

# Policy iteration for stochastic zero-sum games

Marianne Akian

► **To cite this version:**

Marianne Akian. Policy iteration for stochastic zero-sum games. NETCO 2014, 2014, Tours, France.  
<hal-01024097>

**HAL Id: hal-01024097**

**<https://hal.inria.fr/hal-01024097>**

Submitted on 15 Jul 2014

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

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

# Policy iteration for stochastic zero-sum games

*Marianne Akian*

INRIA Saclay - Île-de-France and CMAP, École Polytechnique

NETCO 2014  
23-27 June 2014, Tours

Joint work with Stéphane Gaubert, see arXiv:1310.4953

# Hamilton-Jacobi-Bellman-Isaacs equations

The stationary equation:

$$-H(x, Dv(x)) = 0, x \in X \subset \mathbb{R}^d \quad + \text{ a boundary condition,}$$

$$H(x, p) = \min_{a \in \mathcal{A}(x)} \max_{b \in \mathcal{B}(x)} [f(x, a, b) \cdot p + g(x, a, b)], \quad x \in X, \text{ and } p \in \mathbb{R}^d,$$

is the dynamic programming equation satisfied by the (upper) value function of the zero-sum game problem:

$$v(x) = \inf_{(\alpha_t)_{t \geq 0}} \sup_{(\beta_t)_{t \geq 0}} \int_0^\infty g(x_t, \alpha_t, \beta_t) dt,$$

where  $\dot{x}_t = f(x_t, \alpha_t, \beta_t)$ , for all  $t \geq 0$ , and  $\inf$  and  $\sup$  are taken over nonanticipating strategies of the first and second player (where the second player knows the current action of the first player).

Example: pursuit evasion games.

Discretization with a monotone scheme (for instance a Kushner scheme)

⇒

$v = F(v)$ , the fixed point equation of the dynamic programming or Shapley operator  $F$  of a discrete time zero-sum two player stochastic game problem with finite state space.

**Same for:** Discounted problems, Optimal stopping time problems, Stochastic games.

# Discrete time and state zero-sum stochastic games

Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be defined by:

$$[F(v)]_i := \min_{a \in \mathcal{A}_i} \max_{b \in \mathcal{B}_i} \left( \sum_{j \in [n]} M_{ij}^{ab} v_j + r_i^{ab} \right), \quad i \in [n],$$

with  $M_{ij}^{ab} \geq 0$  for all  $i, j \in [n], a \in \mathcal{A}_i, b \in \mathcal{B}_i$ .

The map  $F$  is the *dynamic programming or Shapley operator* of a **discrete time zero-sum two player game problem with perfect information on the finite state space  $\mathcal{X} := [n] := \{1, \dots, n\}$** , with:

$\mathcal{A}_i, \mathcal{B}_i$  sets of actions of the 1st, 2nd player MIN, MAX, when in state  $i$   
 $r_i^{ab}$  reward paid by MIN to MAX, at each time

$$M_{ij}^{ab} := \gamma_i^{ab} P_{ij}^{ab} \geq 0$$

$\gamma_i^{ab} := \sum_{j \in [n]} M_{ij}^{ab} \geq 0$  discount factor ( $< 1$  or  $\leq 1$  or  $= 1$ )

$P_{ij}^{ab}$  transition probability from  $i$  to  $j$  ( $\sum_{j \in [n]} P_{ij}^{ab} = 1$ ).

# Discrete time and state zero-sum stochastic games

Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be defined by:

$$[F(v)]_i := \min_{a \in \mathcal{A}_i} \max_{b \in \mathcal{B}_i} \left( \sum_{j \in [n]} M_{ij}^{ab} v_j + r_i^{ab} \right), \quad i \in [n],$$

with  $M_{ij}^{ab} \geq 0$  for all  $i, j \in [n]$ ,  $a \in \mathcal{A}_i$ ,  $b \in \mathcal{B}_i$ .

Denote  $\gamma_i^{ab} := \sum_{j \in [n]} M_{ij}^{ab}$ .

Then

- $F$  is order preserving:  $u \leq v \Rightarrow F(u) \leq F(v)$ , for all  $u, v \in \mathbb{R}^n$ ;
- if  $\gamma_i^{ab} \leq 1$  for all  $i \in [n]$  and  $a \in \mathcal{A}$ ,  $b \in \mathcal{B}$ , then  $F$  is additively sub-homogeneous:  $F(\lambda + u) \leq \lambda + F(u)$ , for all  $\lambda \geq 0$  and  $u \in \mathbb{R}^n$
- thus  $F$  is sup-norm nonexpansive.
- If  $\gamma_i^{ab} = 1$  for all  $i \in [n]$  and  $a \in \mathcal{A}$ ,  $b \in \mathcal{B}$ , then  $F$  is additively homogeneous:  $F(\lambda + u) = \lambda + F(u)$ , for all  $\lambda \in \mathbb{R}$  and  $u \in \mathbb{R}^n$ .

Let the value function of the game **with infinite horizon** be given by:

$$v_x = \inf_{(\alpha_k)_{k \geq 0}} \sup_{(\beta_k)_{k \geq 0}} \mathbb{E} \left[ \sum_{k=0}^{\infty} \left( \prod_{\ell=0}^{k-1} \gamma_{X_\ell}^{\alpha_\ell, \beta_\ell} \right) r_{X_k}^{\alpha_k, \beta_k} \mid X_0 = x \right],$$

where  $\alpha_k$  and  $\beta_k$  are possible strategies of both players of the game (at time  $k$ ), and  $X_k \in [n]$  is the state process of the game satisfying  $P(X_{k+1} = j \mid X_k = i, \alpha_k = a, \beta_k = b) = P_{ij}^{ab}$ .

If  $\gamma_x^{a,b} \leq \bar{\gamma} < 1$ , then  **$F$  is a sup-norm contraction**:

$$\|F(v) - F(w)\|_\infty \leq \bar{\gamma} \|v - w\|_\infty,$$

and  $v$  is the unique solution of

$$v = F(v).$$

Moreover the optimal actions in  $F(v)$  give the optimal stationary strategies of the game.

# Solving stationary dynamic programming equations

Problem: compute  $v \in \mathbb{R}^n$  such that  $F(v) = v$ , when such a solution is unique, and bound the complexity of this computation.

When  $\gamma_i^{ab} \leq \bar{\gamma} < 1$  for all  $i \in [n]$  and  $a \in \mathcal{A}$ ,  $b \in \mathcal{B}$ , then

- Then, the *value iterations* coincide with fixed point iterations:  $v^{k+1} = F(v^k)$ , and with the finite horizon approximations with  $T = k$  and  $\varphi = v^0$ . They converge geometrically towards  $v$  with factor  $\bar{\gamma}$ :

$$\lim_{k \rightarrow \infty} \|v^k - v\|^{1/k} \leq \bar{\gamma} .$$

- However, the value iteration algorithm is only pseudopolynomial.
- Also the existence of a polynomial algorithm is an open problem.
- What about the policy iteration?



# Policy iterations for discounted games

Assume:  $\mathcal{A}_i$  and  $\mathcal{B}_i$  are finite sets, and

Denote by  $\Sigma := \{\sigma : i \in [n] \mapsto \sigma_i \in \mathcal{A}_i\}$  and  $\Delta := \{\delta : i \in [n] \mapsto \delta_i \in \mathcal{B}_i\}$  the sets of policies,

and for  $\sigma \in \Sigma$  and  $\delta \in \Delta$ , define the matrices and vectors:

$$M^{(\sigma\delta)} = (M_{ij}^{\sigma_i\delta_j})_{ij=1,\dots,n}, \quad \text{and } r^{(\sigma\delta)} = (r_i^{\sigma_i\delta_i})_{i=1,\dots,n},$$

and the affine maps

$$F^{(\sigma\delta)}(v) = M^{(\sigma\delta)}v + r^{(\sigma\delta)}, \quad v \in \mathbb{R}^n.$$

Then,  $F$  can be written as:

$$F(v) = \min_{\sigma \in \Sigma} F^{(\sigma)}(v), \quad \text{with } F^{(\sigma)}(v) := \max_{\delta \in \Delta} F^{(\sigma\delta)}(v), \quad v \in \mathbb{R}^n,$$

where minima and maxima are for the partial order of  $\mathbb{R}^n$ .

The maps  $F^{(\sigma\delta)}$ ,  $F^{(\sigma)}$  and  $F$  are all order preserving and contracting for the sup-norm with contraction factor  $\bar{\gamma}$ .

**Important:** the infimum and supremum are attained because the sets  $\{F^{(\sigma)}(v) \mid \sigma \in \Sigma\}$  and  $\{F^{(\sigma\delta)}(v) \mid \delta \in \Delta\}$  are rectangular.

# Policy iterations for discounted games

(Howard, 1960) for 1-player games, (Denardo, 1967) for 2-player games.

## Using operators:

Given an initial policy  $\sigma^0 \in \Sigma$ , apply successively the two following steps for  $s \geq 0$  until  $\sigma^{s+1} = \sigma^s$ :

- 1 Compute the fixed point  $v^s$  of  $F(\sigma^s)$ ;
- 2 Improve the policy: choose an optimal policy for  $v^s$ , that is  $\sigma^{s+1} \in \Sigma$  such that  $F(v^s) = F(\sigma^{s+1})(v^s)$  with  $\sigma^{s+1} = \sigma^s$  as soon as this is possible.

Step 1 is solved by using Policy iteration for the (one-player) game with fixed policy  $\sigma^s$ , which constructs  $v^{s,l}$  and  $\delta^{s,l}$  from  $\delta^{s,0}$ .

# Policy iterations for discounted games

(Howard, 1960) for 1-player games, (Denardo, 1967) for 2-player games.

**With control terminology:**

Given an initial policy  $\sigma^0 \in \Sigma$ , apply successively the two following steps for  $s \geq 0$  until  $\sigma^{s+1} = \sigma^s$ :

- 1 Compute the value  $v^s$  of the game with fixed policy  $\sigma^s$ , that is the solution of  $v = F(\sigma^s)(v)$ ;
- 2 Improve the policy: choose an optimal policy for  $v^s$ , that is  $\sigma^{s+1} \in \Sigma$  such that  $F(v^s) = F(\sigma^{s+1})(v^s)$  or equivalently:

$$\sigma_i^{s+1} \in \operatorname{argmin}_{a \in \mathcal{A}_i} \left\{ \max_{b \in \mathcal{B}_i} \left( \sum_{j \in [n]} M_{ij}^{ab} v_j^s + r_i^{ab} \right) \right\}, \quad i \in [n],$$

with  $\sigma^{s+1} = \sigma^s$  as soon as this is possible.

# Policy iterations for discounted games

(Howard, 1960) for 1-player games, (Denardo, 1967) for 2-player games.

**Simplex algorithm for 1-player games with Dantzig pivoting:**

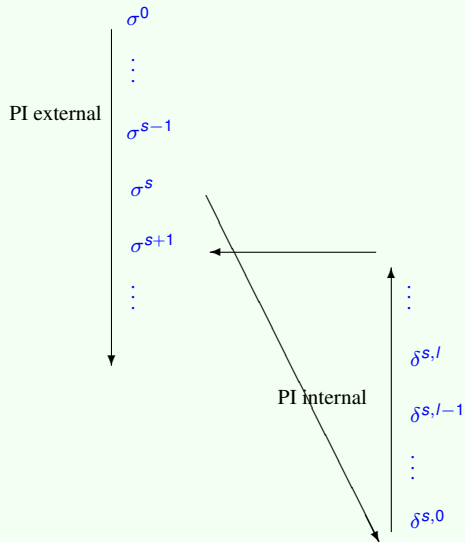
Given an initial policy  $\sigma^0 \in \Sigma$ , apply successively the two following steps for  $s \geq 0$  until  $\sigma^{s+1} = \sigma^s$ :

- 1 Compute the value  $v^s$  of the game with fixed policy  $\sigma^s$ , that is the solution of  $v = F(\sigma^s)(v)$ ;
- 2 Improve the policy: choose a policy  $\sigma^{s+1} \in \Sigma$  such that

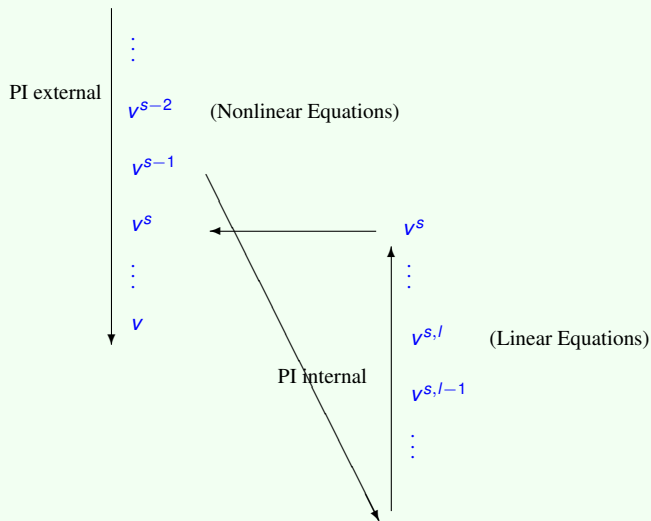
$$\sigma_i^{s+1} \in \operatorname{argmin}_{a \in A_i} \left\{ \max_{b \in B_i} \left( \sum_{j \in [n]} M_{ij}^{ab} v_j^s + r_i^{ab} \right) \right\}, \quad i \in [n],$$

for one  $i$  such that  $(F(\sigma^s)(v^s) - F(v^s))_i$  is maximal.

# Policy iterations for discounted games



# Policy iterations for discounted games



# Policy iterations for discounted games: monotone convergence

- The sequence  $(v^s)_{s \geq 0}$  is nonincreasing;
- Hence, the sequence  $(\sigma^s)_{s \geq 0}$  does not visit the same policy two times, until it becomes stationary;
- So the sequence  $(v^s)_s$  is stationary after a finite time (at most  $\#\Sigma$ ), and converges towards the solution  $v$  of  $v = F(v)$ .

# Policy iterations for discounted games: monotone convergence

- The sequence  $(v^s)_{s \geq 0}$  is nonincreasing;
- Hence, the sequence  $(\sigma^s)_{s \geq 0}$  does not visit the same policy two times, until it becomes stationary;
- So the sequence  $(v^s)_s$  is stationary after a finite time (at most  $\# \Sigma$ ), and converges towards the solution  $v$  of  $v = F(v)$ .
  
- When  $s$  is fixed, the sequence  $(v^{s,l})_l$  is nondecreasing;
- Hence, the sequence  $(\delta^{s,l})_l$  does not visit the same policy two times, until it becomes stationary;
- So the sequence  $(v^{s,l})_l$  is stationary after a finite time (at most  $\# \Delta$ ), and converges towards the solution  $v^s$  of  $v = F^{(\sigma^s)}(v)$ .



# Policy iterations for discounted games: well known properties

- The Policy iterations converge faster than the value iterations: for all  $s \geq 0$ ,  $v \leq v^{s+1} \leq F(v^s) \leq v^s$ , so  $v \leq v^s \leq F^s(v^0) \leq v^0$ .
- If the discount factor is uniformly bounded by some constant  $\bar{\gamma} < 1$ , then for the sup-norm, we have:

$$\|v^{s+1} - v\| \leq \|F(v^s) - v\| \leq \bar{\gamma} \|v^s - v\| .$$

- For 1-player games with an infinite number of actions and under regularity conditions, Policy iterations coincide with the Newton algorithm, and have a *super-linear convergence*.
- However, in general, the number of (external) iterations is bounded by  $\#\Sigma \geq 2^n$  if  $\#\mathcal{A}_i \geq 2$  for all  $i \in [n]$ .

# Policy iterations for discounted games: recent results

- (Friedmann, 2009) showed a 2-player deterministic game problem with  $\gamma \simeq 1$  and an exponential number of iterations.
- (Fearnley, 2010) and (Andersson, 2009) showed the same for a 1-player stochastic game.

# Policy iterations for discounted games: recent results

(Ye, 2011) showed that Policy iteration algorithm and Simplex algorithm solve 1-player discounted games with fixed discount factor  $\gamma < 1$  in *strongly polynomial* time.

(Hansen, Miltersen and Zwick, 2011) extended and improved this result to Policy iteration algorithm for 2-player games. They show that the number of iterations  $s_{\max}$  (to obtain stationarity) satisfies:

$$s_{\max} \leq (m + 1) \left( 1 + \frac{\log(n^2 / (1 - \gamma))}{-\log(\gamma)} \right) = \mathcal{O}\left(\frac{m}{1 - \gamma} \log \frac{n}{1 - \gamma}\right),$$

with  $m =$  the *total number of actions*: the number of  $(i, a, b)$  with  $i \in [n]$ ,  $a \in \mathcal{A}_i$  and  $b \in \mathcal{B}_i$ .

(Feinberg, Huang, 2013): Same for a one-player game with mean-payoff, and a state  $i_0$  such that  $P_{i,i_0}^a \geq 1 - \gamma$ , for all  $i \in [n]$ ,  $a \in \mathcal{A}_i$ .

Question: What remains true when the discount factors  $\gamma_i^{ab}$  are not uniformly bounded by a constant  $< 1$ ?  
or for games with mean-payoff?

## Theorem (A., Gaubert, arXiv:1310.4953)

Let us fix  $0 < \lambda < 1$ . The policy iteration algorithm for the class of 2-player games satisfying

$$r(M^{(\sigma\delta)}) \leq \lambda \quad \forall \sigma \in \Sigma, \delta \in \Delta$$

is strongly polynomial. More precisely, the number of external iterations  $S_{\max}$  satisfies:

$$S_{\max} \leq (m_1 - n) \left(1 + \left\lfloor \frac{\log(1 - \lambda)}{\log(\lambda)} \right\rfloor\right) = \mathcal{O}\left(\frac{m_1 - n}{1 - \lambda} \log \frac{1}{1 - \lambda}\right),$$

with  $m_1 =$  the total number of actions of the first player: the number of  $(i, a)$  with  $i \in [n]$  and  $a \in \mathcal{A}_i$ .

*Proof.* • Adapt the proof of (Hansen, Miltersen and Zwick, 2011) by using sup-norms instead of  $l_1$  norms and the nonlinear maps  $F^{(\delta)}$  to obtain the above bound when the discount factors are  $\leq \lambda$ . A similar bound is obtained by (Scherrer, 2013) in the one-player case with fixed discount factor.

- Using nonlinear spectral theory, show that for all  $\lambda < \mu < 1$ , there exists  $\varphi \in \mathbb{R}^n$  such that  $\varphi_i > 0$ ,  $i \in [n]$ , and  $M^{(\sigma^\delta)}\varphi \leq \mu\varphi$ .
- Let  $G(v) = \varphi^{-1}F(\varphi v)$  with  $\varphi v = (\varphi_i v_i)_{i \in [n]}$ . Then  $G$  is the dynamic programming operator of a game with discount factors  $\leq \mu$ , and the sequence of policies  $\sigma^s$  for  $F$  and  $G$  are the same, so is  $s_{\max}$ .
- Equivalently,  $F$  is contracting on  $\mathbb{R}^n$  with contraction factor  $\mu$ , for the weighted sup-norm  $\|\cdot\|_\varphi$  defined by:

$$\|v\|_\varphi := \max_{i \in [n]} \left| \frac{v_i}{\varphi_i} \right| \quad \forall v \in \mathbb{R}^n .$$

- Take the infimum of the bound over all  $\mu$ .



*Proof.* • Adapt the proof of (Hansen, Miltersen and Zwick, 2011) by using sup-norms instead of  $l_1$  norms and the nonlinear maps  $F^{(\delta)}$  to obtain the above bound when the discount factors are  $\leq \lambda$ . A similar bound is obtained by (Scherrer, 2013) in the one-player case with fixed discount factor.

- Using nonlinear spectral theory, show that for all  $\lambda < \mu < 1$ , there exists  $\varphi \in \mathbb{R}^n$  such that  $\varphi_i > 0$ ,  $i \in [n]$ , and  $M^{(\sigma\delta)}\varphi \leq \mu\varphi$ . ← details
- Let  $G(v) = \varphi^{-1}F(\varphi v)$  with  $\varphi v = (\varphi_i v_i)_{i \in [n]}$ . Then  $G$  is the dynamic programming operator of a game with discount factors  $\leq \mu$ , and the sequence of policies  $\sigma^s$  for  $F$  and  $G$  are the same, so is  $s_{\max}$ .
- Equivalently,  $F$  is contracting on  $\mathbb{R}^n$  with contraction factor  $\mu$ , for the weighted sup-norm  $\|\cdot\|_\varphi$  defined by:

$$\|v\|_\varphi := \max_{i \in [n]} \left| \frac{v_i}{\varphi_i} \right| \quad \forall v \in \mathbb{R}^n .$$

- Take the infimum of the bound over all  $\mu$ .

## Definition (Nonlinear spectral radii (Nussbaum, Mallet-Paret, 1998))

Let  $h$  be a nonlinear continuous positively homogenous map on a closed convex cone  $C$  of  $\mathbb{R}^n$  ( $h(\lambda v) = \lambda h(v)$  for all  $\lambda > 0$  and  $v \in C$ ):

- The *cone eigenvalue spectral radius* of  $h$ ,  $\hat{r}_C(h)$ , is the maximal modulus of an eigenvalue of  $h$  in  $C$ , where  $\lambda$  is an eigenvalue associated to  $v \in C \setminus \{0\}$  if  $h(v) = \lambda v$ .
- The *Collatz-Wielandt number*  $cw_C(h)$  is the infimum of the super-eigenvalues of  $h$ , where  $\lambda > 0$  is a super-eigenvalue if there exists  $v$  in the interior of  $C$  such that  $h(v) \leq \lambda v$ .
- The *Bonsall's spectral radius* of  $h$  is defined as:

$$r_C(h) := \inf_{k \geq 1} \|h^k\|_C^{1/k}, \quad \text{with} \quad \|h\|_C := \sup_{x \in C, \|x\|=1} \|h(x)\| ,$$

for any given norm  $\|\cdot\|$  on  $\mathbb{R}^n$ .

Theorem (Nussbaum, LAA 1986, also (A., Gaubert, Nussbaum, arXiv 2011))

For a continuous, positively homogenous, order preserving selfmap  $h$  of  $C = \mathbb{R}_+^n$ , all the above spectral radius notions of  $h$  coincide:

$$\begin{aligned} r(h) &= \inf_{k \geq 1} \|h^k\|_{\mathbb{R}_+^n}^{1/k} \\ &= \max\{\lambda \in \mathbb{R} \mid \exists v \in \mathbb{R}_+^n \setminus \{0\}, h(v) = \lambda v\} \\ &= \inf\{\lambda > 0 \mid \exists v \in (\mathbb{R}_+^*)^n, h(v) \leq \lambda v\} \end{aligned}$$

Proposition (A. Gaubert, Nussbaum, arXiv 2011)

Assume that  $h$  and  $h_\pi$  are continuous, positively homogenous, order preserving selfmaps of  $\mathbb{R}_+^n$ , for all  $\pi \in \Pi$ , and that  $h(v) = \max_{\pi \in \Pi} h_\pi(v)$  for all  $v \in \mathbb{R}_+^n$ , then

$$r(h) = \max_{\pi \in \Pi} r(h_\pi) .$$

Applying the proposition to  $h(v) := \max_{\sigma \in \Sigma} \max_{\delta \in \Delta} (M^{(\sigma\delta)} v)$ , we get that  $r(h) \leq \lambda < \mu$  and so by the theorem, there exists  $\varphi \in (\mathbb{R}_+^*)^n$  such that  $M^{(\sigma\delta)}\varphi \leq h(\varphi) \leq \mu\varphi$ , for all  $\sigma \in \Sigma$ ,  $\delta \in \Delta$ .



Consider the value function of the game **with mean-payoff**:

$$\eta_x = \inf_{(\alpha_k)_{k \geq 0}} \sup_{(\beta_k)_{k \geq 0}} \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[ \sum_{k=0}^{T-1} r_{x_k}^{\alpha_k, \beta_k} \mid X_0 = x \right].$$

Let  $F$  be the dynamic programming operator such that  $\gamma_i^{ab} \equiv 1$ .  $F$  is additively homogeneous. We say that  $v \in \mathbb{R}^n$  is an (*nonlinear additive*) *eigenvector* or *biais* of  $F$  with *eigenvalue*  $\rho \in \mathbb{R}$  if  $F(v) = \rho + v$ .

- If  $\rho$  exists, then  $\eta_x = \rho$  for all  $x \in [n]$ .
- If all the matrices  $M^{(\sigma\delta)}$  are irreducible, then  $\rho$  exists and the eigenvector  $v$  is unique up to an additive constant.
- Other existence results of  $\rho$ : **Bather, 1973, Gaubert, Gunawardena, 2001.**

# Policy iterations for “irreducible” mean-payoff games

(Hoffman and Karp, 1966) We have to solve  $\rho + v = F(v)$ .

Using operators:

Given an initial policy  $\sigma^0 \in \Sigma$ , apply successively the two following steps for  $s \geq 0$  until  $\sigma^{s+1} = \sigma^s$ :

- 1 Compute the additive eigenvalue and eigenvector  $\rho^s$  and  $v^s$  of  $F(\sigma^s)$ , that is the solution of  $\rho + v = F(\sigma^s)(v)$ ;
- 2 Improve the policy: choose an optimal policy for  $v^s$ , that is  $\sigma^{s+1} \in \Sigma$  such that  $F(v^s) = F(\sigma^{s+1})(v^s)$  with  $\sigma^{s+1} = \sigma^s$  as soon as this is possible.

Step 1 is solved by using Policy iteration for the (one-player) game with fixed policy  $\sigma^s$ , which constructs  $\rho^{s,l}$ ,  $v^{s,l}$  and  $\delta^{s,l}$  from  $\delta^{s,0}$ .

# Policy iterations for “irreducible” mean-payoff games

(Hoffman and Karp, 1966) We have to solve  $\rho + v = F(v)$ .

With control terminology:

Given an initial policy  $\sigma^0 \in \Sigma$ , apply successively the two following steps for  $s \geq 0$  until  $\sigma^{s+1} = \sigma^s$ :

- 1 Compute the value  $\rho^s$  and the biases  $v^s$  of the game with fixed policy  $\sigma^s$ , that is the solution of  $\rho + v = F(\sigma^s)(v)$ ;
- 2 Improve the policy: choose an optimal policy for  $v^s$ , that is  $\sigma^{s+1} \in \Sigma$  such that  $F(v^s) = F(\sigma^{s+1})(v^s)$  or equivalently:

$$\sigma_i^{s+1} \in \operatorname{argmin}_{a \in A} \left\{ \max_{b \in B} \left( \sum_{j \in [n]} M_{ij}^{ab} v_j^s + r_i^{ab} \right) \right\}, \quad i \in [n],$$

with  $\sigma^{s+1} = \sigma^s$  as soon as this is possible.

# Policy iterations for “irreducible” mean-payoff games: monotone convergence

- The sequence  $(\rho^s)_{s \geq 0}$  is nonincreasing;
- If  $\rho^s = \rho^{s+1}$ , then  $v^s - v^{s+1}$  is constant and  $v^s = v$ .
- Hence, the sequence  $(\sigma^s)_{s \geq 0}$  does not visit the same policy two times, until it becomes stationary;
- So the sequence  $(\rho^s, v^s)_s$  is stationary after a finite time (at most  $\#\Sigma$ ), up to an additive constant, and converges towards the solution  $(\rho, v)$  of  $\rho + v = F(v)$ .

# Policy iterations for “irreducible” mean-payoff games: monotone convergence

- The sequence  $(\rho^s)_{s \geq 0}$  is nonincreasing;
  - If  $\rho^s = \rho^{s+1}$ , then  $v^s - v^{s+1}$  is constant and  $v^s = v$ .
  - Hence, the sequence  $(\sigma^s)_{s \geq 0}$  does not visit the same policy two times, until it becomes stationary;
  - So the sequence  $(\rho^s, v^s)_s$  is stationary after a finite time (at most  $\#\Sigma$ ), up to an additive constant, and converges towards the solution  $(\rho, v)$  of  $\rho + v = F(v)$ .
- 
- When  $s$  is fixed, the sequence  $(\rho^{s,l})_l$  is nondecreasing;
  - If  $\rho^{s,l} = \rho^{s,l+1}$ , then  $v^{s,l} - v^{s,l+1}$  is constant and  $v^{s,l} = v^s$ .
  - Hence, the sequence  $(\delta^{s,l})_l$  does not visit the same policy two times, until it becomes stationary;
  - So the sequence  $(\rho^{s,l}, v^{s,l})_l$  is stationary after a finite time (at most  $\#\Delta$ ), and converges towards the solution  $(\rho^s, v^s)$  of  $\rho + v = F^{(\sigma^s)}(v)$ .

For a Markov matrix  $M$  and states  $i, j$ , denote:

$$\mathcal{T}_{ij}(M) = \mathbb{E}[\inf\{k \geq 1 \mid X_k = j\} \mid X_0 = i] ,$$

the expected first return (or hitting) time in state  $j$ , starting from  $i$ .  
Note that  $\mathcal{T}_{i_0}(M) < +\infty$  for all  $i \in [n]$  if and only if  $M$  has a unique recurrent (final) class and  $i_0$  belongs to it.

**Theorem (A., Gaubert, arXiv:1310.4953)**

*Let us fix  $K > 0$  and a state  $i_0$ . The policy iteration algorithm for the class of 2-player mean-payoff games such that*

$$\mathcal{T}_{i_0}(M^{(\sigma\delta)}) \leq K \quad \forall \sigma \in \Sigma, \delta \in \Delta, i \in [n]$$

*is strongly polynomial. More precisely, the number of external iterations  $S_{\max}$  satisfies:*

$$S_{\max} \leq (m_1 - n) \left(1 + \left\lfloor \frac{\log(K)}{\log(K/(K-1))} \right\rfloor\right) = \mathcal{O}((m_1 - n)K \log K),$$

*with  $m_1 =$  the total number of actions of the first player.*

*Sketch of the proof.* • Let  $\varphi \in (\mathbb{R}_+^*)^n$  be defined by:

$$\varphi_i = \max_{\sigma \in \Sigma} \max_{\delta \in \Delta} \mathcal{T}_{ii_0}(M^{(\sigma\delta)}).$$

- Let  $Q^{(\sigma\delta)}$  be obtained from  $M^{(\sigma\delta)}$  by putting its  $i_0$ th column to zero. Then  $\varphi = 1 + \max_{\sigma \in \Sigma} \max_{\delta \in \Delta} (Q^{(\sigma\delta)}\varphi)$ .
- Let  $N^{(\sigma\delta)}$  be obtained from  $M^{(\sigma\delta)}$  by replacing its  $i_0$ th column by the nonnegative vector  $(\varphi - 1 - Q^{(\sigma\delta)}\varphi)/\varphi_{i_0}$ .
- $N^{(\sigma\delta)}$  has nonnegative entries and satisfies:

$$N^{(\sigma\delta)}\varphi = \varphi - 1 \leq \lambda\varphi \quad \text{with } \lambda = 1 - 1/K \Rightarrow r(N^{(\sigma\delta)}) \leq \lambda.$$

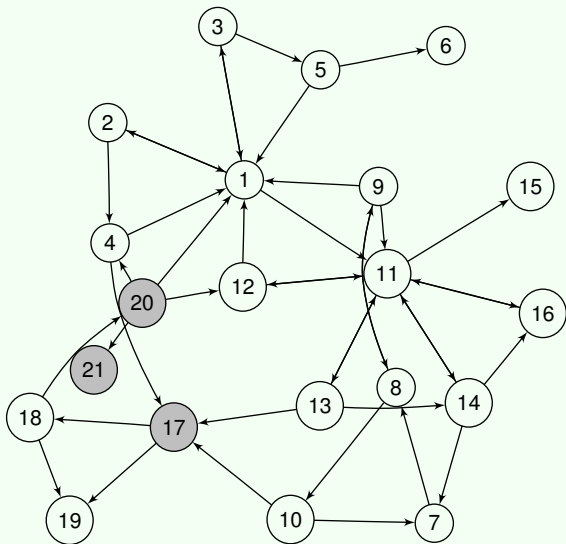
- Then the map

$$G(v) = \min_{\sigma \in \Sigma} \max_{\delta \in \Delta} (N^{(\sigma\delta)}v + r^{(\sigma\delta)}), \quad v \in \mathbb{R}^n$$

satisfies the assumptions of the theorem for discounted games.

- If  $v_{i_0} = 0$ , then  $\rho + v = F(v) \Leftrightarrow \rho\varphi + v = G(\rho\varphi + v)$ .
- Hence, the sequences of policies  $\sigma^s$  and  $\delta^{s,l}$  for  $F$  and  $G$  are the same.

## Example: Spammer vs. Web search engine



Nodes = web pages

Arcs = hyperlinks

⊙21 : spammer page

⊙1 : non controlled page.

Associated Markov matrix

$S$ :  $S_{ij} = 1/N_i$  if  $(i, j)$  is an hyperlink,  $S_{ij} = 0$  otherwise;  $N_i$  = number of hyperlinks from  $i$ .

The PageRank is the invariant measure  $\pi$  of  $S$ .



- Let  $\mathbf{v}$  be the preference probability vector of the Web search engine
- Let  $\alpha$  be a damping factor: the probability for a Web surfer to use the Web search engine.
- Usually, one replaces  $\mathbf{S}$  by  $\alpha\mathbf{S} + (1 - \alpha)\mathbf{1}\mathbf{v}$ ,  $\mathbf{1} = (1 \cdots 1)^T$ .
- Similar to consider the Markov matrix of the Web with the Web search engine:  $M = \begin{bmatrix} 0 & \mathbf{v} \\ \alpha\mathbf{1} & (1-\alpha)\mathbf{S} \end{bmatrix}$ .
- If  $r$  is an instantaneous reward such that  $r_i = 1$  for  $i = s$  and 0 otherwise, then the mean-payoff is the PageRank (frequency of visit)  $\pi_s$  of the spammer site  $s$ .
- Optimizing the spammer site is a 1-player game with mean-payoff (see for instance (Fercoq, A., Bouhtou, Gaubert, IEEE TAC 2013)).

### A zero-sum game problem:

- $\sigma \in \Sigma$  is the policy of the Web search engine, it controls  $\mathbf{v}$  and wants to minimize the PageRank of the spammer site;
- $\delta \in \Delta$  is the policy of the spammer, it controls the rows of  $\mathbf{S}$  with index in his site, and wants to maximize its PageRank.
- All final classes of  $M^{(\sigma\delta)}$  contain state 1 (the Web search engine).

In the general case, we need to apply Policy iterations for multichain mean-payoff games,...

and to find a complexity result.

## Related recent results for 1-player discounted games

- (Post, Ye, 2012) show that the simplex algorithm for deterministic MDP (1-player games) is strongly polynomial independently of the discount factor: it stops after  $\mathcal{O}(n^5 m^2 \log^2 n)$  iterations, where  $m$  is the number of possible actions by state (thus  $m_1 = nm$ ).
- (Scherrer, 2013) generalizes this result to stochastic MDP which satisfy a bound which may be seen (and is equal when the discount factor  $\gamma$  tends to 1) as a bound  $\tau_r$  on the expected first return time to recurrent states and a bound  $\tau_t$  on the expected exit time from transient states. Under these conditions the simplex algorithm stops after  $\mathcal{O}(n^3 m^2 \tau_r \tau_t \log^2(n \tau_r \tau_t))$ .
- (Scherrer, 2013) shows a similar result for Policy Iteration algorithm for stochastic MDP (1-player games), when the set of transient states is independent of the strategy. Under these conditions the Policy iteration algorithm stops after  $n(m-1)(\lceil \tau_r \log(n \tau_r) \rceil + \lceil \tau_t \log(n \tau_t) \rceil)$  iterations.
- However, this assumption implies that the recurrent classes are independent of the strategy.

## Theorem (A., Gaubert, 2014)

Let us fix  $K > 0$  and a state  $i_0$ . The policy iteration algorithm for the class of 2-player discounted games with fixed discount factor,  $M^{(\sigma\delta)} = \gamma P^{(\sigma\delta)}$  with  $\gamma < 1$ , such that

$$\mathcal{T}_{i_0}(P^{(\sigma\delta)}) \leq K \quad \forall \sigma \in \Sigma, \delta \in \Delta, i \in [n]$$

is strongly polynomial. More precisely, the number of external iterations  $S_{\max}$  satisfies:

$$S_{\max} \leq (m_1 - n) \left(1 + \left\lfloor \frac{\log(K)}{\log(K/(K-1))} \right\rfloor\right) = \mathcal{O}((m_1 - n)K \log K),$$

with  $m_1 =$  the total number of actions of the first player.

Hence the bound does not depend on  $\gamma$ .

For a Markov matrix  $M$ , a state  $i$  and set  $C$  of states, denote:

$$\mathcal{T}_{iC}(M) = \mathbb{E}[\inf\{k \geq 1 \mid X_k \in C\} \mid X_0 = i] ,$$

the expected first return (or hitting) time in set  $C$ , starting from  $i$ .

**Theorem (A., Gaubert, 2014)**

*Let us fix  $K > 0$  and a subset  $C$  of states with cardinality  $s$ . The policy iteration algorithm for the class of 2-player multichain mean-payoff games such that for all  $\sigma \in \Sigma$ ,  $\delta \in \Delta$ , each final class of  $M^{(\sigma\delta)}$  contains exactly one element of  $C$  and*

$$\mathcal{T}_{iC}(M^{(\sigma\delta)}) \leq K \quad \forall i \in [n]$$

*is strongly polynomial. More precisely, the number of external iterations  $S_{\max}$  satisfies:*

$$S_{\max} \leq (m_1 - n) \left(1 + \left\lfloor \frac{\log(sK)}{\log(sK/(sK - 1))} \right\rfloor\right) = \mathcal{O}((m_1 - n)sK \log(sK)),$$

*with  $m_1 =$  the total number of actions of the first player.*

# Multichain mean-payoff games

- In general,  $F$  may not have additive eigenvalue and eigenvector, that is  $\rho$  and  $v$  such that  $\rho + v = F(v)$ .
- If the action spaces  $\mathcal{A}_i$  and  $\mathcal{B}_i$  are finite for all  $i \in [n]$ , then  $F$  is *polyhedral*, and since it is also nonexpansive, by the **Kohlberg (1980)** theorem, there exist  $\eta$  and  $v$  in  $\mathbb{R}^n$  such that

$$F(t\eta + v) = (t + 1)\eta + v, \text{ for } t \text{ large enough.}$$

- $(\eta, v)$  is called an *invariant half-line*.
- Then  $\eta$  is the value of the game with mean-payoff.
- Moreover, there exist  $\hat{F}$  and  $\hat{F}_\eta$  such that  $(\eta, v)$  is an invariant half-line if and only if it satisfies the system:

$$\begin{cases} \eta = \hat{F}(\eta) , \\ \eta + v = \hat{F}_\eta(v) . \end{cases}$$

- However  $v$  is not unique.

# Policy iterations for multichain mean-payoff games

Construct a sequence of policies  $\sigma^s$ , values  $\eta^s$  and biases  $v^s$ .

They were introduced and proved to converge by

- (Howard, 1960) and (Denardo and Fox, 1968) for 1-player multichain mean-payoff games,
- (Vöge and Jurdziński, 2000) for parity games,
- (Cochet-Terrasson, Gaubert, Gunawardena, 1998 and 1999), (Bjorklund, Sandberg, Vorobyov, 2004), (Jurdziński, Paterson, Zwick, 2006) for 2-player deterministic games,
- (Cochet-Terrasson and Gaubert, 2006), (A., Cochet-Terrasson, Detournay, and Gaubert, arXiv:1208.0446, and CDC 2013), (Detournay, PIGAMES library, 2012), (Bourque, Raghavan, preprint, 2012) for general multichain 2-player stochastic games. (Detournay, 2012).

To avoid cycling, one need to add some constraints on  $v^s$ , for instance:

- fix the value  $v_i^s = 0$  at one point  $i$  of each final class of  $M(\sigma^\delta)$  (Howard, and Denardo and Fox, for one-player games);
- by a nonlinear projection (Cochet-Terrasson and Gaubert);

and to choose optimal policies in a conservative way

## Summary:

- The policy iteration algorithm for discounted games is strongly polynomial when restricted to the class of games such that *the spectral radii of all  $M^{(\sigma\delta)}$  are bounded by  $\lambda < 1$* . This result is invariant by diagonal scaling.
- The policy iteration algorithm for ergodic mean-payoff games is strongly polynomial when restricted to the class of ergodic games such that *the expected first return (or hitting) time in some fixed state  $i_0$  of the Markov chain associated to any  $M^{(\sigma\delta)}$  and initial state is bounded by  $K < \infty$* .
- Same result for *discounted games*.
- Same result for multichain mean-payoff games, when  $i_0$  is replaced by a set of states  $C$ , and each recurrence class contains exactly one element of  $C$ .

## Open:

- Is the policy iteration algorithm for multichain stochastic games strongly polynomial, under some more general constraints on the  $M^{(\sigma\delta)}$  (only)?