The Operator Approach to Entropy Games

Marianne Akian[1] · Stéphane Gaubert[1] · Julien Grand-Clément[2] · Jérémie Guillaud[3]

Abstract

Entropy games and matrix multiplication games have been recently introduced by Asarin et al. They model the situation in which one player (Despot) wishes to minimize the growth rate of a matrix product, whereas the other player (Tribune) wishes to maximize it. We develop an operator approach to entropy games. This allows us to show that entropy games can be cast as stochastic mean payoff games in which some action spaces are simplices and payments are given by a relative entropy (Kullback-Leibler divergence). In this way, we show that entropy games with a fixed number of states belonging to Despot can be solved in polynomial time. This approach also allows us to solve these games by a policy iteration algorithm, which we compare with the spectral simplex algorithm developed by Protasov.

Keywords Stochastic games · Shapley operators · Policy iteration · Perron eigenvalues · Risk sensitive control

✉ Stéphane Gaubert
Stephane.Gaubert@inria.fr

Marianne Akian
Marianne.Akian@inria.fr

Julien Grand-Clément
jg3728@columbia.edu

Jérémie Guillaud
jeremie.guillaud@inria.fr

[1] Inria and CMAP, École polytechnique, CNRS, Palaiseau, France

[2] IEOR Department, Columbia University, New York, NY, USA

[3] Inria Paris, Paris, France

1 Introduction

1.1 Entropy Games and Matrix Multiplication Games

Entropy games have been introduced by Asarin et al. [2]. They model the situation in which two players with conflicting interests, called "Despot" and "Tribune", wish to minimize or to maximize a topological entropy representing the freedom of a half-player, "People". Entropy games are special "matrix multiplication games", in which two players alternatively choose matrices in certain prescribed sets; the first player wishes to minimize the growth rate of the infinite matrix product obtained in this way, whereas the second player wishes to maximize it. Whereas matrix multiplication games are hard in general (computing joint spectral radii is a special case), entropy games correspond to a tractable subclass of multiplication games, in which the matrix sets have the property of being invariant by row interchange, the so called independent row uncertainty (IRU) assumption, sometimes also called *row-wise* or *rectangularity* assumption. In particular, Asarin et al. showed in [2] that the problem of comparing the value of an entropy game to a given rational number is in NP ∩ coNP, giving to entropy games a status somehow comparable to other important classes of games with an unsettled complexity, including mean payoff games, simple stochastic games, or stochastic mean payoff games, see [6] for background.

Another motivation to study entropy games arises from risk sensitive control [1, 16, 17]: as we shall see, essentially the same class of operators arise in the latter setting. A recent application of entropy games to the approximation of the joint spectral radius of nonnegative matrices (without making the IRU assumption) can be found in [21]. Other motivations originate from symbolic dynamics [32, Chapter 1.8.4].

1.2 Contribution

We first show that entropy games, which were introduced as a new class of games, are equivalent to a class of zero-sum mean payoff stochastic games with perfect information, in which some action spaces are simplices, and the instantaneous payments are given by a Kullback-Leibler entropy. Hence, entropy games fit in a classical class of games, with a "nice" payment function over infinite action spaces.

To do so, we introduce a slightly more expressive variant of the model of Asarin et al. [2], in which the initial state is prescribed (the initial state is chosen by a half-player, People, in the original model). This may look like a relatively minor extension, so we keep the name "entropy game" for it, but this extension is essential to develop an operator approach and derive consequences from it. We show that the main results known for stochastic mean payoff games with finite actions space and perfect information, namely the existence of the value and the existence of optimal positional strategies, are still valid for entropy games (Theorems 10 and 9). This is derived from a model theory approach of Bolte, Gaubert, and Vigeral [9], together with the observation that the dynamic programming operators of entropy games are definable in the real exponential field. Then, a key ingredient is the proof of existence of Blackwell optimal policies, as a consequence of o-minimality, see Theorem 8. Another consequence of the operator approach is the existence of Collatz-Wielandt

optimality certificates for entropy games, Theorem 13. When specialized to the one player case, this leads to a convex programming characterization of the value, Corollary 14, which can also be recovered from a characterization of Anantharam and Borkar [1].

Our main result, Theorem 16, shows that entropy games in which Despot has a fixed number of significant states (states with a nontrivial choice) can be solved *strategically* in polynomial time, meaning that optimal (stationary) strategies can be found in polynomial time. Thus, entropy games are somehow similar to stochastic mean payoff games, for which an analogous fixed-parameter tractability result holds (by reducing the one player case to a linear program). This approach also reveals a fundamental asymmetry between the players Despot and Tribune: our approach does not lead to a polynomial bound if one fixes the number of states of Tribune. In our proof, o-minimality arguments allow a reduction from the two-player to the one-player case (Theorem 9). Then, the one-player case is dealt with using several ingredients: ellipsoid method, separation bounds between algebraic numbers, and results from Perron-Frobenius theory.

The operator approach also allows one to obtain practically efficient algorithms to solve entropy games. In this way, the classical policy iteration of Hoffman-Karp [23] can be adapted to entropy games. We report experiments showing that when specialized to one player problems, policy iteration yields a speedup by one order of magnitude by comparison with the "spectral simplex" method recently introduced by Protasov [38].

Let us finally complete the discussion of related works. The formulation of entropy games in terms of "classical" mean payoff games in which the payments are given by a Kullback-Leibler entropy builds on known principles in risk sensitive control [1, 17]. It can be thought as a version for two player problems of the Donsker-Varadhan characterization of the Perron-eigenvalue [15]. The latter is closely related to the log-convexity property of the spectral radius established by Kingman [27]. A Donsker-Varadhan type formula for risk sensitive problems, which can be applied in particular to Despot-free player entropy games, has been recently obtained by Anantharam and Borkar, in a wider setting allowing an infinite state space [1]. In a nutshell, for Despot-free problems, the Donsker-Varadhan formula appears to be the (convex-analytic) dual of the Collatz-Wielandt formula. Chen and Han [13] developed a related convex programming approach to solve the entropy maximization problem for Markov chains with uncertain parameters. We also note that the present Collatz-Wielandt approach, building on [5], yields an alternative to the approach of [2] using the "hourglass alternative" of [28] to produce concise certificates allowing one to bound the value of entropy games. By comparison with [2], a essential difference is the use of o-minimality arguments: these are needed because we study the more precise version of the game, in which the initial state is fixed. Indeed, a counter example of Vigeral shows that the mean payoff may not exist in such cases without an o-minimality assumption [46], whereas the existence of the mean payoff holds universally (without restrictions of an algebraic nature on the Shapley operator) if one allows one player to choose the initial state, see e.g. Proposition 2.12 of [3]. Finally, the identification of tractable subclasses of matrix multiplication games can be traced back at least to the work of Blondel and Nesterov [11].

2 Entropy Games

An entropy game $\mathcal{E}$ is a perfect information game played on a finite directed weighted graph G. There are 2 players, "Despot", "Tribune", and a half-player with a non-deterministic behavior, "People". The set of nodes of the graph is written as the disjoint union $D \cup T \cup P$, where D, T and P represent sets of states in which Despot, Tribune, and People play. We assume that the set of arcs E is included in $(D \times T) \cup (T \times P) \cup (P \times D)$, meaning that Despot, Tribune, and People alternate their actions. A *weight* m_{pd}, which is a positive real number, is attached to every arc $(p, d) \in P \times D$. All the other arcs in E have weight 1. An initial state, $\bar{d} \in D$, is known to the players. A token, initially in node $\bar{d}$, is moved in the graph according to the following rule. If the token is currently in a node d belonging to D, then, Despot chooses an arc $(d, t) \in E$ and moves the token to a node t. Similarly, if the token is currently in a node $t \in T$, Tribune chooses an arc $(t, p) \in E$ and moves the token to node p. Finally, if the token is in a node $p \in P$, People chooses an arc $(p, d') \in E$ and moves the token to a node $d' \in D$. We will assume that every player has at least one possible action in each state in which it is his or her turn to play. In other words, for all $d \in D$, the set of actions $\{(d, t) \in E\}$ must be nonempty, and similar conditions apply to $t \in T$ and $p \in P$.

A *history* of the game consists of a finite path in the directed graph G, starting from the initial node $\bar{d}$. The *number of turns* of this history is defined to be the length of this path, each arc counting for a length of one third. The *weight* of a history is defined to be the product of the weights of the arcs arising on this path. For instance, a history $(d_0, t_0, p_0, d_1, t_1, p_1, d_2, t_2)$ where $d_i \in D$, $t_i \in T$ and $p_i \in P$, makes 2 and 1/3 turn, and its weight is $m_{p_0 d_1} m_{p_1 d_2}$.

A *strategy* of Player Despot is a map δ which assigns to every history ending in some node d in D an arc of the form $(d, t) \in E$. Similarly, a *strategy* of Player Tribune is a map τ which assigns an arc $(t, p) \in E$ to every history ending with a node t in T. The strategy δ is said to be *positional* if it only depends on the last node d which has been visited and eventually of the number of turns. Similarly, the strategy τ is said to be *positional* if it only depends on t and eventually of the number of turns. These strategies are in addition *stationary*, if they do not depend on the number of turns.

For every integer k, we define as follows the *game in horizon k* with initial state $\bar{d}$, $\mathcal{E}_{\bar{d}}^k$. We assume that Despot and Tribune play according to the strategies δ, τ. Then, People plays in a nondeterministic way. Therefore, the pair of strategies δ, τ allows for different histories. The payment received by Tribune, in k turns, is denoted by $R_{\bar{d}}^k(\delta, \tau)$. It is defined as the sum of the weights of all the paths of the directed graph G of length k with initial node $\bar{d}$ determined by the strategies δ and τ: each of these paths corresponds to different successive choices of People, leading to different histories allowed by the strategies δ, τ. The payment received by Despot is defined to be the opposite of $R_{\bar{d}}^k(\delta, \tau)$, so that the game in horizon k is zero-sum. In that way, the payment $R_{\bar{d}}^k$ measures the "freedom" of People, Despot wishes to minimize it whereas Tribune wishes to maximize it.

We say that the game $\mathcal{E}_{\bar{d}}^k$ in horizon k with initial state $\bar{d}$ *has the value* $V_{\bar{d}}^k$ if for all $\epsilon > 0$, there is a strategy δ_ϵ^* of Despot such that for all strategies τ of Tribune,

$$\epsilon + V_{\bar{d}}^k \geqslant R_{\bar{d}}^k(\delta_\epsilon^*, \tau) \ , \tag{1}$$

and similarly, there is a strategy τ_ϵ^* of Tribune such that for all strategies δ of Despot,

$$R_{\bar{d}}^k(\delta, \tau_\epsilon^*) \geqslant V_{\bar{d}}^k - \epsilon \ . \tag{2}$$

The strategies δ_ϵ^* and τ_ϵ^* are said to be ϵ-optimal. In other words, Despot can make sure his loss will not exceed $V_{\bar{d}}^k + \epsilon$ by playing δ_ϵ^*, and Tribune can make sure to win at least $V_{\bar{d}}^k - \epsilon$ by playing τ_ϵ^*. The strategies δ^* and τ^* are optimal if they are 0-optimal, i.e., if we have the saddle point property:

$$R_{\bar{d}}^k(\delta, \tau^*) \geqslant R_{\bar{d}}^k(\delta^*, \tau^*) = V_{\bar{d}}^k \geqslant R_{\bar{d}}^k(\delta^*, \tau) \ , \tag{3}$$

for all strategies δ, τ of Despot and Tribune. If the value $V_{\bar{d}}^k$ exists for all choices of the initial state $\bar{d}$, we define the *value vector* of the family of games $(\mathcal{E}_d^k)_{d \in D}$ in horizon k, to be $V^k := (V_d^k)_{d \in D} \in \mathbb{R}^D$.

We now define the *infinite horizon game* $\mathcal{E}_{\bar{d}}^\infty$, in which the payment received by Tribune is given by

$$R_{\bar{d}}^\infty(\delta, \tau) := \limsup_{k \to \infty} (R_{\bar{d}}^k(\delta, \tau))^{1/k}$$

and the payment received by Despot is the opposite of the latter payment. (The choice of limsup is somehow arbitrary, we could choose liminf instead without affecting the results which follow.) The *value* $V_{\bar{d}}^\infty$ of the infinite horizon game $\mathcal{E}_{\bar{d}}^\infty$, and the optimal strategies in this game, are still defined by a saddle point condition, as in (1), (2) and (3), the payment $R_{\bar{d}}^k(\delta, \tau)$ being now replaced by $R_{\bar{d}}^\infty(\delta, \tau)$.

We denote by $V^\infty = (V_d^\infty)_{d \in D} \in \mathbb{R}^D$ the *value vector* of the infinite horizon games $(\mathcal{E}_d^\infty)_{d \in D}$.

We associate to the latter games the dynamic programming operator $F : \mathbb{R}^D \to \mathbb{R}^D$, such that, for all $X \in \mathbb{R}^D$, and $d \in D$,

$$F_d(X) = \min_{(d,t) \in E} \max_{(t,p) \in E} \sum_{(p,d') \in E} m_{pd'} X_{d'} \ . \tag{4}$$

To relate this operator with the value of the above finite or infinite horizon games, we shall interpret these games as zero-sum stochastic games with expected multiplicative criteria. The one-player case was studied in particular by Howard and Matheson under the name of risk-sensitive Markov decision processes [24] and by Rothblum under the name of multiplicative Markov decision processes, see for instance [40].

For any node $p \in P$, we denote by $E_p := \{(p,d) \in E\}$ the set of actions available to People in state p, and we denote by q_p the probability measure on E_p obtained by normalizing the restriction of the weight function m to E_p: $q_{pd} = m_{pd}/\gamma(p)$ with $\gamma(p) = \sum_{(p,d') \in E_p} m_{pd'}$. Then, F can be rewritten as

$$F_d(X) = \min_{(d,t) \in E} \max_{(t,p) \in E} \left(\gamma(p) \sum_{(p,d') \in E} q_{pd'} X_{d'} \right) \ .$$

A pair of strategies δ and τ of both players, determine the stochastic process $(D_k, T_k, P_k)_{k \geqslant 0}$ with values in $D \times T \times P$, such that $P(D_{k+1} = d' \mid H) = q_{pd'}$ for all $d' \in D$ and all histories H having $k - 1/3$ turns and ending in $p \in P$, and such that the transitions from D to T and T to D are deterministicaly determined by the strategies δ and τ respectively as in the above description of the entropy games $\mathcal{E}$. Then, the payoff of the entropy game with horizon k starting in $\bar{d}$, $\mathcal{E}_{\bar{d}}^k$, is equal to the following expected multiplicative/ risk-sensitive criterion:

$$R_{\bar{d}}^k(\delta, \tau) = \mathbb{E}\left(\gamma(P_0) \cdots \gamma(P_{k-1}) \mid D_0 = \bar{d}\right) \ .$$

Proposition 1 *The value of the entropy game in horizon k with initial state d, $\mathcal{E}_d^k$, does exists. The value vector V^k of this game is determined by the relations $V^0 = e$, $V^k = F(V^{k-1})$, $k = 1, 2, \ldots$, where e is the unit vector $(1, ..., 1)^\top$ of $\mathbb{R}^D$. Moreover, there exist optimal strategies for Despot and Tribune that are positional.*

Proof This result follows from a classical dynamic programming argument. Indeed, in the one player case, that is when there is only one choice of δ or one choice of τ, that is when the operator F contains only a "min" or a "max", the game is in the class of Markov Decision Problems with multiplicative criterion and the Dynamic Programming Principle has already been proved in this setting in [24, 40], see also [47, Th. 1.1, Chap 11]. This shows that the game has a value which satifies $V^k = F(V^{k-1})$ and $V^0 = e$, and that an optimal strategy is obtained using these equations. For instance for a "max" (when Despot has only one choice), Tribune chooses any action (t, p) attaining the maximum in

$$\max_{(t,p) \in E} \sum_{(p,d') \in E} m_{pd'} V_{d'}^{k-1} = \max_{(t,p) \in E} \left(\gamma(p) \sum_{(p,d') \in E} q_{pd'} V_{d'}^{k-1}\right) \ .$$

The resulting strategy τ^* is positional and it is optimal among all strategies τ. A similar result holds for a "min", leading to a positional strategy δ^* for Despot.

Let us now consider the general two-player case. Define the sequence of vectors V^k by

$$V_d^k = \min_{(d,t) \in E} \max_{(t,p) \in E} \sum_{(p,d') \in E} m_{pd'} V_{d'}^{k-1} \ . \tag{5}$$

with $V_d^0 = 1$, for all $d \in D$. We construct candidate strategies δ^* and τ^*, depending on the current position and number of turns, as follows. In state d, if there remains k turns to be played, Despot selects an action (d, t) achieving the minimum in (5). We denote by $\delta^*(d, k)$ the value of t such that (d, t) is selected. In state t, if there remains $k - 1/3$ turns to be played, Tribune chooses any action (t, p) attaining the maximum in

$$\max_{(t,p) \in E} \sum_{(p,d') \in E} m_{pd'} V_{d'}^{k-1} \ .$$

Now, if Player Despot plays according to δ^*, we obtain a reduced one player game. It follows from the same dynamic programming principle as above (applied here to

time dependent transition probabilities q and factors $\gamma(\cdot)$) that the value vector $V^{\delta^*,k}$ of this reduced game in horizon k does exist and satisfies the recursion

$$V_d^{\delta^*,k} = \max_{(\delta^*(d,k),p)\in E} \left(\gamma(p) \sum_{(p,d')\in E} q_{pd'} V_{d'}^{\delta^*,k-1} \right),$$

with $V_d^{\delta^*,0} = 1$, for all $d \in D$. Since $V^{\delta^*,k}$ is the value, we have $V_d^{\delta^*,k} \geqslant R_d^k(\delta^*, \tau)$ for all strategies τ of Tribune. Noting that $V_d^{\delta^*,k} = V_d^k$ by definition of δ^*, we deduce that Despot, by playing δ^*, can guarantee that his loss in the horizon k game starting from state d will not exceed V_d^k. A dual argument shows that by playing τ^*, Tribune can guarantee that his win will be at least V_d^k.   □

Example 1 Consider the entropy game whose graph and dynamic programming operator are given by:
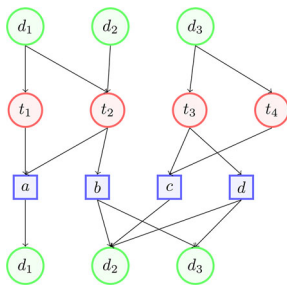
$$F_1(X) = \min\big(X_1, \max(X_1, X_2 + X_3)\big),$$

$$F_2(X) = \max\big(X_1, X_2 + X_3\big),$$

$$F_3(X) = \min\big(\max(X_2, X_2 + X_3), X_2\big).$$

For readability, the states of Despot are shown twice on the picture. Here, $D = \{d_1, d_2, d_3\}$, $T = \{t_1, t_2, t_3, t_4\}$, $P = \{a, b, c, d\}$, $E = \{(d_1, t_1), (d_1, t_2), (d_2, t_2),$ $(d_3, t_3), (d_3, t_4), (t_1, a), (t_2, a), (t_2, b), (t_3, c), (t_3, d), (t_4, c), (a, d_1), (b, d_2), (b, d_3),$ $(c, d_2), (d, d_2), (d, d_3)\}$, and all the weights are equal to 1, i.e., $m_{pd_i} = 1$ for all $p \in P$ and $1 \leqslant i \leqslant 3$ such that $(p, d_i) \in E$.

One can check that $V^k = (1, \phi_{k+1}, \phi_k)$, where $\phi_0 = \phi_1 = 1$ and $\phi_{k+2} = \phi_k + \phi_{k+1}$ is the Fibonacci sequence. As an application of Theorem 10 below, it can be checked that the value vector of this entropy game is $V^\infty = (1, \varphi, \varphi)$ where $\varphi := (1 + \sqrt{5})/2$ is the golden mean.

3 Stochastic Mean Payoff Games with Kullback-Leibler Payments

We next show that entropy games are equivalent to a class of stochastic mean payoff games in which some action spaces are simplices, and payments are given by a Kullback-Leibler divergence.

To the entropy game $\mathcal{E}$, we associate a stochastic zero-sum game with Kullback-Leibler payments, denoted $\mathcal{KL}$ and defined as follows, referred to as "Kullback-Leibler game" for brevity. This new game is played by the same players, Despot, and Tribune, on the same weighted directed graph G (so with same sets E, P, D and

same weight function m). The nondeterministic half-player, People, will be replaced by a standard probabilistic half-player, Nature.

For any node $p \in P$, recalling that $E_p := \{(p, d) \in E\}$ is the set of actions available to People in state p, we denote by Δ_p the set of probability measures on E_p. Therefore, an element of Δ_p can be identified to a vector $\vartheta = (\vartheta_{pd})_{(p,d) \in E_p}$ with nonnegative entries and sum 1. The admissible actions of Despot and Tribune in the states $d \in D$ and $t \in T$ are the same in the game $\mathcal{KL}$ and in the entropy game $\mathcal{E}$. However, the two games have different rules when the state $p \in P$ belongs to the set of People's states. Then, Tribune is allowed to play again, by selecting a probability measure $\vartheta \in \Delta_p$; in other words, Tribune plays twice in a row, selecting first an arc $(t, p) \in E$, and then a measure $\vartheta \in \Delta_p$. Then, Nature chooses the next state d according to probability ϑ_{pd}, and Tribune receives the payment $-S_p(\vartheta; m)$, where $S_p(\vartheta; m)$ is the relative entropy or Kullback-Leibler divergence between ϑ and the measure obtained by restricting the weight function m to E_p:

$$S_p(\vartheta; m) := \sum_{(p,d) \in E_p} \vartheta_{pd} \log(\vartheta_{pd}/m_{pd}) \ . \tag{6}$$

Therefore, using the notations of Section 2, we get that

$$S_p(\vartheta; m) = -\log \gamma(p) + \sum_{(p,d) \in E_p} \vartheta_{pd} \log(\vartheta_{pd}/q_{pd})$$

is minimal when the chosen probability distribution ϑ on E_p is equal to the probability distribution q_p of the transitions from state p in the stochastic game defined in Section 2. Recall that relative entropy is related to information theory and statistics [30]. An interesting special case arises when $m \equiv 1$, as in [2], thus q_p is the uniform distribution on E_p. Then, $S_p(\vartheta; m) = S_p(\vartheta) := \sum_{(p,d) \in E_p} \vartheta_{pd} \log \vartheta_{pd}$ is nothing but the Shannon entropy of ϑ.

A history in the game $\mathcal{KL}$ now consists of a finite sequence $(d_0, t_0, p_0, \vartheta_0, d_1, t_1, p_1, \dots)$, which encodes both the states and actions which have been chosen. A strategy δ of Despot is still a function which associates to a history ending in a state in d an arc (d, t) in E. A strategy of Tribune has now two components (τ, π), τ is a map which assigns to a history ending in a state in t an arc $(t, p) \in E$, as before, whereas π assigns to the same history and to the next state p chosen according to τ a probability measure on Δ_p.

To each history corresponds a path in G, obtained by ignoring the occurrences of probability measures. For instance, the path corresponding to the history $h = (d_0, t_0, p_0, \vartheta_0, d_1, t_1, p_1)$ is $(d_0, t_0, p_0, d_1, t_1, p_1)$. Again, the number of turns of a history is defined as the length of this path, each arc counting for $1/3$. So the number of turns of h is 1 and $2/3$. Choosing strategies δ and (τ, π) of both players and fixing the initial state $d_0 = \bar{d}$ determines a probability measure on the space of histories h. We denote by

$$r_{\bar{d}}^k(\delta, (\tau, \pi)) := -\mathbb{E}\left(S_{p_0}(\vartheta_0; m) + \cdots + S_{p_{k-1}}(\vartheta_{k-1}; m)\right)$$

the expectation of the payment received by Tribune, in k turns, with respect to this measure, where S_p is as in (6) and m is the weight function of the graph of the

game. We denote by $v_{\bar{d}}^k$ the *value* of the game in horizon k, with initial state $\bar{d}$, and we denote by $v^k = (v_d^k)_{d \in D}$ the *value vector*. As in the case of entropy games, we shall use subscripts and superscripts to indicate special versions of the game, e.g., $\mathcal{KL}_d^k$ refers to the game in horizon k with initial state d. Note also our convention to use lowercase letters (as in v_d^k) to refer to the game with Kullback-Leibler payments, whereas we used uppercase letters (as in V_d^k) to refer to the entropy game.

It will be convenient to consider more special games in which the actions of one of the players are restricted. We will call *policy* of Despot a stationary positional strategy of this player, i.e., a map which assigns to every node $d \in D$ a node $\delta(d) = t \in T$ such that $(d, t) \in E$. Similarly, we will call *policy* of Tribune a map which assigns to every node $t \in T$ a node $\tau(t) = p \in P$ such that $(t, p) \in E$. Observe, in this definition of policy, the symmetry between Despot and Tribune, while the game is asymetric: the policy τ is not enough to determine a positional strategy of Tribune, because the probability distribution at every state $p \in P$ is not specified by the policy τ. The set of policies of Despot and Tribune are denoted by $\mathcal{P}_D$ and $\mathcal{P}_T$, respectively.

If one fixes a policy δ of Despot, we end up with a reduced game $\mathcal{KL}^k(\delta, *)$ in which only Tribune has actions. We denote by $v^k(\delta, *) = (v_d^k(\delta, *))_{d \in D} \in \mathbb{R}^D$ the value vector of this game in horizon k. Similarly, if one fixes a policy τ of Tribune, we obtain a reduced game denoted by $\mathcal{KL}^k(*, (\tau, *))$, in which Despot plays when the state is in D, Tribune selects an action according to the policy τ when the state is in T, and Tribune plays when the state is in P. The value vector of this reduced game is denoted by $v^k(*, (\tau, *)) = (v_d^k(*, (\tau, *)))_{d \in D} \in \mathbb{R}^D$. We also denote by $v^k(\delta, (\tau, *)) = (v_d^k(\delta, (\tau, *)))_{d \in D} \in \mathbb{R}^D$ the value of the reduced game in which both policies δ of Despot and τ of Tribune are fixed, which means that only Tribune plays when the state is in P. The systematic character of notation used here should be self explanatory: the symbol $*$ refers to the actions which are not fixed by the policy.

We also consider the *infinite horizon* or *mean payoff* game $\mathcal{KL}^\infty$, in which the payment of Tribune is now

$$r_{\bar{d}}^\infty(\delta, (\tau, \pi)) := \limsup_{k \to \infty} k^{-1} r_{\bar{d}}^k(\delta, (\tau, \pi)) \ .$$

For $0 < \alpha < 1$, we also consider the *discounted game* $^\alpha\mathcal{KL}$ with a discount factor α, in which the payment of Tribune is

$$^\alpha r_{\bar{d}}(\delta, (\tau, \pi)) := -\mathbb{E}\left(S_{p_0}(\vartheta_0; m) + \alpha S_{p_1}(\vartheta_1; m) + \alpha^2 S_{p_2}(\vartheta_2; m) + \cdots \right)$$

The value of the mean payoff game is denoted by $v_{\bar{d}}^\infty$, whereas the value of the discounted game is denoted by $^\alpha v_{\bar{d}}$. As above, we denote by $\mathcal{KL}^\infty(\delta, *)$ and $\mathcal{KL}^\infty(*, (\tau, *))$ the games restricted by the choice of policies δ, τ, and use an analogous notation for the corresponding value vectors. For instance, $^\alpha v(*, (\tau, *))$ refers to the value vector of the game $^\alpha\mathcal{KL}(*, (\tau, *))$ with a discount factor α. We define the notion of value, as well as the notion of optimal strategies, by saddle point conditions, as in Section 2.

The following dynamic programming principle entails that the value of the stochastic game with Kullback-Leibler payments in horizon k is the log of the value of the entropy game.

Proposition 2 *The value vector $v^k = (v_d^k)_{d \in D}$ of the Kullback-Leibler game in horizon k does exist. It is determined by the relations $v^0 = 0$, $v^k = f(v^{k-1})$, $k = 1, 2, \ldots$, where*

$$f_d(x) = \min_{(d,t) \in E} \max_{(t,p) \in E} \log \left(\sum_{(p,d') \in E} m_{pd'} \exp(x_{d'}) \right) , \qquad (7)$$

and we have $v_d^k = \log V_d^k$.

In order to prove Proposition 2, we recall the following classical result in convex analysis showing that the "log-exp"function is the Legendre-Fenchel transform of Shannon entropy.

Lemma 3 *The function $x \mapsto \log(\sum_{1 \leqslant i \leqslant n} e^{x_i})$ is convex and it satisfies*

$$\log \left(\sum_{1 \leqslant i \leqslant n} e^{x_i} \right) = \max \sum_{1 \leqslant i \leqslant n} \vartheta_i(x_i - \log \vartheta_i); \quad \vartheta_i \geqslant 0, \ 1 \leqslant i \leqslant n, \ \sum_{1 \leqslant i \leqslant n} \vartheta_j = 1 \ .$$

This result is mentioned in [42], Example 11.12. This convexity property is a special instance of the general fact that the log of the Laplace transform of a positive measure is convex (which follows from the Cauchy-Schwarz inequality), whereas the explicit expression as a maximum follows from a straightforward computation (apply Lagrange multipliers rule).

Proof of Proposition 2 For a zero-sum game with finite horizon and additive criterion, the existence of the value is a standard fact, proved in a way similar to Proposition 1. The value vector v^k satisfies the following dynamic programming equation

$$v_d^k = \min_{(d,t) \in E} \max_{(t,p) \in E} \max_{\vartheta \in \Delta_p} \left(-S_p(\vartheta; m) + \langle \vartheta, v^{k-1} \rangle \right) , \qquad (8)$$

where $\langle \vartheta, x \rangle = \sum_{(p,d') \in E_p} \vartheta_{pd'} x_{d'}$ for $x \in \mathbb{R}^D$, and $v_d^0 = 0$. By Lemma 3,

$$\log \left(\sum_{(p,d') \in E} m_{pd'} \exp(x_{d'}) \right) = \log \left(\sum_{(p,d') \in E_p} \exp(x_{d'} + \log m_{pd'}) \right)$$

$$= \max_{\vartheta \in \Delta_p} \sum_{(p,d') \in E_p} \vartheta_{pd'}(x_{d'} + \log m_{pd'} - \log \vartheta_{pd'})$$

$$= \max_{\vartheta \in \Delta_p} \left(-S_p(\vartheta; m) + \langle \vartheta, x \rangle \right)$$

and so, (8) can be rewritten as $v^k = f(v^{k-1})$ where f is given by (7). Observe that the operator f is the conjugate of the operator F of the original entropy game: $f = \log \circ F \circ \exp$. It follows that $v^k = f^k(v^0) = \log F^k(V^0) = \log V^k$, where for a vector $Y \in (\mathbb{R}_+^*)^D$ the notation '$\log(Y)$' denotes the vector $(\log(Y_i))_{1 \leqslant i \leqslant D}$, and $\exp := \log^{-1}$.   □

The map f arising in (7) is obviously order preserving and it commutes with the addition of a constant, meaning that $f(x+\lambda e) = f(x)+\lambda e$ where e is the unit vector $(1, ..., 1)^\top$ of $\mathbb{R}^D$, and $\lambda \in \mathbb{R}$. Any map with these two properties is nonexpansive in the sup-norm, meaning that $\|f(x) - f(y)\|_\infty \leqslant \|x - y\|_\infty$, see [14]. Hence, the map $x \mapsto f(x\alpha)$ has a unique fixed point. For discounted games, the existence of the value and of optimal positional strategies is a known fact:

Proposition 4 *The discounted game $^\alpha\mathcal{KL}$ with discount factor $0 < \alpha < 1$ has a value and it admits optimal strategies that are positional and stationary. The value vector $^\alpha v$ is the unique solution of $^\alpha v = f(^\alpha v\alpha)$.*

Proof The existence and the characterization of the value are standard results, see e.g. the discussion in [36]. It is also known that the optimal strategies are obtained by selecting actions of the player attaining the minimum and maximum when evaluating every coordinate of $f(^\alpha v\alpha)$, in a way similar to the proof of Proposition 1, V^{k-1} there being replaced by $^\alpha v$. Since $^\alpha v$ does not depend on the number of turns, the optimal strategies are also stationary. $\square$

Nonexpansive maps can be considered more generally with respect to an arbitrary norm. In this setting, the issue of the existence of the limit of $v^k/k = f^k(v^0)/k$ as $k \to \infty$, and of the limit of $(1 - \alpha)(^\alpha v)$, as $\alpha \to 1^-$, where $^\alpha v$ is the unique fixed point of $x \mapsto f(x\alpha)$, has received much attention. The former limit is sometimes called *escape rate* vector. Nonexpansiveness implies that the set of accumulation points of the sequence v^k/k is independent of the choice of v^0, but it does not suffice to establish the existence of the limit; some additional "tameness" condition on the map f is needed. Indeed, a result of Neyman [36], using a technique of Bewley and Kohlberg [10], shows that the two limits $\lim_{k\to\infty} f^k(v^0)/k$ and $\lim_{\alpha\to 1^-}(1-\alpha)^\alpha v$ do exist and coincide if f is semi-algebraic. More generally, Bolte, Gaubert and Vigeral [9] showed that the same limits still exist and coincide if the nonexpansive mapping f is definable in an o-minimal structure. A counter example of Vigeral shows that the latter limit may not exist, even if the action spaces are compact and the payment and transition probability functions are continuous, so the o-minimality assumption is essential in what follows [46].

In order to apply this result, let us recall the needed definitions, referring to [44, 45] for background. An *o-minimal structure* consists, for each integer n, of a family of subsets of $\mathbb{R}^n$. A subset of $\mathbb{R}^n$ is said to be *definable* with respect to this structure if it belongs to this family. It is required that definable sets are closed under the Boolean operations, under every projection map (elimination of one variable) from $\mathbb{R}^n$ to $\mathbb{R}^{n-1}$, and under the lift, meaning if $A \subset \mathbb{R}^n$ is definable, then $A \times \mathbb{R} \subset \mathbb{R}^{n+1}$ and $\mathbb{R} \times A \subset \mathbb{R}^{n+1}$ are also definable. It is finally required that when $n = 1$, definable subsets are precisely finite unions of intervals. A function f from $\mathbb{R}^n$ to $\mathbb{R}^k$ is said to be *definable* if its graph is definable.

An important example of o-minimal structure is the *real exponential field* $\mathbb{R}_{\mathrm{alg,exp}}$. The definable sets in this structure are the *subexponential sets* [45], i.e., the images under the projection maps $\mathbb{R}^{n+k} \to \mathbb{R}^n$ of the *exponential sets* of $\mathbb{R}^{n+k}$, the latter

being sets of the form $\{x \mid P(x_1, \ldots, x_{n+k}, e^{x_1}, \ldots, e^{x_{n+k}}) = 0\}$ where P is a real polynomial. A theorem of Wilkie [48] implies that $\mathbb{R}_{\text{alg,exp}}$ is o-minimal, see [45]. Observe in particular that the set $\{x \in \mathbb{R}^2 \mid x_1 \leqslant x_2\}$ is definable in this structure, being the projection of $\{x \in \mathbb{R}^3 \mid x_2 - x_1 = x_3^2\}$. Using the o-minimal character of this structure, this implies that definable maps are stable by the operations of pointwise maximum and minimum. We deduce the following key fact.

Fact 5 *The dynamic programming operator f of the Kullback-Leibler game, defined by* (7)*, is definable in the real exponential field.*

Theorem 6 ([9]) *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be nonexpansive in any norm, and suppose that f is definable in an o-minimal structure. Then,*

$$\lim_{k \to \infty} f^k(0)/k$$

does exists, and it coincides with

$$\lim_{\alpha \to 1^-} (1 - \alpha)^{\alpha} v \ .$$

Corollary 7 *Let $v^k = (v_d^k)_{d \in D}$ be the value vector in horizon k of the stochastic game with Kullback-Leibler payments, $\mathcal{KL}^k$, and for $0 < \alpha < 1$, let $^\alpha v$ denote the value vector of the discounted game $^\alpha\mathcal{KL}$ with discount factor $0 < \alpha < 1$. Then $\lim_{k \to \infty} v^k/k$ does exist and it coincides with $\lim_{\alpha \to 1^-}(1 - \alpha)(^\alpha v)$.*

Proof We already noted that the map f in (7) is nonexpansive in the sup-norm. It is definable in the real exponential field. So Theorem 6 can be applied to it. □

Corollary 7 will allow us to establish the existence of the value of the mean pay-off game, and to obtain optimal strategies, by considering the discounted game, for which, as noted in Proposition 4, the existence of the value and of optimal policies are already known.

Let us recall that a strategy in a discounted game is said to be *Blackwell optimal* if it is optimal for all discount factors sufficiently close to one. The existence of Blackwell optimal positional strategies is a basic feature of perfect information zero-sum stochastic games with finite action spaces (see [39, Chap. 10] for the one-player case, the two-player case builds on similar ideas, e.g. [18, Lemma 26]). We next show that this result has an analogue for entropy games. To get a Blackwell type optimality result, we need to restrict to a setting with finitely many positional strategies. Recall that $\mathcal{P}_D$ (resp. $\mathcal{P}_T$) denotes the set of policies of Despot (resp. Tribune). We also recall our notation $v^\infty(\delta, *)$ for the value of the mean payoff game $\mathcal{KL}^\infty(\delta, *)$ in which Despot plays according to the policy δ.

We define the *projection* of a pair of strategies $(\delta, (\tau, \pi))$ in the game $\mathcal{KL}$ to be the strategy (δ, τ) in the game $\mathcal{E}$. In the present setting, it is appropriate to say that a pair of policies $(\delta, \tau) \in \mathcal{P}_D \times \mathcal{P}_T$ is *Blackwell optimal* if there is a real number $0 < \alpha_0 < 1$ such that, for all $\alpha \in (\alpha_0, 1)$, (δ, τ) is the projection of a pair of optimal strategies $(\delta, (\tau, \pi))$ in the discounted game $^\alpha\mathcal{KL}$.

Theorem 8 *The family of discounted Kullback-Leibler games $({}^{\alpha}\mathcal{KL})_{\alpha \in (0,1)}$ has positional Blackwell optimal strategies.*

Proof For all $\alpha \in (0, 1)$, the *discounted* game has positional optimal strategies $\delta^*, (\tau^*, \pi^*)$. This follows from the standard dynamic programming argument mentioned in the proofs of Proposition 1 and 4, noting that $\delta^*(d)$ is obtained by choosing any $t \in T$ such that $(d, t) \in E$ attains the minimum in the expression

$$
{}^{\alpha}v_d = \min_{(d,t)\in E} \max_{(t,p)\in E} \max_{\vartheta \in \Delta_p} \left(-S_p(\vartheta; m) + \langle \vartheta, \alpha({}^{\alpha}v) \rangle \right) .
$$

Similarly, $\tau^*(t)$ is chosen to be any $p \in P$ such that $(t, p) \in E$ attains the maximum in

$$
\max_{(t,p)\in E} \max_{\vartheta \in \Delta_p} \left(-S_p(\vartheta; m) + \langle \vartheta, \alpha({}^{\alpha}v) \rangle \right) ,
$$

and $\pi^*(p)$ is chosen to be the unique action ϑ attaining the maximum in

$$
\max_{\vartheta \in \Delta_p} \left(-S_p(\vartheta; m) + \langle \vartheta, \alpha({}^{\alpha}v) \rangle \right)
$$

(observe that the function to be maximized is strictly concave and continuous on Δ_p, and that Δ_p is compact and convex, so the maximum is achieved at a unique point).

By definition of the value and of optimal strategies, we have, for all strategies δ and (τ, π) of Despot and Tribune respectively,

$$
{}^{\alpha}r_d(\delta^*, (\tau, \pi)) \leqslant {}^{\alpha}v_d = {}^{\alpha}r_d(\delta^*, (\tau^*, \pi^*)) \leqslant {}^{\alpha}r_d(\delta, (\tau^*, \pi^*)) , \tag{9}
$$

which is equivalent to

$$
{}^{\alpha}v_d = {}^{\alpha}r_d(\delta^*, (\tau^*, \pi^*)) = {}^{\alpha}v_d(\delta^*, *) = {}^{\alpha}v_d(*, (\tau^*, \pi^*)) . \tag{10}
$$

Specializing the first inequality in (9) to $\tau = \tau^*$, and bounding above the last term, we deduce that, for all for all strategies δ and τ of Despot and Tribune respectively, we have

$$
{}^{\alpha}v(\delta^*, (\tau, *)) \leqslant {}^{\alpha}v_d = {}^{\alpha}v(\delta^*, (\tau^*, *)) \leqslant {}^{\alpha}v(\delta, (\tau^*, *)) , \tag{11}
$$

where ${}^{\alpha}v_d(\delta, (\tau, *))$ is the value of the reduced discounted 1-player discounted game ${}^{\alpha}\mathcal{KL}(\delta, (\tau, *))$ starting at $d \in D$, in which the (not necessarily positional) strategies δ of Despot and τ of Tribune are fixed. The inequalities (11) can be specialized in particular to policies $\delta \in \mathcal{P}_D$ and $\tau \in \mathcal{P}_T$. Then, by Proposition 4, ${}^{\alpha}v(\delta, (\tau, *))$ is the unique fixed point of the self-map $x \mapsto {}^{\tau}f_d^{\delta}(x\alpha)$ of $\mathbb{R}^D$, where ${}^{\tau}f^{\delta}$ is the dynamic programming operator given by

$$
{}^{\tau}f_d^{\delta}(x) = \log \left(\sum_{(\tau \circ \delta(d), d') \in E} m_{\tau \circ \delta(d), d'} \exp(x_{d'}) \right) . \tag{12}
$$

It follows that the map $\alpha \mapsto {}^{\alpha}v(\delta, (\tau, *))$ is definable in the real exponential field $\mathbb{R}_{\mathrm{alg,exp}}$. (To see this, observe that, by Fact 5, the set $\{(x, y) \mid x = {}^{\tau}f_d^{\delta}(y)\} \times \mathbb{R}$ is definable in this structure; then, taking the intersection of this set with the definable sets $\{(x, y, \alpha) \mid y_d = x_d \alpha\}$, for $d \in D$, and projecting the intersection keeping only

the x and α variables, we obtain a definable set which is precisely the graph of the map $\mapsto {}^\alpha v(\delta, (\tau, *)))$.

For all $(\bar{\delta}, \bar{\tau}) \in \mathcal{P}_D \times \mathcal{P}_T$, let $I(\bar{\delta}, \bar{\tau})$ denote the set of $\alpha \in (0, 1)$ such that

$$ {}^\alpha v(\bar{\delta}, (\tau, *)) \leqslant {}^\alpha v(\bar{\delta}, (\bar{\tau}, *)) \leqslant {}^\alpha v(\delta, (\bar{\tau}, *)) \tag{13} $$

holds for all $(\delta, \tau) \in \mathcal{P}_D \times \mathcal{P}_T$. Since the saddle point property (11) holds for all α (δ^* and τ^* depend on α, of course), we have

$$ \cup_{(\bar{\delta}, \bar{\tau})} I(\bar{\delta}, \bar{\tau}) = (0, 1) \ . \tag{14} $$

Observe that the set $I(\bar{\delta}, \bar{\tau})$ is a subset of $\mathbb{R}$ definable in the real exponential field, which is o-minimal. It follows that $I(\bar{\delta}, \bar{\tau})$ is a finite union of intervals. Hence, (14) provides a covering of $(0, 1)$ by finitely many intervals, and so, one of the sets $I(\bar{\delta}, \bar{\tau})$ must include an interval of the form $(1 - \epsilon, 1)$.

To show that the policies $\bar{\delta}, \bar{\tau}$ obtained in this way are Blackwell optimal, it remains to show that if $(\bar{\delta}, \bar{\tau})$ satisfies (13) for some α, then it is the projection of a pair of optimal strategies $(\bar{\delta}, (\bar{\tau}, \bar{\pi}))$ in the discounted game ${}^\alpha\mathcal{KL}$. For this, we shall apply the existence of optimal strategies that are positional and the resulting (10) and (11) to the reduced games ${}^\alpha\mathcal{KL}(\bar{\delta}, *)$ and ${}^\alpha\mathcal{KL}(*, (\bar{\tau}, *))$, respectively.

The first game leads to the existence of positional stationary strategies τ^1, π^1 of Tribune such that, for all $d \in D$,

$$ {}^\alpha v_d(\bar{\delta}, *) = {}^\alpha r_d(\bar{\delta}, (\tau^1, \pi^1)) = {}^\alpha v(\bar{\delta}, (\tau^1, *)) \ . $$

Then, using (13), we get that ${}^\alpha v(\bar{\delta}, (\bar{\tau}, *)) \geqslant {}^\alpha v(\bar{\delta}, (\tau^1, *)) = {}^\alpha v_d(\bar{\delta}, *) \geqslant {}^\alpha v(\bar{\delta}, (\bar{\tau}, *))$, hence the equality ${}^\alpha v(\bar{\delta}, (\bar{\tau}, *)) = {}^\alpha v_d(\bar{\delta}, *)$.

The second one leads to the existence of positional stationary strategies δ^2, π^2 of Despot and Tribune respectively such that, for all $d \in D$,

$$ {}^\alpha v_d(*, (\bar{\tau}, *)) = {}^\alpha r_d(\delta^2, (\bar{\tau}, \pi^2)) = {}^\alpha v_d(\delta^2, (\bar{\tau}, *)) = {}^\alpha v_d(*, (\bar{\tau}, \pi^2)) \ . $$

Then, using (13), we deduce that ${}^\alpha v_d(\bar{\delta}, (\bar{\tau}, *)) \leqslant {}^\alpha v_d(\delta^2, (\bar{\tau}, *)) = {}^\alpha v_d(*, (\bar{\tau}, \pi^2)) \leqslant {}^\alpha v_d(\bar{\delta}, (\bar{\tau}, \pi^2)) \leqslant {}^\alpha v_d(\bar{\delta}, (\bar{\tau}, *))$, hence the equality ${}^\alpha v_d(\bar{\delta}, (\bar{\tau}, *)) = {}^\alpha v_d(*, (\bar{\tau}, \pi^2)) = {}^\alpha v_d(\bar{\delta}, (\bar{\tau}, \pi^2))$. With the equality proved with the first game, this leads to ${}^\alpha v_d(\bar{\delta}, *) = {}^\alpha v(\bar{\delta}, (\bar{\tau}, *)) = {}^\alpha v_d(\bar{\delta}, (\bar{\tau}, \pi^2)) = {}^\alpha v_d(*, (\bar{\tau}, \pi^2))$. This shows that $(\bar{\delta}, (\bar{\tau}, \pi^2))$ is a pair of optimal strategies for the discounted game ${}^\alpha\mathcal{KL}$. Since $(\bar{\delta}, \bar{\tau})$ is its projection, we get that it is Blackwell optimal. □

Theorem 9 *The value v^∞ of the stochastic mean payoff game with Kullback-Leibler payments does exist, and it coincides with $\lim_{k \to \infty} v^k/k = \lim_{\alpha \to 1^-}(1 - \alpha)({}^\alpha v)$. For all $(\delta, \tau) \in \mathcal{P}_D \times \mathcal{P}_T$, the same properties hold for the values $v^\infty(\delta, *)$ and $v^\infty(*, (\tau, *))$ of the reduced games in which Despot plays according to δ when the state is in D and Tribunes plays according to τ when the state is in T, respectively. Moreover, the Blackwell optimal strategies $(\delta^*, \tau^*) \in \mathcal{P}_D \times \mathcal{P}_T$ of Corollary 8 satisfy, for all $d \in D$,*

$$ v_d^\infty = v_d^\infty(\delta^*, (\tau^*, *)) = v_d^\infty(\delta^*, *) = v_d^\infty(*, (\tau^*, *)) \ . \tag{15} $$

In particular,

$$v_d^\infty = \min_{\delta \in \mathcal{P}_D} v_d^\infty(\delta, *) = \max_{\tau \in \mathcal{P}_T} v_d^\infty(*, (\tau, *)) \ .$$

Proof We already noted in Corollary 7 that

$$\lim_{k \to \infty} v^k/k = \lim_{\alpha \to 1^-} (1 - \alpha)(^\alpha v) \ . \tag{16}$$

Moreover, it is shown in [9, Corollary 2, (iii)], as a consequence of a theorem of Mertens and Neyman [34], that this limit coincides with the value of the game. These results rely on the definable and sup-norm nonexpansive character of the dynamic programming operator f. The dynamic programming operator f^δ, associated to the reduced game determined by the strategy $\delta \in \mathcal{P}_D$ can be written as

$$f_d^\delta(x) := \max_{(\delta(d), p) \in E} \log \left(\sum_{(p, d') \in E} m_{pd'} \exp(x_{d'}) \right) \ . \tag{17}$$

It is definable and sup-norm nonexpansive, hence the same conclusions apply to the game $\mathcal{KL}(\delta, *)$, i.e,

$$\lim_{k \to \infty} v^k(\delta, *)/k = \lim_{\alpha \to 1^-} (1 - \alpha)(^\alpha v(\delta, *)) \tag{18}$$

is the value of the game $\mathcal{KL}^\infty(\delta, *)$. We argue in the same way for the game $\mathcal{KL}(*, (\tau, *))$, noting that the associated dynamic programming operator is now

$$^\tau f_d(x) := \min_{(d, t) \in E} \log \left(\sum_{(\tau(t), d') \in E} m_{\tau(t)d'} \exp(x_{d'}) \right) \ . \tag{19}$$

which is still definable and sup-norm nonexpansive. Hence,

$$\lim_{k \to \infty} v^k(*, (\tau, *))/k = \lim_{\alpha \to 1^-} (1 - \alpha)(^\alpha v(*, (\tau, *))) \tag{20}$$

is the value of the game $\mathcal{KL}^\infty(*, (\tau, *))$.

Let δ^*, τ^* denote the positional Blackwell optimal strategies constructed in Corollary 8. By definition, there is an interval $(\alpha_0, 1)$ such that for all $\alpha \in (\alpha_0, 1)$, there exists π^* depending on α such that

$$^\alpha v = ^\alpha v(\delta^*, (\tau^*, \pi^*)) = ^\alpha v(\delta^*, *) = ^\alpha v(*, (\tau^*, \pi^*))$$

which by (11) leads to

$$^\alpha v = ^\alpha v(\delta^*, *) = ^\alpha v(*, (\tau^*, *)) = ^\alpha v(\delta^*, (\tau^*, *)) \ . \tag{21}$$

Multiplying these expressions by $(1 - \alpha)$, passing to the limit, and using (16), (18) and (20), we obtain (15). $\qquad\square$

Remark 1 Corollary 9 shows that Player Despot has an optimal positional strategy δ^* in the mean payoff Kullback-Leibler game. It also shows that in the same game, the actions of Player Tribune at states $t \in T$ can be chosen according to the optimal positional strategy τ^*. This theorem does not imply, however, that at every state $p \in$

P, the optimal action $\vartheta \in \Delta_p$ can be chosen optimally according to a positional strategy π. Indeed, the proof by an o-minimality argument uses in an essential way the fact that there are finite actions spaces at every states d and t, whereas Δ_p is infinite. We leave for further investigation the question of the existence of such a positional strategy, noting that it is not needed in the application to entropy games.

Remark 2 It is shown in [9, Corollary 2, (iii)], as a consequence of a theorem of Mertens and Neyman [34], that a stochastic game with a definable Shapley operator has a *uniform value*, a property which is stronger than the mere existence of the value. Loosely speaking, a stochastic game with initial state d is said to have a uniform value v_d^∞ if both players can almost guarantee v_d^∞ provided that the length of the k-stage game is large enough. In the present setting, we get the following property: for any $\epsilon > 0$, there is a couple of strategies of $(\delta, (\tau, \pi))$ and a time K such that, for every $k \geqslant K$, every starting state d and every strategies δ' and (τ', π'),

$$r_d^k(\delta, \tau', \pi')/k \leqslant v_d^\infty + \epsilon, \qquad r_d^k(\delta', \tau, \pi)/k \geqslant v_d^\infty - \epsilon \ .$$

4 Application to the Entropy Game Model

4.1 Existence of Optimal Positional Strategies in the Entropy Game

As an application of Corollary 9, we obtain the existence of optimal positional strategies in the entropy game model of Section 2.

Theorem 10 *The infinite horizon entropy game has a value and it has optimal positional strategies, namely the Blackwell optimal strategies $(\delta^*, \tau^*) \in \mathcal{P}_D \times \mathcal{P}_T$ of Corollary 8. Moreover, for all initial states d,*

$$V_d^\infty = \lim_{k \to \infty} \left(V_d^k \right)^{1/k} \ .$$

Proof By Proposition 1, $V^k = F(V^{k-1})$ where F is as in (4). Moreover, $F = \exp \circ f \circ \log$ is the conjugate of the dynamic programming operator f of the Kullback-Leibler game introduced in (7). Corollary 7 shows that $v^\infty = \lim_{k \to \infty} v^k/k$ does exists. It follows that $V_d^\infty := \lim_{k \to \infty} (V_d^k)^{1/k} = \exp(v_d^\infty)$ does exist for all $d \in D$.

Let (δ^*, τ^*) denote the Blackwell optimal strategies given by Corollary 8. We showed in Corollary 9 that $v^\infty = v^\infty(\delta^*, *)$, and by Corollary 7, we have

$$v^\infty(\delta^*, *) = \lim_{k \to \infty} v^k(\delta^*, *)/k \ .$$

Using the dynamic programming principle for finite horizon 1-player games, or equivalently, by applying Proposition 2 to the reduced finite horizon game $\mathcal{KL}^k(\delta^*, *)$, we obtain that

$$v^k(\delta^*, *)/k = (f^{\delta^*})^k(0)/k \ ,$$

where for all δ, f^δ is the dynamic programming operator defined in (17) associated to the reduced game $\mathcal{KL}(\delta, *)$. Consider now the conjugate $F^\delta := \exp \circ f^\delta \circ \log$, so that

$$F_d^\delta(X) := \max_{(\delta(d), p) \in E} \left(\sum_{(p, d') \in E} m_{pd'} X_{d'} \right) . \tag{22}$$

By Proposition 1, for all strategies τ of Tribune, non necessarily positional, and all initial states $d \in D$, we have

$$R_d^k(\delta^*, \tau) \leqslant [(F^{\delta^*})^k(e)]_d .$$

Applying all the above equalities and inequalities, we deduce that

$$\limsup_{k \to \infty} (R_d^k(\delta^*, \tau))^{1/k} \leqslant \lim_{k \to \infty} [(F^{\delta^*})^k(e)]_d^{1/k} = \exp(v_d^\infty) = V_d^\infty ,$$

so Player Despot can guarantee his loss does not exceed V_d^∞ by playing δ^*.

Let us now consider the reduced infinite horizon or finite horizon entropy and Kullback-Leibler games in which the strategy of Tribune is fixed and equal to τ^*. By the same arguments as above, we show that the positional strategy τ^* guarantees to Player Tribune to win at least V_d^∞ in the entropy game. Indeed, applying successively Corollary 9 and Proposition 2, we deduce

$$v^\infty = v^\infty(*, (\tau^*, *)) = \lim_{k \to \infty} v^k(*, (\tau^*, *))/k = \lim_{k \to \infty} (^{\tau^*} f)^k(0)/k ,$$

where, for all τ, $^\tau f$ is as in (19). Then, considering the conjugate $^\tau F := \exp \circ^\tau f \circ \log$, so that, for all τ,

$$^\tau F_d(X) := \min_{(d, t) \in E} \left(\sum_{(\tau(t), d') \in E} m_{\tau(t)d'} X_{d'} \right) , \tag{23}$$

and applying Proposition 2 and 1, we deduce that

$$V_d^\infty = \lim_{k \to \infty} (^{\tau^*} F_d)^k(e)^{1/k} \leqslant \liminf_{k \to \infty} R_d^k(\delta, \tau^*)^{1/k} ,$$

for all strategies δ of Despot, non necessarily positional, and all initial states $d \in D$. So Player Tribune can win at least V_d^∞ by playing τ^*. $\qquad\square$

4.2 Comparison with the Original Entropy Game Model

The original entropy game model of Asarin et al. [2] is a zero-sum game defined in a way similar to Section 2, up to a technical difference: in their model, the initial state is not prescribed. The payment of Tribune in horizon k, instead of being $R_d^k(\delta, \tau)$, is the quantity $\bar{R}^k(\delta, \tau)$, defined now as the sum of weights of all paths of length k starting at a node in D and ending at a node in D. Hence, $\bar{R}^k(\delta, \tau) = \sum_{d \in D} R_d^k(\delta, \tau)$. The payment of Tribune can be defined in their game as follows $\bar{R}^\infty(\delta, \tau) = \limsup_{k \to \infty} (\bar{R}^k(\delta, \tau))^{1/k}$. This game is denoted by $\mathcal{E}^{\mathrm{orig}, \infty}$, we denote by $\bar{V}^\infty$ the value of this game, which is shown to exist in [2].

Note that in the initial model in [2], the weights $m_{pd'}$ are equal to 1. The generalization to weighted entropy games, in which the weights $m_{pd'}$ are integers is

discussed in Section 6 of [2]. The case in which the weights $m_{pd'}$ take rational values can be reduced to the latter case by multiplying all the weights by an integer factor. Therefore, we will ignore the restriction that $m_{pd'} = 1$ in our definition of $\mathcal{E}^{\mathrm{orig},\infty}$ and will refer to the entropy game model with rational weights as the entropy game model. The following observation follows readily from Theorem 10.

Proposition 11 *The value of the original entropy game $\mathcal{E}^{orig,\infty}$ considered by Asarin et al. [2] (with a free initial state), coincides with the maximum of the values of the games $\mathcal{E}_d^\infty$, taken over all initial states $d \in D$:*

$$\bar{V}^\infty = \max_{d \in D} V_d^\infty \ .$$

Example 2 Proposition 11 is illustrated by the game of Example 1. In the original model of [2], the value, defined independently of the initial state, is $(1 + \sqrt{5})/2$, whereas our model associates to the initial state d_1 a value 1 which differs from the values of d_2 and d_3.

In [2], entropy games were compared with matrix multiplication games. We present here this correspondence in the case of general weights $m_{pd'}$. Given policies $\delta \in \mathcal{P}_D$ and $\tau \in \mathcal{P}_T$, let $A(\delta) \in \mathbb{R}^{D \times T}$ and $B(\tau) \in \mathbb{R}^{T \times D}$ be such that $A(\delta)_{dt} = 1$ if $t = \delta(d)$ and 0 otherwise, and $B(\tau)_{td} = m_{\tau(t)d}$ if $(\tau(t), d) \in E$ and 0 otherwise, for all $(d, t) \in D \times T$. We shall think of $A(\delta)$ and $B(\tau)$ as rectangular matrices. Then $\bar{R}^k(\delta, \tau) = \|(A(\delta)B(\tau))^k\|_1$, where for any $A \in \mathbb{R}^{D \times D}$, A^k denotes its kth power and $\|A\|_1 = \sum_{dd'} |A_{dd'}|$ its ℓ^1 norm. From this, one deduces that $\bar{R}^\infty(\delta, \tau) = \rho(A(\delta)B(\tau))$, where $\rho(A)$ denotes the spectral radius of the matrix A. Moreover, let $\mathcal{A}$ and $\mathcal{B}$ denote the sets of all matrices of the form $A(\delta)$ and $B(\tau)$ respectively, and let $\mathcal{AB}$ be the set of all matrices AB with $A \in \mathcal{A}$ and $B \in \mathcal{B}$. The sets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{AB}$ are subsets of matrices $\mathcal{M}$ satisfying the property that all elements of $\mathcal{M}$ have same dimension and if $\mathcal{M}_i$ is the set of ith rows of the elements of $\mathcal{M}$, then $\mathcal{M}$ is the set of matrices the ith row of which belongs to $\mathcal{M}_i$. Such a property defines the notion of IRU matrix sets (for independent row uncertainty sets) in [2]. The following property proved in [2] is the analogue of Theorem 9, V_d^∞ being replaced by $\bar{V}^\infty$:

$$\bar{V}^\infty = \min_{A \in \mathcal{A}} \max_{B \in \mathcal{B}} \rho(AB) = \max_{B \in \mathcal{B}} \min_{A \in \mathcal{A}} \rho(AB) \ . \tag{24}$$

A more general property is proved in [5, Section 8], as a consequence of the Collatz-Wielandt theorem (see Corollary 13 below).

Remark 3 Our approach shows that entropy games reduce to Kullback-Leibler games, which are stochastic mean payoff games (with compact action spaces), Asarin et al. [2] remarked that the special *deterministic* entropy games, in which People has only one possible action in each state, can be re-encoded as deterministic mean payoff games. This can also be recovered from our approach: in this deterministic case, the simplices Δ_p are singletons in the Kullback-Leibler game, and the entropy

function vanishes, so the Kullback-Leibler game degenerates to a deterministic mean payoff game.

5 Applying the Collatz-Wielandt Theorem to Entropy Games

The classical Collatz-Wielandt formulæ provide variational characterizations of the spectral radius $\rho(M)$ of a nonnegative matrix $M \in \mathbb{R}^{D \times D}$:

$$\rho(M) = \inf\{\lambda > 0 \mid \exists X \in \text{int } \mathbb{R}_+^D, \ MX \leqslant \lambda X\}$$
$$= \max\{\lambda \geqslant 0 \mid \exists X \in \mathbb{R}_+^D \setminus \{0\}, \ MX = \lambda X\} \ ,$$
$$= \max\{\lambda \geqslant 0 \mid \exists X \in \mathbb{R}_+^D \setminus \{0\}, \ MX \geqslant \lambda X\} \ ,$$

where $\mathbb{R}_+^D$ denotes the nonnegative orthant of $\mathbb{R}^D$, and int $\mathbb{R}_+^D$ its interior, i.e, the set of positive vectors. The infimum is not always attained in the first line, whereas by writing "max", we mean that the suprema are always attained.

This has been extended to non-linear, order preserving and continuous self-maps of the standard positive cone by Nussbaum [37], see also [3, 5, 19, 22, 31]. Recall that a self-map of $\mathbb{R}_+^D$ is said to be *order preserving* if $u \leqslant v \Rightarrow F(u) \leqslant F(v)$ for all $u, v \in \mathbb{R}_+^D$, the relation $\leqslant$ being understood entrywise. This map is *positively homogeneous of degree 1* if $F(\alpha u) = \alpha u$, for all $\alpha > 0$ and $u \in \mathbb{R}_+^D$.

Theorem 12 *Let F be a continuous, order preserving, and positively homogeneous of degree 1 self-map of $\mathbb{R}_+^D$. Then the following quantities coincide*

$$\lim_{k \to \infty} \max_{d \in D}[F^k(e)]_d^{1/k} \tag{25}$$

$$\inf\{\lambda > 0 \mid \exists X \in int\ \mathbb{R}_+^D, \ F(X) \leqslant \lambda X\} \tag{26}$$

$$\max\{\lambda \geqslant 0 \mid \exists X \in \mathbb{R}_+^D \setminus \{0\}, \ F(X) = \lambda X\} \ , \tag{27}$$

$$\max\{\lambda \geqslant 0 \mid \exists X \in \mathbb{R}_+^D \setminus \{0\}, \ F(X) \geqslant \lambda X\} \ , \tag{28}$$

Proof The existence of (25) and the fact it coincides with (26) is proved in [19, Prop 1]. The fact that (26) and (27) coincide is proved in [37, Theorem 3.1]. The fact that (27) and (28) coincide is proved in [3, Lemma 2.8]. □

The common value of the expressions in Theorem 12 is called the non-linear spectral radius of F.

We showed in Proposition 11 that the value of the original entropy game $\mathcal{E}^{\text{orig}}$ is precisely

$$\bar{V}^\infty = \max_{d \in D} V_d^\infty = \lim_{k \to \infty} \max_{d \in D}[F^k(e)]_d^{1/k}$$

where F is the dynamic programming operator defined in (4). This operator is continuous, order preserving, and homogeneous of degree one. Hence, we get as an immediate corollary of Theorem 12:

Corollary 13 *The value $\bar{V}^\infty$ of the original entropy game $\mathcal{E}^{orig}$ (with a free initial state) coincides with any of the following expressions*

$$\inf\{\lambda > 0 \mid \exists X \in int\, \mathbb{R}_+^D,\ F(X) \leqslant \lambda X\} \tag{29}$$

$$\max\{\lambda > 0 \mid \exists X \in \mathbb{R}_+^D \setminus \{0\},\ F(X) = \lambda X\} \tag{30}$$

$$\max\{\lambda > 0 \mid \exists X \in \mathbb{R}_+^D \setminus \{0\},\ F(X) \geqslant \lambda X\}\ , \tag{31}$$

where F is the dynamic programming operator (4).

The Collatz-Wielandt formulæ of Theorem 13 are helpful to establish strong duality results, like (24). Note that (24) is weaker than Theorem 13 since it does not imply the existence of a nonlinear vector whereas (30) of Theorem 13 does. See also [3] for an application to mean payoff games and tropical geometry. Our main interest here lies in the following application of (29). We say that a state d of Despot is *significant* if the set of actions of Despot in this state, $\{(d, t) \in E\}$, has at least two elements (i.e., Despot has to make a choice in this state). We say that an entropy game is *Despot-free* if the Despot player does not have any significant state. A Despot-free game is essentially a one (and half) player problem, since the minimum term in the corresponding dynamic programming operator (4) vanishes. Indeed, for each $d \in D$, there is a unique node t such that $(d, t) \in E$, and we define the map $\sigma : D \to T$ by $\sigma(d) = t$. The following corollary, which follows from Corollary 13 by making the change of variables $\mu = \log \lambda$ and $x = \log X$, is also a special case of a result of Anantharam and Borkar [1].

Corollary 14 *The logarithm of the value of a Despot-free original entropy game is given by the value of the optimization problem*

$$\inf \mu$$
$$(\mu, x) \in \mathbb{R} \times \mathbb{R}^D, \textit{ satisfying}$$
$$\mu + x_d \geqslant \log\left(\sum_{d' \in D} m_{p,d'} e^{x_{d'}}\right) \textit{ for all } d \in D,\, p \in P \textit{ such that } (\sigma(d), p) \in E\ .$$
$$\tag{32}$$

Observe that the latter expression is the value of an optimization problem in which the variables are μ and $x = (x_d)_{d \in D}$, the objective function is the linear form $(\mu, x) \mapsto \mu$, and the feasible set is convex. Hence, this will lead us to a polynomial time decision procedure in the Despot free case, which we develop in the next section.

6 Polynomial time Solvability of Entropy Games with a Few Significant Despot Positions

By *solving strategically* an entropy game, we mean finding a pair of optimal policies. We assume from now that the weights $m_{p,d}$ are integers. Since policies are combinatorial objects, solving strategically the game is a well posed problem in the Turing

(bit) model of computation. Once optimal policies are known, the value of the game, which is an algebraic number, can be obtained as the Perron root of an associated integer matrix. Our first main result is the following.

Theorem 15 *Despot-free entropy games can be solved strategically in polynomial time.*

This will be proved in Section 7, by combining several ingredients: a reduction to the irreducible case, an application of the ellipsoid method, and separation bounds for algebraic numbers.

We will also show the following generalization of Theorem 15.

Theorem 16 *Entropy games in which Despot has a fixed number of significant states can be solved strategically in polynomial time.*

7 Proof of Theorem 15 and Theorem 16

We start by considering a Despot-free game. We decompose the proof of Theorem 15 in several steps, corresponding to different subsections.

7.1 Reduction to the Irreducible Case

First, we associate to a Despot-free entropy game a projected directed graph $\bar{G}$, with node set D and an arc $d \to d'$ if there is a path (d, t, p, d') in the original directed graph G. We say that the game is *irreducible* if $\bar{G}$ is strongly connected. Recall that we assumed that any node has a successor, so that strongly connected components are all non trivial (not reduced to a node with no edge).

Lemma 17 *The value of an irreducible Despot-free entropy game is independent of the initial state. Moreover, there is a vector $U \in \operatorname{int} \mathbb{R}_+^D$ and a scalar $\lambda^* > 0$ such that $F(U) = \lambda^* U$, and λ^* coincides with the value of any initial state in this game.*

Proof The non-linear Perron-Frobenius theorem in [19] provides a sufficient condition for the existence of a positive eigenvector of an order preserving positively homogeneous self-map F of the interior of the cone. It suffices to check that a certain directed graph $G(F)$ associated to F is strongly connected. Specialized to the present setting, this directed graph is defined as follows: the node set is D, and there is an arc from $d \to d'$ if

$$\lim_{s \to \infty} F_d(se_{d'}) = +\infty \ ,$$

where $e_{d'} = (0, \dots, 0, 1, 0, \dots, 0)^\top$ is the d'th vector of the canonical basis of $\mathbb{R}^D$. By considering the explicit form of F, we see that $G(F)$ is precisely the directed graph $\bar{G}$. Hence, by Theorem 2 of [19], there exists a vector $U \in \operatorname{int} \mathbb{R}_+^D$ and a scalar

$\lambda^* > 0$ such that $F(U) = \lambda^* U$. It follows from Theorem 13 and from the fact that the value of the original entropy game is $\max_{d \in D} V_d^\infty$ that $V_d^\infty \leqslant \lambda^*$ holds for all $d \in D$.

We next show that the other inequality holds. Recall that $V_d^\infty = \lim_{k \to \infty} [F^k(e)]_d^{1/k}$ is the value of the entropy game with initial state d, where $e = (1, ..., 1)^\top \in \mathbb{R}_+^D$.

Since D is finite, there is a positive scalar $\beta > 0$ such that $e \geqslant \beta U$ (indeed, take $\beta := (\max_{d \in D} U_d)^{-1}$). Using the order preserving and positively homogeneous character of the dynamic programming operator F, we get

$$F^k(e) \geqslant F^k(\beta U) = \beta F^k(U) = \beta (\lambda^*)^k U$$

and so, using that U is positive,

$$V_d^\infty = \lim_{k \to \infty} [F^k(e)]_d^{1/k} \geqslant \lim_{k \to \infty} (\beta (\lambda^*)^k U_d)^{1/k} = \lambda^* \ .$$

Hence, $V_d^\infty = \lambda^*$ holds for all $d \in D$.   □

Thanks to this lemma, we will speak of "value", without mentioning the initial state, when the entropy game is irreducible.

Every strongly connected component C of $\bar{G}$ with set of nodes $D_C \subset D$ yields a reduced game, in which the set of states of Despot is D_C, and Tribune only selects actions such that the next state of Despot will remain in D_C for at least one action chosen by People. Moreover, People chooses only actions so that the next state remains in D_C. By definition of $\bar{G}$, this reduced game is irreducible. We denote it by $\mathcal{E}[C]$. The following elementary observation allows us to reduce the general Despot-free case to the irreducible Despot-free case.

Lemma 18 *In a Despot-free entropy game, the value of a state d is the maximum of the value of the irreducible games $\mathcal{E}[C]$ corresponding to the different strongly connected components C of $\bar{G}$ to which d has access.*

Proof This follows from a more precise of Zijm, Theorem 5.1 in [50], which determines the asymptotic expansion of $F^k(e)$ as $k \to \infty$. However, the present lemma is more elementary. Alternatively, one can note that the operator $f := \log \circ F \circ \exp$ is precisely in the class of operators considered in [18, Section 4]. The lemma is a special case of Theorem 29 there.   □

Therefore, from now on, we make the following assumption.

Assumption 19 *The game is Despot-free and irreducible.*

We also make the following assumption.

Assumption 20 *The weights $m_{p,d}$ are integers.*

The case in which the weights are rational numbers reduces to this one (multiplying all weights by a common denominator does not change optimal positional strategies).

7.2 Reduction to a Well Conditioned Convex Programming Problem

Our strategy, to prove Theorem 15 when the game is irreducible is to apply the ellipsoid method to the convex programming formulation (32). To do so, we must replace this formulation by another convex program whose feasible set is included in a ball $B_2(a, R)$, (the Euclidean ball with center a and radius R), and contains a Euclidean ball $B_2(a, r)$, where $\log(R/r)$ is polynomially bounded in the size of the input. The following key lemma allows us to do so. It bounds the non-linear eigenvalue and eigenvector of the dynamic programming operator F, which have been shown to exist in Lemma 17. We set $W := \max_{(p,d) \in E} m_{p,d}$ and $n := |D|$.

Lemma 21 *Suppose the game is Despot-free and irreducible. Then, the value λ of the game is such that $1 \leqslant \lambda \leqslant nW$. Moreover, there exists a vector $U \in \operatorname{int} \mathbb{R}^n_+$ such that $F(U) = \lambda U$, and*

$$1 \leqslant U_d \leqslant \lambda^{n-1} \ , \forall d \in D \ . \tag{33}$$

Proof (1) The fact that $\lambda \leqslant nW$ follows from the first Collatz-Wielandt formula (29), which implies that $\lambda \leqslant \max_d F_d(e)$, where e is the unit vector of $\mathbb{R}^D$. We have $\max_d F_d(e) \leqslant nW$. Similarly, the last Collatz-Wielandt formula (31) implies that $\lambda \geqslant \min_d F_d(e)$. Since we assumed that every node in G has at at least one successor, we have $F_d(e) \geqslant 1$ for all $d \in D$, and so $\lambda \geqslant 1$.

(2) Let (d, d') be an arc in $\bar{G}$, corresponding to a path of length 1 in G, of the form (d, t, p, d'). Then, $m_{pd'} U_{d'} \leqslant F_d(U) = \lambda U_d$ holds, with $\lambda \leqslant nW$ and $m_{pd'}$ integer. In particular, $U_{d'} \leqslant \lambda U_d$ holds. Since the game is irreducible, any two vertices of $\bar{G}$ are connected by a path of length at most $n - 1$. It follows that $U_{d'}/U_d \leqslant \lambda^{n-1}$ holds for all $d, d' \in D$. We may assume that the minimal entry of U is equal to 1, by dividing U by this minimal entry. Then, $U_d \leqslant \lambda^{n-1}$ holds for all d. $\qquad \square$

We denote by $\mathcal{K}$ the set of pairs $(u, \mu) \in \mathbb{R}^D \times \mathbb{R} \simeq \mathbb{R}^{n+1}$, such that

$$f(u) \leqslant \mu e + u \tag{34a}$$

$$0 \leqslant u_d \leqslant (n-1)\lceil \log(nW) \rceil, \forall d \in D, \tag{34b}$$

$$0 \leqslant \mu \leqslant \lceil \log(nW) \rceil + 2 \ , \tag{34c}$$

where $\lceil t \rceil$ denotes the smallest integer greater than or equal to t, and f is given by (7), recalling that e denotes the unit vector of $\mathbb{R}^D$. Recall that $W \geqslant 1$ since this is an integer, and that if $n \neq 1$, then $(n-1)\lceil \log(nW) \rceil \geqslant 1$. By combining Corollary 14, Lemma 17 and Lemma 21, we arrive at the following result.

Proposition 22 *The value of a Despot-free irreducible entropy game coincides with the exponential of the value of the convex program:*

$$\min \mu, \ (u, \mu) \in \mathscr{K} \ , \tag{35}$$

where $\mathscr{K}$ is defined by (7.2).

Proof If $(u, \mu) \in \mathscr{K}$, then it satisfies (32), so by Corollary 14, it is not smaller than the logarithm of the value of the game. Hence, the value of (35) is an upper bound for the logarithm of the value of the game. Now, if we take λ to be equal to the value of the game, we know by Lemma 21 and Lemma 17 that $1 \leqslant \lambda \leqslant nW$ and that we can find a vector $U \in \operatorname{int} \mathbb{R}^n_+$ such that $F(U) = \lambda U$, and the bounds (33) on U hold. Then, setting $u := \log(U)$ and $\mu := \log \lambda$, we get that $(u, \mu) \in \mathscr{K}$. It follows that the exponential of the value of (35) coincides with the value of the game.

Finally, the convexity of $\mathscr{K}$ follows from the convexity of every coordinate map of f, which is an immediate consequence of Lemma 3. □

We denote by $B_2(a, r)$ the Euclidean ball with center a and radius r. The sup-norm ball with the same radius and center is denoted by $B_\infty(a, r)$. We have the following lemma.

Lemma 23 *Let $a = ((1/2)e, \lceil \log(nW) \rceil + 3/2) \in \mathbb{R}^D \times \mathbb{R}$, and let*

$$r := 1/3, \qquad R := \sqrt{n+1}((n-1)\log(nW) + n + 1) \ . \tag{36}$$

Then,

$$B_2(a, r) \subset \mathscr{K} \subset B_2(a, R) \ .$$

Proof Any point (u, μ) in $B_\infty(a, r)$ satisfies $(1/2 - r)e \leqslant u \leqslant (1/2 + r)e$ and $\lceil \log(nW) \rceil + 3/2 - r \leqslant \mu \leqslant \lceil \log(nW) \rceil + 3/2 + r$. Since $n \neq 1$, we get that $(n-1)\lceil \log(nW) \rceil \geqslant 1$, and since $r \leqslant 1/2$, we obtain that (u, μ) satisfies the box constraints (34b) and (34c) defining $\mathscr{K}$. Since f is order preserving and commutes with the addition of a constant,

$$
\begin{aligned}
f(u) &\leqslant f((1/2 + r)e) = (1/2 + r)e + f(0) \\
&\leqslant (1/2 + r + \lceil \log(nW) \rceil)e \leqslant (1/2 + r + \lceil \log(nW) \rceil)e + (u + (r - 1/2)e) \\
&= (2r + \lceil \log(nW) \rceil)e + u \leqslant (3r - 1 + \mu)e + u
\end{aligned}
$$

and so $f(u) \leqslant \mu e + u$ as soon as $r \leqslant 1/3$. We deduce that $B_2(a, 1/3) \subset B_\infty(a, 1/3) \subset \mathscr{K}$.

Moreover, since $\mathscr{K}$ is included in a box of width $\ell = (n-1)\lceil \log(nW) \rceil + 2$, for any choice of $a' \in \mathscr{K}$, $\mathscr{K}$ is included in the sup-norm ball $B_\infty(a', \ell)$, and so, in the euclidean ball $B_2(a', \ell\sqrt{n+1})$. It follows that $\mathscr{K} \subset B_2(a, R)$. □

7.3 Construction of a Polynomial Time Weak Separation Oracle

We shall solve Problem (35) by the ellipsoid method [20]. The latter needs the following notions.

Definition 1 Let $\mathcal{K}$ denote a convex body in $\mathbb{R}^q$. A *weak separation oracle* for $\mathcal{K}$ is a procedure, taking as input a rational number $\nu > 0$ and a rational vector $y \in \mathbb{R}^q$, which concludes one of the following: (i) asserting that y is at Euclidean distance at most ν from $\mathcal{K}$; (ii) finding an *approximate separating half-space of precision ν*, i.e., a linear form $\phi : x \mapsto c \cdot x$, with $c \in \mathbb{R}^q$, of Euclidean norm at least 1, such that for every $x \in \mathcal{K}$,

$$\phi(x) \leqslant \phi(y) + \nu \ .$$

Let us now recall the main complexity result about the ellipsoid method [20]. To do so, we denote by $\langle r \rangle$ the number of bits needed to code an object r, under the standard binary encoding. For instance, if r is an integer, $\langle r \rangle := \lceil \log_2(r) \rceil + 1$, if $r = p/q$ is a rational, $\langle r \rangle := \langle p \rangle + \langle q \rangle$, if $r = (r_i)$ is a rational vector, $\langle r \rangle := \sum_i \langle r_i \rangle$, and if ψ is a linear form with rational coefficients over $\mathbb{R}^q$, $\psi(x) = \sum_i r_i x_i$, for $x \in \mathbb{R}^q$, then $\langle \psi \rangle = \sum_i \langle r_i \rangle$. Here and after, the notion of length of an input refers to the binary encoding.

The ellipsoid method can be applied to solve the following problem consisting in finding an approximate minimum of precision ϵ of a linear form ψ with rational coefficients over a convex body $\mathcal{K} \subset \mathbb{R}^q$. This means looking for a vector x^* such that $d_2(x^*, \mathcal{K}) \leqslant \epsilon$ and $\psi(x^*) \leqslant \min_{x \in \mathcal{K}} \psi(x) + \epsilon$, where d_2 denotes the Euclidean distance. We assume that we know a vector $a \in \mathcal{K}$ with rational coordinates, and rational numbers $0 < r < R$ such that

$$B(a, r) \subset \mathcal{K} \subset B(a, R) \ .$$

In that case, the size of the input of the approximate minimization problem is measured by $\langle \psi \rangle + \langle a \rangle + \langle r \rangle + \langle R \rangle + \langle \epsilon \rangle$.

It is shown in [20] that if the convex set $\mathcal{K}$ admits a polynomial time weak separation oracle, the ellipsoid method computes an ϵ-approximate solution of the minimization problem in a time polynomial in the size of the input. Specialized to the present setting, and taking into account the polynomial estimates for $\log r$ and $\log R$ in Lemma 23, we get the following result.

Theorem 24 (Corollary of [20, Th. 3.1]) *Suppose that the set $\mathcal{K}$ defined by* (7.2) *admits a weak separation oracle which runs in polynomial time in the bitsize of the input and in the bitsize of the game. Then, the ellipsoid method returns an approximate optimal solution of precision ϵ of Problem* (35)*, i.e., a vector (u, μ) such that that $d_2((u, \mu), \mathcal{K}) \leqslant \epsilon$ and μ does not exceed the value of Problem* (35) *by more than ϵ, in a time that is polynomial in $\langle \epsilon \rangle + |E| + \log W$.*

Recall that $|\cdot|$ denotes the cardinality of a set, in particular $|E|$ denotes the number of arcs of G.

The following result allows us to apply Theorem 24 to Problem (35).

Proposition 25 *The convex set $\mathcal{K}$ defined by* (7.2) *admits a weak separation oracle which runs in polynomial time.*

To show this proposition, we need a series of arguments. Some of these arguments, like the next lemma, are standard, whereas other arguments require some transparent but rather technical bookkeeping, exploiting the non-expansive character of f to control the approximation errors.

Lemma 26 *Let $\epsilon > 0$ and t be rational numbers, and assume first that $t \leqslant 0$. Then, a rational approximation of absolute precision ϵ of $\exp(t)$ can be computed in a time that is polynomial in $\langle t \rangle$ and $\langle \epsilon \rangle$. Assume now that $t > 0$. Then, a rational approximation of absolute precision ϵ of $\log(t)$ can be computed in a time that is polynomial in $\langle t \rangle$ and $\langle \epsilon \rangle$.*

Proof It is shown in [7] that the conclusion is true when the input belongs to a fixed compact subset of the intervals $(-\infty, 0]$, in the case of exp, or of $(0, \infty)$, in the case of log. The fact that the same property still holds for the whole intervals $(-\infty, 0]$ and $(0, \infty)$ follows from the range reduction techniques [35]. $\square$

Lemma 27 *Let x be a vector in $\mathbb{R}^D$ with rational entries, and let $\epsilon > 0$ be a rational number. An approximation of $f(x)$ with a sup-norm error not exceeding ϵ can be obtained in polynomial time in $\langle x \rangle + \langle \epsilon \rangle + |E| + \log W$.*

Proof We have

$$f_d(x) = \max_{(\sigma(d), p) \in E} \log\left(\sum_{(p,d') \in E} m_{pd'} \exp(x_{d'}) \right) .$$

Hence, it suffices to check that for every $p \in P$, the value

$$h(x) := \log\left(\sum_{(p,d') \in E} m_{pd'} \exp(x_{d'}) \right)$$

can be approximated with a precision ϵ within a polynomial time. We set $\bar{x} := \max_{d' \in D} x_{d'}$, and make the change of variables $x_{d'} = \bar{x} + \tilde{x}_{d'}$, so that

$$h(x) = \bar{x} + \log t, \ t := \sum_{(p,d') \in E} m_{pd'} \exp(\tilde{x}_{d'})$$

We observe that $1 \leqslant t$ and that log has Lipschitz constant 1 over $[1, \infty)$. Hence, to evaluate $h(x)$ with a precision ϵ, it suffices to compute an approximation $\tilde{t}$ of t with precision $\epsilon/2$, which can be done in polynomial time thanks to Lemma 26, and then to approximate $\log \tilde{t}$ with precision $\epsilon/2$, which can also be done in polynomial time by the same lemma. $\square$

Proof of Proposition 25 Let $\nu > 0$. Our aim is to check whether a given pair $(\bar{\nu}, \bar{\mu})$ is at distance at most ν from $\mathcal{K}$. Since we already showed that we can get an

approximation of f in polynomial time, the proof will be a matter of routine book-keeping (except perhaps the use of the subdifferential of f to construct a separating halfspace).

We denote by $\epsilon > 0$ a rational number, $\epsilon \leqslant 1$, which we shall fix in the course of the proof.

We provide the announced separation oracle. We first check that every box con-straint, as well as the non-linear constraints $f_d(\bar{v}) \leqslant \bar{v}_d + \bar{\mu}$, for $d \in D$, are satisfied up to a precision ϵ, which can be done in polynomial time in $\langle v \rangle + \langle \epsilon \rangle + |E| + \log W$ thanks to Lemma 27.

(i) If these constraints are satisfied up to a precision ϵ, we have $-\epsilon \leqslant \bar{v}_d \leqslant (n-1)\lceil \log(nW) \rceil + \epsilon$, $-\epsilon \leqslant \bar{\mu} \leqslant \lceil \log(nW) \rceil + 2 + \epsilon$, and $f(\bar{v}) \leqslant (\bar{\mu} + \epsilon)e + \bar{v}$. Setting $\tilde{v}_d = \min(\max(\bar{v}_d, 0), \lceil (n-1)\log(nW) \rceil)$, we get that $\|\bar{v} - \tilde{v}\|_\infty \leqslant \epsilon$ with $\tilde{v}$ satisfying the constraint (34b). Using the fact that f is nonexpansive in the sup-norm, we deduce that $f(\tilde{v}) \leqslant \tilde{\mu}e + \tilde{v}$, where $\tilde{\mu} = \bar{\mu} + 3\epsilon$. Moreover, $0 \leqslant \tilde{\mu} \leqslant \lceil \log(nW) \rceil + 2 + 4\epsilon$. Now, since f is also convex, for all $t \in [0, 1]$, we have

$$f(t\tilde{v}) \leqslant tf(\tilde{v}) + (1-t)f(0) \leqslant \tilde{\mu}'e + t\tilde{v} \;,$$

with $\tilde{\mu}' = t\tilde{\mu} + (1-t)\lceil \log(nW) \rceil$. Hence, taking $t = 1/(1+2\epsilon)$, we get that $\tilde{\mu}'$ satisfies the constraint (34c), whereas $\tilde{v}' = t\tilde{v}$ still satisfies the constraint (34b), so that $(\tilde{v}', \tilde{\mu}')$ belongs to $\mathcal{K}$. Using $t \leqslant 2\epsilon$, we also have $\|(\bar{v}, \bar{\mu}) - (\tilde{v}', \tilde{\mu}')\|_\infty \leqslant \epsilon L$, with $L = (5 + 2(n-1)\lceil \log(nW) \rceil)$, implying that $d_2((\bar{v}, \bar{\mu}), \mathcal{K}) \leqslant L\sqrt{n+1}\epsilon$. Hence, we shall require that

$$\epsilon \leqslant \epsilon_1 := \frac{\nu}{(5 + 2(n-1)\lceil \log(nW) \rceil)\sqrt{n+1}} \;,$$

to make sure that $d_2((\bar{v}, \bar{\mu}), \mathcal{K}) \leqslant \nu$.

(ii) Assume now that one of the box constraints is violated by more than ϵ. Then, one of the linear forms $(v, \mu) \mapsto \pm v_d$ or $(v, \mu) \mapsto \pm\mu$ provides a separating half-space, and the norm of this linear form is 1. Assume finally that all the box constraints are satisfied up to ϵ, and that one of the non-linear constraints is violated by more than ϵ. Let us write this constraint as

$$g(v, \mu) := \log\left(\sum_{d' \in D} m_{pd'} \exp(v_{d'}) \right) - v_d - \mu \leqslant 0$$

for some $d \in D$ and $(\sigma(d), p) \in E$, so that $g(\bar{v}, \bar{\mu}) > \epsilon$. Since g is convex, the differential ϕ of g at point $(\bar{v}, \bar{\mu})$ satisfies

$$\phi(v - \bar{v}, \mu - \bar{\mu}) \leqslant g(v, \mu) - g(\bar{v}, \bar{\mu}) \leqslant -\epsilon$$

for all $(v, \mu) \in \mathcal{K}$, i.e.,

$$\phi(v, \mu) \leqslant \phi(\bar{v}, \bar{\mu}) - \epsilon, \; \forall(v, \mu) \in \mathcal{K}$$

showing that ϕ is a separating half-space. However, we need an approximate half-space given by a linear form with rational coefficients, which we next construct by approximating ϕ.

To do so, we first compute the differential of g at point $(\bar{v}, \bar{\mu})$. This is the linear form

$$\phi : (x, y) \in \mathbb{R}^D \times \mathbb{R} \mapsto \sum_{d'} x_{d'} m_{pd'} \exp(\bar{v}_{d'}) / \left(\sum_{d''} m_{pd''} \exp(\bar{v}_{d''}) \right) - x_d - y \ .$$

The maximum of $\bar{v}$ can be subtracted to every coordinate of $\bar{v}$ without changing this linear form. Then, by Lemma 26, the coefficients of this linear form can be approximated in polynomial time. It follows that we can compute an approximation $\tilde{\phi}$ of ϕ of precision ϵ in polynomial time in $\langle \bar{v} \rangle + \langle \epsilon \rangle$. Observe that the coefficient of the variable y in the linear form ϕ is always equal to -1. Hence, the approximate linear form $\tilde{\phi}$ can be chosen with the same coefficient, and then, $\tilde{\phi}$ is of norm at least 1.

Since any element (v, μ) of $\mathcal{K}$ satisfies the box constraints (34b) and (34c), whereas $(\bar{v}, \bar{\mu})$ satisfies these constraints up to $\epsilon \leqslant 1$, we get that

$$|(v, \mu) - (\bar{v}, \bar{\mu})|_\infty \leqslant M := (n - 1)\lceil \log(nW) \rceil) + 3 \ ,$$

hence

$$|\tilde{\phi}(v - \bar{v}, \mu - \bar{\mu}) - \phi(v - \bar{v}, \mu - \bar{\mu})| \leqslant M(D + 1)\epsilon, \forall (v, \mu) \in \mathcal{K} \ .$$

So it suffices that

$$\epsilon \leqslant \epsilon_2 := \frac{\nu}{M(D + 1)}$$

to make sure that $\tilde{\phi}$ defines an approximate separating half-space of precision ν.

To summarize, it suffices to take $\epsilon = \min(\epsilon_1, \epsilon_2)$ in the previous analysis, so that the conditions of Definition 1 are satisfied. Moreover, for this choice, all the computations take a polynomial time in $\langle \nu \rangle + \langle \bar{v} \rangle + \langle \bar{\mu} \rangle$, and the size $|E| + \log W$ of the description of the game. $\qquad \square$

7.4 Using Separation Bounds Between Algebraic Numbers to Compare Policies

It follows from Theorem 24 that we can compute in polynomial time an approximate solution of Problem (35) with precision ϵ. We next show that it is possible to choose ϵ with a polynomial number of bits, in such a way that this approximate solution allows us to identify an optimal policy. We shall actually prove a version of this result in the more general two-player case. This extended version, stated as Corollary 29, will apply both to the Despot-free case, and to the case of entropy games with a fixed number of significant states, see Section 7.6. To prove it, we rely on separation bounds for algebraic numbers.

Theorem 28 [41] *Let p be a univariate polynomial of degree n with integer coefficients, possibly with multiple roots. Let S be the sum of the absolute values of its coefficients. Then, the distance between any two distinct roots of p is at least*

$$(2n^{\frac{n}{2}+2}(S + 1)^n)^{-1} \ .$$

To any given pair (δ, τ) of policies, one can associate a directed sub-graph $G(\delta, \tau)$ of G, obtained, by erasing for all $d \in D$, every arc in $\{(d, t) \in E\}$ except $\{d, \delta(d)\}$, and similarly, by erasing for all $t \in T$, every arc in $\{(t, p) \in E\}$ except $\{t, \tau(t)\}$. The dynamic programming operator of the game associated to this sub-graph $G(\delta, \tau)$ coincides with the conjugate ${}^\tau F^\delta \tau := \exp \circ^\tau f^\delta \circ \log$ of (12) and is equal to the linear operator with matrix ${}^\tau M^\delta \in \mathbb{N}^{D \times D}$:

$$ {}^\tau F^\delta : \mathbb{R}^D \to \mathbb{R}^D, \; X \mapsto {}^\tau M^\delta X \;, \tag{37} $$

where the entry (d, d') of ${}^\tau M^\delta$ is equal to $m_{\tau \circ \delta(d), d'}$ when $(\tau \circ \delta(d), d') \in E$ and to zero otherwise. Then, the value of the entropy game starting in state d, $R_d^\infty(\delta, \tau)$ coincides with the maximum of the Perron-roots (that is the positive eigenvalues which coincide with the spectral radii) of the submatrices of ${}^\tau M^\delta$ with nodes in a strongly connected component to which d has access in the graph $G(\delta, \tau)$. The value of the original entropy game coincides with the Perron-root of ${}^\tau M^\delta$.

Corollary 29 *There exists a rational function $(n, W) \mapsto \eta_{sep}(n, W) > 0$ such that for every two different pairs (δ, τ) and (δ', τ') which yield different values of the entropy game, these two values differ by at least $\eta_{sep}(n, W)$ and*

$$ \eta_{sep}(n, W) \geqslant \exp(-poly(n + \log W)) $$

where the polynomial inside the exponential is independent of the input.

Proof Given a pair (δ, τ) of policies, the values of the entropy game are eigenvalues of the matrix $A = {}^\tau M^\delta \in \mathbb{N}^{D \times D}$. Observe that the entries of A are integers bounded by W. Let f_A be the characteristic polynomial of A. The coefficient of the monomial of degree $n - k$ in f_A is the sum of the C_n^k principal minors of A of size k. By Hadamard's inequality, each absolute value of these minors is at most $(\sqrt{n} W)^k$, and so, every coefficient of f_A has an absolute value bounded by $C_n^k (\sqrt{n} W)^k$ and their sum is $\leqslant (2\sqrt{n} W)^n$. Two different pairs of strategies yield two characteristic polynomials, f_A and f_B, whose product is of degree $2n$ and whose sum of absolute value of coefficients is bounded by the product of such bounds for f_A and f_B, so by $(2\sqrt{n} W)^{2n}$. Therefore, the size S appearing in Theorem 28 is bounded by $(2\sqrt{n} W)^{2n}$. We deduce that if the two pairs of strategies yield distinct values, the distance between these values is at least

$$ \eta_{sep}(n, W) := (2(2n)^{n+2}((2\lceil \sqrt{n} \rceil W)^{2n} + 1)^{2n})^{-1} \;. $$

This number is rational and satisfies

$$ \eta_{sep}(n, W) \geqslant \exp(-poly(n + \log W)), $$

for some polynomial function poly. Since the above lower bound is true for every two pairs of different policies (δ, τ) and (δ', τ'), we obtain the result. $\qquad \square$

Hence, if two policies of Tribune yield different values λ and λ', then, $|\lambda - \lambda'|$ is bounded below by the rational number $\eta_{\text{sep}} > 0$ whose number of bits is polynomially bounded in the size of the input.

7.5 Synthesis of an Optimal Strategy of Tribune from an Approximate Solution of the Convex Program in Proposition 22

To any policy τ of Tribune, we associate a dynamic programming operator $^\tau F$, which is the specialization of the map $^\tau F^\delta$ of previous section to the case where $\delta = \sigma$. This is the self-map of $\mathbb{R}^D$ defined by

$$
^\tau F_d(X) = \sum_{(\tau(\sigma(d)), d') \in E} m_{\tau(\sigma(d))d'} X_{d'} \ .
$$

So $^\tau F(X) = {}^\tau M X$, where $^\tau M = {}^\tau M^\sigma$ is the $|D| \times |D|$ matrix with nonnegative entries equal to $m_{\tau(\sigma(d))d'}$ when $(\tau(\sigma(d)), d') \in E$ and zero otherwise.

7.5.1 The Simpler Situation in Which Every Policy τ of Tribune Yields an Irreducible Matrix

To explain our method, we make first the restrictive assumption that for every policy τ of Tribune, the matrix $^\tau M$ is irreducible. In particular, we can take an optimal policy τ^*. By a standard result of Perron-Frobenius theory [12], $^{\tau^*}M$ has a left eigenvector π with positive entries, associated to the spectral radius $\lambda^{\tau^*} := \rho(^{\tau^*}M)$, called Perron root. Hence, $\pi^{\tau^*}M = \lambda^{\tau^*}\pi$. Since τ^* is optimal, $\lambda^{\tau^*} = \lambda^*$, where λ^* is the value of the entropy game starting from any node $d \in D$, see Lemma 17. Moreover, by applying Lemma 21 to the linear map $U \mapsto (^{\tau^*}M)^T U$, where ^{T} denotes the transposition, we deduce that $\pi_d/\pi_{d'} \leqslant (nW)^{n-1}$.

For any rational number $\epsilon > 0$, the ellipsoid algorithm, applied to the optimization problem of Proposition 22, yields in polynomial time a vector u and a scalar μ such that $\mu \leqslant \log \lambda^* + \epsilon$ and $d_2((u, \mu), \mathcal{K}) \leqslant \epsilon$. So there exists $(\tilde{u}, \tilde{\mu}) \in \mathcal{K}$ such that $\|u - \tilde{u}\|_\infty \leqslant \epsilon$ and $|\mu - \tilde{\mu}| \leqslant \epsilon$. Since $(\tilde{u}, \tilde{\mu}) \in \mathcal{K}$, and $\log \lambda^*$ is the value of (35), we deduce that $\log \lambda^* \leqslant \tilde{\mu} \leqslant \mu + \epsilon$, so $\lambda^* \exp(-\epsilon) \leqslant \exp(\mu) \leqslant \lambda^* \exp(\epsilon)$. Using (7.2), and assuming that $\epsilon \leqslant 1$, we deduce that $u_d - u_{d'} \leqslant (n-1)\lceil \log(nW) \rceil + 2\epsilon \leqslant (n-1)(\log(nW) + 1) + 2$, for all $d, d' \in D$. Using the nonexpansivity of f, we also obtain that $f(u) \leqslant f(\tilde{u}) + \epsilon e \leqslant \tilde{u} + (\tilde{\mu} + \epsilon)e \leqslant (\mu + 3\epsilon)e + u \leqslant (\log \lambda^* + 4\epsilon)e + u$. Taking $U := (U_d)_{d \in D}$ with $U_d := \exp(u_d)$, we get $F(U) \leqslant \lambda^* \exp(4\epsilon)U$ and $U_d/U_{d'} \leqslant (enW)^{n-1}e^2$.

We choose any policy $\underline{\tau}$ such that $F(U) = {}^{\underline{\tau}}M U$. Therefore,

$$
\underline{\tau}(\sigma(d)) \in \operatorname*{argmax}_{\tau \in \mathcal{P}_T} \sum_{(\tau(\sigma(d)), d') \in E} m_{\tau(\sigma(d))d'} U_{d'} \ .
$$

We claim that $\underline{\tau}$ is optimal if ϵ is sufficiently small.

To show the latter claim, we observe that $^{\tau^*}MU \leqslant F(U)$. For all $d \in D$,

$$
\begin{aligned}
0 \leqslant \pi_d(\lambda^* \exp(4\epsilon)U_d - F_d(U)) & \leqslant \pi_d(\lambda^* \exp(4\epsilon)U_d - (^{\tau^*}MU)_d) \\
& \leqslant \sum_{d' \in D} \pi_{d'}(\lambda^* \exp(4\epsilon)U_{d'} - (^{\tau^*}MU)_{d'}) \\
& = \pi(\lambda^* \exp(4\epsilon)U - {}^{\tau^*}MU) \\
& = \lambda^*(\exp(4\epsilon) - 1)\pi U \ .
\end{aligned}
\tag{38}
$$

Using $\pi_d/\pi_{d'} \leqslant (nW)^{n-1}$ and $U_d/U_{d'} \leqslant (enW)^{n-1}e^2$, we deduce that $\pi U \leqslant \pi_d U_d(1 + (n-1)(enW)^{2(n-1)}e)$, so $F(U) \geqslant \underline{\lambda}U$, where $\underline{\lambda} := \lambda^*[1 - (\exp(4\epsilon) - 1)(n-1)e(enW)^{2(n-1)}]$. In view of the formula of $\underline{\lambda}$, we can choose $\epsilon > 0$, with a polynomially bounded number of bits, such that $\underline{\lambda} > \lambda^* - \cdot_{\text{sep}}$. Since, $^{\underline{\tau}}MU \geqslant \underline{\lambda}U$, we have $\rho(^{\underline{\tau}}M) \geqslant \underline{\lambda}$ and so $\rho(^{\underline{\tau}}M) > \lambda^* - \eta_{\text{sep}}$. Since λ^* is the maximum of the values of all the policies, $\rho(^{\underline{\tau}}M) \leqslant \lambda^*$. By definition of the separation parameter η_{sep} given in Corollary 29, this implies that $\rho(^{\underline{\tau}}M) = \lambda^*$, and so the policy $\underline{\tau}$ of Tribune which we just constructed is optimal, showing the claim.

In the preceding argument, the computation (38) may look a bit magic at the first sight, it should become intuitive if one interprets it as an approximate complementary slackness condition for the semi-infinite program of Proposition 22, the invariant measure π playing the role of a Lagrange multiplier.

When some policies τ yield a *reducible* matrix $^\tau M$, the synthesis of the optimal policy $\underline{\tau}$ still exploits the same idea with an additional technicality, since we can only guarantee that the inequality $F_d(U) \geqslant \underline{\lambda}U_d$ is valid for every state d such that $\pi_d > 0$. We explain the more technical argument in the next section.

7.5.2 Synthesis of an Optimal Strategy of Tribune, in General

Recall that if M is a reducible nonnegative matrix, a *class* of M is a strongly connected component of the directed graph of M, and that this class is *basic* if the $B \times B$ submatrix of M, denoted by M_{BB}, has Perron root $\rho(M)$. It is known [12] that M has always a basic class, and a nonnegative left eigenvector associated with $\rho(M)$. Moreover, choosing a basic class B which is final among the basic classes of M, that is such that the set S of nodes $d' \in D$ that are reachable in the directed graph of M starting from some node in the basic class B does not contain any node of another basic class, then there exists a nonnegative left eigenvector π so that its support $\{d \mid \pi_d \neq 0\}$ coincides with S. We shall assume that π, S and B satisfy these properties, for $M = {}^{\tau^*}M$ corresponding to an optimal policy τ^*. We set $N := D \setminus B$, and for any $D \times D$ matrix M, any vector $u \in \mathbb{R}^D$, and any subsets F and G of D, we denote by M_{FG} the $F \times G$ submatrix of M and by v_F the vector of $\mathbb{R}^F$ given by $v_F := (v_d)_{d \in F}$. Since B is a basic class and B has access to any element of S, we get that no element of $S \setminus B$ has access to an element of B, and since π equals zero outside S, we get that the restriction of $\pi^{\tau^*}M$ to B equals $\pi_B{}^{\tau^*}M_{BB}$ and

so $\pi_B{}^{\tau^*}M_{BB} = \lambda^*\pi_B$. The same computation as in Section 7.5.1 restricted to the elements $d \in B$ now gives

$$
\begin{aligned}
0 \leqslant \pi_d(\lambda^* \exp(4\epsilon)U_d - F_d(U)) &\leqslant \pi_d(\lambda^* \exp(4\epsilon)U_d - ({}^{\tau^*}MU)_d) \\
&\leqslant \pi_B(\lambda^* \exp(4\epsilon)U_B - ({}^{\tau^*}M_{BB}U_B + {}^{\tau^*}M_{BN}U_N)) \\
&= \lambda^*(\exp(4\epsilon) - 1)\pi_B U_B - \pi_B{}^{\tau^*}M_{BN}U_N \ .
\end{aligned}
$$

The bounds on U obtained in Section 7.5.1 are still valid, and the ones of $\pi_d/\pi_{d'}$ are valid only for $d, d' \in B$ using $\pi_B{}^{\tau^*}M_{BB} = \lambda^*\pi_B$ and the irreducibility of ${}^{\tau^*}M_{BB}$. We deduce that $F_d(U) \geqslant \underline{\lambda}U_d$ for all $d \in B$, for the same $\underline{\lambda}$ as in Section 7.5.1. Moreover, $\pi_B{}^{\tau^*}M_{BN}U_N \leqslant \lambda^*(\exp(4\epsilon) - 1)\pi_B U_B$. We define $\epsilon' := \lambda^*(\exp(4\epsilon) - 1)en(enW)^{2(n-1)}$, so that ${}^{\tau^*}M_{dN}U_N \leqslant \epsilon'U_d$ for all $d \in B$, where ${}^{\tau^*}M_{dN}$ is the dth line of the matrix ${}^{\tau^*}M_{BN}$.

We first choose, any policy $\underline{\tau}$ and set B' such that ${}^{\underline{\tau}}M_{dD}U \geqslant \underline{\lambda}U_d$ and ${}^{\underline{\tau}}M_{dN'}U_{N'} \leqslant \epsilon'U_d$, for all $d \in B'$, with $N' = D \setminus B'$. We know from the above analysis that there is always at least one policy $\underline{\tau}$ and set B' with this property (namely τ^* and $B' = B$). Moreover, such a policy and set can be obtained by the following algorithm. Indeed, let us start from any policy $\underline{\tau}$ such that $F(U) = {}^{\underline{\tau}}MU$, and choose B' as the set of $d \in D$ such that ${}^{\underline{\tau}}M_{dD}U \geqslant \underline{\lambda}U_d$. Then, $B \subset B'$ since ${}^{\underline{\tau}}M_{dD}U = F_d(U)$. At each step of the algorithm, one applies the following operations to each $d \in B'$ in some order: set $N' = D \setminus B'$ and check if ${}^{\underline{\tau}}M_{dN'}U_{N'} \leqslant \epsilon'$. If this does not hold, change $\tau(\sigma(d))$ to any action so that ${}^{\underline{\tau}}M_{dD}U \geqslant \underline{\lambda}U_d$ and ${}^{\underline{\tau}}M_{dN}U_N \leqslant \epsilon'$. If this is impossible, then eliminate d from B' and continue. Then stop at any step in which B' does not change. Since the cardinality of B' decreases by one at each step to which one does not stop and $B' \supset B$, we get that the algorithm stops after at most n iterations and needs at most n^2 products of a matrix by a vector, so it takes a polynomial time. Moreover at each step and so at the end of the algorithm, we have $B' \supset B$ and $N' \subset N$.

We deduce that ${}^{\underline{\tau}}M_{B'B'}U_{B'} \geqslant (\underline{\lambda} - \epsilon')U_{B'}$, showing that $\rho({}^{\underline{\tau}}M) \geqslant \rho({}^{\underline{\tau}}M_{B'B'}) \geqslant (\underline{\lambda} - \epsilon')$. In view of the formula of $\underline{\lambda}$ and ϵ' we can always choose $\epsilon > 0$, with a polynomially bounded number of bits, such that $\underline{\lambda} - \epsilon' > \lambda^* - \eta_{\text{sep}}$. Hence, $\rho({}^{\underline{\tau}}M) = \rho({}^{\underline{\tau}}M_{B'B'}) = \lambda^*$, since $\rho({}^{\underline{\tau}}M)$ is an eigenvalue of ${}^{\underline{\tau}}M$. We also deduce from ${}^{\underline{\tau}}M_{B'B'}U_{B'} \geqslant (\underline{\lambda} - \epsilon')U_{B'}$ that every state $d \in B'$ has value λ^*. Finally, since the game is irreducible, we can always replace $\underline{\tau}(\sigma(d))$ for $d \notin B'$ to make B' accessible from any initial state, so that the policy $\underline{\tau}$ is optimal. This concludes the proof of Theorem 15.

7.6 Derivation of Theorem 16 from Theorem 15

By Corollary 9,

$$
V_d^\infty = \min_{\delta \in \mathcal{P}_D} V_d^\infty(\delta, *)
$$

Observe that $|\mathcal{P}_D| \leqslant |E|^s$, where s is the number of significant states for despot, hence, if s is fixed, this minimum involves a polynomial number of terms. Thanks to the separation bound given in Corollary 29, it suffices to compute an approximation of each $V_d^\infty(\delta, *)$ for some $\epsilon > 0$ such that $\log \epsilon$ is polynomially bounded in the

size of the input, to make sure that a policy δ achieving the minimum in the above expression, in which every term $V_d^\infty(\delta, *)$ is replaced by its approximate value, is an optimal policy.

8 Multiplicative Policy Iteration Algorithm and Comparison with the Spectral Simplex Method of Protasov

We now consider the question of solving entropy games in practice.

8.1 Algorithms

The equivalence between entropy games and some special class of stochastic mean payoff games, through logarithmic glasses (see Section 3), will allow us to adapt classical algorithms for one or two player zero sum games, such as the value iteration and the policy iteration algorithm. We next present a multiplicative version of the policy iteration algorithm, which follows by adapting policy iteration ideas for two player games by Hoffman and Karp [23], with "multiplicative" policy iteration techniques of Howard and Matheson [24], Rothblum [40] and Sladky [43], The latter "multiplicative" policy iteration techniques apply to the Despot-free case. For clarity, we shall explain first policy iteration in the special Despot-free case: this is more transparent, and this also will allow us to interpret Protasov's spectral simplex method [38] as a variant of policy iteration. The newer part here is the two player case, which is dealt with in Algorithm 9.

We assume that $D = T = \{1, \ldots, |T|\}$ and σ is the identity in (32) . Let $^\tau F$ and $^\tau M$, $\tau \in \mathcal{P}_T$, be defined as in the previous section. If $^\tau M$ is irreducible, in particular if all its entries are positive, $^\tau M$ has an eigenvector $X^\tau > 0$, associated to the Perron root $\lambda^\tau := \rho(^\tau M)$. Moreover, X^τ is unique up to a multiplicative constant and is called a Perron eigenvector. If all the matrices $^\tau M$, $\tau \in \mathcal{P}_T$ are irreducible, one can construct a multiplicative version of the policy iteration, Algorithm 1.

Algorithm 1 Multiplicative policy Iteration for Despot-free entropy games.

1: Initialize $k = 1$, τ^0, $\tau^1 \neq \tau^0$ randomly.
2: **while** $\tau^k \neq \tau^{k-1}$ **do**
3:　　Compute the Perron root λ^{τ^k} and a Perron eigenvector X^{τ^k} of $^{\tau^k}M$.
4:　　Compute a new policy τ^{k+1} such that, for all $t \in T$,

$$\tau^{k+1}(t) \in \operatorname*{argmax}_{p \in P,\ (t,p) \in E} \sum_{t' \in T,\ (p,t') \in E} m_{p,t'} X_{t'}^{\tau^k} ,$$

　　and set $\tau^{k+1}(t) = \tau^k(t)$ if this choice is compatible with the former condition.
5:　　$k \leftarrow k + 1$
6: **end while**
7: **return** the optimal policy τ^k, the Perron root λ^{τ^k} and Perron eigenvector X^{τ^k} of $^{\tau^k}M$.

The following result shows that Algorithm 1 does terminate. The proof relies on a multiplicative version of the classical strict monotonicity argument in policy iteration, which was already used by Howard and Matheson [24], Rothblum [40] and Sladky [43]. We reproduce the short proof for completeness.

Proposition 30 *Consider Algorithm 1, where the computations are performed in exact arithmetics and all the matrices $^\tau M$, $\tau \in \mathcal{P}_T$ are supposed to be irreducible. Then, the sequence λ^{τ^k} is increasing as long as $\tau^k \neq \tau^{k-1}$. Moreover, the algorithm ends after a finite number of iterations k, and λ^{τ^k} is the value of the game (at any initial state).*

Proof The property that λ^{τ^k} is increasing uses the general property that for an irreducible nonnegative matrix M, and for a positive vector u, $Mu \geqslant \lambda u$ (component-wise) and $Mu \neq \lambda u$ implies $\rho(M) > \lambda$. Then, the algorithm terminates since the number of policies τ is a finite set, and so $\{\lambda^\tau \mid \tau \in \mathcal{P}_T\}$ is finite. When $\tau^k = \tau^{k-1}$, we get $F(X^{\tau^k}) = \lambda^{\tau^k} X^{\tau^k}$, and Lemma 17 shows that λ^{τ^k} is the value of the game (at any initial state).   □

Algorithm 1 has a dual version, in which maximization is replaced by minimization, in order to solve the Tribune-free setting of entropy games. For this dual version of Algorithm 1, the sequence λ^{τ^k} is decreasing as long as $\tau^k \neq \tau^{k-1}$. This uses the property (dual to the previous one) that for an irreducible nonnegative matrix M, and for a positive vector u, $Mu \leqslant \lambda u$ and $Mu \neq \lambda u$ implies $\rho(M) < \lambda$. Then the algorithm terminates as for the primal version.

In practice, Algorithm 1 can only be implemented in an approximate way. A bottleneck in this algorithm is the computation of the Perron root and Perron eigenvector. The later can be computed by standard double precision algorithms, like the QR method. The latter method requires $O(D^3)$ flops, where D is the size of the matrix $^\tau M$ associated to a policy τ. (Note that such complexity estimates in an informal "floating point" arithmetic model are meaningful only for well conditioned instances, in contrast with the Turing-model complexity estimates that we derived unconditionnally in Section 7.) One may also use a more scalable algorithm, like the power algorithm. In fact, we shall see in Algorithm 5 that the power idea can be applied directly and in a simpler way to solve the non-linear equation, avoiding the recourse to policy iteration. So it is not clear that the power algorithm will be the best choice to compute the eigenpair in situations in which Algorithm 1 is competitive. In the experiments which follow, we used the QR method in Algorithm 1.

In [38], Protasov introduced the Spectral Simplex Algorithm. His algorithm is a variant of Algorithm 1 in which at every iteration the policy is improved only at *one* state, which is the first state t such that $F_t(X^{\tau^k}) > \lambda^{\tau^k} X_t^{\tau^k}$. We shall also consider another version of Algorithm 1, in which we also change the policy at only one state t, but we choose it in order to maximize the expression $F_t(X^{\tau^k}) - \lambda^{\tau^k} X_t^{\tau^k}$. We shall refer to this algorithm as "Spectral Simplex-D" since this is analogous to Dantzig's pivot rule in the original simplex method [49].

The Spectral Simplex Algorithm introduced by Protasov in [38] is described in the Despot-free setting in Algorithm 2.

Algorithm 2 Spectral simplex algorithm [38].

1: Initialize $k = 1$, $\tau^0, \tau^1 \neq \tau^0$ randomly.
2: **while** $\tau^k \neq \tau^{k-1}$ **do**
3:      Compute the Perron root λ^{τ^k} and a Perron eigenvector X^{τ^k} of $^{\tau^k}M$.
4:      Set $\tau^{k+1}(t) = \tau^k(t)$ for $t \in T = \{1, \ldots, |T|\}$.
5:      bool = true, $t = 1$.
6:      **while** (bool) and $(t \leq |T|)$ **do**
7:          **if** $^{\tau^k}F_t(X^{\tau^k}) \neq \max_{p \in P, (t,p) \in E} \sum_{t' \in T, (p,t') \in E} m_{p,t'} X_{t'}^{\tau^k}$ **then**
8:          Choose

$$\tau^{k+1}(t) \in \underset{p \in P, (t,p) \in E}{\arg\max} \sum_{t' \in T, (p,t') \in E} m_{p,t'} X_{t'}^{\tau^k} \ .$$

9:             bool = false.
10:          **end if**
11:          $t \leftarrow t + 1$.
12:      **end while**
13:      $k \leftarrow k + 1$
14: **end while**
15: **return** the first policy τ^k such that $\tau^{k-1} = \tau^k$, the Perron root λ^{τ^k} and a Perron eigenvector X^{τ^k} of the matrix $^{\tau^k}M$.

The Spectral Simplex Algorithm with Dantzig update is shown in Algorithm 3, again in the Despot-free setting.

Algorithm 3 Spectral simplex algorithm for despot free game - dantzig update.

1: Initialize $k = 1$, $\tau^0, \tau^1 \neq \tau^0$ randomly.
2: **while** $\tau^k \neq \tau^{k-1}$ **do**
3:      Compute the Perron root λ^{τ^k} and a Perron eigenvector X^{τ^k} of $^{\tau^k}M$.
4:      Set $\tau^{k+1}(t) = \tau^k(t)$ for $t = 1, \ldots, |T|$.
5:      Compute the vector

$$(\epsilon_t)_{1 \leq t \leq |T|} = (F_t(X^{\tau^k}) - \lambda^{\tau^k} X_t^{\tau^k})_{1 \leq t \leq |T|}.$$

6:      Choose $t^* \in \arg\max_{1 \leq t \leq |T|} \epsilon_t$.
7:      Choose

$$\tau^{k+1}(t^*) \in \arg \max_{p \in P, (t^*,p) \in E} \sum_{t' \in T, (p,t') \in E} m_{p,t'} X_{t'}^{\tau^k} \ .$$

8:      $k \leftarrow k + 1$
9: **end while**
10: **return** the first policy τ^k such that $\tau^{k-1} = \tau^k$, the Perron root λ^{τ^k} and a Perron eigenvector X^{τ^k} of the matrix $^{\tau^k}M$.

One can also adapt the Algorithms 2 and 3 to the Tribune-free setting, again using minimization instead of maximization and replacing $^{\tau}F$ by F^δ.

8.2 Numerical Experiments

We next report numerical experiments in the case of Despot-free in order to compare Protasov's spectral simplex algorithm (with and without the improvement of Dantzig' pivot rule) with the multiplicative policy iteration algorithm (Algorithm 1). In the log-log Figs. 1 and 2, these algorithms are respectively named "Policy Iteration", "Spectral Simplex" and "Spectral Simplex-D".

We constructed random Despot-free instances in which $D = T$ has cardinal n, and every coordinate of the operator is of the form $F_t(X) = \max_{1 \leqslant p \leqslant m} \sum_{t'} A_{tt'}^p X_{t'}$, where $(A_{tt'}^p)$ is a 3-dimensional tensor whose entries are independent random variables drawn with the uniform law on $\{1, \ldots, 15\}$. Remember that m is an integer which represents the number of possible different actions per state t. All the results shown on the figures are the average made over 30 simulations, they concern the situation in which one of the two parameters m, n is kept constant while the other one increases. The time is given in seconds. The computations were performed on Matlab R2016a, using an Intel(R) Core(TM) i7-6500 CPU @ 2.59GHz processor with 12,0Go of RAM.

In both figures, Spectral Simplex-D appears to be more efficient than the Spectral Simplex algorithm with its original rule. However both algorithms are experimentally outperformed by policy iteration, by one to two order of magnitude, when $n \to \infty$, whereas when $m \to \infty$ (for constant parameter n), the performance of the three algorithms seem to deteriorate at the same rate.

9 Two-player Entropy Games: Multiplicative Policy Iteration and Power Algorithm

Let us now consider the general two-player case. For $\delta \in \mathcal{P}_D$ and $\tau \in \mathcal{P}_T$, let F^δ (resp. $^\tau F^\delta$) be the dynamic programming operator of the game in which the strategy
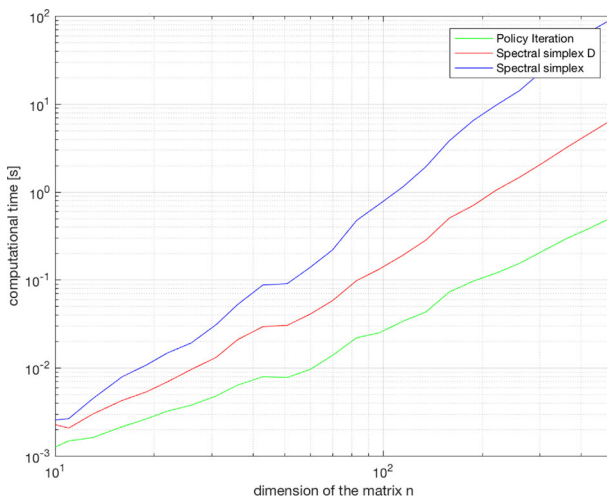

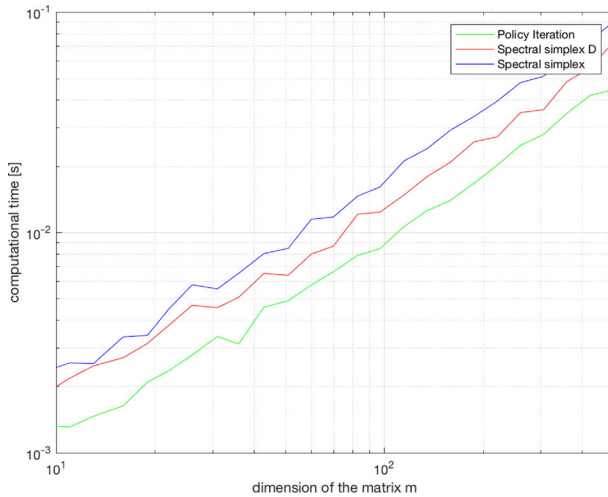**Fig. 1** Performance of Algorithms 1, 2, and 3 for different $n = 10, \ldots, 500$

Fig. 2 Performance of Algorithms 1, 2, and 3 for different $m = 10, ..., 500$

of Despot, δ, is fixed (resp. the strategies of Despot and of Tribune, δ and τ are fixed), see (22) (resp. (37)). We assume here that the matrices ${}^{\tau}M^{\delta}$ of the linear operators ${}^{\tau}F^{\delta}$, $\delta \in \mathcal{P}_D$, $\tau \in \mathcal{P}_T$, see (37), are irreducible.

Then, the Hoffman-Karp's idea [23] is readily adapted to the multiplicative setting: in the following algorithm, a sequence δ^k is constructed in a similar way as τ^k in the dual version of Algorithm 1, except that in Step 3, λ^{δ^k} and X^{δ^k} are computed by applying Algorithm 1 to the dynamic programming operator F^{δ^k} in which the strategy of Despot is fixed to δ^k. We call this the *multiplicative* Hoffman-Karp algorithm. ta It can also be viewed as an "exact" version of the policy iteration algorithm of Hoffman and Karp [23] for f.

Algorithm 4 Policy iteration for two-player entropy games.

1: Initialize $k = 1$, δ^0, $\delta^1 \neq \delta^0$ randomly.
2: **while** $\delta^k \neq \delta^{k-1}$ **do**
3:     Apply Algorithm 1 to F^{δ^k}. This returns a policy $\tau^k \in \mathcal{P}_T$, the Perron root λ^{δ^k} and the Perron eigenvector X^{δ^k} of $M^{\delta^k}\tau^k$.
4:     Compute a new policy δ^{k+1} such that, for all $d \in D$,

$$\delta^{k+1}(d) \in \underset{t \in T, \ (d,t) \in E}{\operatorname{argmin}} \ \max_{(t,p) \in E} \left(\sum_{(p,d') \in E} m_{pd'} X_{d'}^{\delta^k} \right),$$

    taking $\delta^{k+1}(d) = \delta^k(d)$ if it belongs to the latter argmin.
5:     $k \leftarrow k + 1$
6: **end while**
7: **return** the first policies (δ^k, τ^k) such that $\delta^k = \delta^{k-1}$, and the Perron root λ^{δ^k} and the Perron eigenvector X^{δ^k} of M^{δ^k, τ^k}.

A similar proof as the one of Proposition 30 shows that Algorithm 4, implemented in exact arithmetics, terminates and is correct under the previous assumption that for any pair of policies of the two players, the associated transition matrix is irreducible. Indeed, as for the dual version of Algorithm 1, λ^{δ^k} is decreasing as long as $\delta^k \neq \delta^{k-1}$. Then, since again the set of policies of Despot is finite, the algorithm ends.

To have an additional point of comparison, we used a power type algorithm, more precisely, the projective version of the Krasnoselski-Mann iteration, proposed in [21]. The original Krasnoselski-Mann iteration was developed in [29, 33], its analysis was extended and refined by Ishikawa [26] and Baillon and Bruck [8]. Recall that

$$F_d(X) = \min_{(d,t)\in E} \max_{(t,p)\in E} \sum_{(p,d')\in E} m_{pd'} X_{d'} \ ,$$

and let us define

$$H_d(X) = (X_d G_d(X))^{1/2}, \ \text{where} \, G_d(X) = \frac{F_d(X)}{(\prod_{d'\in E} F_{d'}(X))^{1/|D|}}.$$

Every fixed point of H is an eigenvector $X \in R_+^D$ of F such that $\prod_{d'\in E} X_d = 1$, indeed

$$
\begin{aligned}
H(X) = X &\iff (X_d G_d(X))^{1/2} = X_d, \forall d \in D \\
&\iff G_d(X) = X_d, \forall d \in D \\
&\iff \frac{F_d(X)}{(\prod_{d'} F_{d'}(X))^{1/|D|}} = X_d, \forall d \in D \\
&\iff F_d(X) = \lambda X_d, \forall d \in D, \ \text{and} \ \prod_{d'\in E} X_d = 1 \ ,
\end{aligned}
$$

for some $\lambda > 0$ (since then $\lambda = (\prod_{d'} F_{d'}(X))^{1/|D|}$). It can be shown as a corollary of a general result of Ishikawa [26] concerning nonexpansive mappings in Banach spaces that Algorithm 5 does converge if F has an eigenvector in the interior of the cone. We refer the reader to [21] for more details on the analysis of the projective Krasnoselski-Mann iteration. We use Hilbert's projective metric $d_H(X, Y) = \|\log(X) - \log(Y)\|_H$ (with $\|X\|_H = \max_d X_d - \min_d X_d$) to test the approximate termination.

Algorithm 5 Power algorithm for two-player entropy games.

1: Initialize $k = 0$, $V^0 = e \in \mathbb{R}^D$, $V^1 = F(X^0)$.
2: **while** $d_H(V^{k+1}, V^k) > \epsilon$ **do**
3: $\quad V^k \leftarrow V^{k+1}$
4: $\quad V^{k+1} \leftarrow H(V^k)$
5: $\quad k \leftarrow k + 1$
6: **end while**
7: **return** the first vector V^k such that $d_H(V^{k+1}, V^k) \leq \epsilon$.
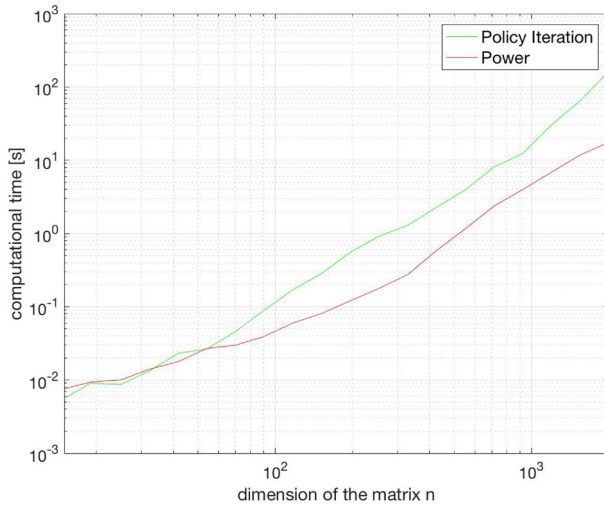

Fig. 3 Performance of Algorithms 4 and 5 for $n = 10, ..., 2000$

The log-log Figs. 3 and 4 show the performances of the power algorithm and the two-player policy iteration algorithm. The computation were performed on the same computer as in the previous section; the time is still given in seconds. The policy iteration algorithm outperforms the power algorithm asymptotically for large number of actions m, whereas for large number of states n, the power algorithm is more efficient. The experimentally observed efficiency of the –naive– power algorithm actually reveals that the random instances we considered are relatively "easy".
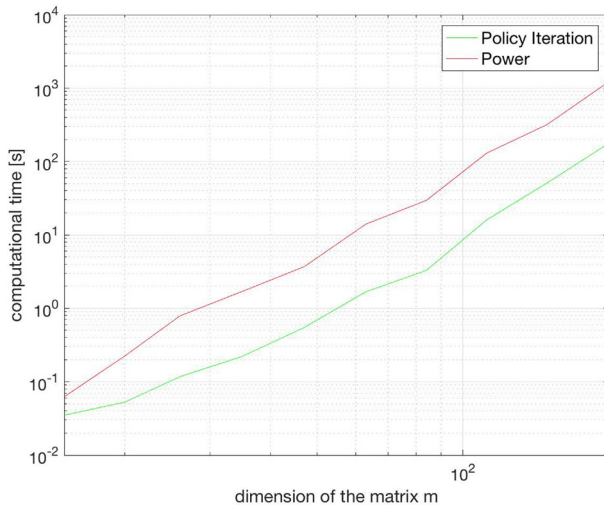
Fig. 4 Performance of Algorithms 4 and 5 for $m = 10, ..., 200$

10 Concluding Remarks

We developed an operator approach for entropy games, relating them with risk sensitive control via non-linear Perron-Frobenius theory. This leads to a theoretical result (polynomial time solvability of the Despot-free case), and this allows one to adapt policy iteration to these games. Several issues concerning policy iteration in the spectral setting remains unsolved. A first issue is to understand what kind of approximate eigenvalue algorithms are best suited. A second issue is to identify significant classes of entropy games on which the Hoffman-Karp type policy iteration algorithm can be shown to run in polynomial time (compare with [25, 49] in the case of Markov decision processes). In view of the asymmetry between Despot and Tribune, one may expect that Tribune-free entropy games are at least as hard as deterministic mean payoff games, it would be interesting to confirm that this is the case.

References

1. Anantharam, V., Borkar, V.S.: A variational formula for risk-sensitive reward. SIAM J. Contro Optim. **55**(2), 961–988 (2017). arXiv:1501.00676
2. Asarin, E., Cervelle, J., Degorre, A., Dima, C., Horn, F., Kozyakin, V.: Entropy games and matrix multiplication games. In: 33rd Symposium on Theoretical Aspects of Computer Science, STACS, Orléans, France, pp. 11:1–11:14 (2016)
3. Akian, M., Gaubert, S., Guterman, A.: Tropical polyhedra are equivalent to mean payoff games. Int. J. Algebra Comput. **22**(1), 125001 (43 pages) (2012)
4. Akian, M., Gaubert, S., Grand-Clément, J., Guillaud, J.: The Operator Approach to Entropy Games. In: Vollmer, H., Vallée, B. (eds.) 34th Symposium on Theoretical Aspects of Computer Science (STACS 2017), volume 66 of Leibniz International Proceedings in Informatics (LIPIcs), pp. 6:1–6:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl (2017)
5. Akian, M., Gaubert, S., Nussbaum, R.: A Collatz-Wielandt characterization of the spectral radius of order-preserving homogeneous maps on cones. arXiv:1112.5968 (2011)
6. Andersson, D., Miltersen, P.B.: The complexity of solving stochastic games on graphs. In: Proceedings of ISAAC'09, number 5878 in LNCS, pp 112–121. Springer (2009)
7. Borwein, J.M., Borwein, P.B.: On the complexity of familiar functions and numbers. SIAM Rev. **30**(4), 589–601 (1988)
8. Baillon, J.B., Bruck, R.E.: Optimal rates of asymptotic regularity for averaged nonexpansive mappings. In: Tan, K.K. (ed.) Proceedings of the Second International Conference on Fixed Point Theory and Applications, pp. 27–66. World Scientific Press (1992)
9. Bolte, J., Gaubert, S., Vigeral, G.: Definable zero-sum stochastic games. Math. Oper. Res. **40**(1), 171–191 (2014)
10. Bewley, T., Kohlberg, E.: The asymptotic theory of stochastic games. Math. Oper. Res. **1**(3), 197–208 (1976)
11. Blondel, V.D., Nesterov, Y.: Polynomial-time computation of the joint spectral radius for some sets of nonnegative matrices. SIAM J. Matrix Anal. **31**(3), 865–876 (2009)
12. Berman, A., Plemmons, R.J.: Nonnegative matrices in the mathematical sciences. Academic Press, New York (1994)
13. Chen, T., Han, T.: On the complexity of computing maximum entropy for markovian models. In: 34th International Conference on Foundation of Software Technology and Theoretical Computer Science, FSTTCS 2014, pp. 571–583, New Delhi (2014)

14. Crandall, M.G., Tartar, L.: Some relations between non expansive and order preserving maps. Proc. AMS **78**(3), 385–390 (1980)
15. Donsker, M.D., Varadhan, R.: On a variational formula for the principal eigenvalue for operators with maximum principle. Proc. Nat. Acad. Sci. USA **72**(3), 780–783 (1975)
16. Fleming, W.H., Hernández-Hernández, D.: Risk-sensitive control of finite state machines on an infinite horizon. I SIAM J. Control Optim. **35**(5), 1790–1810 (1997)
17. Fleming, W.H., Hernández-Hernández, D.: Risk-sensitive control of finite state machines on an infinite horizon. II. SIAM J. Control Optim. **37**(4), 1048–1069 (electronic) (1999)
18. Gaubert, S., Gunawardena, J.: A non-linear hierarchy for discrete event dynamical systems. In: Proceedings of the Fourth Workshop on Discrete Event Systems (WODES98), pp. 249–254. IEEE, Cagliari (1998)
19. Gaubert, S., Gunawardena, J.: The Perron-Frobenius theorem for homogeneous, monotone functions. Trans. AMS **356**(12), 4931–4950 (2004)
20. Grötschel, M., Lovász, L., Schrijver, A.: The ellipsoid method and its consequences in combinatorial optimization. Combinatorica **1**(2), 169–197 (1981)
21. Gaubert, S., Stott, N.: A convergent hierarchy of non-linear eigenproblems to compute the joint spectral radius of nonnegative matrices. In: Proceedings of the 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS2018), Hong Kong (2018)
22. Gaubert, S., Vigeral, G.: A maximin characterization of the escape rate of nonexpansive mappings in metrically convex spaces. Math Proc. Camb. Phil. Soc. **152**, 341–363 (2012)
23. Hoffman, A.J., Karp, R.M.: On nonterminating stochastic games. Manag. Sci. J. Inst. Manag. Sci. Appl. Theory Ser. **12**, 359–370 (1966)
24. Howard, R.A., Matheson, J.E.: Risk-sensitive markov decision processes. Manag. Sci. **18**(7), 356–369 (1972)
25. Hansen, T.D., Miltersen, P.B., Zwick, U.: Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. In: Innovations in Computer Science 2011, pp. 253–263. Tsinghua University Press (2011)
26. Ishikawa, S.: Fixed points and iteration of a nonexpansive mapping in a Banach space. Proc. Amer. Math. Soc. **59**(1), 65–71 (1976)
27. Kingman, J.F.C.: A convexity property of positive matrices. Quart. J. Math. Oxford Ser. **2**(12), 283–284 (1961)
28. Kozyakin, V.: Hourglass alternative and the finiteness conjecture for the spectral characteristics of sets of non-negative matrices. Linear Algebra Appl. **489**, 167–185 (2016)
29. Krasnosel'skiĭ, M.A.: Two remarks on the method of successive approximations. Uspekhi Matematicheskikh Nauk **10**, 123–127 (1955)
30. Kullback, S.: Information theory and statistics. Dover Publications, Inc., Mineola (1997). Reprint of the second (1968) edition
31. Lemmens, B., Lins, B., Nussbaum, R., Wortel, M.: Denjoy-Wolff theorems for Hilbert's and Thompson's metric spaces. J. d'Anal. Math. **134**, 671–718 (2018)
32. Lothaire, M.: Applied combinatorics on words. Cambridge, New York (2005)
33. Mann, W.R.: Mean value methods in iteration. Proc. Amer. Math. Soc. **4**, 506–510 (1953)
34. Mertens, J.-F., Neyman, A.: Stochastic games. Internat. J. Game Theory **10**(2), 53–66 (1981)
35. Müller, J.M.: Elementary functions: algorithms and implementation. Birkhaüser, Cambridge (2005)
36. Neyman, A.: Stochastic games and nonexpansive maps. In Stochastic games and applications (Stony Brook, NY, 1999), volume 570 of NATO Sci. Ser. C Math. Phys. Sci., pp. 397–415. Kluwer Acad. Publ., Dordrecht (2003)
37. Nussbaum, R.D.: Convexity and log convexity for the spectral radius. Linear Algebra Appl. **73**, 59–122 (1986)
38. Protasov, V.Yu.: Spectral simplex method. Math. Program. **156**(1-2, Ser. A), 485–511 (2016)
39. Puterman, M.L.: Markov decision processes. Wiley, New York (2005)
40. Rothblum, U.G.: Multiplicative markov decision chains. Math. Oper. Res. **9**(1), 6–24 (1984)
41. Rump, S.M.: Polynomial minimum root separation. Math. Comput. **145**(33), 327–336 (1979)
42. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
43. Sladký, K.: On Dynamic Programming Recursions for Multiplicative Markov Decision Chains, pp. 216–226. Springer, Berlin (1976)
44. van den Dries, L.: Tame topology and o-minimal structures, volume 248 of London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge (1998)

45. van den Dries, L.: o-minimal structures and real analytic geometry. In: Current developments in mathematics, 1998 (Cambridge, MA), pp. 105–152. Int. Press, Somerville (1999)
46. Vigeral, G.: A zero-sum stochastic game with compact action sets and no asymptotic value. Dyn. Games Appl. **3**(2), 172–186 (2013)
47. Whittle, P.: Optimization over time, I. Wiley, New York (1982)
48. Wilkie, A.J.: Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. J. Amer. Math. Soc. **9**(4), 1051–1094 (1996)
49. Ye, Y.: The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. Math. Oper. Res. **36**(4), 593–603 (2011)
50. Zijm, W.H.M.: Asymptotic expansions for dynamic programming recursions with general nonnegative matrices. J. Optim. Theory Appl. **54**(1), 157–191 (1987)