same. Therefore, if player 1 can guarantee v in Γ_1, he can also guarantee v in Γ. Next, consider Γ_1. Change the definition of a state as follows:[3] the state is now a four-tuple consisting of the current state in the old sense, the previous state, the action played by player 1, and the signal received by player 1 (the state space is thus $S \times A \times M^1 \times S$), and let player 1 be informed only of the current state. Thus, we let Γ_2 be the game with state space $S \times A \times M^1 \times S$, action sets A and B, and whose transition and reward functions are deduced from those of Γ_1 in the natural way. In Γ_2, player 1 observes only the current state, while player 2 has full monitoring. Any strategy of player $i = 1, 2$ in Γ_1 can be mimicked by a strategy in Γ_2, and conversely. Thus, if player 1 can guarantee v in Γ_2 he can also guarantee v in Γ_1, hence in Γ.

This construction shows that we can assume that if $q(t \mid s, a, b), q(t \mid s, a', b') > 0$ then $a = a'$. Put otherwise, different actions of player 1 taken at a given state necessarily lead to different states. We will use this assumption in Lemma 12 below.

3 Preliminaries

We start by stating few simple results that are used in the sequel.

Fact 1: for $a, b > 0$ one has

$$\left|1 - \frac{a}{b}\right| \leq \varepsilon \Rightarrow \left|1 - \frac{b}{a}\right| \leq 2\varepsilon, \text{ provided } \varepsilon \leq 1.$$

Fact 2: For $\varepsilon \in (0, 1/3)$, one has

$$\left|\frac{a/b}{A/B} - 1\right| < 3\varepsilon \text{ whenever } \left|\frac{a}{b} - 1\right| < \varepsilon \text{ and } \left|\frac{A}{B} - 1\right| < \varepsilon.$$

Lemma 5 *Let $\zeta > 0$ and $K \in \mathbf{N}$. For $i = 1, ..., K$, let $x_i, y_i \geq 0$ and $n_i \in \mathbf{N}$. Set $n := \sum_{i=1}^{K} n_i$, and define $x := \frac{1}{n}\sum_{i=1}^{K} n_i x_i$, $y := \frac{1}{n}\sum_{i=1}^{K} n_i y_i$. Assume that for each i,*

$$n_i \max\{x_i, y_i\} \geq \zeta n \max\{x, y\} \Rightarrow |x_i - y_i| \leq \zeta x_i. \tag{3}$$

Then

$$|x - y| \leq K\zeta \max\{x, y\}.$$

If moreover $K\zeta \leq 1$ then $|x - y| \leq 2K\zeta x$.

Proof. By (3) $n_i |x_i - y_i| \leq \zeta n \max\{x, y\}$ for each i. The first assertion follows by summation over i. The second assertion follows by **Fact 1**. ■

[3] This is reminiscent of the combinatorial form of Mertens, see chapter IV in Mertens, Sorin and Zamir (1994)

4 The Proof

4.1 Overview

We fix $\bar{\varepsilon} > 0$ once and for all. Our goal is to construct a strategy σ that guarantees v up to $\bar{\varepsilon}$. We will use the strategy suggested by Mertens and Neyman (1981, Section 3, see also RSVa, Section 4, Case 1). This strategy plays in blocks. The size of each block depends on the play prior to that block. During block k, player 1 plays a stationary strategy x_{λ_k} that maximizes the right hand side of (2) up to $\varepsilon\lambda_k$, where ε is sufficiently small. Mertens and Neyman (1981) use λ_{k-1} and the average payoff in the block $k-1$ to determine λ_k. Here the average payoff is not observed, hence player 1 needs to estimate it.

4.2 Definition of the strategy

In this section we define a strategy for player 1. In the rest of the paper we prove that this strategy guarantees v. The strategy depends on two parameters: $\alpha \in (0,1)$ and $Z \in \mathbf{R}$.

There is a semi-algebraic function $\lambda \mapsto \varepsilon(\lambda)$ such that $\lim_{\lambda\to 0} \varepsilon(\lambda) = 0$, $\lim_{\lambda \to 0} v_\lambda^{\varepsilon(\lambda)}(s) = v(s)$ for every $s \in S$, and that $1-d \in (0, \frac{1}{2}]$ is the degree of $\lambda \mapsto \varepsilon(\lambda)$. For notational simplicity, we write v_λ and $\tilde{r}_\lambda$ instead of the more cumbersome $v_\lambda^{\varepsilon(\lambda)}$ and $\tilde{r}_\lambda^{\varepsilon(\lambda)}$.

Fix a constant $\eta_0 \leq \frac{2}{100}$ and $\varepsilon > 0$ sufficiently small so that $\varepsilon < \bar{\varepsilon}/(4|S|^2|B|)$ and $\varepsilon(3+2/\eta_0) \leq \bar{\varepsilon}$. For $\lambda > 0$ and $s \in S$, we let $x_\lambda^s \in \Delta(A)$ be a mixed action that satisfies:

$$\lambda \tilde{r}_\lambda(s, x_\lambda^s, y) + (1-\lambda)\mathbf{E}\left[v_\lambda \mid s, x_\lambda^s, y\right] \geq v_\lambda(s) - \varepsilon\lambda, \quad \forall y \in \Delta(B). \tag{4}$$

Fix $1 < \alpha' < 1/\alpha$. Define two functions $\lambda : (0,+\infty) \to (0,1)$ and $L : (0,+\infty) \to \mathbf{N}$ by:

$$\begin{aligned} \lambda(z) &= z^{-\alpha'}, \text{ and} \\ L(z) &= \lceil \lambda^{-\alpha}(z) \rceil. \end{aligned}$$

Observe that $\lim_{z\to\infty} \lambda(z) = 0$, $\lim_{z\to\infty} L(z) = +\infty$ and $\lim_{z\to\infty} \lambda(z)L(z) = 0$.

Let $(\widehat{r}_k)_{k\in\mathbf{N}}$ be a $[0,1]$-valued process defined on the set of plays. We explicitly define the process $(\widehat{r}_k)_{k\in\mathbf{N}}$ in the next section. Define recursively processes (z_k), (L_k) and (B_k) by the formulas

$$\begin{aligned} z_0 &= Z, B_0 = 1, \\ \lambda_k &= \lambda(z_k), L_k = L(z_k), B_{k+1} = B_k + L_k, \\ z_{k+1} &= \max\left\{ Z, z_k + \lambda_k \left(L_k \widehat{r}_k - \sum_{B_k \leq n < B_{k+1}} w_{\lambda_k}(s_n) \right) + \frac{\varepsilon}{2} \right\}. \end{aligned}$$

So that this definition makes sense, $\widehat{r}_k$ should depend only on the play before stage B_{k+1}. In our construction it depends only on the sequence of signals observed in block k; that is, the signals between stages B_k and $B_{k+1}-1$.

Let $\sigma(\alpha, Z)$ be the strategy that plays in each stage n the mixed action $x_{\lambda_k}(\mathbf{s}_n)$, where $k \in \mathbf{N}$ satisfies $B_k \leq n < B_{k+1}$.

4.3 Structure results

For $C \subset S$, we denote by $e_C := \min\{n > 0 \mid \mathbf{s}_n \notin C\}$ the first exit time from C. By convention, the minimum over an empty set is $+\infty$.

Theorem 6 *There exists: (i) a partition $\mathcal{D}$ of S, (ii) two real numbers $\alpha_1 \in (d, 1)$ and $\alpha_2 \in (0, 1 - \alpha_1)$, (iii) a non empty subset $C^* \subseteq S$ which is a union of some elements of $\mathcal{D}$, and (iv) a constant $M > 0$, such that*

- *For every atom Ω of $\mathcal{D}$, and every $s \in \Omega$, there is a pure stationary strategy $y(s)$ and $\lambda_0 \in (0, 1)$, such that, for every $s' \in \Omega$ and every $\lambda < \lambda_0$,*

$$\mathbf{E}_{s',x_\lambda,y(s)}[e_{\Omega\setminus\{s\}}] \leq 1/\lambda^{\alpha_1} \text{ and } \mathbf{P}_{s',x_\lambda,y(s)}\left(e_\Omega = e_{\Omega\setminus\{s\}}\right) \leq \lambda^{\alpha_2}.$$

- *For every $\alpha \in (d, 1)$, there exists $\lambda_1 \in (0, 1)$ such that for every τ, every initial state s' and every $\lambda < \lambda_1$, (a) if $s' \notin C^*$,*

$$\mathbf{E}_{s',x_\lambda,\tau}\left[\#\left\{n < \min\left\{1/\lambda^\alpha, \min\{j > 1 \mid s_j \in C^*\}\right\} \mid \mathbf{s}_n, \mathbf{s}_{n+1} \text{ belong to different atoms of } \mathcal{D}\right\}\right] \leq M.$$

(b) If $s' \in C^*$ and $s' \in D \in \mathcal{D}$,

$$\mathbf{E}_{s',x_\lambda,\tau}[\#\left\{n < 1/\lambda^\alpha \mid \mathbf{s}_n \in D, \mathbf{s}_{n+1} \notin D\right\}] \leq M\lambda^{1-\alpha}.$$

Observe that α_1 and α_2 may be arbitrarily close to 1 and 0 respectively.

The first condition says that if player 2 plays appropriately, the play can move in each element of $\mathcal{D}$ rather fast, while keeping the probability to leave that element small. The second condition says that whatever player 2 plays, (a) the number of visits to different elements of $\mathcal{D}$ until reaching the set C^* is uniformly bounded, and (b) the expected number of exits from elements of $\mathcal{D}$ that are subsets of C^* is extremely small. (a) and (b) together imply that the expected number of visits to different elements of $\mathcal{D}$ in $1/\lambda^\alpha$ stages is uniformly bounded.

Since the proof of this Theorem is quite involved, and it requires different tools that what we use elsewhere, it is postponed to the appendix.

Atoms of the partition $\mathcal{D}$ are called *communicating sets*. This structure result is best understood when particularized to two polar cases. Consider first the case where player 2 does not exist (equivalently, he has only one action per state). Then, under x_λ, the sequence of states follows a Markov chain where the transition function depends on λ. This Markov chain admits the usual partition into transient states and recurrent sets. Moreover, since $\lambda \to x_\lambda$ is semi-algebraic, this partition is independent of λ, for λ close enough to zero. The partition $\mathcal{D}$ refines it, by somehow requiring that the mean recurrence time within an atom is not too small. This setup has been extensively study under the name of Markov chains with rare transitions (see, e.g., Catoni (1999)).

Consider now the opposite polar case, where $q(\cdot \mid s, x_\lambda, b)$ is independent of λ. The above structure result then relates to Markov Decision Processes. It is a natural generalization of the partition into recurrent states and transient states to the case where one agent can affect transitions. For more in that case, we refer to Rosenberg, Solan and Vieille (2002c).

Lemma 7 *For each atom Ω of $\mathcal{D}$ and every $s, t \in \Omega$, $v_\lambda(s) \geq v_\lambda(t) - \lambda^{\alpha_2} - 2\lambda^{1-\alpha_1}$, for λ close enough to zero.*

Proof. Let $s,t\in\Omega$ and $\lambda>0$ sufficiently small be given. Starting from s, we define a strategy for player 2 by: play $y(t)$ until $e_{\Omega\setminus t}$ (that is, either the play leaves Ω or reaches t); afterwards play a λ-discounted optimal strategy.[4] Since payoffs are non-negative and below one, this strategy guarantees a payoff of at most $\mathbf{E}_{s,x_\lambda,\tau}[(1-(1-\lambda)^{e_{\Omega\setminus t}})]+v_\lambda(t)+\mathbf{P}_{s,x_\lambda,\tau}(e_{\Omega\setminus t}=e_\Omega)$. By Jensen's inequality, and by Theorem 6,

$$\mathbf{E}_{s,x_\lambda,\tau}[(1-\lambda)^{e_{\Omega\setminus t}}]\geq(1-\lambda)^{\mathbf{E}_{s,x_\lambda,\tau}[e_{\Omega\setminus t}]}\geq(1-\lambda)^{1/\lambda^{\alpha_1}}\geq 1-2\lambda^{1-\alpha_1}$$

for λ close enough to zero. Since $\mathbf{P}_{s,x_\lambda,\tau}(e_{\Omega\setminus t}=e_\Omega)\leq\lambda^{\alpha_2}$ the result follows. ■

Define a process $(\widetilde{I}_n)$ by

$$\widetilde{I}_n=1 \text{ if and only if } \mathbf{s}_n \text{ and } \mathbf{s}_{n+1} \text{ belong to different elements of } \mathcal{D},$$

where $\mathcal{D}$ is the partition over S given in Theorem 6.

Define a process (I_k) by

$$I_k=\sum_{B_k\leq n<B_{k+1}}\widetilde{I}_n.$$

That is, the number of visits to different elements in the partition of $\mathcal{D}$ in block k.

Lemma 8 *For every initial state s and every τ,*

$$\mathbf{E}_{s,\sigma(\alpha,Z),\tau}[\sum_{j\leq k}I_k]\leq M+M\times\mathbf{E}_{s,\sigma(\alpha,Z),\tau}[\sum_{j\leq k}\lambda_jL_j],$$

where M is a fixed constant.

Proof. I_k is the number of entries into elements of $\mathcal{D}$ during block k. Define $I_k^*=\sum_{B_k\leq n<B_{k+1},\mathbf{s}_{n+1}\in C^*}\widetilde{I}_n$ as the number of entries into elements of $\mathcal{D}$ which are subsets of C^* during block k. By (a) in the second assertion of Theorem 6, $\mathbf{E}[\sum_{j\leq k}I_j]\leq M+M\times\mathbf{E}[\sum_{j\leq k}I_j^*]$. By (b) in the second assertion of Theorem 6, $\mathbf{E}[I_k^*]\leq\lambda_k^{1-\alpha}\leq\lambda_kL_k$. ■

4.4 Definition of $\widehat{r}_k$

4.4.1 Partition into Subsets

The value of $\widehat{r}_k$ depends only on the sequence of signals received during block k. For notational simplicity, we drop the subscript k : we thus write L instead of L_k, λ instead of λ_k, etc. We also relabel the stages of block k from 1 to L, so that $B_{k+1}=L+1$.

We divide the block into $n_B:=\lfloor L^{1-\alpha}\rfloor$ subblocks of length $L':=\lfloor L^\alpha\rfloor$, and the remaining stages (at most $\lceil L^\alpha\rceil+\lceil L^{1-\alpha}\rceil$). We set

$$\widehat{r}=\frac{1}{n_B}\sum_{p=1}^{n_B}\widehat{\rho}_p,$$

where the value of $\widehat{\rho}_p$ depends only on the sequence of signals received during sub-block p. Once again for notational simplicity, we drop the subscript p.

[4] The existence of an optimal strategy follows from Lemma 3 in RSVa.

4.4.2 Notations

Here we define few r.v.s that will be used all through the paper. We restrict ourselves to a single sub-block.

Denote $\mathbf{N}_{s,a,b\to t} = |\{n \le L' \mid (\mathbf{s}_n, \mathbf{a}_n, \mathbf{b}_n, \mathbf{s}_{n+1}) = (s,a,b,t)\}|$. This is the number of stages in the sub-block where the actions (a,b) were chosen at state s, and the subsequent state was t.

We define $\mathbf{N}_{s,a,b} = \sum_{t\in S} \mathbf{N}_{s,a,b\to t}$, $\mathbf{N}_{s,a} = \sum_{b\in B} \mathbf{N}_{s,a,b}$, $\mathbf{N}_{s,b} = \sum_{a\in A} \mathbf{N}_{s,a,b}$, and $\mathbf{N}_s = \sum_{b\in B, a\in A} \mathbf{N}_{s,a,b}$. Define $\mathbf{N}_{s,a\to t} = \sum_{b\in B} \mathbf{N}_{s,a,b\to t}$, and $\mathbf{N}_{s,b\to t} = \sum_{a\in A} \mathbf{N}_{s,a,b\to t}$. Finally, define $\mathbf{N}_{s\to t} = \sum_{a\in A} \mathbf{N}_{s,a\to t} = \sum_{b\in B} \mathbf{N}_{s,b\to t}$.

We define the *empirical transition function* given different data as follows.

$$\mathbf{q}(t \mid s,a,b) = \frac{\mathbf{N}_{s,a,b\to t}}{\mathbf{N}_{s,a,b}}, \quad \mathbf{q}(t \mid s,a) = \frac{\mathbf{N}_{s,a\to t}}{\mathbf{N}_{s,a}},$$

$$\mathbf{q}(t \mid s,b) = \frac{\mathbf{N}_{s,b\to t}}{\mathbf{N}_{s,b}}, \text{ and } \mathbf{q}(t \mid s) = \frac{\mathbf{N}_{s\to t}}{\mathbf{N}_s}.$$

These quantities are defined whenever the denominator does not vanish. The *empirical play* is the stationary strategy $\overline{\mathbf{y}}$ where $\overline{\mathbf{y}}^s(b) := \frac{\mathbf{N}_{s,b}}{\mathbf{N}_s}$.

For $s \in S$, we let $\overline{A}(s) \subseteq A$ be the set of actions that are relevant for the indistinguishability relation:

$$\overline{A}(s) = \left\{ a \in A \mid x^s_\lambda(a) \ge \frac{\lambda}{\varepsilon(\lambda)} \text{ for every } \lambda \text{ close to zero} \right\}.$$

4.4.3 The estimator

For every $s \in S$, let $\widehat{\mathbf{y}}^s$ minimize

$$\max\{\|\mathbf{q}(\cdot \mid s,a) - q(\cdot \mid s,a,y)\|, a \in \overline{A}(s)\}, \tag{5}$$

among $y \in \Delta(B)$, and define

$$\widehat{\rho} = \frac{1}{L'} \sum_{s\in S} \mathbf{N}_s \widetilde{r}_\lambda(s, x_\lambda, \widehat{\mathbf{y}}^s).$$

The stationary strategy $\widehat{\mathbf{y}} = (\widehat{\mathbf{y}}^s)_{s\in S}$ is a good estimator of the strategy used by player 2, in the sense that it provides the best fit to the observed one-step transitions. However, it fails to take into account the temporal structure of the transitions, for instance cycles that may exist in the history. Thus, there may exist a stationary strategy of a higher order (cyclic Markov strategy) that would fit much more nicely to the observed signals. This is an important issue that is discussed in detail in Rosenberg et *al.*

4.5 Applying the technique of Mertens and Neyman

The main result of this section is Proposition 9 below. This Proposition, together with Theorem 33 below and Lemma 8, imply that σ guarantees $v - 3\bar{\varepsilon}$.

Define $I = \sum_{n\le L} I_n$.

Proposition 9 *There exists α and Z such that the strategy $\sigma(\alpha, Z)$ satisfies the following, for every strategy τ and every initial state $\overline{s} \in S$.*

C1 $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\frac{1}{L}\sum_{n=1}^{L} r(\mathbf{s}_n,\mathbf{a}_n,\mathbf{b}_n)\right] \geq \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\widehat{r}\right] - \overline{\varepsilon}.$

C2 $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\lambda L\widehat{\mathbf{r}} + \beta I + (1-\lambda L)v_\lambda(\mathbf{s}_{L+1})\right] \geq v_\lambda(\overline{s}) - \overline{\varepsilon}\lambda L$, *where* $\beta \in (0, \varepsilon/12M)$, *and* M *is the constant given by Lemma 8.*

We let the initial state $\overline{s} \in S$ and the strategy τ of player 2 be fixed. By section 2.4 and Kuhn's Theorem, it is enough to consider pure strategies τ, that depend only on the sequence of states visited so far.

The proof of **C1** appear in section 5. The proof of **C2**, which is substantially more involved, appears in section 6.

4.6 Fixing parameters

We here offer specific values for Z and α. There are few conditions to be satisfied. Some are related to the data of the game, and some to standard large deviations inequalities.

Fix $\eta < \overline{\varepsilon}/(6|S \times A \times B|)$ that satisfies Lemma 1 w.r.t. $\delta = \varepsilon$, and set $C_1 = 201$ and $C_2 = 220|S|^{|S|}$. Fix $\beta > 0$ sufficiently small such that $\beta < \eta/20|B|$.

Fix $\omega \in (0, 1/9)$ sufficiently small so that $d + 3\omega < 1$.[5] Apply Theorem 6 and get the constants $\alpha_1 > 1 - \omega/2$ and $\alpha_2 < \omega/2$. By Lemma 7, for every $D \in \mathcal{D}$ (where $\mathcal{D}$ is the partition given by Theorem 6), every two states $s, s' \in D$, and every $\lambda > 0$ sufficiently small, $|v_\lambda(s) - v_\lambda(s')| < \lambda^\omega$.

Fix $\psi \in (0, \alpha_2)$, and apply Theorem 17 with ψ and β to obtain ξ, κ, δ_2 and α_0. Choose $\delta, \delta_1 \in (0, \min\{\omega, \delta_2\})$, and $a_3 > 0$.

Choose $\alpha \in (\alpha_0, 1)$ sufficiently large so that the following inequalities hold. (A.i) $-\omega/\alpha + \alpha\delta_1 < -1/\alpha + \alpha$, (A.ii) $-\omega/\alpha + \alpha\delta < -1/\alpha + \alpha$, (A.iii) $\alpha\psi - \alpha_2/\alpha < -1/\alpha + \alpha$. Observe that since $\alpha \in (0, 1)$, we have (A.iv) $-1/\alpha + \alpha < 0$.

Define a function $L' : (0, +\infty) \to \mathbf{N}$ as follows.

$$L'(z) = \lfloor L^\alpha(z) \rfloor.$$

Observe that $\lim_{z\to\infty} L'(z) = +\infty$.

Choose Z_0 sufficiently large so that (a) Theorem 17 holds for every $z \geq Z_0$ (w.r.t. α), (b) Proposition 11 holds for every $z \geq Z_0$, (c) $\frac{2+L^\alpha(z)+L^{1-\alpha}(z)}{L(z)} < \frac{\overline{\varepsilon}}{3}$ for every $z \geq Z_0$, and (d) finitely many inequalities of the form $K_1 L(z)^{d_1} \geq K_2 L(z)^{d_2}$ hold for every $z \geq Z_0$, where $K_1, K_2 > 0$ and $d_1 < d_2$.

From now on, we fix a block, and so we omit the dependence on z from all variables.

4.7 Typical histories

We here define a set of histories that are typical, in a sense close to the meaning assigned to it in information theory, see e.g. Cover and Thomas (1991).

[5] Recall that d is defined in section 4.2.

Definition 10 *$T^{L'}_{\beta,\delta_2}$ is the event consisting of histories of length L' such that, for every $s \in S$, every $C \subseteq S$ and every $b \in B$,*

$$\mathbf{N}_{s,b} \max \{\mathbf{q}(C \mid s, b), q(C \mid s, x_\lambda, b)\} \geq L'^{\delta_2} \Rightarrow \left| 1 - \frac{\mathbf{q}(C \mid s, b)}{q(C \mid s, x_\lambda, b)} \right| \leq \beta.$$

When no confusion may arise, we simple denote this event by T_{β,δ_2}.

Thus, T_{β,δ_2} is the set of histories along which empirical transitions in the first subblock accurately reflect the expected transitions given the empirical play of player 2.

By Theorem 3.5 in Rosenberg, Solan and Vieille (2002b, henceforth RSVb),

$$\mathbf{P}_{\overline{s},x_\lambda,\tau}(T_{\beta,\delta_2}) \geq 1 - \varepsilon\lambda L'.$$

The set T_{β,δ_2} contains only histories of length $L'+1$. It will later be convenient to speak of the first time when the current history fails to be typical. We therefore introduce a slight extension of the previous notion. Given $n \in \mathbf{N}, s \in S, b \in B$ we let $\mathbf{N}^n_s = |\{j < n, \mathbf{s}_j = s\}|$, and define the analogous counters $\mathbf{N}^n_{s,b}$ and $\mathbf{N}^n_{s,b\to t}$ accordingly. Also, we let $\mathbf{q}_n(t \mid s, b) := \frac{\mathbf{N}^n_{s,b\to t}}{\mathbf{N}^n_{s,b}}$ denote the empirical transition function in the first $n-1$ stages. We define T^n_{β,δ_2} to be the event consisting of histories of length n such that, for every $s \in S$, $C \subset S$ and $b \in B$,

$$\mathbf{N}^n_{s,b} \max \{\mathbf{q}_n(C \mid s, b), q(C \mid s, x_\lambda, b)\} \geq L'^{\delta_2} \Rightarrow \left| 1 - \frac{\mathbf{q}_n(C \mid s, b)}{q(C \mid s, x_\lambda, b)} \right| \leq \beta.$$

Let $\theta_1 = \inf \left\{ n \leq L', \mathbf{h}_n \notin T^n_{\beta,\delta_2} \right\}$. Theorem 3.5 in RSVb extends to the next result.

Proposition 11 *Provided λ is sufficiently small, for each τ, and $s \in S$,*

$$\mathbf{P}_{s,x_\lambda,\tau}(\theta_1 \leq L'+1) \leq \varepsilon\lambda L'.$$

5 Proof of C1

This section contains the proof of **C1** in Proposition 9.

Lemma 12 *On the event T_{β,δ_2}, the next implication holds for every $s \in S$ and every $a \in \overline{A}(s)$:*

$$\mathbf{N}_s > \varepsilon L' \Rightarrow \|q(\cdot \mid s, a, \overline{\mathbf{y}}^s) - \mathbf{q}(\cdot \mid s, a)\| < \eta/2.$$

Proof. Let $s \in S$ and $a \in \overline{A}(s)$ be given, and assume that $\mathbf{N}_s > \varepsilon L'$.

We first prove that

$$\left| x^s_\lambda(a) - \frac{\mathbf{N}_{s,a}}{\mathbf{N}_s} \right| \leq 2|B|\beta x_\lambda(a). \tag{6}$$

Denote by $S_a = \{s \in S \mid q(t \mid s, a, b) > 0 \text{ for some } b \in B\}$. Recall that for every $a' \neq a$ and every $b \in B$, $q(S_a \mid s, a', b) = 0$ (see Section 2.4).

Observe that

$$x^s_\lambda(a) = q(S_a \mid s, x_\lambda, \overline{\mathbf{y}}) = \frac{1}{\mathbf{N}_s} \sum_{b \in B} \mathbf{N}_{s,b} q(S_a \mid s, x^s_\lambda, b), \tag{7}$$

and

$$\frac{\mathbf{N}_{s,a}}{\mathbf{N}_s} = \mathbf{q}(S_a \mid s) = \frac{1}{\mathbf{N}_s} \sum_{b \in B} \mathbf{N}_{s,b} \mathbf{q}(S_a \mid s, b). \tag{8}$$

On the event T_{β,δ_2} we have

$$\begin{aligned} &\mathbf{N}_{s,b} \max\{q(S_a \mid s, x_\lambda, b), \mathbf{q}(S_a \mid s, b)\} \geq L'^{\delta_2} \\ &\quad \Rightarrow |q(S_a \mid s, x_\lambda, b) - \mathbf{q}(S_a \mid s, b)| \leq \beta q(S_a \mid s, x_\lambda, b). \end{aligned} \tag{9}$$

By the assumption and since $\delta_2 > \delta_1$, $L'^{\delta_2}/\mathbf{N}_s < \frac{1}{\varepsilon L'^{1-\delta_2}} < \beta x^s_\lambda(a)$, hence (6) follows by (7), (8), (9) and Lemma 5.

Fix now $t \in S_a$.

Observe that $\mathbf{q}(t \mid s, b) = \frac{\mathbf{N}_{s,a,b \to t}}{\mathbf{N}_{s,b}}$, and $q(t \mid s, x_\lambda, b) = x^s_\lambda(a) q(t \mid s, a, b)$. Therefore, on the event T_{β,δ_2} we have

$$\begin{aligned} &\mathbf{N}_{s,b} \max\left\{\frac{\mathbf{N}_{s,a,b \to t}}{\mathbf{N}_{s,b}}, x^s_\lambda(a) q(t \mid s, a, b)\right\} \geq L'^{\delta_2} \\ &\quad \Rightarrow \left|\frac{\mathbf{N}_{s,a,b \to t}}{\mathbf{N}_{s,b}} - x^s_\lambda(a) q(t \mid s, a, b)\right| \leq \beta x^s_\lambda(a) q(t \mid s, a, b). \end{aligned} \tag{10}$$

Note that $\sum_{b \in B} \mathbf{N}_{s,b} \frac{\mathbf{N}_{s,a,b \to t}}{\mathbf{N}_{s,b}} = \mathbf{N}_{s,a \to t}$, and $\sum_{b \in B} \mathbf{N}_{s,b} x^s_\lambda(a) q(t \mid s, a, b) = \mathbf{N}_s x^s_\lambda(a) q(t \mid s, a, \overline{\mathbf{y}})$. Since $L'^{\delta_2}/\mathbf{N}_s < \frac{1}{\varepsilon L'^{1-\delta_2}} < \beta x^s_\lambda(a) q(t \mid s, a, b)$, by applying Lemma 5 we obtain

$$|\mathbf{N}_{s,a \to t} - \mathbf{N}_s x^s_\lambda(a) q(t \mid s, a, \overline{\mathbf{y}}^s)| \leq 2\beta |B| \mathbf{N}_s x^s_\lambda(a). \tag{11}$$

By (6) and **Fact 1**, $\left|1 - \frac{\mathbf{N}_s x_\lambda(a)}{\mathbf{N}_{s,a}}\right| \leq 4|B|\beta$. Therefore, dividing (11) by $\mathbf{N}_{s,a}$ and since $2|B|\beta < 1$ we get

$$\begin{aligned} |\mathbf{q}(t \mid s, a) - q(t \mid s, a, \overline{\mathbf{y}})| &= \left|\frac{\mathbf{N}_{s,a \to t}}{\mathbf{N}_{s,a}} - q(t \mid s, a, \overline{\mathbf{y}})\right| \\ &\leq 4|B|\beta + (1 + 4|B|\beta) 2|B|\beta < \eta/2, \end{aligned}$$

where the last inequality holds by the choice of β.

Fix $(a, b) \in A \times B$. Observe that $\mathbf{N}_{s,a,b} = \sum_{t \in S} \mathbf{N}_{s,b} \frac{\mathbf{N}_{s,a,b \to t}}{\mathbf{N}_{s,b}}$, and $\mathbf{N}_{s,b} x_\lambda(a) = \sum_{t \in S} \mathbf{N}_{s,b} x_\lambda(a) q(t \mid s, a, b)$. Therefore, (10) and Lemma 5 imply in addition

$$|\mathbf{N}_{s,a,b} - \mathbf{N}_{s,b} x_\lambda(a)| \leq 2\beta |S| \mathbf{N}_{s,a,b}, \tag{12}$$

provided $\max\{\mathbf{N}_{s,a,b}, \mathbf{N}_{s,b} x_\lambda(a)\} \geq L'^{\delta_2}/\beta$. ■

Corollary 13 *One has*

$$\left|\mathbf{E}_{\overline{s}, x_\lambda, \tau}\left[\widehat{\rho}\right] - \mathbf{E}_{\overline{s}, x_\lambda, \tau}\left[\frac{1}{L'} \sum_{s \in S} \mathbf{N}_s \widetilde{r}_\lambda(s, x_\lambda, \overline{\mathbf{y}})\right]\right| \leq 4|S|\varepsilon.$$

Proof. Let $s \in S$ be arbitrary. By Lemma 12 and the choice of η,

$$|\widetilde{r}_\lambda(s, x_\lambda, \overline{\mathbf{y}}) - \widetilde{r}_\lambda(s, x_\lambda, \widehat{\mathbf{y}})| \leq \varepsilon \text{ on } \left\{\mathbf{N}_s > \varepsilon L'\right\} \cap T_{\beta,\delta_2}. \tag{13}$$

Whenever $\mathbf{N}_s \leq \varepsilon L'$ we have $\frac{\mathbf{N}_s}{L'} \leq \varepsilon$. Hence, by (13), and Proposition 11,

$$\left|\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\frac{\mathbf{N}_s}{L'}(\widetilde{r}_\lambda(s, x_\lambda, \overline{\mathbf{y}}) - \widetilde{r}_\lambda(s, x_\lambda, \widehat{\mathbf{y}})\right]\right| \leq 4\varepsilon.$$

The result follows by summation over s. ■

Next, we show that the average payoff in the sub-block, $\mathbf{E}_{\overline{s},x_\lambda,\tau}[\overline{r}_{L'}] = \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\frac{1}{L'}\sum_{n=1}^{L'} r(\mathbf{s}_n, \mathbf{a}_n, \mathbf{b}_n)\right]$ is close to $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\frac{1}{L'}\sum_{s\in S}\mathbf{N}_s r(s, x_\lambda, \overline{\mathbf{y}})\right]$.

Lemma 14 *One has*

$$\left|\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\overline{r}_{L'}\right] - \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\frac{1}{L'}\sum_{s\in S}\mathbf{N}_s r(s, x_\lambda, \overline{\mathbf{y}})\right]\right| \leq 2\varepsilon|S| \times |A| \times |B| < \frac{\overline{\varepsilon}}{3}. \tag{14}$$

Proof. The left hand side in (14) is equal to

$$\mathbf{E}_{\overline{s},x_\lambda(a),\tau}\left[\frac{1}{L'}\sum_{s,a,b}(\mathbf{N}_{s,a,b} - x_\lambda(a)\mathbf{N}_{s,b})r(s,a,b)\right].$$

The result follows from (12), since for every a, b, whenever $\mathbf{N}_s \leq \varepsilon L'$ or $\max\{\mathbf{N}_{s,a,b}, \mathbf{N}_{s,b}x_\lambda(a)\} \leq \varepsilon L'$ we have $\mathbf{N}_{s,a,b} - x_\lambda(a)\mathbf{N}_{s,b} < \varepsilon L'$, and since $\overline{\varepsilon} > 6|A \times B \times S|\varepsilon$. ■

If Z_0 is sufficiently large, $\left|\overline{r}_L - \overline{r}_{L'\lfloor L^{1-\alpha}\rfloor}\right| \leq \frac{\overline{\varepsilon}}{3}$. Since $r \geq \widetilde{r}_\lambda$, assertion **C1** in Proposition 9 follows from Corollary 13, Lemma 14, and by summation over the different sub-blocks.

6 Proof of C2

6.1 A decomposition

We denote by Ω the communicating set that contains the initial state $\overline{s}$. Given a bounded stopping time $\theta \leq e_\Omega$, we denote by $\mathcal{H}_\theta$ the finite σ-algebra of events up to θ. Each atom of $\mathcal{H}_\theta$ can be identified with a history of finite length. We denote by $H(\theta)$ those atoms such that $\theta < e_\Omega$.

Given $h = (s_1, a_1, b_1, ..., s_k) \in H(\theta)$, we now define an auxiliary probability distribution $\mathbf{P}^h$ on histories of length k. Let $(z_n)_{n\leq k}$ be the nonhomogenous Markov chain defined over S as follows. For each n, either $z_n = s_n$ or $z_n \notin \Omega$. States in $\overline{\Omega}$ are absorbing. If $z_n = s_n$, $z_{n+1} = s_{n+1}$ with probability $q(\Omega \mid s_n, x_\lambda^{s_n}, \tau(s_1, ..., s_n))$, and, for every $s \in \overline{\Omega}$, $z_{n+1} = s$ with probability $q(s \mid s_n, x_\lambda^{s_n}, \tau(s_1, ..., s_n))$. In words, (z_n) follows h up to stage $k+1$ with the option of leaving Ω and be absorbed before. We let $\mathbf{P}^h$ be the law of (z_n) and we denote by $\mathbf{E}^h$ the expectation w.r.t. $\mathbf{P}^h$. Next, we set $p(h) = \mathbf{P}_{\overline{s},x_\lambda,\tau}(h) = \prod_{n=1}^{k-1} q(s_{n+1} \mid s_n, x_\lambda^{s_n}, \tau(s_1, ..., s_n))$ and

$$\omega(h) = \mathbf{P}^h(e_\Omega \leq k) = 1 - \prod_{n=1}^{k-1} q(\Omega \mid s_n, x_\lambda^{s_n}, \tau(s_1, ..., s_n)).$$

$\omega(h)$ may be interpreted as a probability that exit from Ω occurs along h.

The following lemma follows by elementary algebraic manipulations.

Lemma 15 *For each $\mathcal{H}_\theta$-measurable random variable X, one has*

$$\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[X\right] = \sum_{h\in H(\theta)} \frac{p(h)}{1-\omega(h)}\mathbf{E}^h\left[X\right],$$

where $0/0$ *is defined to be* 0.

The next lemma provides a rewriting of $\mathbf{E}^h\left[v_\lambda(\mathbf{s}_k)\right]$, valid for typical sequences along which the probability of exit from Ω is close to zero.

For simplicity, given a history h, we denote by N^h_s the value of the r.v. $\mathbf{N}_s$ at h. We define $N^h_{s,b}$ and $q^h(t\cdot \mid s,b)$ similarly.

Lemma 16 *Let $h = (s_1, a_1, b_1, ..., s_k)$ be a history of length k, such that $s_l \in \Omega$ for each $l \leq k$. Assume that $\omega(h) < \beta$ and $h \in T^k_{\beta,\delta_2}$. Then*

$$\left|\mathbf{E}^h\left[v_\lambda(\mathbf{s}_k)\right] - \left(v_\lambda(\overline{s}) + \sum_{s\in\Omega} N^h_s\left(\mathbf{E}_{q(\cdot\mid s,x_\lambda,\overline{y})}\left[v_\lambda\right] - v_\lambda(s)\right)\right)\right| \leq 2\varepsilon\lambda L' + \beta C|S|\omega(h).$$

Proof. For notational convenience, we shall establish the lower bound on $\mathbf{E}^h\left[v_\lambda(\mathbf{s}_k)\right]$, all arguments being symmetric. Since $\omega(h) < \beta$ and by Lemma 19 in RSVa,

$$|\omega(h) - \pi| \leq C\omega(h)^2, \text{ where } \pi = \sum_{j=1}^{k-1} q(\overline{\Omega} \mid s_j, x_\lambda, b_j), \tag{15}$$

and by Corollary 20 in RSVa,

$$\left|\mathbf{P}^h(e_\Omega \leq k, \mathbf{s}_{e_\Omega} = t) - \sum_{j=1}^{k-1} q(t \mid s_j, x_\lambda, b_j)\right| \leq C\omega(h)^2, \text{ for } t \notin \Omega. \tag{16}$$

By (15) and (16), and since $\omega(h) < \beta$,

$$\left|\mathbf{E}^h\left[v_\lambda(\mathbf{s}_k)\right] - \left((1-\pi)v_\lambda(s_k) + \sum_{j=1}^{k-1}\mathbf{E}_{q(\cdot\mid s_j,x_\lambda,b_j)}\left[v_\lambda 1_{\overline{\Omega}}\right]\right)\right| \leq C\beta\left|S\right|\omega(h). \tag{17}$$

Next, we rewrite $(1-\pi)v_\lambda(s_k)$. By the definition of ω, $|v_\lambda(s_j) - v_\lambda(s_k)| \leq \lambda^\omega$ for $j \leq k-1$. Therefore,

$$\begin{aligned}
(1-\pi)v_\lambda(s_k) &= v_\lambda(s_k) + \sum_{j=1}^{k-1} v_\lambda(s_k)\left(q(\Omega \mid s_j, x_\lambda, b_j) - 1\right) \\
&\geq v_\lambda(s_k) + \sum_{j=1}^{k-1}\left(v_\lambda(s_{j+1})q(\Omega \mid s_j, x_\lambda, b_j) - v_\lambda(s_{j+1})\right) - \lambda^\omega\sum_{j=1}^{k-1} q(\overline{\Omega} \mid s_j, x_\lambda, b_j) \\
&= v_\lambda(\overline{s}) + \sum_{j=1}^{k-1}\left(v_\lambda(s_{j+1})q(\Omega \mid s_j, x_\lambda, b_j) - v_\lambda(s_j)\right) - \pi\lambda^\omega, \qquad (18)
\end{aligned}$$

where the last equality holds since $\sum_{j=1}^{k-1} v_\lambda(s_{j+1}) = \sum_{j=1}^{k-1} v_\lambda(s_j) + v_\lambda(s_k) - v_\lambda(\overline{s})$.

We plug (18) into (17) to get

$$\begin{aligned}\mathbf{E}^h\left[v_\lambda(\mathbf{s}_k)\right] \geq v_\lambda(\overline{s}) + \sum_{j=1}^{k-1}\Big(&q(\Omega \mid s_j, x_\lambda, b_j) v_\lambda(s_{j+1}) + \mathbf{E}_{q(\cdot|s_j,x_\lambda,b_j)}\left[v_\lambda 1_{\overline{\Omega}}\right] - v_\lambda(s_j)\Big) \\ &- \lambda^\omega - \beta C\left|S\right|\omega(h). \end{aligned} \tag{19}$$

Note that the summation Σ on the right-hand side of (19) coincides with

$$\Sigma = \sum_{s\in\Omega, b\in B} N^h_{s,b}\left(\mathbf{E}_{\widetilde{q}(\cdot|s,b)}\left[v_\lambda(\cdot)\right] - v_\lambda(s)\right), \tag{20}$$

where

$$\widetilde{q}(t \mid s, b) = \begin{cases} q(t \mid s, x_\lambda, b) \text{ for } t \notin \Omega \\ q(\Omega \mid s, x_\lambda, b) q^h(t \mid s, b) \text{ for } t \in \Omega \end{cases}.$$

Observe that $\sum_{s\in\Omega,b\in B} N^h_{s,b} q^h(t \mid s, b) = |\{2 \leq t \leq k, s_n = t\}|$ and $\sum_{t\in\Omega,b\in B} N^h_{s,b} q^h(t \mid s, b) = |\{1 \leq t \leq k-1, s_n = s\}|$. In particular,

$$\sum_{s,t\in\Omega,b\in B} N^h_{s,b} q^h(t \mid s, b)(v_\lambda(t) - v_\lambda(s)) \leq |S|\lambda^\omega. \tag{21}$$

By (15) and the definition of ω,

$$\begin{aligned}\sum_{s,t\in\Omega,b\in B} N^h_{s,b}(q^h(t \mid s, b) - \widetilde{q}(t \mid s, b))(v_\lambda(t) - v_\lambda(s)) = \\ = \sum_{s,t\in\Omega,b\in B} N^h_{s,b} q^h(t \mid s, b) q(\overline{\Omega} \mid s, x_\lambda, b))(v_\lambda(t) - v_\lambda(s)) \\ \leq |S|\lambda^\omega \sum_{s\in\Omega,b\in B} N^h_{s,b} q(\overline{\Omega} \mid s, x_\lambda, b) \leq |S|\pi\lambda^\omega. \end{aligned} \tag{22}$$

By (21) and (22),

$$\sum_{s\in\Omega,b\in B} N^h_{s,b}\widetilde{q}(t \mid s, b)(v_\lambda(t) - v_\lambda(s)) \leq 2|S|\lambda^\omega. \tag{23}$$

For the time being, we fix $s \in \Omega$ and $b \in B$, and we define

$$\Delta_{s,b} = N^h_{s,b}\left(\mathbf{E}_{\widetilde{q}(\cdot|s,b)}\left[v_\lambda\right] - \mathbf{E}_{q(\cdot|s,x_\lambda,b)}\left[v_\lambda\right]\right).$$

Let $\Omega_0 = \left\{t \in \Omega, N^h_{s,b}\max\{\widetilde{q}(t \mid s, b), q(t \mid s, x_\lambda, b)\} \geq L'^{\delta_1}\right\}$, and $\Omega_1 = \Omega \backslash \Omega_0$.

For $t \in \Omega_1$, one has

$$N^h_{s,b}\max\{\widetilde{q}(t \mid s, b), q(t \mid s, x_\lambda, b)\}\left|v_\lambda(t) - v_\lambda(s)\right| < L'^{\delta_2}\lambda^\omega. \tag{24}$$

Consider $t \in \Omega_0$. In particular, $N^h_{s,b} \geq L'^{\delta_2}$. Since $q^h(t \mid s, b) \geq \widetilde{q}(t \mid s, b)$, one has $N^h_{s,b}\max\{q^h(t \mid s, b), q(t \mid s, x_\lambda, b)\} \geq L'^{\delta_2}$. Since $h \in T^k_{\beta,\delta_2}$ we have $\left|1 - \frac{q^h(t|s,b)}{q(t|s,x_\lambda,b)}\right| \leq \beta$. By (15), and since $\omega(h) < \beta$, $\sum_{s\in\Omega,b\in B} N^h_{s,b} q(\overline{\Omega} \mid s, x_\lambda, b) = \pi \leq \beta + C\beta^2$. Hence,

$$\left|\frac{\widetilde{q}(t \mid s, b)}{q^h(t \mid s, b)} - 1\right| = q(\overline{\Omega} \mid s, x_\lambda, b) \leq \frac{\beta + C\beta^2}{N_{s,b}} \leq \frac{2\beta}{L'^{\delta_2}} \leq \beta, \tag{25}$$

where the last inequality holds by provided Z_0 is sufficiently large. By **Fact 2**, this yields $\left|1 - \frac{\widetilde{q}(t|s,b)}{q(t|s,x_\lambda,b)}\right| \leq 3\beta$, and by **Fact 1** we get

$$\left|1 - \frac{q(t \mid s, x_\lambda, b)}{\widetilde{q}(t \mid s, b)}\right| \leq 6\beta. \tag{26}$$

Since $q(t \mid s, x_\lambda, b) = \widetilde{q}(t \mid s, b)$ for each $t \notin \Omega$, one has, by (26) and (24)

$$\begin{aligned}
\Delta_{s,b} &= N^h_{s,b} \sum_{t\in\Omega} (\widetilde{q}(t \mid s, b) - q(t \mid s, x_\lambda, b)) v_\lambda(t) \\
&= N^h_{s,b} \sum_{t\in\Omega} \left(\widetilde{q}(t \mid s, b) - q(t \mid s, x_\lambda, b)\right) \left(v_\lambda(t) - v_\lambda(s)\right) \\
&\geq - |S| \, L'^{\delta_1} \lambda^\omega - 6\beta N^h_{s,b} \sum_{t\in\Omega_0} \widetilde{q}(t \mid s, b)(v_\lambda(t) - v_\lambda(s)) \\
&\geq - |S| \, (6\beta + 1) L'^{\delta_1} \lambda^\omega - 6\beta N^h_{s,b} \sum_{t\in\Omega} \widetilde{q}(t \mid s, b)(v_\lambda(t) - v_\lambda(s)).
\end{aligned} \tag{27}$$

By (27) and (23),

$$\sum_{s\in\Omega, b\in B} \Delta_{s,b} \geq - |S|^2 \, |B| \, (6\beta + 1) L'^{\delta_1} \lambda^\omega - 12\beta |S|^2 \lambda^\omega \geq -\varepsilon \lambda L' \tag{28}$$

Therefore, by plugging (28) into (20) and (20) into (19), one obtains by (A.ii)

$$\mathbf{E}^h \left[v_\lambda(\mathbf{s}_k)\right] - \left(v_\lambda(\overline{s}) + \sum_{s\in\Omega, b\in B} N_{s,b} \left(\mathbf{E}_{q(\cdot|s,x_\lambda,b)} \left[v_\lambda\right] - v_\lambda(s) \right) \right) \geq -2\varepsilon\lambda L' - \beta C \, |S| \, \omega(h).$$

The reverse inequality follows similarly. ■

6.2 The case of a single communicating set

In RSVb we study the case where there is a *single* communicating set Ω. Recall that in this case, for every state $t \in \Omega$ there is a stationary strategy $y(t)$ such that, for every initial state $s \in \Omega$, $\mathbf{E}_{s,x_\lambda,y(t)}[e_{\Omega\setminus\{t\}}] \leq \lambda^{-\alpha_1}$.

We consider there a history $h \in T_{\beta,\delta_2}$, and denote by $\overline{y}^h$ the empirical transition function along h. Our goal is to construct a strategy τ such that (i) in most stages the mixed action played is $\overline{y}$, and (ii) under (x_λ, τ), with high probability the realized number of visits to each state s is close to N^h_s.

To be more precise, we define a partition $\{S_1, S_2, \ldots, S_K\}$ that satisfies certain desirable properties. S_K contains all states in S that are visited infrequently. Define $n_k = \sum_{s\in S_k} N^h_s$ as the number of visits to S_k along h. We then define a strategy τ that has $K-1$ phases. Phase k lasts n_k stages, and in that phase we approximate the number of visits in S_k. Phase k is divided into rounds: we first follow $\overline{y}^h$ until the play leaves S_k, we then follow $y(t)$, for an appropriately chosen $t \in S_k$, until the play reaches t, and so on. The properties of the partition guarantee that (a) the number of stages the play remains out of S_k in phase k is relatively small compared to the number of stages it spends in S_k, and (b) with high probability, the number of times each state in S_k is visited is close to N^h_s.

We now state the result of RSVb formally.

Let $h \in T_{\beta,\delta_2}$. For $\xi > 0$ let $\Omega^h = \Omega^h_\xi = \left\{s \in \Omega, N^h_s \geq L'^{\xi}\right\}$ denote the set of frequently visited states along h. In RSVb we construct a sequence of stopping times $(u_l)_{l\geq 0}$ and a sequence of S-valued r.v.s $(\widetilde{\mathbf{s}}_l)_{l\geq 0}$ such that

1. $\widetilde{\mathbf{s}}_l$ is $\mathcal{H}_{u_l}$-measurable, for each l.
2. $u_0 = 0$ and $u_{l+1} = \inf\{n > u_l : \mathbf{s}_n = \widetilde{s}_l\}$ for each even l.
3. $\mathbf{s}_n \in \Omega^h$ for each $u_l \leq n < u_{l+1}$ for each odd l.

Those two sequences depend only on the history h, and not on the transitions in the game.

We then define a strategy τ^h as follows:

- For l even, τ^h coincides with $y(\widetilde{\mathbf{s}}_l)$ from u_l to u_{l+1}.
- For l odd, τ^h coincides with $\overline{y}^h$ from u_l to u_{l+1}.

If exit from Ω occurs, τ^h coincides with a stationary strategy that minimizes at every state s the quantity $\mathbf{E}_{q(\cdot|s,x_\lambda,\cdot)}[v_\lambda]$. Observe that the sequences $(u_l)_{l\geq 0}$ and $(\widetilde{\mathbf{s}}_l)_{l\geq 0}$, and therefore also τ^h, depend on ξ.

Set $\mathbf{M}^*_s = \{n \leq L' | \mathbf{s}_n = s, u_l \leq n < u_{l+1}$ for some odd $l\}$, and $\mathbf{N}^*_s = |\mathbf{M}^*_s|$. Set $\mathbf{M}^* = \cup_{s\in S}\mathbf{M}^*_s$. Set also $\widetilde{\mathbf{M}}_s = \{n \leq L' | \mathbf{s}_n = s, u_l \leq n < u_{l+1}$ for some even $l\}$, and $\widetilde{\mathbf{N}}_s = \left|\widetilde{\mathbf{M}}_s\right|$

The results from Section 3 in RSVb translates in our setup into the following.

Theorem 17 *For every $\psi \in (0, 1-\alpha_1)$ and every $\beta > 0$ sufficiently small there exist $\xi, \kappa, \delta_2, \alpha_0 \in (0,1)$ and for every $\alpha \in (\alpha_0, 1)$ there exists $Z_0 \in \mathbf{N}$ such that for every $z \geq Z_0$ and every history $h \in T^{L'(z)}_{\beta,\delta_2}$ the strategy τ^h satisfies the following.*

- $\mathbf{E}_{\overline{s},x_{\lambda(z)},\tau^h}[B] \leq L'^{\psi}(z)$ *where* $B = \sup\{l \geq 0, u_l \leq L'\}$.
- *For each* $s \in \Omega^h$, $\mathbf{P}_{\overline{s},x_{\lambda(z)},\tau^h}(\left|\mathbf{N}^*_s - N^h_s\right| > 56C_1\beta N^h_s) \leq \frac{1}{L'^{\kappa}(z)}$.

Remark 18 *The statement we provide here is different than the one in RSVb in two respects. First, in RSVb transitions are independent of λ, whereas in our case they are. However, for the result in RSVb it is sufficient to require that $L'^{1-\psi}$ is higher than the rate of mixing of the Markov decision process induced by x_λ, which is bounded by L'^{α_1/α^2}. Second, in RSVb there is an additional parameter ε, which can always be set to* $\min\{\beta/56C_1, 1/20C_1^2|S|^2\}$.

In our framework, there may be several communicating sets. We will therefore modify the transition function $q(\cdot \mid s, x_\lambda, b)$ by conditioning on remaining in Ω.

6.3 On typical histories

In this section we prove the next result.

Proposition 19 *For each $h \in T_{\beta,\delta_2}$, such that (i) $\omega(h) < \beta$, and (ii) the play along h remains in the same communicating set, there exists a strategy τ^h such that*

$$\left| \mathbf{E}^h \left[\widehat{\rho} \right] - \mathbf{E}_{\overline{s},x_\lambda,\tau^h} \left[\frac{1}{L'} \sum_{n=1}^{L'} \widetilde{r}_\lambda(\mathbf{s}_n, x_\lambda, \mathbf{y}_n) \right] \right| \leq \overline{\varepsilon}/4, \text{ and}$$

$$\mathbf{E}^h \left[v_\lambda(\mathbf{s}_{L'+1}) \right] \geq \mathbf{E}_{\overline{s},x_\lambda,\tau^h} \left[v_\lambda(\mathbf{s}_{L'+1}) \right] - (8 + |S|)\varepsilon\lambda L' - \lambda^\omega - 6\beta C\omega(h).$$

The rest of the section is devoted to the proof of this Proposition. We first construct a strategy τ^h. We then prove that it satisfies the requirements.

Let Ω be the communicating set that contains $s_1 = \overline{s}$. Recall that for every $t \in \Omega$ there is a stationary strategy $y(t)$ such that $\mathbf{E}_{s,x_\lambda,y(t)}[e_{\Omega \setminus t}] \leq 1/\lambda^{\alpha_1}$ and $\mathbf{P}_{s,x_\lambda,y(t)}(e_\Omega = e_{\Omega\setminus t}) \leq \lambda^{\alpha_2}$.

To invoke Theorem 17, we now modify the transitions by conditioning on remaining in Ω. For $s \in \Omega$, let $B(s) = \{b \in B, q(\Omega \mid s, x_\lambda, b) > 0\}$. For $b \in B(s)$, define $\widetilde{q}_\lambda(\cdot \mid s, b)$ to be the distribution $q(\cdot \mid s, x_\lambda, b)$, conditional on Ω: $\widetilde{q}_\lambda(t \mid s, b) = q(t \mid s, x_\lambda, b)/q(\Omega \mid s, x_\lambda, b)$ for every $t \in \Omega$. For $b \notin B(s)$, the definition of $\widetilde{q}_\lambda(\cdot \mid s, b) \in \Delta(\Omega)$ is arbitrary. For every $y \in \Delta(B)$ define $\widetilde{q}_\lambda(\cdot \mid s, y) = \sum_{b \in B} y_b \widetilde{q}_\lambda(\cdot \mid s, b)$. Let $V_s \subseteq \Delta(S)$ be the convex hull of the finite set $\{\widetilde{q}_\lambda(\cdot \mid s, b), b \in B\}$.

Let $(u_l)_{l \geq 0}$ and $(\widetilde{\mathbf{s}}_l)_{l \geq 0})$ be the sequences defined in section 6.2 w.r.t. $\widetilde{q}_\lambda$. Define a strategy τ^h as follows. If the play ever leaves Ω, τ^h is defined arbitrarily. As long as the play remains in Ω, τ^h is defined as follows.

- For l even, τ^h coincides with $y(\widetilde{\mathbf{s}}_l)$ from u_l to u_{l+1}.
- For l odd, τ^h coincides with $\overline{y}^h$ from u_l to u_{l+1}.

Define $\mathbf{M}^*$ and $\mathbf{N}^*_s$ as in Section 6.2.

Lemma 20 *The following implication holds on $T_{\beta,\delta_2} \cap \{\omega(h) < \beta\}$, for each $s, t \in \Omega$:*

$$N^h_s \max\{\mathbf{q}(t \mid s), \widetilde{q}_\lambda(t \mid s, \overline{y}^h)\} \geq L'^\delta$$
$$\Rightarrow \left| 1 - \frac{\mathbf{q}(t \mid s)}{\widetilde{q}_\lambda(t \mid s, \overline{y}^h)} \right| \leq 3\beta \left| B \right|. \quad (29)$$

Proof. Let $s, t \in \Omega$, $b \in B$ be given, and assume $N^h_s \max\{\mathbf{q}(t \mid s), \widetilde{q}_\lambda(t \mid s, \overline{y}^h)\} \geq L'^\delta$. Since $h \in T_{\beta,\delta_2}$, if $N^h_{s,b} \max\{\mathbf{q}(t \mid s, b), q(t \mid s, x_\lambda, b)\} \geq L'^{\delta_2}$, then $\left| 1 - \frac{\mathbf{q}(t|s,b)}{q(t|s,x_\lambda,b)} \right| \leq \beta$. By Lemma 5 and since $\delta_2 > \delta$ we have

$$\left| 1 - \frac{\mathbf{q}(t \mid s)}{q(t \mid s, x_\lambda, \overline{y}^h)} \right| \leq \beta \left| B \right|. \quad (30)$$

We conclude by following closely the proof of (25). Since $\omega(h) < \beta$, we likewise obtain $N^h_s \mathbf{q}(\overline{\Omega} \mid s) \leq \beta$. Hence, by **Fact 1,**

$$\left| 1 - \frac{\widetilde{q}_\lambda(t \mid s, \overline{y}^h)}{q(t \mid s, x_\lambda, \overline{y}^h)} \right| = \left| 1 - \frac{1}{\mathbf{q}(\Omega \mid s, \overline{y}^h)} \right| \leq 2\beta. \quad (31)$$

The result follows by (30) and (31), using **Fact 2.** ■

For every initial state $s \in \Omega$ and every strategy τ, let $\widetilde{\mathbf{P}}_{s,\tau}$ be the probability distribution over plays under the pair of strategies (x_λ, τ), when the transition rule is $\widetilde{q}_\lambda$ (rather than q). Let $\widetilde{\mathbf{E}}_{s,\tau}$ be the corresponding expectation operator. Recall that, for $t \in \Omega$, the stationary strategy $y(t)$ satisfies $\mathbf{E}_{s,x_\lambda,y(t)}\left[e_{\Omega\backslash t}\right] \le \frac{1}{\lambda^{\alpha_1}}$ and $\mathbf{P}_{s,x_\lambda,y(t)}(e_\Omega = e_{\Omega\backslash t}) \le \lambda^{\alpha_2}$. Under $\widetilde{\mathbf{P}}_{s,\tau}$, the expected time to leave $\Omega\backslash t$ is higher since the play bounces back to Ω in case $\mathbf{s}_{e_{\Omega\backslash t}} \notin \Omega$. Letting $M := \sup_{s\in\Omega} \widetilde{\mathbf{E}}_{s,y(t)}\left[e_{\Omega\backslash t}\right]$, one has the following identity

$$M \le \mathbf{E}_{\overline{s},x_\lambda,y(t)}\left[M\mathbf{1}_{e_\Omega = e_{\Omega\backslash t}} + e_{\Omega\backslash t}\right], \text{ hence } M \le \frac{1}{\lambda^{\alpha_1}(1-\lambda^{\alpha_2})} \le \frac{2}{\lambda^{\alpha_1}}.$$

Lemma 21 *On $T_{\beta,\delta_2} \cap \{\omega(h) < \beta\}$ one has*

$$\mathbf{P}_{\overline{s},x_\lambda,\tau^h}(e_\Omega - 1 \notin \mathbf{M}^*) \le \varepsilon\lambda L'. \tag{32}$$

Proof. Let an even l be given. Since $\mathbf{s}_n \in \Omega^h$ for every n such that $u_l \le n < u_{l+1}$ for l odd, and since $\mathbf{P}_{\overline{s},x_\lambda,y(t)}(e_\Omega < \min\{u_{l+1}, L'+1\} \mid e_\Omega \ge u_l) < \lambda^{\alpha_2}$, it follows by (A.iii) that

$$\mathbf{P}_{\overline{s},x_\lambda,\tau^h}(u_l < e_\Omega \le \min(u_{l+1}, L'+1) \text{ for some even } l) \le \lambda^{\alpha_2} \times \widetilde{\mathbf{E}}_{\overline{s},x_\lambda,\tau^h}\left[B\right] \le \lambda^{\alpha_2} L'^{\psi}.$$

and (32) follows. ■

Proposition 22 *On $T_{\beta,\delta_2} \cap \{\omega(h) < \beta\}$ one has*

$$\mathbf{E}^h\left[v_\lambda(\mathbf{s}_{L'+1})\right] - \mathbf{E}_{\overline{s},x_\lambda,\tau^h}\left[v_\lambda(\mathbf{s}_{L'+1})\right] \ge -(11 + \frac{2}{\eta_0} + |S|)\varepsilon\lambda L' - 6\beta C\omega(h) + \varepsilon\,|S|\,\lambda L' - \lambda^\omega - \lambda\,|S|\,L'^{\xi}.$$

Proof. We shall apply Theorem 17 to the transitions $\widetilde{q}_\lambda$. By Lemma 20, h is $(3\beta\,|B|\,, \delta)$-typical. We enrich the probability space to allow for the following construction. Whenever $n \in \mathbf{M}^*$ (so that $\mathbf{s}_n \in \Omega^h$), a random device selects a (fictitious) state $\mathbf{s} \in S$, according to $q(\cdot|\mathbf{s}_n, x_\lambda, \overline{y}^h)$. Denote by $\widehat{\mathbf{P}}_{\overline{s},\tau^h}$ the probability distribution over the larger probability space. Thus, the marginal of $\widehat{\mathbf{P}}_{\overline{s},\tau^h}$ over the set of regular histories coincides with $\widetilde{\mathbf{P}}_{\overline{s},\tau^h}$. For $n \le L'$, set $Z_n = \mathbf{1}_{\mathbf{s}\notin\Omega}$ and $Y_n^t = \mathbf{1}_{\mathbf{s}=t}$ for $t \notin \Omega$. For $n > L'$, set $Y_n^t = 0$. Let $\overline{\theta} = \inf\{n \in \mathbf{M}^* : Z_n = 1\}$. By construction and (32),

$$\left|\mathbf{P}_{\overline{s},x_\lambda,\tau^h}(e_\Omega - 1 \in \mathbf{M}^*, \mathbf{s}_{e_\Omega} = t) - \widehat{\mathbf{P}}_{\overline{s},\tau^h}(Y_{\overline{\theta}}^t = 1)\right| \le \varepsilon\lambda L'. \tag{33}$$

Define the final state $\widehat{\mathbf{s}}_{L'+1}$ by

$$\widehat{\mathbf{s}}_{L'+1} = \left\{ \begin{array}{c} t \text{ if } Y_{\overline{\theta}}^t = 1 \\ \mathbf{s}_{L'+1} \text{ otherwise} \end{array} \right. .$$

Thus, it is the true final state if the auxiliary random device never chooses to exit from Ω, and the exit state otherwise. By (33),

$$\left|\mathbf{E}_{\overline{s},x_\lambda,\tau^h}\left[v_\lambda(\mathbf{s}_{L'+1})\right] - \widehat{\mathbf{E}}_{\overline{s},\tau^h}\left[v_\lambda(\widehat{\mathbf{s}}_{L'+1})\right]\right| \le \varepsilon\,|S|\,\lambda L' + \lambda^\omega. \tag{34}$$

We shall use a decomposition formula for $\widehat{\mathbf{E}}_{\overline{s},\tau^h}\left[v_\lambda(\widehat{\mathbf{s}}_{L'+1})\right]$ that is similar to Lemma 15. Given a sequence $\widetilde{h} = (\widetilde{s}_1, ..., \widetilde{s}_{L'}, \widetilde{s}_{L'+1})$, we let $\widehat{\mathbf{P}}^{\widetilde{h}}_{\overline{s},\tau^h}$ be the law of the process that evolves as in section 6.1: at stage l, the process moves to $\widetilde{s}_{l+1}$ and the auxiliary device chooses a state according to $q(\cdot|\widetilde{s}_l, x_\lambda, \overline{y}^h)$. It is straightforward to check that

$$\widehat{\mathbf{E}}_{\overline{s},\tau^h}\left[v_\lambda(\widehat{\mathbf{s}}_{L'+1})\right] = \sum_{\widetilde{h}} \widehat{\mathbf{P}}_{\overline{s},\tau^h}(\widetilde{h}) \widehat{\mathbf{E}}^{\widetilde{h}}_{\overline{s},\tau^h}\left[v_\lambda(\widehat{\mathbf{s}}_{L'+1})\right]. \tag{35}$$

Let T_1 be the event consisting of histories such that

$$\left|\mathbf{N}^*_s - N^h_s\right| \le \beta N^h_s \text{ for each } s \in \Omega^h. \tag{36}$$

By Theorem 17, and since $\frac{1}{L'^\kappa} < \varepsilon\lambda L' < \beta$, $\widehat{\mathbf{P}}_{\overline{s},x_\lambda,\tau^h}(T_1) \ge 1 - \varepsilon\lambda L'$.

Let $\widetilde{h} \in T_1$. Plainly,

$$\widetilde{\omega}(\widetilde{h}) := \widehat{\mathbf{P}}^{\widetilde{h}}_{\overline{s},\tau^h}(\overline{\theta} \le L') = 1 - \prod_{l\in\mathbf{M}^*} q(\Omega|\widetilde{s}_l, x_\lambda, \overline{y}) \tag{37}$$

and, for $t \notin \Omega$,

$$\widehat{\mathbf{P}}^{\widetilde{h}}_{\overline{s},\tau^h}(Y^t_{\overline{\theta}} = 1) = 1 - \sum_{l\in\mathbf{M}^*} q(t|\widetilde{s}_l, x_\lambda, \overline{y}) \prod_{p\in\mathbf{M}^*, p<l} q(\Omega|\widetilde{s}_p, x_\lambda, \overline{y}).$$

Since $\widetilde{h} \in T_1$, one has by (36)

$$\widetilde{\omega}(\widetilde{h}) \le \sum_{s\in\Omega^h} \mathbf{N}^*_s q(\overline{\Omega}|s, x_\lambda, \overline{y}) \le (1+\beta) \sum_{s\in\Omega^h} N^h_s q(\overline{\Omega}|s, x_\lambda, \overline{y}) \le C(1+\beta)^2 \omega(h) \le C\beta(1+\beta)^2.$$

By Lemma 18 in RSVa,

$$\left|\widetilde{\omega}(\widetilde{h}) - \sum_{s\in\Omega^h} \mathbf{N}^*_s q(\overline{\Omega}|s, x_\lambda, \overline{y})\right| \le C\widetilde{\omega}(\widetilde{h})^2.$$

By the triangle inequality,

$$\begin{aligned}\left|\widetilde{\omega}(\widetilde{h}) - \sum_{s\in\Omega^h} N^h_s q(\overline{\Omega}|s, x_\lambda, \overline{y})\right| &\le C\widetilde{\omega}(\widetilde{h})^2 + \beta C\omega(h)(1+\beta) \\ &\le \beta C\omega(h)\left(1 + \beta + (1+\beta)^2\right).\end{aligned} \tag{38}$$

We next use the following variant of Lemma 16.

Lemma 23 *Assume $\widetilde{h} \in T_{\beta,\delta_2}$ and $\widetilde{\omega}(\widetilde{h}) < \beta$. Then*

$$\begin{aligned}&\left|\widehat{\mathbf{E}}^{\widetilde{h}}_{\overline{s},\tau^h}\left[v_\lambda(\widehat{\mathbf{s}}_{L'+1})\right] - \left(v_\lambda(\overline{s}) + \sum_{s\in\Omega^h} \mathbf{N}^*_s\left(\mathbf{E}_{q(\cdot|s,x_\lambda,\overline{y}^h)}\left[v_\lambda\right] - v_\lambda(s)\right)\right)\right| \\ &\qquad \le \sum_{s\in\Omega} \widetilde{\mathbf{N}}_s q(\overline{\Omega}|s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}}) + 2\varepsilon\lambda L' + \beta C|S|\widetilde{\omega}(\widetilde{h}),\end{aligned}$$

where $\overline{\mathbf{y}}^{\widetilde{h}}_s$ is the empirical distribution of player 2's moves in the stages of $\widetilde{\mathbf{M}}_s$.

Proof. We establish only the lower bound on $\widehat{\mathbf{E}}^{\widetilde{h}}_{\overline{s},\tau^h}[v_\lambda(\widehat{\mathbf{s}}_{L'+1})]$, since all arguments are symmetric. The proof follows closely the proof of Lemma 16. We limit ourselves to pointing out the main differences. One minor difference is that b_j is replaced by $\overline{y}$. This saves all the summations that involve $b \in B$. In the first summations, $q(\overline{\Omega}|s_j, x_\lambda, \overline{y}^h)$ is replaced by 0 whenever $j \notin \mathbf{M}^*$. The first important difference arises in the definition of $\widetilde{q}(\cdot|s)$. For $s \in \Omega^h$, we adapt the previous definition to

$$\widetilde{q}(t \mid s) = \begin{cases} q(t \mid s, x_\lambda, \overline{y}) \text{ for } t \notin \Omega \\ q(\Omega \mid s, x_\lambda, \overline{y}) q^h(t \mid s) \text{ for } t \in \Omega \end{cases}.$$

For the visits to s corresponding to the stages in $\widetilde{\mathbf{M}}_s$, we set $\widetilde{q}(\cdot|s) = q^h(\cdot|s)$. This modification affects the estimate on Δ_s (after inequality (26)), which now becomes

$$\begin{aligned}\Delta_s &= \widetilde{\mathbf{N}}_s \sum_{t\in S} (\widetilde{q}(t \mid s) - q(t \mid s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}})) v_\lambda(t) \\ &= \widetilde{\mathbf{N}}_s \sum_{t\in S} \left(\widetilde{q}(t \mid s) - q(t \mid s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}})\right) (v_\lambda(t) - v_\lambda(s)) \\ &\geq -\left|S\right| L'^{\delta_1} \lambda^\omega - \widetilde{\mathbf{N}}_s q(\overline{\Omega}|s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}}) - 6\beta \widetilde{\mathbf{N}}_s \sum_{t\in\Omega_0} \widetilde{q}(t \mid s)(v_\lambda(t) - v_\lambda(s))\end{aligned}$$

The remaining part is identical. ■

We now conclude the proof of Proposition 22. For $\widetilde{h} \in T_1 \cap T_{\beta,\delta_2}$, one has, by Lemma 23 and (36),

$$\begin{aligned}&\left| \widehat{\mathbf{E}}^{\widetilde{h}}_{\overline{s},\tau^h}[v_\lambda(\widehat{\mathbf{s}}_{L'+1})] - \left(v_\lambda(\overline{s}) + \sum_{s\in\Omega^h} N^h_s \left(\mathbf{E}_{q(\cdot|s,x_\lambda,\overline{y})}[v_\lambda] - v_\lambda(s)\right)\right)\right| \\ &\qquad \leq 2\varepsilon\lambda L' + \beta C|S|\widetilde{\omega}(\widetilde{h}) + \beta\lambda L' + \sum_{s\in\Omega} \widetilde{\mathbf{N}}_s q(\overline{\Omega}|s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}}) \\ &\qquad \leq 3\varepsilon\lambda L' + 5\beta C|S|\omega(h) + \sum_{s\in\Omega} \widetilde{\mathbf{N}}_s q(\overline{\Omega}|s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}}). \qquad (39)\end{aligned}$$

We next apply Lemma 16. For $s \notin \Omega^h$, one has $N^h_s \left(\mathbf{E}_{q(\cdot|s,x_\lambda,\overline{y})}[v_\lambda] - v_\lambda(s)\right) \geq -(1+\varepsilon)\lambda N^h_s$. Therefore, by (39),

$$\mathbf{E}^h[v_\lambda(\mathbf{s}_{L'+1})] \geq \widehat{\mathbf{E}}^{\widetilde{h}}_{\overline{s},\tau^h}[v_\lambda(\widehat{\mathbf{s}}_{L'+1})] - 5\varepsilon\lambda L' - 6\beta C|S|\omega(h) - \sum_{s\in\Omega} \widetilde{\mathbf{N}}_s q(\overline{\Omega}|s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}}). \qquad (40)$$

Since $\widetilde{\omega}(\widetilde{h}) \leq 2C\beta$ for each $\widetilde{h} \in T_1$, and by Lemma 21, one has for each set T of histories,

$$\mathbf{P}_{\overline{s},x_\lambda,\tau^h}(T_1 \cap T) \geq (1 - 2C\beta - \varepsilon\lambda L')\widetilde{\mathbf{P}}_{\overline{s},\tau^h}(T_1 \cap T). \qquad (41)$$

Choosing for T the set $\overline{T}_{\beta,\delta_2}$ (of non-typical histories) and using $\mathbf{P}_{\overline{s},x_\lambda,\tau^h}(\overline{T}_{\beta,\delta_2}) \leq \varepsilon\lambda L'$, one gets $\widetilde{\mathbf{P}}_{\overline{s},\tau^h}(T_1 \cap \overline{T}_{\beta,\delta_2}) \leq 2\varepsilon\lambda L'$. Since $\widetilde{\mathbf{P}}_{\overline{s},\tau^h}(T_1) \geq 1 - \varepsilon\lambda L'$, this yields $\widetilde{\mathbf{P}}_{\overline{s},\tau^h}(T_1 \cap T_{\beta,\delta_2}) \geq 1 - 3\varepsilon\lambda L'$.

We shall proceed below to the summation of (40) over $\widetilde{h}$. It is convenient to enrich further the probability space and to add a random device that, for each stage $n \notin \mathbf{M}^*$, takes the value 1 with probability $q(\overline{\Omega}|s_n, x_\lambda, b_n)$. Let $\widehat{\omega}(\widetilde{h})$ the $\widehat{\mathbf{P}}^{\widetilde{h}}_{\overline{s},\tau^h}$-probability that this device takes the value 1 before stage $\overline{\theta}$. Thus, $\sum_{\widetilde{h}} \widetilde{\mathbf{P}}_{\overline{s},\tau^h}(\widetilde{h})\widehat{\omega}(\widetilde{h}) = \mathbf{P}_{\overline{s},x_\lambda,\tau^h}(e_\Omega - 1 \notin \mathbf{M}^*) \leq \varepsilon\lambda L'$. In particular,

$$\widetilde{\mathbf{P}}_{\overline{s},\tau^h}(\widetilde{h} : \widehat{\omega}(\widetilde{h}) > \frac{\eta_0}{2}) \leq \frac{2}{\eta_0}\varepsilon\lambda L'. \tag{42}$$

Next, observe that for each history $\widetilde{h}$ such that $\widetilde{\omega}(\widetilde{h}) \leq 2\beta$ and $\widehat{\omega}(\widetilde{h}) \leq \frac{\eta_0}{2}$, one has $\sum_{s\in\Omega} \widetilde{\mathbf{N}}_s q(\overline{\Omega}|s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}}) \leq 2\widehat{\omega}(\widetilde{h})$, by the choice of η_0 and Lemma 19 in RSVa. Since $\widetilde{\omega}(\widetilde{h}) \leq 2\beta$ on $T_1 \cap T_{\beta,\delta_2}$, by (41) and (42), one obtains

$$\widehat{\mathbf{E}}_{\overline{s},\tau^h}\left[\sum_{s\in\Omega} \widetilde{\mathbf{N}}_s q(\overline{\Omega}|s, x_\lambda, \overline{\mathbf{y}}^{\widetilde{h}})\right] \leq (3 + \frac{2}{\eta_0})\varepsilon\lambda L' + 2\widehat{\mathbf{E}}_{\overline{s},\tau^h}\left[\widehat{\omega}(\widetilde{h})\right] \leq (5 + \frac{2}{\eta_0})\varepsilon\lambda L'.$$

The result follows. ■

Lemma 24 *On $T_{\beta,\delta_2} \cap \{\omega(h) < \beta\}$ one has*

$$\left|\mathbf{E}_{\overline{s},x_\lambda,\tau^h}\left[\frac{1}{L'}\sum_{n=1}^{L'} \widetilde{r}_\lambda(\mathbf{s}_n, x_\lambda, \mathbf{y}_n)\right] - \frac{1}{L'}\sum_{s\in\Omega} N^h_s \widetilde{r}_\lambda(s, x_\lambda, \overline{y})\right| \leq 3\omega(h) + 3\beta + \varepsilon\lambda L'.$$

Proof. Using the notations of the proof of Lemma 21, one has

$$\left|\sum_{n=1}^{L'} \widetilde{r}_\lambda(\mathbf{s}_n, x_\lambda,\mathbf{y}_n) - \sum_{s\in\Omega} N^h_s \widetilde{r}_\lambda(s, x_\lambda, \overline{y})\right| \leq 3\beta L' \text{ on } T_1. \tag{43}$$

Since

$$\left|\widetilde{\mathbf{E}}_{\overline{s},\tau^h}\left[\sum_{n=1}^{L'} \widetilde{r}_\lambda(\mathbf{s}_n, x_\lambda,\mathbf{y}_n)\right] - \mathbf{E}_{\overline{s},x,\tau^h}\left[\sum_{n=1}^{L'} \widetilde{r}_\lambda(\mathbf{s}_n, x_\lambda,\mathbf{y}_n)\right]\right| \leq L'\mathbf{P}_{\overline{s},x_\lambda,\tau^h}(e_\Omega \leq L' + 1)$$

and since $\widetilde{\mathbf{P}}_{\overline{s},x_\lambda,\tau^h}(T_1) \geq 1 - \varepsilon\lambda L'$, the result follows from (43), using Lemma 21. ■

6.4 Proof of C2

The following is a corollary of Proposition 19 and the $\varepsilon\lambda$-optimality of x_λ.

Corollary 25 *For each $h \in T_{\beta,\delta_2}$ such that $\omega(h) < \beta$ one has*

$$\mathbf{E}^h\left[\lambda L'\widehat{\rho} + (1 - \lambda L')v_\lambda(\mathbf{s}_{L'+1})\right] \geq v_\lambda(\overline{s}) - \left(\frac{\overline{\varepsilon}}{4} + \varepsilon\right)\lambda L' - C\beta\omega(h).$$

Define

$$\theta_1 = \inf\left\{n \geq 1, \mathbf{h}_n \notin T^n_{\beta,\delta_2}\right\}, \theta_2 = L' + 1, \theta_3 = \inf\left\{n \geq 1, \omega(\mathbf{h}_n) \geq \beta\right\},$$

and $\theta = \min\{\theta_1, \theta_2, \theta_3, e_\Omega\}$. Set $Z_1 = \lambda\theta\widetilde{\rho} + (1 - \lambda\theta)v_\lambda(\mathbf{s}_{L'+1})$, where $\widetilde{\rho} = \widehat{\rho}$ if $\theta = \theta_2$, and $\widetilde{\rho} = 1$ otherwise. Recall that $H(\theta)$ is the set of atoms of $\mathcal{H}_\theta$ such that $\theta < e_\Omega$.

Lemma 26 *For each $h = (s_1, b_1, ..., s_{k+1}) \in H(\theta)$, one has*

$$\mathbf{E}^h [Z_1] \geq v_\lambda(\overline{s}) - 2\beta C \omega(h) - \left(\varepsilon + \frac{\overline{\varepsilon}}{4}\right) \lambda L' - 2(1-\omega(h))\mathbf{1}_{h \notin T^{k+1}_{\beta,\delta_2}}. \tag{44}$$

Proof. Case 1: h is such that $\theta = \theta_1 < \theta_3$. In that case, $h \notin T^{k+1}_{\beta,\delta_2}$, and $1-\omega(h) \geq 1-\beta \geq 1/2$. Therefore, $2(1-\omega(h))\mathbf{1}_{h \notin T^{k+1}_{\beta,\delta_2}} \geq 1$, and (44) holds since $Z_1 \geq 0$ and $v_\lambda(\overline{s}) \leq 1$.

Case 2: h is such that $\theta = \theta_2 < \min\{\theta_1, \theta_3\}$. Thus, $h \in T^{L'+1}_{\beta,\delta_2} = T_{\beta,\delta_2}$ and $\omega(h) < \beta$. The claim then follows by Corollary 25.

Case 3: h is such that $\theta = \theta_3$.

Let $h_k = (s_1, b_1, \ldots, s_k)$ the restriction of h to the first k stages. By definition of θ_3, one has $\omega(h_k) < \beta$ and $h_k \in T^k_{\beta,\delta_2}$. By Lemma 16, one has

$$\mathbf{E}^{h_k} [v_\lambda(\mathbf{s}_k)] \geq v_\lambda(\overline{s}) + \sum_{s \in \Omega} N^{h_k}_s \left(\mathbf{E}_{q(\cdot|s,x_\lambda,\overline{y})} [v_\lambda] - v_\lambda(s)\right) - 2\varepsilon\lambda L' - \beta C \omega(h).$$

By the $\varepsilon\lambda$-optimality of x_λ (see (4)),

$$\mathbf{E}_{q(\cdot|s,x_\lambda,b)} [v_\lambda] - v_\lambda(s) \geq -\lambda(1+\varepsilon), \text{ for every } s \in \Omega,$$

so that

$$\mathbf{E}^{h_k} [v_\lambda(\mathbf{s}_k)] \geq v_\lambda(\overline{s}) - (1+3\varepsilon)\lambda L' - \beta C|S|\omega(h_k). \tag{45}$$

We now compare $\mathbf{E}^h [v_\lambda(\mathbf{s}_{k+1})]$ to $\mathbf{E}^{h_k} [v_\lambda(\mathbf{s}_k)]$. Plainly,

$$\begin{aligned}\mathbf{E}^h [v_\lambda(\mathbf{s}_{k+1})] &= \mathbf{E}^{h_k} [v_\lambda(\mathbf{s}_k)] + (1-\omega(h_k))\left(\mathbf{E}_{q(\cdot|s_k,x_\lambda,b_k)} [v_\lambda] - v_\lambda(s_k)\right) \\ &\geq \mathbf{E}^{h_k} [v_\lambda(s_k)] - \lambda(1+\varepsilon). \end{aligned} \tag{46}$$

Finally, note that $\mathbf{E}^h [Z_1] \geq \mathbf{E}^h [v_\lambda(\mathbf{s}_{L'+1})] - \lambda L'$, hence by (45) and (46),

$$\mathbf{E}^h [Z_1] \geq v_\lambda(\overline{s}) - 2\lambda L' - \beta C|S|\omega(h) \geq v_\lambda(\overline{s}) - 2\beta C\omega(h),$$

where the second inequality obtains since $\omega(h) \geq \beta$ and (A.iv). ■

For notational ease, set $Y_j := \lambda L' \widehat{\rho}_j + (1-\lambda L')v_\lambda(\mathbf{s}_{(j+1)L'+1})$.

Corollary 27 *One has*

$$\mathbf{E}_{\overline{s},x_\lambda,\tau} [Y_1] \geq v_\lambda(\overline{s}) - \left(5\varepsilon + \frac{\overline{\varepsilon}}{4}\right) \lambda L' - 2\beta(C+1)\mathbf{P}_{\overline{s},x_\lambda,\tau}(e_\Omega \leq L'+1). \tag{47}$$

Proof. By summing over h, one obtains, by Lemmas 15 and 26 and Proposition 11,

$$\begin{aligned}\mathbf{E}_{\overline{s},x_\lambda,\tau} [Z_1] &\geq v_\lambda(\overline{s}) - 2\beta C\mathbf{P}_{\overline{s},x_\lambda,\tau}(e_\Omega \leq L'+1) - \left(\varepsilon + \frac{\overline{\varepsilon}}{4}\right) \lambda L' \\ &\quad - 2\mathbf{P}_{\overline{s},x_\lambda,\tau}(\theta_1 \leq L'+1) \\ &\geq v_\lambda(\overline{s}) - 2\beta C\mathbf{P}_{\overline{s},x_\lambda,\tau}(e_\Omega \leq L'+1) - \left(3\varepsilon + \frac{\overline{\varepsilon}}{4}\right) \lambda L'. \end{aligned} \tag{48}$$

We now compare $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Z_1\right]$ and $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_1\right]$, by discussing according to θ.

Since $\widehat{\rho}=1$ whenever $\theta=e_\Omega$, by using the $\varepsilon\lambda$-optimalilty of x_λ, one has

$$\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_1 \mid \mathcal{H}_\theta\right] \geq \lambda\theta\widehat{\rho}+(1-\lambda\theta)v_\lambda(s_\theta)-\varepsilon\lambda L' \text{ on the event } \theta=e_\Omega.$$

By Proposition 11, the probability of $\theta=\theta_1$ is at most $\varepsilon\lambda L'$, and $Z_1=Y_1$ if $\theta=\theta_2$. It remains to compare $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_1 1_E\right]$ and $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Z_1 1_E\right]$, where $E=\{\theta=\theta_3<\min(\theta_1,\theta_2,e_\Omega)\}$. Since $\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_1|\mathcal{H}_\theta\right] \geq \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Z_1|\mathcal{H}_\theta\right]-\lambda(1+\varepsilon)L'$ on E, one obtains

$$\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_1 1_E\right] \geq \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Z_1 1_E\right]-2\lambda L'\mathbf{P}_{\overline{s},x_\lambda,\tau}(E).$$

On the other hand,

$$\begin{aligned}\mathbf{P}_{\overline{s},x_\lambda,\tau}(E) &\leq \sum_{h\in H(\theta),\theta=\theta_3} p(h) \leq \sum_{h\in H(\theta),\theta=\theta_3} \frac{p(h)}{1-\omega(h)}\\ &\leq \frac{1}{\beta}\sum_{h\in H(\theta),\theta=\theta_3} \frac{p(h)}{1-\omega(h)}\omega(h) \leq \frac{1}{\beta}\sum_{h\in H(\theta)} \frac{p(h)}{1-\omega(h)}\omega(h)\\ &\leq \frac{1}{\beta}\mathbf{P}_{\overline{s},x_\lambda,\tau}(e_\Omega \leq L'+1).\end{aligned}$$

By (A.iv) $\lambda L' \leq \beta^2$, hence

$$\begin{aligned}\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_1 1_E\right] &\geq \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Z_1 1_E\right]-\frac{2\lambda L'}{\beta}\mathbf{P}_{\overline{s},x_\lambda,\tau}(e_\Omega \leq L'+1)\\ &\geq \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Z_1 1_E\right]-2\beta\mathbf{P}_{\overline{s},x_\lambda,\tau}(e_\Omega \leq L'+1).\end{aligned}$$

It follows that

$$\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_1\right] \geq \mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Z_1\right]-2\beta\mathbf{P}_{\overline{s},x_\lambda,\tau}(e_\Omega \leq L'+1)-2\varepsilon\lambda L'.$$

The result follows, using (48). ■

Set $\chi_n=1$ if $\mathbf{s}_n$ and $\mathbf{s}_{n+1}$ belong to different communicating sets, and $\chi_n=0$ otherwise. Inequality (47) extends to

$$\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[Y_{j+1} \mid \mathcal{H}^1_{jL'+1}\right] \geq v_\lambda(\mathbf{s}_j)-\left(5\varepsilon+\frac{\overline{\varepsilon}}{4}\right)\lambda L'-2\beta(C+1)\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\sum_{n=jL'+1}^{(j+1)L'}\chi_n \mid \mathcal{H}^1_{jL'+1}\right].$$

By taking expectations and summing over k, and since $\beta=\varepsilon\lambda\overline{L}$, this yields

$$\mathbf{E}_{\overline{s},x_\lambda,\tau}\left[\lambda L\widehat{r}+(1-\lambda L)v_\lambda(\mathbf{s}_{L+1})\right] \geq v_\lambda(\overline{s})-\left(5\varepsilon+\frac{\overline{\varepsilon}}{4}\right)\lambda L-2\varepsilon M_\alpha(C+1)\lambda\overline{L},$$

and **C2** follows by the choice of ε.

A Proof of Theorem 6

A.1 Markov chains with rare transitions

We recall few basic facts on Markov chains with rare transitions. For more details, we refer to Catoni (1999). These tools have already been of use in the analysis of stochastic games, see for instance Vieille (2000) and Solan and Vieille (2002).

Let (S, v) be a irreducible Markov chain with semi-algebraic transitions; that is, $v = (v_\lambda)_{\lambda>0}$ is a family of irreducible transition functions over S such that for every $s, t \in S$, the function $\lambda \mapsto v_\lambda(t \mid s)$ is a semi-algebraic function of λ.

Denote by $\overline{e}_C := e_{\overline{C}}$ the first hitting time of C. For each $\lambda > 0$ we denote by $\mathbf{Q}_s^{v_\lambda}(\cdot \mid C)$ the distribution of s_{e_C} (the exit law) and by $X_s^{v_\lambda}(C) = \mathbf{E}_s\left[e_C\right]$ the expected exit time. Observe that for fixed C and s, the functions $Q_s^v : \lambda \mapsto \mathbf{Q}_s^{v_\lambda}(\cdot \mid C)$ and $X_s^v : \lambda \mapsto X_s^{v_\lambda}(C)$ are semi-algebraic.

For every semi-algebraic function f we denote by d_f its degree.

An important notion is that of a cycle. The definition we provide below differs from the standard one (see Catoni (1999, Definition 6), or Vieille (2000, Lemma 6)), but is easily seen to be equivalent.

Definition 28 *A subset C of S is a* cycle *if (i) $d_{X_s^v(C)}$ is independent of $s \in C$ (it is denoted by $d_{X^v(C)}$), and (ii) $-d_{X_s^v(C)} > -d_{X_s^v(D)}$ for every proper subset D of C and every $s \in D$.*

Since (s, v) is irreducible, S is a cycle. By convention, each singleton is a cycle. The set of all cycles w.r.t. (S, v) is denoted by $\mathcal{C}(v)$.

If C is a cycle, then $\lim_{\lambda\to0} \mathbf{Q}_s^{v_\lambda}(\cdot \mid C)$ is independent of $s \in C$ (see Catoni (1999, Proposition 9), or Vieille (2000, Lemma 5)). The limit is denoted by $\mathbf{Q}^v(\cdot \mid C)$. The *boundary* of a cycle C is the set $B(C) = \operatorname{supp}(\mathbf{Q}^v(\cdot \mid C))$, and is disjoint of C.

The set of cycles, ordered by inclusion, has a tree structure: if two cycles intersect, then one of them is a subset of the other (see Catoni (1999, Proposition 7) or Vieille (2002)).

The set of all cycles is denoted by $\mathcal{C}(v)$, and, to emphasize their relation to v, we call them *v-cycles*. Finally, we denote $\mathcal{C}^*(v) = \{C \in \mathcal{C}(v) \mid -d_{X^v(C)} \geq 1\}$.

A.2 Maximal communicating sets

Every stationary strategy y of player 2 defines naturally an irreducible Markov chain (S, y) with semi-algebraic transitions:

$$y_\lambda(t \mid s) = (1 - \lambda^2)q(t \mid s, x_\lambda^s, y^s) + \lambda^2/|S|.$$

We identify each stationary strategy with the corresponding Markov chain.

Since the number of *pure stationary* strategies is finite, there exists $\alpha_0 < 1$ such that for every pure stationary strategy y and every y-cycle C, $-d_{X^y(C)} < \alpha_0$ or $-d_{X^y(C)} \geq 1$.

Let $\alpha_2 \in (0, 1 - \alpha_0)$ be sufficiently small so that $\alpha_2 < d_{X_s^y(C)} - d_{X^y(C)}$ for every pure stationary strategy y, every y-cycle C, and every $s \in C$.

Let $\alpha_1 \in (1 - \alpha_2, 1)$ be arbitrary.

Let G be the non-directed graph with vertex set S, for which (s, s') is an edge of G if and only if there exists a pure stationary strategy y and a y-cycle C such that (i) $s, s' \in C$, and (ii) $-d_{X^y(C)} < \alpha$.

Definition 29 *The* maximal communicating sets *(MC sets for short) are the connected components of G. The class of all MC sets is $\mathcal{MC}$.*

The following lemma states that there is a pure stationary strategy y such that for every MC-set C and every proper subset D of C, the process reaches D in much fewer than $1/\lambda$ stages. It proves the first assertion of Theorem 6.

Lemma 30 *Let $C \in \mathcal{MC}$, and let D be a proper subset of C. There exists pure stationary strategy y, such that for every $s \in C$ and every $\lambda > 0$ sufficiently small,*

1. $\mathbf{E}_{s,x_\lambda,y}[e_{\overline{C}\cup D}] < 1/\lambda^{\alpha_1}$*, and*
2. $\mathbf{P}_{s,x_\lambda,y}\left[e_{\overline{C}} < e_D\right] < \lambda^{\alpha_2}$*.*

Proof. Let $C_1, \ldots, C_p = D$ be a collection of subsets of C such that (i) For $i < p$, $C_i \in \mathcal{C}(y_i)$, for some pure stationary strategy y_i, (ii) For every $i < p$ there is j, $i < j \leq p$, such that $C_i \cap C_j \neq \emptyset$, (iii) $\cup_{i=1}^p C_p = C$. Define y as follows. (a) If $s \in D$, y^s is defined arbitrarily. (b) For $s \in C \setminus D \subseteq \cup_{i=1}^{p-1} C_i$, let j_s be the maximal index such that $s \in C_{j_s}$. Then $y^s = y^s_{j_i}$.

By definition, and since $\alpha_1 > \alpha \geq d_{X(C_i)}$, $i = 1, \ldots, p-1$, $\mathbf{E}_{s,x_\lambda,y}[e_{\overline{C}} \cup (\cup_{i=j_s+1}^p C_j)] \leq 1/|S|\lambda^{\alpha_1}$ for every λ sufficiently small. Claim (1) follows.

By Aldous and Fill (2002, Corollary II.10), and by the choice of α_2,

$$\mathbf{P}_{s,x_\lambda,y}(e_{\overline{C}} < e_{C_{j_s}}) \leq \lambda^{\alpha_2}/|S|.$$

Claim (2) follows. ■

A.3 Number of visits to MC sets

In this section we prove the second assertion of Theorem 6.

We set $C^* = \cap_y \mathcal{C}^*(y)$, where the intersection is over all pure stationary strategies y. C^* contains all subsets C such that, whatever player 2 plays, the expected time to leave C is at least $O(1/\lambda)$.

For every $L \in \mathbf{N}$ denote by

$$W(L) = |\{n < L \mid s_n \in C, s_{n+1} \notin C \text{ for some } MC\text{-set } C\}|$$

the number of exits from MC-sets.

The following two lemmas are proven for pure stationary strategies y of player 2. It is not difficult to deduce that they hold for every strategy τ of player 2 as well.

Lemma 31 *There exist a constant M and $\lambda_0 > 0$ such that*

$$\mathbf{E}_{s,x_\lambda,y}\left[W(\min_{C\in\mathcal{C}^*(y)} \overline{e}_C)\right] \leq M,$$

for every pure stationary strategy y, every initial state s, and every $\lambda \in (0, \lambda_0)$.

Proof. Fix a pure stationary strategy y of player 2. We introduce a directed graph G with vertex set $\mathcal{C}(y)$, and with an edge $C \to C'$ if and only if $C \notin \mathcal{C}^*(y)$ and $B(C) \cap C' \neq \emptyset$ (recall that $B(C)$ is the principal boundary of C). By definition of $\mathcal{C}(y)$, this graph has the following property: given $C \notin \mathcal{C}^*(y)$, there is a path joining C to $\mathcal{C}^*(y)$. Therefore, the probability that $W(\min_{C\in\mathcal{C}^*(y)} \overline{e}_C) \leq |S|$ is strictly positive, for each $s \in S$. The result follows. ■

Lemma 32 *There exists a constant $M > 0$ such that for every $\alpha \in (1 - \alpha_2, 1)$, every initial state s, every pure stationary strategy y, and every $\lambda > 0$ sufficiently small,*

$$\mathbf{E}_{s,x_\lambda,y}[\#\{n \leq 1/\lambda^\alpha \mid \mathbf{s}_n, \mathbf{s}_{n+1} \text{ belong to different MC-sets, and } \mathbf{s}_n \in C^*] \leq M\lambda^{1-\alpha}.$$

Proof. Fix a pure stationary strategy y and a set $C \in \mathcal{C}^*(y)$. To prove the claim, it is sufficient to show that for some constant $M > 0$,

$$\mathbf{P}_{s,x_\lambda,y}(e_C < 1/\lambda^\alpha) \leq M\lambda^{1-\alpha}, \tag{49}$$

for every λ sufficiently small and every $s \in C$. Indeed, (49) implies that the expected number of exits from C until stage $1/\lambda^\alpha$ is bounded by $3\lambda^{1-\alpha}/(1-\lambda^{1-\alpha})^2$.

Let $q > 0$, and assume that there is $s \in C$ such that $\mathbf{P}_{s,x_\lambda,y}(e_C < 1/\lambda^\alpha) \geq q$. We show that $q \leq M\lambda^{1-\alpha}$, for some constant M. Since $C \in \mathcal{C}^*(y)$, $\mathbf{E}_{t,x_\lambda,y}[e_{C\setminus\{s\}}] < 1/\lambda^{1-\alpha_2}$, hence by Markov inequality $\mathbf{P}_{t,x_\lambda,y}(e_{C\setminus\{s\}} \geq 1/\lambda^\alpha) \leq \lambda^{\alpha-(1-\alpha_2)}$ for every $t \in C$. In particular, $\mathbf{P}_{t,x_\lambda,y}(e_C < 2/\lambda^\alpha) \geq q(1-\lambda^{\alpha-(1-\alpha_2)})$, so that

$$a/\lambda \leq \mathbf{E}_{t,x_\lambda,y}[e_C] \leq \frac{2}{\lambda^\alpha q(1-\lambda^{\alpha-(1-\alpha_2)})},$$

where M is determined by the leading non-zero coefficient of the expansion of $d_{X^y(C)}$. It follows that $q \leq 3a\lambda^{1-\alpha_1}$, as desired. ■

B Reminder on zero-sum games

The purpose of this section is to provide a slight modification of a result due to Mertens and Neyman (1981, hereafter MN), that we will use. We let $\lambda \longmapsto w_\lambda$ be a $\mathbf{R}^S$-valued semi-algebraic function, and $w = \lim_{\lambda\to 0} w_\lambda$.

Let $\varepsilon > 0$, $Z \geq 0$ and two functions $\lambda : (0,+\infty) \to (0,1)$ and $L : (0,+\infty) \to \mathbf{N}$ be given. Set $\delta = \varepsilon/48$. Assume that the following conditions are satisfied for every $z \geq Z$, every $|\eta| \leq 4$ and every $s \in S$:

C1 $|w_\lambda(s) - w(s)| \leq 4\delta$;

C2 $4L(z) \leq \delta z$;

C3 $|\lambda(z + \eta L(z)) - \lambda(z)| \leq \delta\lambda(z)$

C4 $\left|w_{\lambda(z+\eta L(z))}(s) - w_{\lambda(z)}(s)\right| \leq 4\delta L(z)\lambda(z)$

C5 $\int_Z^\infty \lambda(z)dz \leq 4\delta$.

Mertens and Neyman (1981) note that **C1-C5** hold for Z large enough, in each of the next two cases:

Case 1 $\lambda(z) = z^{-\beta}$ and $L(z) = \lceil \lambda(z)^{-\alpha} \rceil$,[6] where $\alpha \in (0,1)$ and $\beta > 1$ satisfy $\alpha\beta < 1$;

[6] For every $c \in \mathbf{R}$, $\lceil c \rceil$ is the minimal integer greater than or equal to c.

Case 2 $L(z) = 1$ and $\lambda(z) = 1/z(\ln z)^2$.

Let $(\widehat{r}_k)_{k\in\mathbf{N}}$ be a $[0,1]$-valued process defined on the set of plays. Define recursively processes $(z_k), (L_k)$ and (B_k) by the formulas

$$\begin{aligned}
&z_0 = Z, B_0 = 1,\\
&\lambda_k = \lambda(z_k), L_k = L(z_k), B_{k+1} = B_k + L_k,\\
&z_{k+1} = \max\left\{ Z, z_k + \lambda_k \left(L_k \widehat{r}_k - \sum_{B_k \le n < B_{k+1}} w_{\lambda_k}(s_n) \right) + \frac{\varepsilon}{2} \right\}.
\end{aligned}$$

Let (I_k) be an integer-valued process, where I_k is $\mathcal{H}^1_{B_{k-1}}$-measurable that satisfies

$$\mathbf{E}_{s,\sigma,\tau}[\sum_{j\le k} I_j] \le M + M \times \mathbf{E}_{s,\sigma,\tau} \sum_{j\le k} \lambda_j L_j, \qquad \forall s \in S, \sigma, \tau, \tag{50}$$

for some constant M. This process does not appear in Mertens and Neyman's (1981) formulation.

Theorem 33 *Let (σ, τ) be a strategy pair. Assume that for every $k \ge 0$,*

$$\mathbf{E}_{s,\sigma,\tau}\left[\lambda_k L_k \widehat{r}_k + \beta I_k + (1 - \lambda_k L_k) w_{\lambda_k}(s_{B_{k+1}}) | \mathcal{H}^1_{B_k}\right] \ge w_{\lambda_k}(s_{B_k}) - \frac{\varepsilon}{12} \lambda_k L_k, \tag{51}$$

where $\beta \in (0, \delta/4M)$. Then there exists $N_0 \in \mathbf{N}$, independent of (σ, τ), such that the following holds for every $n \ge N_0$:

$$\mathbf{E}_{s,\sigma,\tau}\left[\frac{1}{n}\sum_{p=1}^{n} \widehat{R}_n\right] \ge w(s) - 2\varepsilon, \tag{52}$$

where $\widehat{R}_n = \widehat{r}_k$ whenever $B_k \le n < B_{k+1}$. Moreover,

$$\mathbf{E}_{s,\sigma,\tau}\left[\sum_{k=1}^{\infty} \lambda_k L_k\right] < +\infty. \tag{53}$$

The result also holds when replacing in (51) and (52) $\ge$ by $\le$, and the '+' sign on the right-hand side by a '-' sign.

Proof. This statement differs from the statement in MN through the additional process I_k.

To handle the term βI_k on the left-hand side of (51), it is enough to introduce the following changes in MN. First, add βI_k in Lemmas 3.4 and 3.5 in MN. Second, define Y_k as $Y_k = l_k - t_k + \beta \sum_{j\le k} I_j$. The proof of Proposition 3.6 in MN goes as follows. By definition $Y_k - Y_0 = l_k - t_k - l_0 + t_0 + \beta \sum_{j\le k} I_j$. As payoffs are between 0 and 1, and by (50),

$$\mathbf{E}[Y_k - Y_0] \le 2 + \beta \mathbf{E}[\sum_{j\le k} I_j] \le 3 + \beta M \mathbf{E}[\sum_{j\le k} \lambda_j L_j].$$

On the other hand, by Lemma 3.5 in MN,

$$\mathbf{E}[Y_k - Y_0] \geq \delta E[\sum_{j \leq k} \lambda_j L_j].$$

Since $\delta - \beta M \geq 3\delta/4$, we get $\mathbf{E}[\sum_{j \leq k} \lambda_j L_j] \leq 4/\delta$. Letting k go to infinity,

$$\mathbf{E}[\sum_{j \leq \infty} \lambda_j L_j] \leq 4/\delta.$$

This proves (c) in Proposition 3.6 of MN. Moreover, it implies that $(\mathbf{E}[Y_k])_{k \geq 0}$ is uniformly bounded. By the martingale convergence Theorem (see, e.g., Billingsley (1995, Theorem 35.5)) the submartingale (Y_k) converges a.s. to Y_∞, and $\mathbf{E}[Y_\infty] \leq 4/\delta$. The rest of the proof is as in MN. ■

References

[1] Aldous D.J. and Fill J.A. (2002) Reversible Markov Chains and Random Walks on Graphs, Book in preparation; draft available via homepage http://www.stat.berkeley.edu/users/aldous

[2] Billingsley P. (1995) Probability and Measure, third edition, John Wiley and sons

[3] Catoni O. (1999) Simulated annealing algorithms and Markov chains with rare transitions. Séminaire de Probabilités, XXXIII, 69-119, Lecture Notes in Mathematics, 1709, Springer, Berlin

[4] Coulomb J.M. (1999) Generalized Big-Match, *Math. Oper. Res.*, **24**, 795-816

[5] Coulomb J.M. (2001) Absorbing Games with a Signaling Structure, *Math. Oper. Res.*, **26**, 286-303

[6] Cover T.M. and Thomas J.A. (1991) Elements of Information Theory, Wiley Series in Telecommunications. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York

[7] Mertens J.F. and Neyman A. (1981) Stochastic Games, *Int. J. Game Th.*, **10**, 53-66

[8] Mertens J.F., Sorin S. and Zamir S. (1994) Repeated Games, CORE Discussion Paper 9420-9422

[9] Rosenberg D., Solan E. and Vieille N. (2002a) On the Max-Min Value of Stochastic Games with Imperfect Monitoring, *preprint.*

[10] Rosenberg D., Solan E. and Vieille N. (2002b) Approximating a Sequence of Observations by a simple Process, *preprint.*

[11] Rosenberg D., Solan E. and Vieille N. (2002c) Stochastic Games with a Single Controller and Incomplete Information on One Side, *preprint.*

[12] Solan E. and Vieille N. (2002), Correlated Equilibrium in Stochastic Games, *Games Econ. Behavior*, **38**, 362-399

[13] Vieille N. (2000) Small Perturbations and Stochastic Games, *Israel J. Math.*, **119**, 127-142