Solving Ergodic Markov Decision Processes and Perfect Information Zero-sum Stochastic Games by Variance Reduced Deflated Value Iteration

Marianne Akian*, Stéphane Gaubert*, Zheng Qu†, Omar Saadi*

INRIA and CMAP, École polytechnique*, The University of Hong Kong†

Perfect information stochastic games

Shapley operator $T : \mathbb{R}^n \to \mathbb{R}^n$; for $i \in [n], v \in \mathbb{R}^n$:

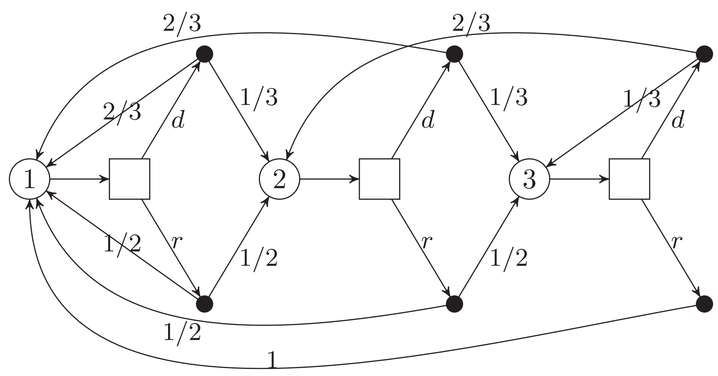
$$T_i(v) = \min_{a \in A_i} \max_{b \in B_{i,a}} \left\{ r_i^{ab} + \gamma_i^{ab} \sum_{j \in [n]} P_{ij}^{ab} v_j \right\} .$$

The value vector $v = (v_i)_{i \in [n]}$ of the infinite horizon discounted game is solution of the fixed point problem $v = T(v)$.

$$R := \max_{(i,a,b) \in E} |r_i^{ab}| \in \mathbb{R}_+, \quad \Gamma := \max_{(i,a,b) \in E} \gamma_i^{ab} \in (0, \infty);$$

$$E := \{(i, a, b) \mid i \in [n], a \in A_i, b \in B_{i,a}\} .$$

One player example



Mean payoff problem: discount factor $\gamma \equiv 1$

If the non-linear eigenproblem

$$\eta e + v = T(v)$$

is solvable for some $\eta \in \mathbb{R}$, $v \in \mathbb{R}^n$, with $e := (1, \dots, 1)^\top \in \mathbb{R}^n$, then the mean payoff vector $\chi(T) := \lim_{k \to \infty} T^k(0)/k$ coincides with ηe, where $e := (1, \dots, 1)^\top \in \mathbb{R}^n$, then $\chi(T) = \eta e$.

Previous complexities: discounted case

Value Iteration (VI) [1]	$O\left(\frac{\lvert E\rvert n \log\left(\frac{R}{(1-\gamma)\epsilon}\right)}{1-\gamma}\right)$
Variance Reduced VI (Sidford et al. [2])	$\tilde{O}\left(\frac{\lvert E\rvert R^2}{(1-\gamma)^4 \epsilon^2} \log(\frac{1}{\delta})\right)$

Table 1: Running times to compute ϵ-approximate solutions.

Deflation and Doob's h-transform ...

Key assumption: there is a distinguished state c to which all other states have access, for all policies of the two players. The maximal first hitting time of c

$$\varphi_i^\star := \max_{\sigma,\tau} \mathbb{E}_{i,\sigma,\tau}[\inf\{k \geq 1 \mid X_k = c\}] < +\infty \ , \qquad (X_0, X_1, \dots) \text{ trajectory determined by policies } \sigma, \tau \text{ of the players}$$

is the smallest solution φ of

$$\varphi_i \geq 1 + \max_{a,b} [P_{(c)i}^{ab} \varphi], \qquad \forall i \in [n], \tag{1}$$

where $P_{(c)} \in \mathbb{R}^{n \times n}$ is the matrix obtained from P by replacing the column c of P with zeros. We suppose an a priori bound of first hitting times $\|\varphi^\star\|_\infty \leq H := \frac{1}{1-\lambda}$ is known.

$$T_i^\varphi(w) = \min_{a \in A_i} \max_{b \in B_{i,a}} \left\{ \varphi_i^{-1} P_i^{ab} \varphi (w - w_c e) + \varphi_i^{-1} r_i^{ab} + w_c (1 - \varphi_i^{-1}) \right\}, \forall i \in [n], \forall w \in \mathbb{R}^n. \tag{2}$$

...reduce the mean payoff problem to the discounted one

Theorem 2. *The mean-payoff problem*

$$\eta e + v = T(v) \tag{3}$$

where $\eta \in \mathbb{R}$ and $v \in \mathbb{R}^n$ with $v_c = 0$ is equivalent to the discounted problem:

$$T^\varphi(w) = w \tag{4}$$

where $w \in \mathbb{R}^n$ is such that $w = \eta + \varphi^{-1} v$. The operator T^φ is a contraction of rate $1 - 1/\|\varphi\|_\infty$ in $\|\cdot\|_\infty$.

Example

In every state i, there are two actions r (reset) and d (drift): $P_{n,1}^r = 1$, $P_{1,1}^d = 2/3$, $P_{1,2}^d = 1/3$, $P_{n,n-1}^d = 2/3$, $P_{n,n}^d = 1/3$ and

$$P_{i,1}^r = P_{i,i+1}^r = 1/2, \ \forall 1 \leq i \leq n-1, \quad P_{i,i-1}^d = 2/3, \quad P_{i,i+1}^d = 1/3, \ \forall 2 \leq i \leq n-1.$$

Choosing $c = 1$, we get $\|\varphi^\star\|_\infty = O(n)$. Note that $c = n$ would lead to a hitting time $\Omega(2^n)$ for the "reset everywhere" policy.

Two stage approach

1. compute an approximation of $\varphi^\star$, this is a stochastic shortest path problem.
2. deduce the mean payoff from Thm. 2 by solving the discounted problem (4).

In the first case, the Shapley operator is also a λ-contraction, but in the weighted sup-norm $\|x\|_{\varphi^\star} := \max_{i \in [n]} |x_i / \varphi_i^\star|$.

Structured input and weighted sup-norm contraction

Both stages can be handled in a unified manner by adapting the method of Sidford et al., to find a fixed point of a Shapley operator of the following structured form

$$\bar{T}_i(w) = \min_{a \in A_i} \max_{b \in B_{i,a}} \left\{ \gamma_i^{ab} P_i^{ab} L w + G_i^{ab}(w) \right\}, w \in \mathbb{R}^n \tag{5}$$

where L (linear) and G_i^{ab} (affine) are sparse, and $\bar{T}$ is $\lambda-$contracting in a weighted norm $\|\cdot\|_\psi$, where $\psi \gg 0$, and $\|w^\star\|_\psi := \max_{i \in [n]} \frac{|w_i^\star|}{\psi_i} \leq W$. We assume we can sample in time $O(1)$ according to every law $P_i^{ab} = (P_{ij}^{ab})_{j \in [n]}$.

Variance reduced deflated value iteration for the Mean Payoff problem

Theorem 3. *With probability $1 - \delta$, the call of Algorithm 4, `SSublinearRandVI`$(\frac{1}{4}, \frac{\delta}{2}, \lambda, 1, 1, 1, \frac{1}{1-\lambda})$ allows to obtain φ satisfying (1) with $\|\varphi\|_\infty \leq 2\|\varphi^\star\|_\infty + 1$. Then, a new call of Algorithm 4, `SSublinearRandVI`$(\epsilon, \frac{\delta}{2}, 1 - \frac{1}{\|\varphi\|_\infty}, R, 1, 1, 1)$ returns $w \in \mathbb{R}^n$ such that $\|w - w^\star\|_\infty \leq \epsilon$. Therefore we obtain $\eta = w_c$ and $v = \varphi(w - w_c e)$ such that $|\eta - \eta^\star| \leq \epsilon$ and $\|v - v^\star\|_\infty \leq \frac{5\epsilon}{1-\lambda}$. The total execution time is*

$$\tilde{O}\left(|E| \left[\frac{R^2}{(1-\lambda)^4 \epsilon^2} + \frac{1}{(1-\lambda)^6} \right] \right) \log(\frac{1}{\delta}) \Bigg) \ .$$

Deflation and h-transform repair the absence of one-shot contraction in relative value iteration

Relative value iteration [3] solves the mean payoff problem by computing the normalized sequence $x^{k+1} = T(x^k) - (T(x^k) \cdot e)/n$. It converges under the demanding assumption that a Dobrushin ergodicity coefficient is < 1, e.g., if there is a Doeblin state δ (such that $\mathbb{P}(X_{k+1} = \delta | X_k = i) \geq \epsilon > 0$ for all i and for all policies). Then, T is an $1 - \epsilon$ contraction in Hilbert's seminorm $\|x\|_H := \max_i x_i - \min_i x_i$. Such assumptions are not needed for the deflation and h-transform method to converge. Note in particular that relative VI fails for "cyclic models".

Variance Reduced Value Iteration of Sidford el al., extended to structured weighted sup-norm contracting Shapley operators

Algorithm 1: Approximate transition with cemetery: `ApxTransC`$(u, M, i, a, b, \epsilon, \delta)$

Data: Vector $u \in \mathbb{R}^n$ and $M \geq 0$ such that $\|u\|_\infty \leq M$. State $i \in [n]$ and actions $a \in A_i$, $b \in B_{i,a}$. Target accuracy $\epsilon \in (0, 1)$ and failure probability $\delta \in (0, 1)$.

$u_0 = 0$;
$m = \lceil \frac{2M^2}{\epsilon^2} \ln(\frac{2}{\delta}) \rceil$;
for $k \in [m]$ **do**
  choose $i_k \in [n] \cup \{0\}$ with probabilities $\mathbb{P}(i_k = j) = \bar{P}_{ij}^{ab}$ for $j \in [n] \cup \{0\}$.
end
return $Y = \frac{1}{m} \sum_{k \in [m]} u_{i_k}$;

Algorithm 2: Structured approximate value operator: `SApxVal`$(w, w_0, x, \epsilon, \delta)$

Data: Current vector $w \in \mathbb{R}^n$ and initial vector $w_0 \in \mathbb{R}^n$. Precomputed offsets: $x \in \mathbb{R}^E$ with $|x_i^{ab} - P_i^{ab} L w_0| \leq \epsilon$ for all $i \in [n], a \in A_i$, $b \in B_{i,a}$. Target accuracy $\epsilon \in (0, 1)$ and failure probability $\delta \in (0, 1)$.

Compute $M = \|L\|_\infty \|w - w_0\|_\infty$;
Compute $u = L(w - w_0)$;
for $i \in [n]$ **do**
  for $a \in A_i$ **do**
    for $b \in B_{i,a}$ **do**
      $\tilde{S}_i^{ab} = x_i^{ab} +$ `ApxTransC`$(u, M, i, a, b, \epsilon, \frac{\delta}{|E|})$;
      $\tilde{Q}_i^{ab} = \gamma_i^{ab} \tilde{S}_i^{ab} + G_i^{ab}(w)$;
    end
    $\tilde{w}_i^a = \max_{b \in B_{i,a}} \tilde{Q}_i^{ab}$, $\tau(i, a) \in \operatorname{argmax}_{b \in B_{i,a}} \tilde{Q}_i^{ab}$;
  end
  $\tilde{w}_i = \min_{a \in A_i} \tilde{w}_i^a$, $\sigma(i) \in \operatorname{argmin}_{a \in A_i} \tilde{w}_i^a$;
end
return $(\tilde{w}, \sigma, \tau)$;

Algorithm 3: Structured sampled randomized VI: `SSampledRandVI`$(w_0, J, \epsilon, \delta)$

Data: Initial vector $w_0 \in \mathbb{R}^n$ and number of iterations $J > 0$. Target accuracy $\epsilon \in (0, 1)$ and failure probability $\delta \in (0, 1)$.

Sample to obtain approximate offsets: $\tilde{x} \in \mathbb{R}^E$ such that with probability $1 - \frac{\delta}{2}$, $|\tilde{x}_i^{ab} - P_i^{ab} L w_0| \leq \epsilon$ for all $i \in [n], a \in A_i$, $b \in B_{i,a}$:
$\tilde{x}_i^{ab} =$ `ApxTransC`$(w_0, \|L\|_\infty \|w_0\|_\infty, i, a, b, \epsilon, \frac{\delta}{2|E|})$;
for $j \in [J]$ **do**
  $(w_j, \sigma_j, \tau_j) =$ `SApxVal`$(w_{j-1}, w_0, x, \epsilon, \frac{\delta}{2J})$;
end
return (w_J, σ_J, τ_J);

Algorithm 4: Structured sublinear randomized VI (5): `SSublinearRandVI`$(\epsilon, \delta, \lambda, W, \Gamma, \|\psi^{-1}\|_\infty, \|\psi\|_\infty)$

Data: Target accuracy $\epsilon \in (0, 1)$ and failure probability $\delta \in (0, 1)$.

Let $K = \lceil \log_2(\frac{\|\psi\|_\infty W}{\epsilon}) \rceil$ and $J = \lceil \frac{1}{1-\lambda} \log(4) \rceil$;
$w_0 = 0$ and $\epsilon_0 = W$;
for $k \in [K]$ **do**
  $\epsilon_k = \frac{\epsilon_{k-1}}{2} = \frac{W}{2^k}$;
  $(w_k, \sigma_k, \tau_k) =$ `SSampledRandVI`$(w_{k-1}, J, \frac{1-\lambda}{4\|\psi^{-1}\|_\infty \Gamma} \epsilon_k, \frac{\delta}{K})$;
end
return (w_K, σ_K, τ_K);

Sublinear complexity (5)

Theorem 1. *With probability $1 - \delta$, the vector w_K returned by Algorithm 4 satisfies $\|w_K - \bar{w}^\star\|_\infty \leq \epsilon$. The algorithm runs in time:*

$$\tilde{O}\left(|E| \Gamma^2 \|\psi\|_\infty^2 \|\psi^{-1}\|_\infty^2 \left[\frac{\|\psi\|_\infty^2 W^2}{(1-\lambda)^2 \epsilon^2} + \frac{1}{(1-\lambda)^3} \right] \|L\|_\infty^2 \log(\frac{1}{\delta}) \right) \ .$$

(Compare with the input size of order $|E|n$.)

References

[1] P. Tseng. Solving H-horizon, stationary Markov decision problems in time proportional to $\log(H)$. *Operations Research Letters*, 9(5):287–297, 1990.

[2] A. Sidford, M. Wang, X. Wu, and Y. Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *29th ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787, 2018.

[3] A. Gupta, R. Jain, and P. W. Glynn. An empirical algorithm for relative value iteration for average-cost MDPs. In *54th IEEE Conference on Decision and Control (CDC)*, pages 5079–5084, Dec 2015.

[4] M. Akian and S. Gaubert. Policy iteration for perfect information stochastic mean payoff games with bounded first return times is strongly polynomial, 2013. arXiv:1310.4953.