UVIP: Model-Free Approach to Evaluate Reinforcement Learning Algorithms

Denis Belomestny

Duisburg-Essen University

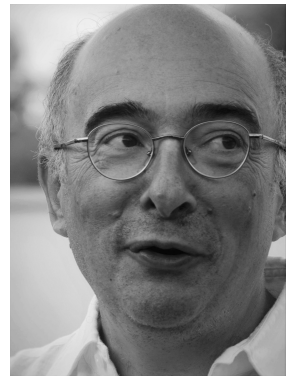
HSE University

July 16, 2021

Joint work with

Sergey Samsonov

Ilya Levin



Eric Moulines

Alexey Naumov

Veronika Zorina

Paper available at https://arxiv.org/abs/2105.02135

Markov Decision Process (MDP)

- X - state space. By $(X_k)_{k\geq 0}$ we denote a sequence of random states.
- A - action space. Let $(A_k)_{k\geq 0}$ be a sequence of random actions.
- Agent's policy π is the distribution on A

$$\pi(a|x) = \mathbb{P}(A_k = a|X_k = x)$$

- Family of Markov transition kernels $(\mathrm{P}^a(x'|x))_{a\in \mathrm{A}}$:

$$\mathrm{P}^a(x'|x) := \mathbb{P}(X_k = x'|X_{k-1} = x, A_{k-1} = a)\,.$$

- (Deterministic) reward $r^a(x) : \mathrm{A} \times \mathrm{X} \mapsto \mathbb{R}$
- At step k in the state $X_k = x$ the agent performs an action $A_k \sim \pi(\cdot|x)$ (suppose $A_k = a$), obtains a reward $r^a(X_k)$ and transits to $X_{k+1} \sim \mathrm{P}^a(\cdot|x)$

Markov Decision Process

Let $\gamma \in (0, 1]$ be the discount factor. Tuple $(\mathrm{X}, \mathrm{A}, \mathrm{P}, \mathrm{r}, \gamma)$ is called Markov Decision Process.

How to measure policy's quality?

Given policies π_1 and π_2, how to compare their quality?

How to measure policy's quality?

Given policies π_1 and π_2, how to compare their quality?

Value function, associated with the policy π, is defined as

$$V^{\pi}(x) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r^{A_k}(X_k) | X_0 = x\right]$$

How can we estimate this quantity?

1. **Monte-Carlo.** Run a series of independent simulations to compute $V^{\pi}(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r^{A_k}(X_k) | X_0 = x\right]$.
2. **Stochastic Approximation.** Use Robbins-Monro procedure to obtain *Temporal difference*-based method (e.g. $TD(0)$, Sutton [1988], Tsitsiklis and Van Roy [1997]). Preliminary estimates are available even at the early stage.

Optimal policy and optimal value function

We call policy $\pi^\star$ an *optimal policy*, if $V^{\pi^\star}(x) \geq V^\pi(x)$ for any $x \in \mathsf{X}$.

The following result holds (see, e.g., Puterman [2014]):

Theorem

When the reward function is bounded, one can always find a *deterministic* Markov policy that is optimal. Moreover, the optimal value function $V^\star$ satisfies the *Bellman optimality equation*:

$$V^\star(x) = \max_{a \in \mathsf{A}} \left[r^a(x) + \gamma \, \mathsf{P}^a \, V^\star(x) \right] .$$

In the formula above $\mathsf{P}^a \, V^\star(x) = \int_{\mathsf{X}} V^\star(x') \, \mathsf{P}^a(dx'|x)$.

Policy quality metrics

How far is the given policy π from the optimal one $\pi^\star$?

- Popular performance metrics is the *total regret* (Azar et al. [2017], Jin et al. [2018]) of the learning algorithm with respect to an optimal policy.
- Given a sequence of policies $(\pi_k)_{k\in\mathbb{N}}$ and a sequence of episodes starting from the states $(x_0^k)_{k\in\mathbb{N}}$, we define the regret after K iterations as

$$\sum_{k=1}^{K} \left[V^\star(x_0^k) - V^{\pi_k}(x_0^k) \right].$$

- Regret bounds are available for many tabular MDP learning algorithms. Yet the regret analysis does not provide much information for the *policy approximation methods*.

Policy quality metrics

How far is the given policy π from the optimal one $\pi^\star$?

- Consider, for example, the family of *policy approximation methods*;
- We approximate $\pi(a|x)$ by the parametric family of policies $\pi_\theta(a|x)$, $\theta \in \Theta \subseteq \mathbb{R}^{d'}$;
- It is hard to quantify the *approximation error*

$$V^\star(x) - \sup_{\theta \in \Theta} V^{\pi_\theta}(x)\,,$$

  moreover, the bounds on approximation error are typically pessimistic and relies upon unknown regularity properties of given MDP;
- Is it a generic way to estimate the *policy error*

$$\Delta_\pi(x) \doteq V^\star(x) - V^\pi(x)\,,$$

  if $\pi^\star$ (and $V^\star$) are unknown?

Value iteration algorithm

- First idea: construct an upper biased estimate of $V^\star(x)$;
- Assume that the transition kernel $(\mathrm{P}^a)_{a\in\mathrm{A}}$ is known;
- The *value iteration* algorithm (Bertsekas and Shreve [1978]): starting from some $V_0(x)$, iterate

$$V_{k+1} = \max_{a\in\mathrm{A}} [r^a(x) + \gamma\, \mathrm{P}^a\, V_k(x)]\ .$$

- An important property: $V_k(x) \geq V^\star(x)$ for any $x \in \mathrm{X}$ and $k \in \mathbb{N}$, provided that $V_0(x) \geq V^\star(x)$.
- Unfortunately, this property is lost if $(\mathrm{P}^a)_{a\in\mathrm{A}}$ is not known and could be only approximated.

Model assumptions

1. Our aim is to construct an (upper biased) estimate of $\Delta_\pi(x)$;
2. We consider infinite-horizon MDPs with discount factor $\gamma < 1$;
3. We can sample from the conditional distribution $\mathrm{P}^a(\cdot|x)$ for any $x \in \mathrm{X}$ and $a \in \mathrm{A}$;
4. Value function $V^\pi(x)$ is known .

Upper solutions concept

Upper solution

We call a function V^{up} an *upper solution* to the Bellman optimality equation, if

$$V^{\mathrm{up}}(x) \geq \max_{a \in \mathsf{A}} \{ r^a(x) + \gamma \, \mathsf{P}^a \, V^{\mathrm{up}}(x) \} \, , \forall x \in \mathsf{X} \, . \tag{1}$$

If $V^{\mathrm{up}}(x)$ satisfies (1), then using the Bellman optimality equation we can show that $V^{\mathrm{up}}(x) \geq V^{\star}(x)$ for any $x \in \mathsf{X}$.

The upper solution concept

- The concept is related to the martingale duality approach in optimal control;
- Rogers [2007];
- Belomestny and Schoenmakers [2018];
- Shar and Jiang [2020] used information relaxation approach to learn (approximate) upper and lower bounds for the optimal action-value function.

How to construct an upper solution?

- In practice we can construct an upper solution as follows. Consider arbitrary *martingale function* $\Phi : \mathsf{X} \mapsto \mathbb{R}$, such that $\mathsf{P}^a \Phi(x) = 0$ for any $a \in \mathsf{A}, x \in \mathsf{X}$;
- Define V^{up} as a solution to the following fixed point equation:

$$V^{\mathrm{up}}(x) = \mathbb{E}[\max_a \{r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))\}]\,,$$

where $Y^{x,a} \sim \mathsf{P}^a(\cdot|x)\,.$

Checking upper solution properties

Indeed, we easily check that V^{up} is an upper solution:

$$\begin{aligned} V^{\mathrm{up}}(x) &= \mathbb{E}[\max_a \{r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))\}] \\ &\geq \max_a \mathbb{E}[r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))] \\ &= \max_a \{r^a(x) + \gamma \, \mathsf{P}^a \, V^{\mathrm{up}}(x)\} \end{aligned}$$

Solving the fixed point problem

- Recall that V^{up} as a solution to the following fixed point equation:

$$V^{\mathrm{up}}(x) = \mathbb{E}[\max_a \{r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))\}]\,;$$

- Consider the iteration process

$$V^{\mathrm{up}}_{k+1}(x) = \mathbb{E}\big[\max_a \{r^a(x) + \gamma(V^{\mathrm{up}}_k(Y^{x,a}) - \Phi(Y^{x,a}))\}\big]$$

- The upper biasedness property is preserved: if $V^{\mathrm{up}}_0(x) \geq V^\star(x)$ for any $x \in \mathsf{X}$, then $V^{\mathrm{up}}_k(x) \geq V^\star(x)$ for any k and x.

UVIP

- Given a policy π to be evaluated and the corresponding value function V^π, we set
$$\Phi_\pi^{x,a}(y) \doteq V^\pi(y) - (\mathrm{P}^a\, V^\pi)(x)\,.$$
- Then we write the update of the Upper value iterative procedure (UVIP) as follows:
$$V_{k+1}^{\mathrm{up}}(x) = \mathbb{E}\big[\max_a\{r^a(x) + \gamma(V_k^{\mathrm{up}}(Y^{x,a}) - \Phi_\pi^{x,a}(Y^{x,a}))\}\big]$$
- We enjoy the same upper biasedness property: $V_k^{\mathrm{up}}(x) \geq V^\star(x)$ for any k and x provided that $V_0^{\mathrm{up}}(x) \geq V^\star(x)$;
- Then we can evaluate the policy π by computing the difference
$$\Delta_{\pi,k}^{\mathrm{up}}(x) \doteq V_k^{\mathrm{up}}(x) - V^\pi(x) \geq \Delta_\pi(x)\,.$$

Self-normalizing property

- Consider the case $\pi = \pi^\star$, then the corresponding martingale function is given by
$$\Phi^{x,a}(y) \doteq V^\star(y) - (\mathrm{P}^a\, V^\star)(x)\,.$$
- The corresponding fixed point equation writes as
$$V^{\mathrm{up}}(x) = \mathbb{E}[\max_a\{r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi^{x,a}(Y^{x,a}))\}] \qquad (2)$$
- It is easy to check that $V^\star(x)$ in this scenario is a solution to (2), indeed,
$$\begin{aligned} V^\star(x) &= \mathbb{E}[\max_a\{r^a(x) + \gamma(V^\star(Y^{x,a}) - V^\star(Y^{x,a}) + (\mathrm{P}^a\, V^\star)(x))\}] \\ &= \max_a\{r^a(x) + \gamma\, \mathrm{P}^a\, V^\star(x)\}\,, \end{aligned}$$
which coincides with Bellman optimality equation.
- Thus $\Delta^{\mathrm{up}}_{\pi,k}(x) \to 0$ as $k \to \infty$.

Approximate UVIP

- Consider the $(k+1)$-th UVIP iteration:

$$V_{k+1}^{\mathrm{up}}(x) = \mathbb{E}\big[\max_{a}\{r^a(x) + \gamma(V_k^{\mathrm{up}}(Y^{x,a}) - \Phi_\pi^{x,a}(Y^{x,a}))\}\big]$$

- We need to approximate an outer expectation and the one-step transition operator P^a.

Tabular UVIP: Monte-Carlo version

Algorithm 1: UVIP

Require: $V^{\pi}, \widehat{V}_0^{\mathrm{up}}, \gamma, \varepsilon, M_1, M_2$

Ensure: V^{up}

for $x \in \mathrm{X}, \mathrm{a} \in \mathrm{A}$ **do**

$\overline{V}(x,a) = M_1^{-1} \sum_{i=1}^{M_1} V^{\pi}(Y_i^{x,a}), \;\; Y_i^{x,a} \sim \mathrm{P}^a(\cdot|x)$

for $y \in \mathrm{X}$ **do**

$\Phi_{\pi}^{x,a}(y) = V^{\pi}(y) - \overline{V}(x,a)$

end for

end for

$k = 1$

while $\| \widehat{V}_k^{\mathrm{up}} - \widehat{V}_{k-1}^{\mathrm{up}} \|_{\mathrm{x}} > \varepsilon$ **do**

for $x \in \mathrm{X}$ **do**

$\widehat{V}_{k+1}^{\mathrm{up}}(x) = M_2^{-1} \sum_{i=1}^{M_2} [\max_a \{ r^a(x) + \gamma (\widehat{V}_k^{\mathrm{up}}(Y_i^{x,a}) - \Phi_{\pi}^{x,a}(Y_i^{x,a})) \}],$
$Y_i^{x,a} \sim \mathrm{P}^a(\cdot|x)$

end for

$k = k + 1$

end while

$V^{\mathrm{up}} = \widehat{V}_k^{\mathrm{up}}$

UVIP in general state space

Assumption A1

We suppose that $(\mathsf{X}, \rho_{\mathsf{X}})$ and $(\mathsf{A}, \rho_{\mathsf{A}})$ are compact metric spaces. Moreover, $\mathsf{X} \times \mathsf{A}$ is equipped with some metric ρ, such that $\rho\big((x, a), (x', a)\big) = \rho_{\mathsf{X}}(x, x')$ for any $x, x' \in \mathsf{X}$ and $a \in \mathsf{A}$.

In this scenario we fix a set of points $\mathsf{X}_{\mathsf{N}} = (\mathsf{x}_1, \ldots, \mathsf{x}_{\mathsf{N}})$ and aim to evaluate $\Delta_\pi(x)$ for $x \in \mathsf{X}_{\mathsf{N}}$.

UVIP in general state space

- Recall that $(k+1)$-th UVIP iteration writes as

$$V_{k+1}^{\mathrm{up}}(x) = \mathbb{E}\big[\max_a\{r^a(x) + \gamma(V_k^{\mathrm{up}}(Y^{x,a}) - \Phi_\pi^{x,a}(Y^{x,a}))\}\big]\,.$$

  In order to update $\hat{V}_{k+1}^{\mathrm{up}}(\mathsf{x}_i)$ the algorithm requires to calculate $\hat{V}_k^{\mathrm{up}}(Y^{\mathsf{x}_i,a})$ with $Y^{\mathsf{x}_i,a} \sim \mathsf{P}^a(\cdot|x)$;
- This means that we need an additional interpolation step to calculate $\hat{V}_k(y)$ at points $y \notin (\mathsf{x}_1, \ldots, \mathsf{x}_N)$;
- For example, for a Lipshitz function $f \in \mathrm{Lip}(L)$ we can use *optimal central interpolant*

$$\begin{aligned} I[f](x) &= (H_f^{\mathrm{low}}(x) + H_f^{\mathrm{up}}(x))/2, \text{ where} \\ H_f^{\mathrm{low}}(x) &= \max_{\ell\in\{1,\ldots,N\}} (f(\mathsf{x}_\ell) - L\rho_{\mathsf{X}}(x, \mathsf{x}_\ell)), \\ H_f^{\mathrm{up}}(x) &= \min_{\ell\in\{1,\ldots,N\}} (f(\mathsf{x}_\ell) + L\rho_{\mathsf{X}}(x, \mathsf{x}_\ell)). \end{aligned}$$

Approximate UVIP

Input: Sample $(\mathrm{x}_1, \ldots, \mathrm{x}_N)$; V^π, $\widetilde{V}_0^{\mathrm{up}}$, M_1, M_2, γ, ε

while $\|\widetilde{V}_k^{\mathrm{up}} - \widetilde{V}_{k-1}^{\mathrm{up}}\|_\infty > \varepsilon$ **do**

  for $a \in \mathrm{A}$ **do**

    for $i \in [N]$ **do**

      for $j \in [M_1 + M_2]$ **do**

        $\widehat{V}_k^{\mathrm{up}}(Y_j^{\mathrm{x}_i,a}) = I[\widetilde{V}_k^{\mathrm{up}}](Y_j^{\mathrm{x}_i,a}),\ Y_j^{\mathrm{x}_i,a} \sim \mathrm{P}^a(\cdot|\mathrm{x}_i)$

      end

      $$\overline{V}^{(i,a)} = M_1^{-1} \sum_{j=1}^{M_1} V^\pi(Y_j^{\mathrm{x}_i,a});$$

    end

  end

  for $i \in [N]$ **do**

    $$\widetilde{V}_{k+1}^{\mathrm{up}}(\mathrm{x}_i) = M_2^{-1} \sum_{j=M_1+1}^{M_1+M_2} \max_{a \in \mathcal{A}} \left\{ r^a(\mathrm{x}_i) + \gamma \left(\widehat{V}_k^{\mathrm{up}}(Y_j^{\mathrm{x}_i,a}) - V^\pi(Y_j^{\mathrm{x}_i,a}) + \overline{V}^{(i,a)} \right) \right\};$$

  end

end

Theoretical assumptions

Assumption A2

There exists a measurable mapping $\psi : \mathsf{X} \times \mathsf{A} \times \mathbb{R}^{\mathsf{m}} \to \mathsf{X}$ such that $Y^{x,a} = \psi(x, a, \xi)$, where ξ is a random variable with values in $\Xi \subseteq \mathbb{R}^m$ and distribution P_ξ on Ξ, that is, $\psi(x, a, \xi) \sim \mathsf{P}^a(\cdot|x)$.

Assumption A3

For some positive constant $R_{\max}$ and all $a \in \mathsf{A}$, $\|r^a\|_{\mathsf{X}} \leq R_{\max}$.

Assumption A4

For some positive constants $L_\psi \leq 1, L_{\max}, L_\pi$ and all $a \in \mathsf{A}$, $\xi \in \Xi$,

$$\mathrm{Lip}_{\rho_{\mathsf{X}}}(r^a(\cdot)) \leq L_{\max}, \quad \mathrm{Lip}_\rho(\psi(\cdot, \cdot, \xi)) \leq L_\psi, \quad \mathrm{Lip}_\rho((V^\pi \circ \psi)(\cdot, \cdot, \xi)) \leq L_\pi \,.$$

Tabular MDP's case

Checking A4

Let $|\mathsf{X}| < \infty$ and $|\mathsf{A}| < \infty$. Consider

$$\rho_{\mathsf{X}}(x, x') = \mathbb{1}_{\{x \neq x'\}}$$
$$\rho((x, a), (x', a')) = \mathbb{1}_{\{(x,a) \neq (x', a')\}} .$$

Then the assumption A4 holds with constants $L_\psi = 1$, $L_{\max} = R_{\max}$, and $L_\pi = R_{\max}/(1 - \gamma)$.

Measuring the distance between $\widehat{V}_k^{\mathrm{up}}$ and V^*

Theorem 1*a*.

Let $|\mathsf{X}|, |\mathsf{A}| < \infty$ and assume A2, A3. Then for any $k \in \mathbb{N}$ and $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \left\|\widehat{V}_0^{\mathrm{up}} - V^*\right\|_{\mathsf{X}} + \|V^{\pi} - V^*\|_{\mathsf{X}} + \sqrt{\frac{\log(|\mathsf{X}||\mathsf{A}|/\delta)}{M_1}}\,.$$

Measuring the distance between $\widehat{V}_k^{\mathrm{up}}$ and V^*

Theorem 1*a*.

Let $|\mathsf{X}|, |\mathsf{A}| < \infty$ and assume A2, A3. Then for any $k \in \mathbb{N}$ and $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \left\|\widehat{V}_0^{\mathrm{up}} - V^*\right\|_{\mathsf{X}} + \|V^{\pi} - V^*\|_{\mathsf{X}} + \sqrt{\frac{\log(|\mathsf{X}||\mathsf{A}|/\delta)}{M_1}}\,.$$

Theorem 1*a* can be generalized under A1 – A4. In this case there is an additional term in the r.h.s, depending on the *covering radius* of the set X_N w.r.t. X, that is,

$$\rho(\mathsf{X}_N, \mathsf{X}) = \max_{x \in \mathsf{X}} \min_{k \in [N]} |x - \mathsf{X}_k|\,.$$

Notations

In case of more general X and A we introduce some additional notations:

- Define $\mathcal{N}(X \times A, \rho, \varepsilon)$ the covering number of the set $X \times A$ w.r.t. metric ρ, that is, the smallest cardinality of an ε-net of $X \times A$ w.r.t. ρ;
- Define D the diameter of $X \times A$, that is,

$$D = \operatorname{diam}(X \times A) = \max_{(x,a),(x',a') \in X \times A} \rho((x,a),(x',a'))\,;$$

- Then $\log \mathcal{N}(X \times A, \rho, \varepsilon)$ is the metric entropy of $X \times A$ and

$$I_{\mathcal{D}} = \int_0^D \sqrt{\log \mathcal{N}\left(X \times A, \rho, u\right)}\, du$$

is the Dudley's integral.

Some examples

- Assume $|\mathsf{X}| < \infty$, $|\mathsf{A}| < \infty$.
- We bypass the approximation step. Consider $\rho((x,a),(x',a')) = \mathbb{1}_{\{(x,a)\neq(x',a')\}}$. Then

$$\mathsf{D} = 1$$

$$I_{\mathcal{D}} = \sqrt{\log(|\mathsf{X}||\mathsf{A}|)}\,.$$

Some examples

- Assume $\mathsf{X} \subseteq [0,1]^{d_\mathsf{X}}$, $|\mathsf{A}| < \infty$
- Let $\rho_\mathsf{X}(x, x') = \|x - x'\|$, $\rho\big((x,a),(x',a')\big) = \|x - x'\| + \mathbb{1}_{\{a \neq a'\}}$. Then

$$\mathsf{D} \leq \sqrt{d_\mathsf{X}} + 1,$$
$$I_\mathcal{D} \lesssim \sqrt{d_\mathsf{X} \log |\mathsf{A}|} + \sqrt{d_\mathsf{X} \log d_\mathsf{X}}.$$

Measuring the distance between $\widehat{V}_k^{\mathrm{up}}$ and V^*

Theorem 1*b*.

Assume A1 – A4 and suppose that $\mathrm{Lip}_{\rho_{\mathsf{X}}}(\widehat{V}_0^{\mathrm{up}}) \le L_0$ with some constant $L_0 > 0$. Then for any $k \in \mathbb{N}$ and $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \|V^{\pi} - V^*\|_{\mathsf{X}} + \frac{I_{\mathcal{D}} + \mathsf{D}\sqrt{\log(1/\delta)}}{\sqrt{M_1}} + \rho(\mathsf{X}_{\mathsf{N}}, \mathsf{X})\,.$$

In case $\mathsf{X} = [0,1]^{\mathsf{d}}$ and $\mathsf{X}_{\mathsf{N}} = \{\mathsf{X}_1, \ldots, \mathsf{X}_{\mathsf{N}}\}$ being a set of N points, uniformly distributed over X, the following bound is available: for any $\delta \in (0,1)$,

$$\rho(\mathsf{X}_{\mathsf{N}}, \mathsf{X}) \lesssim \sqrt{\mathsf{d}_{\mathsf{X}}} \left(\frac{\log(1/\delta)\log \mathsf{N}}{\mathsf{N}}\right)^{1/\mathsf{d}_{\mathsf{X}}}$$

with probability at least $1-\delta$.

Variance of the estimator

We need to assume in addition that $\mathsf{X} \times \mathsf{A}$ is a parametric class with the metric entropy satisfying the following assumption:

Assumption A5

There exist a constant $C_{\mathsf{X},\mathsf{A}} > 1$ such that for any $\varepsilon \in (0, \mathsf{D})$,

$$\log \mathcal{N}(\mathsf{X} \times \mathsf{A}, \rho, \varepsilon) \leq \mathsf{C}_{\mathsf{X},\mathsf{A}} \log(1 + 1/\varepsilon).$$

Variance of the estimator

Theorem 2.

Let us introduce σ_k as an upper bound for $\mathbb{E}^{1/2}[\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}}^2]$, that is,

$$\sigma_k \doteq \gamma^k \left\| \widehat{V}_0^{\mathrm{up}} - V^* \right\|_{\mathsf{X}} + \|V^\pi - V^*\|_{\mathsf{X}} + \frac{I_{\mathcal{D}} + \mathsf{D}}{\sqrt{M_1}} + \rho(\mathsf{X}_{\mathsf{N}}, \mathsf{X}) \,.$$

Let A1 – A5 hold and assume additionally $\mathrm{Lip}_{\rho_{\mathsf{X}}}(\widehat{V}_0^{\mathrm{up}}) \le L_0$ for some $L_0 > 0$. Then

$$\max_{x \in \mathsf{X}} \mathrm{Var}\left[\widehat{V}_k^{\mathrm{up}}(x)\right] \le \mathsf{C}\, \sigma_k^2 \log(\mathrm{e} \vee \sigma_k^{-1}) M_2^{-1} \,,$$

where the constant C depends on $C_{\mathsf{X},\mathsf{A}}, \gamma, L_{\max}, L_\psi, L_\pi, L_0$ and $R_{\max}$.

Confidence intervals for $V^\star$

Theorem 3.

For any $x \in \mathsf{X}_\mathsf{N}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$V^\pi(x) \le V^\star(x) \le \widehat{V}_k^{\mathrm{up}}(x) + \sigma_k \sqrt{\mathsf{C} \log(\mathrm{e} \vee \sigma_k^{-1}) \delta^{-1} M_2^{-1}} \,.$$

Moreover, for k and M_1 large enough,

$$\sigma_k \lesssim \|V^\pi - V^\star\|_\mathsf{X} \,.$$

Numerical results: Chain problem

- Chain is a finite MDP, where agent can move only (`right` and `left`);
- Chain has two terminal states at the ends. For transition to the terminal states agent receives 10 points and episode ends, otherwise the reward is equal to $+1$;
- $p \in (0, 1)$ - noise in the system, i.e. the agent's action A_k at state x is drawn from distribution

$$A_k \sim \begin{cases} \pi(\cdot|x)\,, \text{ with probability } p; \\ \text{uniform distribution over A, with probability } 1 - p\,. \end{cases}$$

Numerical results: Chain problem

- Consider the sequence of policies $(\pi_k)_{k=0}^{15}$, obtained via the value iteration procedure. We evaluate policies π_k by computing $\Delta^{\mathrm{up}}_{\pi,k}(x)$ for some time steps k.

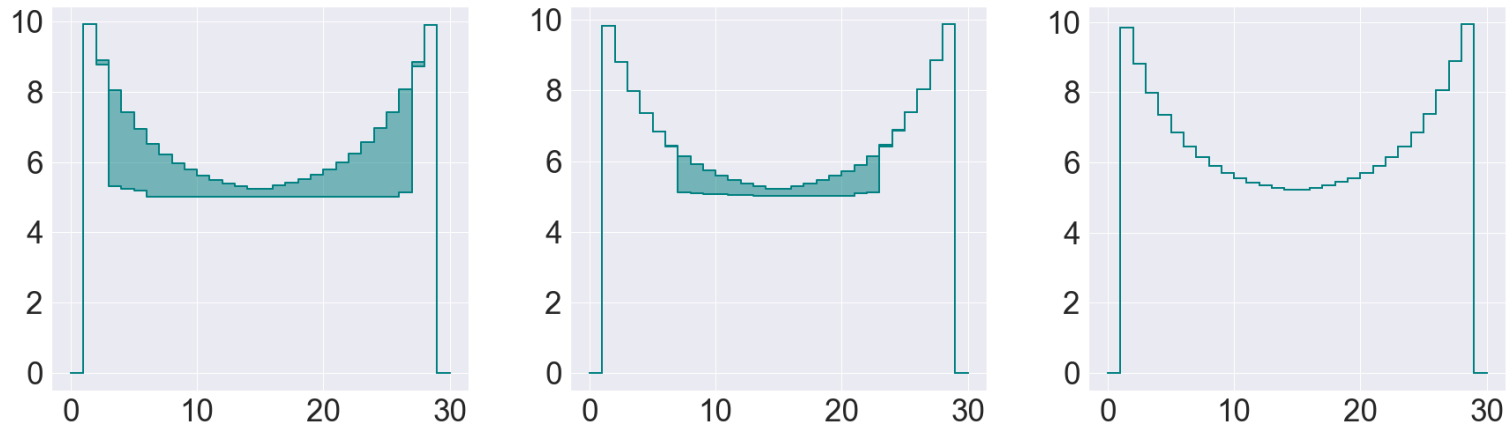


Figure: Chain environment: evaluating π_0, π_5 and π_{15}

Numerical results: Garnet problem

- A Garnet model is specified by a triplet (N_S, N_A, N_B), where N_S and N_A are the number of states and actions, respectively.
- The parameter N_B denotes the branching factor, that is, the number of states reachable from any state-action pair.
- We choose $N_S = 20$, $N_A = 10$, $N_B = 2$.

Numerical results: Garnet problem

- Consider the sequence of policies $(\pi_k)_{k=0}^{15}$, obtained via the value iteration procedure. We evaluate policies π_k by computing $\Delta^{\mathrm{up}}_{\pi,k}(x)$ for some time steps k.

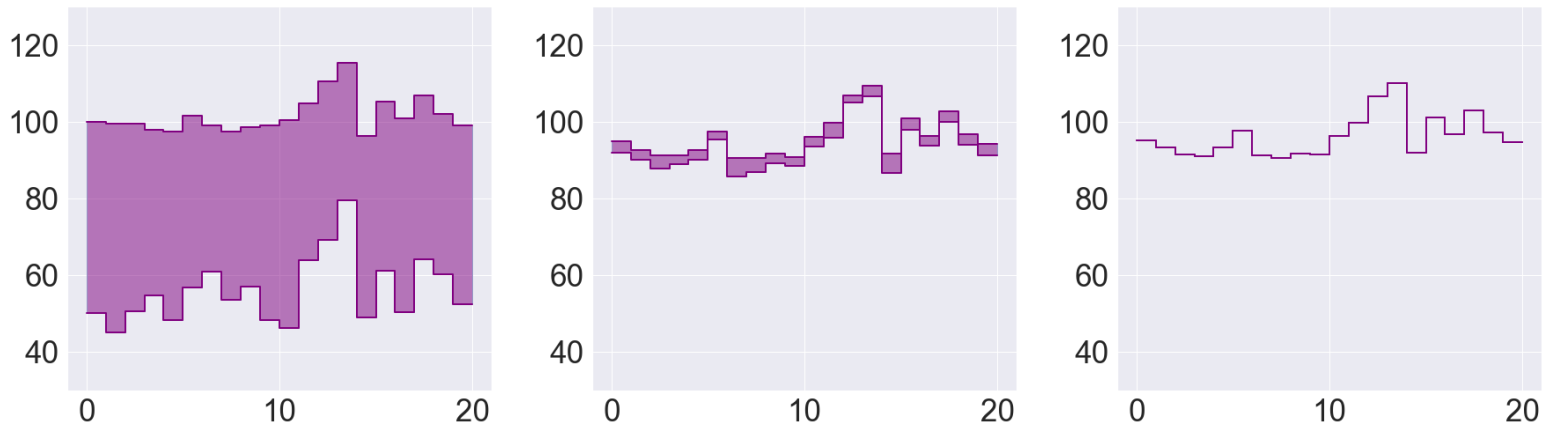


Figure: Garnet environment: evaluating π_0, π_5 and π_{15}

Numerical results: Cartpole

- CartPole is an example of the environment with a finite action space and infinitely large state space. In this environment agent can push cart with pole on it to the left or right direction and the target is to hold the pole up as long as possible.
- Reward equals to 1 is gain every time until failing or the end of episode.
- We apply normally distributed random variable (additional randomness) to the angle.

Numerical results: Cartpole

- LD(Linear Deterministic) policy can be expressed as $\mathbb{I}\{3 \cdot \theta + \dot{\theta} > 0\}$, where θ is an angle between pole and normal to cart.

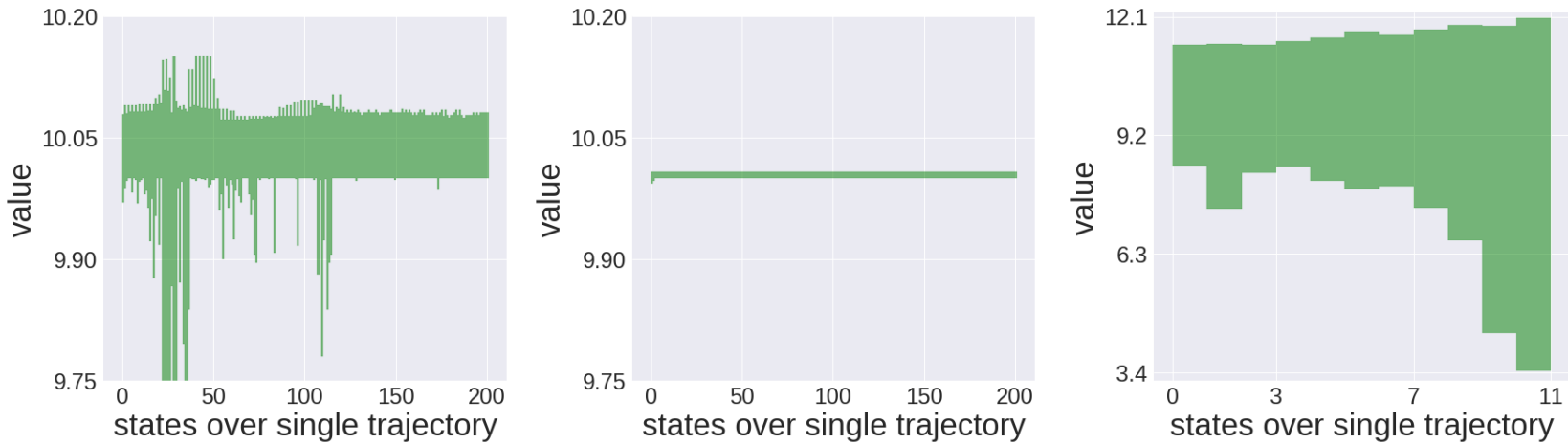


Figure: CartPole environment: evaluating A2C, deterministic policy and random policy.

UVIP: reminder

Require: $V^{\pi}, \widehat{V}_0^{\mathrm{up}}, \gamma, \varepsilon, M_1, M_2$

Ensure: V^{up}

for $x \in \mathsf{X}, \mathsf{a} \in \mathsf{A}$ **do**

$\overline{V}(x,a) = M_1^{-1} \sum_{i=1}^{M_1} V^{\pi}(Y_i^{x,a}), \;\; Y_i^{x,a} \sim \mathsf{P}^a(\cdot|x)$

for $y \in \mathsf{X}$ **do**

$\Phi_{\pi}^{x,a}(y) = V^{\pi}(y) - \overline{V}(x,a)$

end for

end for

$k = 1$

while $\| \widehat{V}_k^{\mathrm{up}} - \widehat{V}_{k-1}^{\mathrm{up}} \|_{\mathsf{X}} > \varepsilon$ **do**

for $x \in \mathsf{X}$ **do**

$\widehat{V}_{k+1}^{\mathrm{up}}(x) = M_2^{-1} \sum_{i=1}^{M_2} [\max_a \{ r^a(x) + \gamma (\widehat{V}_k^{\mathrm{up}}(Y_i^{x,a}) - \Phi_{\pi}^{x,a}(Y_i^{x,a})) \}],$
$Y_i^{x,a} \sim \mathsf{P}^a(\cdot|x)$

end for

$k = k + 1$

end while

$V^{\mathrm{up}} = \widehat{V}_k^{\mathrm{up}}$

Further research directions: online method

- Instead of estimating $\mathrm{P}^a V(x)$ with Monte-Carlo method, consider estimating P^a online;
- Use $\widehat{P}^a V(x)$ instead of $\overline{V}(x, a)$ where

$$\widehat{P}^a(x'|x) = \frac{N(x', a, x)}{N(a, x)}\,, x, x' \in \mathsf{X}, \mathrm{a} \in \mathrm{A}\,;$$

- $N(a, x)$ - number of visits of pair (a, x),
- $N(x', a, x)$ - number of visits to x' by taking action a at state x.

UVIP: reminder

Algorithm 2: UVIP: online modification

Require: $V^{\pi}, \widehat{V}_0^{\mathrm{up}}, \gamma, \varepsilon, M_1, M_2$

Ensure: V^{up}

$k = 1$

while $\| \widehat{V}_k^{\mathrm{up}} - \widehat{V}_{k-1}^{\mathrm{up}} \|_{\mathrm{X}} > \varepsilon$ **do**

for $x \in \mathrm{X}$ **do**

$$\widehat{V}_{k+1}^{\mathrm{up}}(x) = M_2^{-1} \sum_{i=1}^{M_2} \Big[\max_a \Big\{ r^a(x) + \gamma \big(\widehat{V}_k^{\mathrm{up}}(Y_i^{x,a}) - V^{\pi}(Y_i^{x,a}) + \widehat{P}^a V^{\pi}(x) \big) \Big\} \Big],$$

$$Y_i^{x,a} \sim \mathrm{P}^a(\cdot|x)$$

end for

$k = k + 1$

end while

$V^{\mathrm{up}} = \widehat{V}_k^{\mathrm{up}}$

Online method: Garnet

- Consider the sequence of policies $(\pi_k)_{k=0}^{15}$, obtained via the value iteration procedure. We evaluate policies π_k by computing $\Delta^{\mathrm{up}}_{\pi,k}(x)$ for some time steps k.

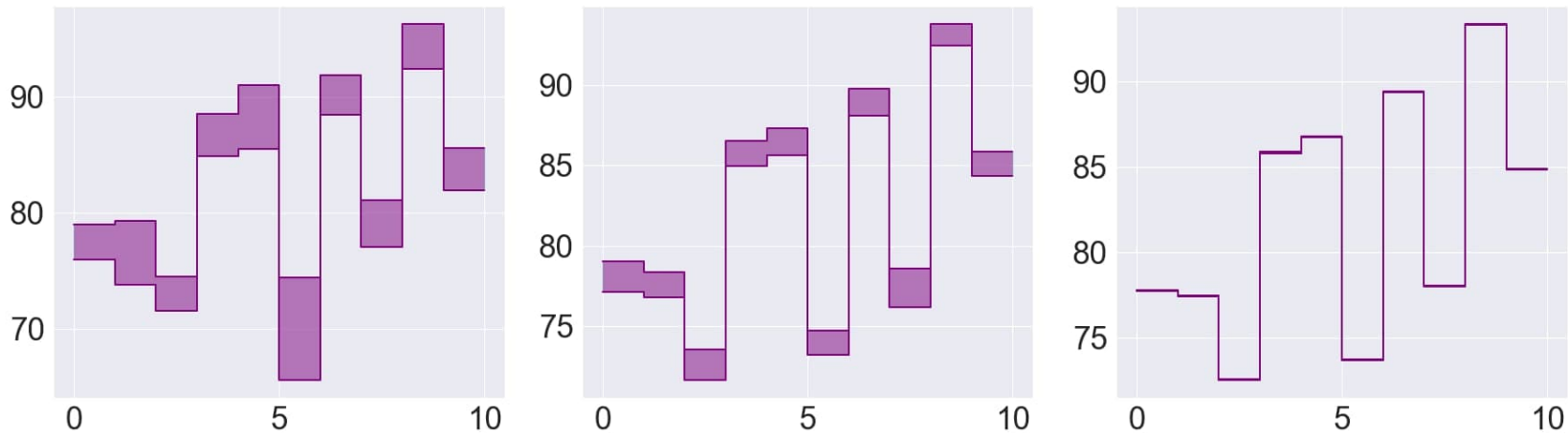


Figure: Garnet environment: evaluating π_0, π_5 and π_{15} with an online method

Extensions: real-time UVIP

Consider now the extensions to finite-horizon case with episodes of length H.

Algorithm 3: Real-Time UVIP

Initialize: $V_t^{0,\mathrm{up}}$.
for $k = 1, 2, \ldots$ **do**
  Initialize X_1^k
  for $t = 0, \ldots, H-1$ **do**
    $A_t^k \in \arg\max_{a \in \mathsf{A}} \{ r^a(X_t^k) + \mathsf{P}^a \, V_{t+1}^{\mathrm{up},k-1}(X_t^k) \}$
    Act with A_t^k and observe X_{t+1}^k
    $V_t^{\mathrm{up},k}(X_t^k) =$
    $\mathbb{E}\big[\max_{a \in \mathsf{A}} \{ r^a(X_t^k) + V_{t+1}^{\mathrm{up},k-1}(Y_{t+1}^{a,k}) - V^\pi(Y_{t+1}^{a,k}) + (\mathsf{P}^a \, V^\pi)(X_t^k) \}\big]$,
    $Y_{t+1}^{a,k} \sim \mathsf{P}^a(\cdot | X_t^k)$.
  end for
end for

In the algorithm above $\mathsf{P}^a \, V^\pi(x)$ can be replaced by $\widehat{P}^a V^\pi(x)$ and the outer expectation can be replaced by its Monte-Carlo estimate.

Paper available at https://arxiv.org/abs/2105.02135

Thank you!

References I

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. 70:263–272, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/azar17a.html.

Denis Belomestny and John Schoenmakers. *Advanced Simulation-Based Methods for Optimal Stopping and Control: With Applications in Finance*. Springer, 2018.

Dimitri P Bertsekas and Steven E Shreve. Stochastic optimal control, volume 139 of mathematics in science and engineering, 1978.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? 31, 2018. URL https://proceedings.neurips.cc/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

L. Rogers. Pathwise stochastic optimal control. *SIAM J. Control and Optimization*, 46:1116–1132, 01 2007. doi: 10.1137/050642885.

Ibrahim El Shar and Daniel Jiang. Lookahead-bounded q-learning. 119:8665–8675, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/shar20a.html.

Richard Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3: 9–44, 08 1988. doi: 10.1007/BF00115009.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997. ISSN 2334-3303. doi: 10.1109/9.580874.