The Operator Approach to Entropy Games*

Marianne Akian[1], Stéphane Gaubert[2], Julien Grand-Clément[3], and Jérémie Guillaud[4]

1 **INRIA and CMAP, École polytechnique, CNRS, Palaiseau, France**
marianne.akian@inria.fr
2 **INRIA and CMAP, École polytechnique, CNRS, Palaiseau, France**
stephane.gaubert@inria.fr
3 **École polytechnique, Palaiseau, France**
julien.grand-clement@polytechnique.edu
4 **École polytechnique, Palaiseau, France**
jeremie.guillaud@polytechnique.edu

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.

1998 ACM Subject Classification G.2.1 Combinatorial Algorithms, F.2.1 Numerical Algorithms and Problems

Keywords and phrases Stochastic games, Shapley operators, policy iteration, Perron eigenvalues, Risk sensitive control

Digital Object Identifier 10.4230/LIPIcs.STACS.2017.6

1 Introduction

1.1 Entropy games and matrix multiplication games

Entropy games have been introduced by Asarin et al. [5]. 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 general 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. In particular, Asarin et al. showed

* The authors were partially supported by the ANR through the MALTHY INS project, and by the PGMO program of FMJH and EDF.

in [5] 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 [4] for background.

Another motivation to study entropy games arises from risk sensitive control [13, 14, 3]: as we shall see, essentially the same class of operators arise in the latter setting. Further motivations originate from symbolic dynamics [21, 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 more expressive variant of the model of Asarin et al [5], called here *extended* entropy games for clarity, in which the initial state is prescribed (the initial state is chosen by one player, People, in the original model). This extension is needed 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, namely the existence of the value and the existence of optimal positional strategies, are still valid for extended entropy games (Theorems 2 and 3). This is derived from a model theory approach of Bolte, Gaubert, and Vigeral [8], together with the observation that the dynamic programming operators of extended entropy games are definable in the real exponential field. Another consequence of the operator approach is the existence of Collatz-Wielandt optimality certificates for entropy games, Theorem 12. When specialized to the one player case, this leads to a convex programming characterization of the value, Corollary 13, which can also be recovered from a characterization of Anantharam and Borkar [3].

This leads us to our main result, Theorem 14, showing that (extended) entropy games in which Despot has a fixed number of significant states (states with a nontrivial choice) can be solved 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 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. The proof relies on several ingredients: ellipsoid method, separation bounds between algebraic numbers, 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 [19] 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 [23].

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 [14, 3]. It can be thought as a version for two player problems of the Donsker-Varadhan characterization of the Perron-eigenvalue [11]. 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 [3]. 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 [10] 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 [2], yields an alternative to the approach of [5] using the "hourglass alternative" of [20] to produce concise certificates allowing one to bound the value of entropy games. Finally, the identification of tractable subclasses of matrix multiplication games can be traced back at least to the work of Blondel and Nesterov [7].

2 Entropy games

2.1 Entropy games with prescribed initial state

An extended entropy game Γ^{eent} 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 nondeterministic 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 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 digraph 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.

For every integer k, we define as follows the *game in horizon* k with initial state $\bar{d}$, $\Gamma^{\mathrm{eent}}(k, \bar{d})$. 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 $R^k_{\bar{d}}(\delta, \tau)$. It is defined as the sum of the weights of all the paths of the digraph 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 $R^k_{\bar{d}}(\delta, \tau)$, so that the game in horizon k is zero-sum. In that way, the payment $R^k_{\bar{d}}$ measures the "freedom" of People, Despot wishes to minimize it whereas Tribune wishes to maximize it.

We say that the game $\Gamma^{\mathrm{eent}}(k, \bar{d})$ in horizon k with initial state $\bar{d}$ *has the value* $V^k_{\bar{d}}$ and that δ^*, τ^* *are optimal strategies* of Despot and Tribune if for all strategies δ, τ of Despot

and Tribune, we have the saddle point property:

$$R_{\bar{d}}^k(\delta,\tau^*) \geqslant R_{\bar{d}}^k(\delta^*,\tau^*) = V_{\bar{d}}^k \geqslant R_{\bar{d}}^k(\delta^*,\tau) \ . \tag{1}$$

If the value $V_{\bar{d}}^k$ exists for all choices of the initial state $\bar{d}$, we define the *value vector* of the game $\Gamma^{\mathrm{eent}}(k,\cdot)$ in horizon k, to be $V^k := (V_d^k)_{d\in D} \in \mathbb{R}^D$.

We now define the *infinite horizon game* $\Gamma^{\mathrm{eent}}(\infty,\bar{d})$, 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 $\Gamma^{\mathrm{eent}}(\infty,\bar{d})$, and the optimal strategies in this game, are still defined by a saddle point condition, as in (1), 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 energy game $\Gamma^{\mathrm{eent}}(\infty,\cdot)$.

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 $d \in D$,

$$F_d(X) = \min_{(d,t)\in E} \max_{(t,p)\in E} \sum_{(p,d')\in E} m_{pd'} X_{d'} \ . \tag{2}$$

The existence of the value for the *finite* horizon game follows from a standard dynamic programming argument.

▶ **Proposition 1.** *The value of the extended entropy game in horizon k, $\Gamma^{eent}(k,\cdot)$, 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 of $\mathbb{R}^D$.*

Recall that a strategy is said to be *positional* or is called a *policy* if the decision taken at a given stage depends only on the last state which has been visited. The following theorem follows from Theorem 9 stated in Section 3, by using an equivalence with a special class of stochastic mean payoff games with infinite actions spaces, through log-glasses.

▶ **Theorem 2.** *The infinite horizon extended entropy game has a value and it has optimal positional strategies. Moreover, for all initial states d, $V_d^\infty = \lim_{k\to\infty}(V_d^k)^{1/k}$.*

The following result is deduced from Theorem 11, using the same technique as for Theorem 2. We denote by $\mathcal{P}_D$ (resp. $\mathcal{P}_T$) the set of policies (i.e., positional strategies) of Despot (resp. Tribune). If one fixes a strategy $\delta \in \mathcal{P}_D$ or $\tau \in \mathcal{P}_T$, we end up with a one player infinite horizon entropy game (a two player game in which either Despot or Tribune has no options), whose value is denoted by $V_d^\infty(\delta,\star)$ (resp. $V_d^\infty(\star,\tau)$). Similarly, if we fix the two strategies, we end up in a game in which only People has options, and the value of this game, denoted by $V_d^\infty(\delta,\tau)$, coincides with $R_{\bar{d}}^\infty(\delta,\tau)$.

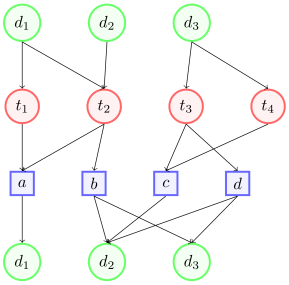
▶ **Theorem 3.** *We have*

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

Moreover, for all $\delta \in \mathcal{P}_D$ and for all $\tau \in \mathcal{P}_T$,

$$V_d^\infty(\delta,\star) = \max_{\tau'\in\mathcal{P}_T} V_d^\infty(\delta,\tau'), \qquad V_d^\infty(\star,\tau) = \min_{\delta'\in\mathcal{P}_D} V_d^\infty(\delta',\tau),$$

▶ **Example 4.** Take $D = \{d_1, d_2, d_3\}$, $T = \{t_1, t_2, t_3, t_4\}$, $P = \{a, b, c, d\}$, $E = \{(d_1, t_1), (d_1, t_2), (d_2, t_2), (d_3, t_3), (d_3, t_4), (t_1, a), (t_2, a), (t_2, b), (t_3, c), (t_3, d), (t_4, c), (a, d_1), (b, d_2), (b, d_3), (c, d_2), (d, d_2), (d, d_3)\}$ and $m_{pd_i} = 1$ for all $p \in P$ and $1 \leqslant i \leqslant 3$ such that $(p, d_i) \in E$. The corresponding graph and dynamic programming operator are given by:



$$F_1(X) = \min\big(X_1, \max(X_1, X_2 + X_3)\big),$$

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

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

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. Hence, by Theorem 2, the value vector of this entropy game is $V^\infty = (1, \omega, \omega)$ where $\omega := (1 + \sqrt{5})/2$.

2.2 The original entropy game model

The original entropy game model of Asarin et al. [5] is a zero-sum game defined in a similar way, up to a 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 $\Gamma^{\mathrm{ent}}(\infty)$, we denote by $\bar{V}^\infty$ the value of this game, which is shown to exist in [5].

Note that in the initial model in [5], 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 [5]. 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 $\Gamma^{\mathrm{ent}}(\infty)$ and will refer to the entropy game model with rational weights as the entropy game model. The next result, which can be deduced from the existence of the value of the extended entropy game (Theorem 2 above), shows that the value of the original entropy game can be recovered from the value vector of the extended one:

▶ **Proposition 5.** *The value of the original entropy game $\Gamma^{ent}(\infty)$ coincides with the maximum of the values of the extended entropy games $\Gamma^{eent}(\infty, d)$, taken over all initial states: $\bar{V}^\infty = \max_{d \in D} V_d^\infty$.*

▶ **Example 6.** This is illustrated by the game of Example 4. In the original model of [5], 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 [5], 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 [5]. The following property proved in [5] is the analogue of Theorem 3, V_d^∞ being replaced by $\bar{V}^\infty$:

$$\bar{V}^\infty = \min_{A\in\mathcal{A}} \max_{B\in\mathcal{B}} \rho(AB) = \max_{B\in\mathcal{B}} \min_{A\in\mathcal{A}} \rho(AB) \ . \tag{3}$$

A more general property is proved in [2, Section 8], as a consequence of the Collatz-Wielandt theorem (see Theorem 12 below).

3 Stochastic mean payoff game with Kullback-Leibler payments

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

To the extended entropy games Γ^{eent}, we associate a family of stochastic zero-sum games with Kullback-Leibler payments, denoted Γ^{kl} and defined as follows. These new games are still played on the weighted digraph G. 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 Δ_p the set of probability measures on E_p. Therefore, an element of Δ_p can be identified to a vector $\nu = (\nu_{p,d})_{(p,d)\in E_p}$ with nonnegative entries and sum 1. The actions of Despot and Tribune in the states $d \in D$ and $t \in T$ are the same in the games Γ^{kl} and in the games Γ^{eent}. However, the two games have different rules when the state is in P, since the nondeterministic half-player, People, is now replaced by a standard probabilistic half-player, Nature. In the game Γ^{kl}, Tribune, who arrived in a state $p \in P$ by choosing first an action in some state $t \in T$, so that $(t,p) \in E$, has to play again in state p, by choosing a probability measure $\nu \in \Delta_p$. Then, Nature chooses the next state d according to probability $\nu_{p,d}$, and Tribune receives the payment $-S_p(\nu;m)$, where $S_p(\nu;m)$ is the relative entropy or Kullback-Leibler divergence:

$$S_p(\nu;m) := \sum_{(p,d)\in E_p} \nu_{pd} \log(\nu_{pd}/m_{pd}) \ .$$

An interesting special case arises when $m \equiv 1$, as in [5]. Then, $S_p(\nu;m) = S_p(\nu) := \sum_{(p,d)\in E_p} \nu_{pd} \log \nu_{pd}$ is nothing but the Shannon entropy of ν.

A history in the game Γ^{kl} now consists of a finite sequence $(d_0,t_0,p_0,\nu_0,d_1,t_1,p_1,\dots)$, 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 $E_d := \{(d,t) \in E\}$. 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 $p = \tau(d)$ 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 $h = (d_0,t_0,p_0,\nu_0,d_1,t_1,p_1)$ 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 $r^k_{\bar{d}}(\delta,(\tau,\pi)) := -\mathbb{E}\big(S_{p_0}(\nu_0;m) + \cdots + S_{p_{k-1}}(\nu_{k-1};m)\big)$ the expectation of the payment received by Tribune, in k turns, with respect to this measure, We also consider the *infinite horizon* or *mean payoff* game $\Gamma^{\mathrm{kl}}(\infty,\bar{d})$, in which the payment of Tribune is now

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

We define the *value* of the game in horizon k, $v^k_{\bar{d}}$, and the value of the infinite horizon game, $v^\infty_{\bar{d}}$, as well as optimal strategies, by saddle point conditions, as in Section 2.1. We have the following dynamic programming principle.

▶ **Proposition 7.** *The value vector $v^k = (v^k_d)_{d\in D}$ in horizon k of the stochastic game with Kullback-Leibler payments, Γ^{kl}, does exists. It is determined by the relations $v^0 = 0$, $v^k = f(v^{k-1})$, $k = 1,2,\ldots$, where*

$$f_d(x) = \min_{(d,t)\in E} \max_{(t,p)\in E} \log\Big(\sum_{(p,d')\in E} m_{pd'} \exp(x_{d'})\Big) \ , \tag{4}$$

and we have $v^k_d = \log V^k_d$.

The explicit form of f in (4) originates from the following expression of the Legendre-Fenchel transform of Shannon entropy, which is a classical result in convex analysis, see e.g. [25].

▶ **Lemma 8.** *The function $x \mapsto \log(\sum_{1\leqslant i\leqslant n} e^{x_i})$ is convex and it satisfies*

$$\log\Big(\sum_{1\leqslant i\leqslant n} e^{x_i}\Big) = \max \sum_{1\leqslant i\leqslant n} \nu_i(x_i - \log\nu_i); \qquad \nu_i \geqslant 0,\ 1\leqslant i\leqslant n,\ \sum_{1\leqslant i\leqslant n}\nu_j = 1 \ .$$

The following result shows that the extended entropy game Γ^{eent} is equivalent to the stochastic mean payoff game Γ^{kl}, through logarithmic glasses. Theorem 2 above is deduced from it. We define the *projection* of a pair of strategy $(\delta,(\tau,\pi))$ in Γ^{kl} to be the strategy (δ,τ) in Γ^{eent}.

▶ **Theorem 9.** *The stochastic mean payoff game with Kullback-Leibler payments, $\Gamma^{kl}(\infty,\cdot)$, has a value. For all states $d\in D$, we have*

$$v^\infty_d = \log V^\infty_d, \qquad v^\infty_d = \lim_{k\to\infty} \frac{1}{k} v^k_d \ .$$

Moreover, the optimal strategies of $\Gamma^{eent}(k)$ are precisely the projections of the optimal strategies of $\Gamma^{kl}(k)$, for all k integer or equal to ∞.

This theorem shows that extended entropy games are particular stochastic mean payoff games (with compact action spaces). Asarin et al. [5] remarked that the special *deterministic* entropy games, in which People has only one possible action in each state, can be reencoded as deterministic mean payoff games. This can also be recovered from our approach: in this deterministic case, the simplices Δ_p are singletons and the entropy function vanishes.

We next sketch the derivation of Theorem 9 from a result of Bolte, Gaubert and Vigeral [8] on the escape rate of nonexpansive mappings that are definable in an o-minimal structure. A map f is *nonexpansive* with respect to a norm $\|\cdot\|$ if $\|f(x)-f(y)\| \leqslant \|x-y\|$. Recall that an o-minimal structure [12, 28] 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* [28], 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 $\{x \mid P(x_1, \dots, x_{n+k}, e^{x_1}, \dots, e^{x_{n+k}}) = 0\}$ where P is a real polynomial. A theorem of Wilkie [29] implies that $\mathbb{R}_{\mathrm{alg,exp}}$ is o-minimal, see [28].

If $f : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive in any norm, given any $0 < \alpha < 1$, we define the *discounted value vector* $z_\alpha \in \mathbb{R}^n$ by $f(\alpha z_\alpha) = z_\alpha$. This vector exists and is unique (apply Banach fixed point theorem to the contraction mapping $x \mapsto f(\alpha x)$).

▶ **Theorem 10** ([8]). *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, the limit $\lim_{k\to\infty} f^k(0)/k$ does exists, and it coincides with the limit $\lim_{\alpha\to 1^-}(1-\alpha)z_\alpha$.*

The vector z_α is nothing but the value of the *discounted variant* of the stochastic game Γ^{kl}, where α is the discount factor. The map f in (4) is nonexpansive in the sup-norm, and it is definable in the real exponential field. So Theorem 10 can be applied to it. The existence of the limit in Theorem 9 is deduced from this result.

A policy 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 policies is a basic feature of perfect information zero-sum stochastic games with finite action spaces (see [24, Chap. 10] for the one-player case, the two-player case builds on similar ideas, e.g. [15, Lemma 26]). It allows one to reduce the mean payoff problem to the discounted problem. We next show that this result has an analogue for entropy games. We shall say that a pair of strategies $(\delta, \tau) \in \mathcal{P}_D \times \mathcal{P}_T$ 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 policies $(\delta, (\tau, \pi))$ in the discounted version of the game Γ^{kl}. The fact that the value of the entropy games commutes with maxima and minima of policies (Theorem 3) is derived by combining Theorem 10 with the following result, whose proof relies, again, on an o-minimality argument.

▶ **Theorem 11.** *The stochastic perfect information game with Kullback-Leibler payments, Γ^{kl}, has Blackwell optimal strategies.*

4 Applying the Collatz-Wielandt theorem to entropy games

The classical Collatz-Wielandt formulæ provide the following characterizations of the spectral radius $\rho(M)$ of a nonnegative matrix M:

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

where $\mathbb{R}^D_+$ denotes the nonnegative orthant of $\mathbb{R}^D$, and $\operatorname{int}\mathbb{R}^D_+$ its interior, i.e, the set of positive vectors. This has been extended to non-linear, order preserving and continuous self-maps of the standard positive cone [22, 2]. In particular, the following result can be derived from the non-linear Collatz-Wielandt formulæ in these works.

▶ **Theorem 12** (Corollary of [2]). *The value $\bar{V}^\infty$ of the original entropy game Γ^{ent} (with a free initial state) coincides with any of the following expressions*

$$\inf\{\lambda > 0 \mid \exists X \in \operatorname{int} \mathbb{R}^D_+,\ F(X) \leqslant \lambda X\} \tag{5}$$
$$\max\{\lambda > 0 \mid \exists X \in \mathbb{R}^D_+ \setminus \{0\},\ F(X) = \lambda X\} \tag{6}$$
$$\max\{\lambda > 0 \mid \exists X \in \mathbb{R}^D_+ \setminus \{0\},\ F(X) \geqslant \lambda X\}\ , \tag{7}$$

where F is the dynamic programming operator (2).

The value of these expressions is called the non-linear spectral radius of F. The Collatz-Wielandt formulæ are helpful to establish strong duality results, like (3), see also [1] for an application to mean payoff games and tropical geometry. Our main interest here lies in the following application of (5). 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 (2) 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 Theorem 12 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 [3].

▶ **Corollary 13.** *The logarithm of the value of a Despot-free entropy game is given by*

$$\inf \mu,\ \mu \in \mathbb{R}, x \in \mathbb{R}^D,$$
$$\mu + x_d \geqslant \log(\sum_{d' \in D} m_{p,d'} e^{x_{d'}}) \text{ for all } d \in D, p \in P \text{ such that } (\sigma(d),p) \in E\ . \tag{8}$$

5 Polynomial time solvability of entropy games with a few significant Despot positions

By *solving strategically* an (extended) 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 main result is the following.

▶ **Theorem 14.** *Despot-free entropy games can be solved strategically in polynomial time.*

We indicate here the main arguments of proof.

Step 1. *Reduction to the irreducible case.* First, we associate to a Despot-free extended entropy game a projected digraph $\bar{G}$, with node set D and an arc $d \to d'$ if there is a path (d,t,p,d') in the original digraph G. We say that the game is *irreducible* if $\bar{G}$ is strongly connected. It is not difficult to see that in a Despot-free extended entropy game, the value of a state d is the maximum of the value of the irreducible games corresponding to the different strongly connected components of $\bar{G}$ to which d has access under some policy of Tribune (this is a special case of a more general known property, see [15, Th.29]). Hence, we will assume that the game is irreducible in the rest of the proof.

The following result is a consequence of the non-linear Perron-Frobenius theorem in [16].

▶ **Lemma 15.** *The value of an irreducible Despot-free extended 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.*

Thanks to this lemma, we will speak of "value" without making explicit the initial state. We set $W := \max_{(p,d)\in E} m_{p,d}$ and $n := |D|$.

Step 2. *Reduction to a convex program with bounded feasible set.* To prove Theorem 14, we apply the ellipsoid method. To do so, we must replace the convex program (8) 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.

▶ **Lemma 16.** *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 for all $d \in D$, $1 \leqslant U_d \leqslant (nW)^{n-1}$.*

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

$$f(u) \leqslant \mu e + u, \qquad 0 \leqslant u_d \leqslant (n-1)\lceil \log(nW) \rceil, \qquad 0 \leqslant \mu \leqslant \lceil \log(nW) \rceil + 2 \ , \tag{9}$$

where $\lceil t \rceil$ denotes the smallest integer greater than or equal to t, and f is given by (4), recalling that e denotes the unit vector of $\mathbb{R}^n$. By combining Corollary 13, Lemma 15 and Lemma 16, we arrive at the following result.

▶ **Proposition 17.** *The value of a Despot-free irreducible entropy game coincides with the exponential of the value of the convex program: $\min \mu$, $(u, \mu) \in \mathcal{K}$. Moreover, $B_2(a, r) \subset \mathcal{K} \subset B_2(a, R)$ where $a = (e, \lceil \log(nW) \rceil + 1) \in \mathbb{R}^D \times \mathbb{R}$, $r := 1/3$, and $R := 2\sqrt{D+1}(n-1)\log(nW)$.*

Step 3. *Show the existence of a polynomial time approximate separation oracle [17] for the program of Proposition 17.* The non-trivial separating half-spaces are obtained by computing the differential of the logarithmic expressions in (8). To do so, we use the fact that the values of the logarithm and exponential function can be approximated in polynomial time [9].

Step 4. *Show that if any two policies of Tribune yield different values λ and λ', then, $|\lambda - \lambda'|$ is bounded below by a rational number $\eta_{sep} > 0$ whose number of bits is polynomially bounded in the size of the input.* This relies on separation results between algebraic numbers [27], since the value of a strategy of Tribune is an eigenvalue of a $n \times n$ matrix with integer coefficients bounded by the number W.

Step 5. *Synthesize an optimal strategy of Tribune from an approximate solution of the program in Proposition 17.*

To any policy τ of Tribune, we associate a dynamic programming operator F^τ, which is the self-map of $\mathbb{R}^D$ defined by $F^\tau_d(X) = \sum_{(\tau(\sigma(d)),d')\in E} m_{\tau(\sigma(d))d'} X_{d'}$. In other words, $F^\tau(X) = M^\tau X$, where $M^\tau = (m_{\tau(\sigma(d))d'})_{d,d'\in D}$ is a $|D| \times |D|$ matrix with nonnegative entries.

To explain our method, we make first the restrictive assumption that for every policy τ, the matrix M^τ is irreducible. In particular, we can take an optimal policy τ^*. By a standard result of Perron-Frobenius theory [6], M^{τ^*} has a left eigenvector π with positive entries, associated to the maximal eigenvalue $\lambda^{\tau^*} := \rho(M^{\tau^*})$, called Perron root. Hence, $\pi M^{\tau^*} = \lambda^{\tau^*}\pi$. Since τ^* is optimal, $\lambda^{\tau^*} = \lambda^*$. Moreover, by applying Lemma 16 to the linear map $U \mapsto (M^{\tau^*})^T U$, 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 17, yields in polynomial time a vector u and a scalar μ such that

$f(u) \leqslant (\log \lambda^* + \epsilon)e + u$ and $\lambda^* \leqslant \exp(\mu) \leqslant \lambda^* \exp(\epsilon)$. Taking $U := (U_d)_{d\in D}$ with $U_d := \exp(u_d)$, we get $F(U) \leqslant \lambda^* \exp(\epsilon)U$. We choose any policy $\underline{\tau}$ such that $F(U) = M^{\underline{\tau}}U$. Therefore, $\underline{\tau}(\sigma(d))$ is chosen to be any term attaining the maximum when evaluating $F_d(U)$. We claim that $\underline{\tau}$ is optimal if ϵ is sufficiently small.

To show the latter claim, we observe that $M^{\tau^*}U \leqslant F(U)$. Moreover, for all $d \in D$, $0 \leqslant \pi_d(\lambda^* \exp(\epsilon)U_d - F_d(U)) \leqslant \pi_d(\lambda^* \exp(\epsilon)U_d - (M^{\tau^*}(U))_d) \leqslant \sum_{d'\in D} \pi_{d'}(\lambda^* \exp(\epsilon)U_{d'} - (M^{\tau^*}U)_{d'}) = \pi(\lambda^* \exp(\epsilon)U - M^{\tau^*}U) = \lambda^*(\exp(\epsilon) - 1)\pi U$. Using $\pi_d/\pi_{d'} \leqslant (nW)^{n-1}$ and $U_d/U_{d'} \leqslant (nW)^{n-1}$, we deduce that $F(U) \geqslant \underline{\lambda}U$, where $\underline{\lambda} := \lambda^*[\exp(\epsilon) - (\exp(\epsilon) - 1)n(nW)^{2(n-1)}]$. Since, $M^{\underline{\tau}}U \geqslant \underline{\lambda}U$, we have $\rho(M^{\underline{\tau}}) \geqslant \underline{\lambda}$. It follows that we can choose $\epsilon > 0$, with a polynomially bounded number of bits, such that $\underline{\lambda} > \lambda^* - \eta_{\text{sep}}$. Moreover, since λ^* is the maximum of the values of all the policies, $\underline{\lambda} \leqslant \lambda^*$. By definition of the separation parameter η_{sep}, this implies that $\underline{\lambda} = \lambda^*$, and so the policy $\underline{\tau}$ of Tribune which we just constructed is optimal, showing the claim.

When some policies τ yield a *reducible* matrix M^τ, 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 $\pi_d > 0$.

Theorem 3, showing that $V_d^\infty = \min_{\delta\in\mathcal{P}_D} V_d^\infty(\delta, \star)$, allows us to solve an entropy game by enumerating the policies $\delta \in \mathcal{P}_D$ and solving the Despot-free entropy game determined by each δ. This leads to an algorithm with execution time $\left(\prod_{d\in D} |E_d|\right) T_{\text{D-free}}$, where the factor $T_{\text{D-free}}$ is polynomial in the size of the input. We have in particular:

▶ **Corollary 18.** *Entropy games in which Despot has a fixed number of significant states can be solved strategically in polynomial time.*

6 Multiplicative policy iteration algorithm and comparison with the spectral simplex method of Protasov

The equivalence between extended entropy games and some special class of stochastic mean payoff games, through logarithmic glasses (see Section 3), allows 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 of Hoffman and Karp [19] and Rothblum [26].

To simplify the presentation, we consider first a Despot-free entropy game. Without loss of generality, we assume that $D = T = \{1, \dots, |T|\}$ and σ is the identity. Let F^τ and M^τ, $\tau \in \mathcal{P}_T$, be defined as in the previous section. If M^τ is irreducible, in particular if all its entries are positive, M^τ has an eigenvector $X^\tau > 0$, associated to the Perron root $\lambda^\tau := \rho(M^\tau)$. Moreover, X^τ is unique up to a multiplicative constant and is called a Perron eigenvector. If all the matrices M^τ, $\tau \in \mathcal{P}_T$ are irreducible, one can construct the following multiplicative version of the policy iteration algorithm.

This algorithm has a dual version, in which maximization is replaced by minimization. Then, the Hoffman-Karp's idea [19] is readily adapted to the multiplicative setting: 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. A variation of the original proof shows that this algorithm, implemented in exact arithmetics, terminates and is correct if for any pair of policies of the two players, the associated transition matrix is irreducible.

In [23], Protasov introduced the Spectral Simplex Algorithm. His algorithm is a variant of Algorithm 1 in which the policy is improved only at *one* state, which is the first state t

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

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 M^{τ^k}.
4: Compute a new policy τ^{k+1} such that, for all $t \in T$,

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

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 M^{τ^k}.

Table 1 Comparing multiplicative policy iteration with spectral simplex.

Number of states	10	20	30	40	50	60	70	80	90	100
Time : Policy Iteration	0.0018	0.0037	0.0057	0.0095	0.0115	0.0141	0.0171	0.0283	0.0308	0.0363
Time : Spectral Simplex-D	0.0026	0.0083	0.0158	0.0317	0.0433	0.0511	0.0797	0.1261	0.1533	0.1950
Time : Spectral Simplex	0.0034	0.0149	0.0350	0.0934	0.1419	0.1615	0.3070	0.5835	0.7418	1.0257
Iterations : Policy Iteration	3	3	3.4	3	3	3	3.2	3.2	3	3.2
Iterations : Spectral Simplex-D	5.6	7.4	10.2	10	11.8	13.4	14.8	14.2	16	17.2
Iterations : Spectral Simplex	15.4	22	40.8	53.4	57.8	83	87	102.2	106.8	122.8

such that $F_t(X^{\tau^k}) > \lambda^{\tau^k} X_t^{\tau^k}$. We also considered another version of Algorithm 1, in which we change the policy at only one state t, which maximizes 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 [30].

We next report numerical experiments in the case of Despot-free entropy game, in order to compare Protasov's spectral simplex algorithm (with the improvement of Dantzig' pivot rule) with the multiplicative Policy Iteration algorithm (Algorithm 1). In Table 1, 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 $F_t(X) = \max_{1\leqslant p\leqslant q} \sum_{t'} A_{tt'}^p X_{t'}$, where $(A_{tt'}^p)$ is a 3-dimensional tensor whose entries are independent random variables drawn with the uniform law in $[0, 1]$. All the results below are the average made over 10 simulations, they concern the situation in which the number of actions q is kept constant, equal to 10, whereas n varies. The time is given in seconds. The number of "iterations" denotes the number of times that the algorithm goes through the main loop, regardless of how many operations are performed inside the loop. The computations were performed on Matlab R2016a, using an Intel(R) Core(TM) i7-6500 CPU @ 2.59GHz processor with 12,0Go of RAM.

Spectral Simplex-D appears to be more efficient than the Spectral Simplex algorithm with its original rule [23]. Both algorithms are experimentally outperformed by policy iteration, by one order of magnitude, when $n \to \infty$.

7 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 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 [30, 18] in the case of Markov decision processes).

Acknowledgments. We thank all the referees for their comments. We thank especially one referee for detailed suggestions and for pointing out reference [10].

References

1 M. Akian, S. Gaubert, and A. Guterman. Tropical polyhedra are equivalent to mean payoff games. *International Journal of Algebra and Computation*, 22(1):125001 (43 pages), 2012. arXiv:0912.2462, doi:10.1142/S0218196711006674.

2 M. Akian, S. Gaubert, and R. Nussbaum. A Collatz-Wielandt characterization of the spectral radius of order-preserving homogeneous maps on cones, 2011. URL: https://arxiv.org/abs/1112.5968.

3 V. Anantharam and V. S. Borkar. A variational formula for risk-sensitive reward, 2015. arXiv:1501.00676.

4 D. Andersson and P. B. Miltersen. The complexity of solving stochastic games on graphs. In *Proceedings of ISAAC'09*, number 5878 in LNCS. Springer, 2009.

5 E. Asarin, J. Cervelle, A. Degorre, C. Dima, F. Horn, and V. Kozyakin. Entropy games and matrix multiplication games. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016, February 17-20, 2016, Orléans, France*, pages 11:1–11:14, 2016. doi:10.4230/LIPIcs.STACS.2016.11.

6 A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Academic Press, 1994.

7 V. D. Blondel and Y. Nesterov. Polynomial-time computation of the joint spectral radius for some sets of nonnegative matrices. *SIAM J. Matrix Anal.*, 31(3):865–876, 2009.

8 J. Bolte, S. Gaubert, and G. Vigeral. Definable zero-sum stochastic games. *Mathematics of Operations Research*, 40(1):171–191, 2014. arXiv:1301.1967, doi:10.1287/moor.2014.0666.

9 J. M. Borwein and P. B. Borwein. On the complexity of familiar functions and numbers. *SIAM Review*, 30(4):589–601, 1988.

10 T. Chen and T. Han. 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, December 15-17, 2014, New Delhi, India*, pages 571–583, 2014. doi:10.4230/LIPIcs.FSTTCS.2014.571.

11 M. D. Donsker and S. R. S. Varadhan. On a variational formula for the principal eigenvalue for operators with maximum principle. *Proc. Nat. Acad. Sci. USA*, 72(3):780–783, 1975.

12 L. van den Dries. *Tame topology and o-minimal structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1998. doi:10.1017/CBO9780511525919.

13 W. H. Fleming and D. Hernández-Hernández. Risk-sensitive control of finite state machines on an infinite horizon. I. *SIAM J. Control Optim.*, 35(5):1790–1810, 1997. doi:10.1137/S0363012995291622.

14 W. H. Fleming and D. Hernández-Hernández. Risk-sensitive control of finite state machines on an infinite horizon. II. *SIAM J. Control Optim.*, 37(4):1048–1069 (electronic), 1999. doi:10.1137/S0363012997321498.

15 S. Gaubert and J. Gunawardena. A non-linear hierarchy for discrete event dynamical systems. In *Proc. of the Fourth Workshop on Discrete Event Systems (WODES98)*, pages 249–254, Cagliari, Italy, 1998. IEE.

16 S. Gaubert and J. Gunawardena. The Perron-Frobenius theorem for homogeneous, monotone functions. *Trans. of AMS*, 356(12):4931–4950, 2004. arXiv:math.FA/0105091, doi:10.1090/S0002-9947-04-03470-1.

17 M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

18 T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. In *Innovations in Computer Science 2011*, pages 253–263. Tsinghua University Press, 2011.

19 A. J. Hoffman and R. M. Karp. On nonterminating stochastic games. *Management Science. Journal of the Institute of Management Science. Application and Theory Series*, 12:359–370, 1966.

20 V. Kozyakin. Hourglass alternative and the finiteness conjecture for the spectral characteristics of sets of non-negative matrices. arXiv:1507.00492, 2015.

21 M. Lothaire. *Applied Combinatorics on Words*. Cambridge, 2005.

22 R. D. Nussbaum. Convexity and log convexity for the spectral radius. *Linear Algebra Appl.*, 73:59–122, 1986.

23 V. Yu. Protasov. Spectral simplex method. *Mathematical Programming*, 2015. doi:10.1007/s10107-015-0905-2.

24 M. L. Puterman. *Markov Decision Processes*. Wiley, 2005.

25 R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer-Verlag, Berlin, 1998. doi:10.1007/978-3-642-02431-3.

26 U. G. Rothblum. Multiplicative markov decision chains. *Mathematics of Operations Research*, 9(1):6–24, 1984.

27 S. M. Rump. Polynomial minimum root separation. *Mathematics of Computation*, 145(33):327–336, 1979.

28 L. van den Dries. o-minimal structures and real analytic geometry. In *Current developments in mathematics, 1998 (Cambridge, MA)*, pages 105–152. Int. Press, Somerville, MA, 1999.

29 A. J. Wilkie. 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. doi:10.1090/S0894-0347-96-00216-0.

30 Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate, 2011. doi:10.1287/moor.1110.0516.