Loop Estimator for Discounted Values in Markov Reward Processes

Falcon Z. Dai   Matthew R. Walter
{dai, mwalter}@ttic.edu

Toyota Technological Institute at Chicago

AAAI 2021

Preliminaries: MRP

Parameters of the Markov reward process

Preliminaries: MRP

Parameters of the Markov reward process

- state space $\mathcal{S} := \{1, \cdots, S\}$.

Preliminaries: MRP

Parameters of the Markov reward process

- state space $\mathcal{S} := \{1, \cdots, S\}$.
- transition probability matrix $\mathbf{P} : \mathcal{S} \times \mathcal{S} \to [0, 1]$.

Preliminaries: MRP

Parameters of the Markov reward process

- state space $\mathcal{S} := \{1, \cdots, S\}$.
- transition probability matrix $\mathbf{P} : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$.
- reward function $r : \mathcal{S} \rightarrow \mathcal{P}([0, r_{\max}])$ and mean rewards as $\bar{\mathbf{r}} : s \mapsto \mathbb{E}[r(s)]$.

Preliminaries: MRP

Parameters of the Markov reward process

- state space $\mathcal{S} := \{1, \cdots, S\}$.
- transition probability matrix $\mathbf{P} : \mathcal{S} \times \mathcal{S} \to [0, 1]$.
- reward function $r : \mathcal{S} \to \mathcal{P}([0, r_{\mathrm{max}}])$ and mean rewards as $\bar{\mathbf{r}} : s \mapsto \mathbb{E}[r(s)]$.

$(X_t, R_t)_{t \geq 0}$ is an MRP.

Preliminaries: MRP

Parameters of the Markov reward process

- state space $\mathcal{S} := \{1, \cdots, S\}$.
- transition probability matrix $\mathbf{P} : \mathcal{S} \times \mathcal{S} \to [0, 1]$.
- reward function $r : \mathcal{S} \to \mathcal{P}([0, r_{\max}])$ and mean rewards as $\bar{\mathbf{r}} : s \mapsto \mathbb{E}[r(s)]$.

$(X_t, R_t)_{t\geq 0}$ is an MRP.
Note that $(X_t)_{t\geq 0}$ is a Markov chain.

Preliminaries: stopping times

As conventions, we denote $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|X_0 = s]$ and $\mathbb{P}_s[\cdot] := \mathbb{P}[\cdot|X_0 = s]$.

Preliminaries: stopping times

As conventions, we denote $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|X_0 = s]$ and $\mathbb{P}_s[\cdot] := \mathbb{P}[\cdot|X_0 = s]$.

- First return time $H_s^+ := \inf\{t > 0 : X_t = s\}$.

Preliminaries: stopping times

As conventions, we denote $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|X_0 = s]$ and $\mathbb{P}_s[\cdot] := \mathbb{P}[\cdot|X_0 = s]$.

- First return time $H_s^+ := \inf\{t > 0 : X_t = s\}$.
- Expected recurrence time $\rho_s := \mathbb{E}_s\left[H_s^+\right]$.

Preliminaries: stopping times

As conventions, we denote $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|X_0 = s]$ and $\mathbb{P}_s[\cdot] := \mathbb{P}[\cdot|X_0 = s]$.

- First return time $H_s^+ := \inf\{t > 0 : X_t = s\}$.
- Expected recurrence time $\rho_s := \mathbb{E}_s\left[H_s^+\right]$.
- Maximal expected hitting time $\tau_s := \max_{s' \in \mathcal{S}} \mathbb{E}_{s'}[H_s^+]$.

Preliminaries: stopping times

As conventions, we denote $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|X_0 = s]$ and $\mathbb{P}_s[\cdot] := \mathbb{P}[\cdot|X_0 = s]$.

- First return time $H_s^+ := \inf\{t > 0 : X_t = s\}$.
- Expected recurrence time $\rho_s := \mathbb{E}_s\left[H_s^+\right]$.
- Maximal expected hitting time $\tau_s := \max_{s' \in \mathcal{S}} \mathbb{E}_{s'}[H_s^+]$.
- The waiting time for the n-th visit be $W_n(s) := \inf\left\{w : n \leq \sum_{t=0}^{w} \mathbb{1}[X_t = s]\right\}$.

Preliminaries: stopping times

As conventions, we denote $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|X_0 = s]$ and $\mathbb{P}_s[\cdot] := \mathbb{P}[\cdot|X_0 = s]$.

- First return time $H_s^+ := \inf\{t > 0 : X_t = s\}$.
- Expected recurrence time $\rho_s := \mathbb{E}_s\left[H_s^+\right]$.
- Maximal expected hitting time $\tau_s := \max_{s' \in \mathcal{S}} \mathbb{E}_{s'}[H_s^+]$.
- The waiting time for the n-th visit be $W_n(s) := \inf\left\{w : n \leq \sum_{t=0}^{w} \mathbb{1}[X_t = s]\right\}$.
- Interarrival times $I_n(s) := W_{n+1}(s) - W_n(s)$.

Problem formulation

Problem formulation

- Discounted value $v(s) := \mathbb{E}_s\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]$.

Problem formulation

- Discounted value $v(s) := \mathbb{E}_s\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]$.
- $v(s)$ satisfies the Bellman equation
  $v(s) = \bar{r}_s + \gamma \sum_{s' \in \mathcal{S}} P_{ss'} v(s')$.

Problem formulation

- Discounted value $v(s) := \mathbb{E}_s \left[\sum_{t=0}^{\infty} \gamma^t R_t \right]$.
- $v(s)$ satisfies the Bellman equation $v(s) = \bar{r}_s + \gamma \sum_{s' \in \mathcal{S}} P_{ss'} v(s')$.
- However, in RL settings, we do not know the MRP parameters and wish to estimate $v(s)$ from a single sample path, i.e., $(X_t, R_t)_{0 \le t \le T}$.

Assumption: reachability

Assumption: reachability

We assume state s is reachable from all states, i.e., $\tau_s < \infty$.

Assumption: reachability

We assume state s is reachable from all states, i.e., $\tau_s < \infty$. Otherwise, we cannot hope for a PAC-style error bound under arbitrarily high probability.

Observation: regenerative structure

- The sub-MRPs starting at different visits to state s are the same as stochastic processes.

Observation: regenerative structure

- The sub-MRPs starting at different visits to state s are the same as stochastic processes.
- Loop γ-discounted rewards $G_n(s) := \sum_{u=0}^{I_n(s)-1} \gamma^u R_{W_n(s)+u}$.

Observation: regenerative structure

- The sub-MRPs starting at different visits to state s are the same as stochastic processes.
- Loop γ-discounted rewards $G_n(s) := \sum_{u=0}^{l_n(s)-1} \gamma^u R_{W_n(s)+u}$.
- Loop γ-discount $\Gamma_n(s) := \gamma^{l_n(s)}$.

Observation: regenerative structure

- The sub-MRPs starting at different visits to state s are the same as stochastic processes.
- Loop γ-discounted rewards $G_n(s) := \sum_{u=0}^{I_n(s)-1} \gamma^u R_{W_n(s)+u}$.
- Loop γ-discount $\Gamma_n(s) := \gamma^{I_n(s)}$.
- $(I_n(s), G_n(s))$ are IID.

Observation: regenerative structure

- The sub-MRPs starting at different visits to state s are the same as stochastic processes.
- Loop γ-discounted rewards $G_n(s) := \sum_{u=0}^{l_n(s)-1} \gamma^u R_{W_n(s)+u}$.
- Loop γ-discount $\Gamma_n(s) := \gamma^{l_n(s)}$.
- $(l_n(s), G_n(s))$ are IID.
- Denote the expected loop γ-discount as $\alpha(s) := \mathbb{E}_s[\Gamma_1(s)]$ and the expected loop γ-discounted rewards as $\beta(s) := \mathbb{E}_s[G_1(s)]$.

Results: loop Bellman equation

Theorem (Loop Bellman equations)

We can relate the state value $v(s)$ to itself

$$v(s) = \beta(s) + \alpha(s)\, v(s). \tag{1}$$

Results: loop Bellman equation

Theorem (Loop Bellman equations)

We can relate the state value $v(s)$ to itself

$$v(s) = \beta(s) + \alpha(s)\, v(s). \tag{1}$$

Define the n-th *loop estimator* for state value $v(s)$

$$\hat{v}_n(s) := \hat{\beta}_n(s)/(1 - \hat{\alpha}_n(s)), \tag{2}$$

where

$$\hat{\alpha}_n(s) := \frac{1}{n}\sum_{i=1}^{n} \gamma^{l_i(s)}$$

and

$$\hat{\beta}_n(s) := \frac{1}{n}\sum_{i=1}^{n} G_i(s)$$

.

Results: sample complexity

Overall approach:

Results: sample complexity

Overall approach:

- Convergence for $\hat{v}_n(s)$ over visits to state s.

$$|\hat{v}_n(s) - v(s)| = O\left(\frac{r_{\max}}{(1-\gamma)^2}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\right).$$

Results: sample complexity

Overall approach:

- Convergence for $\hat{v}_n(s)$ over visits to state s.

$$|\hat{v}_n(s) - v(s)| = O\left(\frac{r_{\max}}{(1-\gamma)^2}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\right).$$

- Lower-bound the visits to s by step T. There are at least $\widetilde{\Omega}(T/\tau_s)$-many visits.

Results: sample complexity

Overall approach:

- Convergence for $\hat{v}_n(s)$ over visits to state s.

$$|\hat{v}_n(s) - v(s)| = O\left(\frac{r_{\max}}{(1-\gamma)^2}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\right).$$

- Lower-bound the visits to s by step T. There are at least $\widetilde{\Omega}(T/\tau_s)$-many visits.
- Convergence over steps.

$$\left|\hat{v}_T(s) - v(s)\right| = \widetilde{O}\left(\frac{r_{\max}}{(1-\gamma)^2}\sqrt{\frac{\tau_s}{T}\log\frac{1}{\delta}}\right).$$

Results: sample complexity

Overall approach:

- Convergence for $\hat{v}_n(s)$ over visits to state s.

$$|\hat{v}_n(s) - v(s)| = O\left(\frac{r_{\max}}{(1-\gamma)^2}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\right).$$

- Lower-bound the visits to s by step T. There are at least $\widetilde{\Omega}(T/\tau_s)$-many visits.
- Convergence over steps.

$$\left|\hat{v}_T(s) - v(s)\right| = \widetilde{O}\left(\frac{r_{\max}}{(1-\gamma)^2}\sqrt{\frac{\tau_s}{T}\log\frac{1}{\delta}}\right).$$

- Convergence of $\hat{v}_T$ under ℓ_∞-norm.

$$\|\hat{\mathbf{v}}_T - \mathbf{v}\|_\infty = \widetilde{O}\left(\frac{r_{\max}}{(1-\gamma)^2}\sqrt{\frac{\max_s \tau_s}{T}\log\frac{S}{\delta}}\right).$$

Proof ideas: lower-bounding the visits

The key steps are

Proof ideas: lower-bounding the visits

The key steps are

Lemma (Exponential concentration of first return times (Lee et al, 2013; Aldous and Fill, 1999))

Given a Markov chain $(X_t)_{t\geq 0}$ defined on a finite state space $\mathcal{S}$, for any state $s \in \mathcal{S}$ and any $t > 0$, we have

$$\mathbb{P}\left[H_s^+ \geq t\right] \leq e \cdot e^{-t/e\tau_s}.$$

Proof ideas: lower-bounding the visits

The key steps are

Lemma (Exponential concentration of first return times (Lee et al, 2013; Aldous and Fill, 1999))

Given a Markov chain $(X_t)_{t \geq 0}$ defined on a finite state space $\mathcal{S}$, for any state $s \in \mathcal{S}$ and any $t > 0$, we have

$$\mathbb{P}\left[H_s^+ \geq t\right] \leq e \cdot e^{-t/e\tau_s}.$$

and then we invert to find a lower bound on visits with the help of Lambert W function.

Open problems

- How to extend this idea to MRPs with large state spaces? Null-recurrence?

Open problems

- How to extend this idea to MRPs with large state spaces? Null-recurrence?
- Is the upper bound of TD obtained under a generative model tight in the Markov setting?

More questions?

- Feel free to contact me during or after the conference: `dai@ttic.edu`
- Join the poster sessions for live Q & A.
- Scan for related resources (paper, code, slides).