Off-Policy Policy Evaluation

(with temporal-difference learning and value-function approximation)

Balázs Csanád Csáji

MDP-RL Course, SZTAKI

25 April, 2006 and 9 May, 2006

Main literature

1. Preecup, D.; Sutton, R. S.; Singh, S.: *Eligibility Traces for Off-Policy Policy Evaluation*; International Conference on Machine Learning (ICML 2000) http://www.cs.ualberta.ca/∼sutton/papers/PSS-00.pdf

2. Preecup, D.; Sutton, R. S.; Dasgupta, S.: *Off-Policy Temporal-Difference Learning with Function Approximation*; International Conference on Machine Learning (ICML 2001) http://www.cs.ualberta.ca/∼sutton/papers/PSD-01.pdf

What is off-policy learning good for?

- Off-policy learning is learning about one way of behaving while actually behaving in another way.
- Classical example: Q-learning – it learns about the optimal policy while taking actions in a more exploratory fashion.
- Note that the original TD(λ) (and its generalization) is on-policy!
- „Off-policy learning is of interest because only one way of selecting actions can be used at any time, but we would like to learn about many different ways of behaving from the single resultant stream of experience.”
- Example: the options framework for temporal abstraction (macro actions)

Notations - Markov Decision Processes

Consider a (finite, discrete time, stationary, fully observable, episodic) Markov Decision Process (MDP), $\mathcal{M} = \langle S, A, \mathcal{A}, p, r, \gamma \rangle$, where

- $S = \{s_0, \ldots, s_N\}$ a set of discrete states. We only consider episodic MDPs, thus, there is an initial state s_0 and a terminal state s_G. (SSP)
- A is a finite set of control actions
- $\mathcal{A} : S \to \mathcal{P}(A)$ is an avaibility function, $\mathcal{P}$ denotes power set
- $p : S \times A \to \Delta(S)$ is a transition function, where $\Delta(S)$ denotes the set of probability distributions over S
- $r : S \times A \times S \to \mathbb{R}$ is an immediate reward function
- $\gamma \in [0, 1]$ is the discount factor

Reminder - action-value functions

A (randomized, stationary, Markov) control policy is $\pi : S \to \Delta(A)$.

The action-value function of a policy is $Q^\pi : S \times A \to \mathbb{R}$, where

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r(S_t, A_t, S_{t+1}) \;\middle|\; S_0 = s, A_0 = a \right],$$

where $A_t \sim \pi(S_t)$, $A_{t+1} \sim p(S_t, A_t)$ and T is random variable denoting the time when the terminal state is reached ($S_T = s_G$).

A policy is called *proper* if it reaches the terminal state with probability one.

The problem to be considered is to estimate Q^π for an arbitrary proper *target policy* π, given all the data is generated by a different proper *behavior policy* b, where b is *soft*, meaning that $\forall s \in S, a \in \mathcal{A}(s) : b(s,a) > 0$.

Importance sampling

The expected value of a random variable X with distribution d has to be estimated from samples drawn from another distribution d'. Note that

$$\mathbb{E}_d[X] = \int x\, d(x)dx = \int x \frac{d(x)}{d'(x)} d'(x)dx = \mathbb{E}_{d'}\left[X \frac{d(X)}{d'(X)}\right].$$

Therefore, estimations of the expected value can be given by

$$\mathbb{E}_d[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i \frac{d(x_i)}{d'(x_i)}, \qquad \text{or} \qquad \mathbb{E}_d[X] \approx \frac{\sum_{i=1}^{n} x_i \frac{d(x_i)}{d'(x_i)}}{\sum_{i=1}^{n} \frac{d(x_i)}{d'(x_i)}},$$

where the x_i are samples selected according to d'. The former is called *importance sampling* and is a consistent and unbiased estimator. The latter is called *weighted importance sampling* and is still consistent but biased.

Importance sampling estimation of Q^π

We can use importance sampling in (first-visit) Monte-Carlo policy evaluation

$$Q^\pi(s,a) \approx \frac{1}{M}\sum_{m=1}^{M} R_m\, w_m,$$

where M is the number of episodes containing state-action pair (s,a), and

$$R_m = \sum_{t=t_m}^{T_m-1} \gamma^{t-t_m} r(s_t^m, a_t^m, s_{t+1}^m),$$

where t_m is the first time when $(s_t^m, a_t^m) = (s,a)$. The sampling weights are

$$w_m = \prod_{t=t_m}^{T_m-1} \frac{\pi(s_t^m, a_t^m)}{b(s_t^m, a_t^m)}.$$

Per-decision algorithms

Monte-Carlo methods above consider complete returns. An estimator that breaks down into rewards could be more efficient and more easily implemented in an incremental, step-by-step basis. For example,

$$Q^\pi(s,a) \approx \frac{1}{M}\sum_{m=1}^{M}\sum_{t=t_m}^{T_m-1}\gamma^{t-t_m} r_t^{(m)} \prod_{i=t_m+1}^{t}\frac{\pi_i^{(m)}}{b_i^{(m)}},$$

where $r_t^{(m)} = r(s_t^m, a_t^m, s_{t+1}^m)$, $b_i^{(m)} = b(s_i^m, a_i^m)$ and $\pi_i^{(m)} = \pi(s_i^m, a_i^m)$.

This estimator is called *per-decision importance sampling estimator* and it is a consistent and unbiased estimator of Q^π (Preecup, Sutton and Singh, 2000).

The construction of the weighted per-decision estimator is straightforward.

Eligibility-traces for per-decision estimate

Now, let us consider an on-line variant of the per-decision importance sampling algorithm that incorporates the ideas of temporal-difference learning, as well. (1) Updating the eligibility-traces for all state-action pairs

$$e_t(s,a) = \begin{cases} 1 & \text{iff } \; t = t_m(s,a) \\ e_{t-1}(s,a)\,\gamma\,\lambda\,\frac{\pi(s_t,a_t)}{b(s_t,a_t)} & \text{otherwise} \end{cases},$$

where $\lambda \in [0,1]$ is an eligibility trace decay factor. (2) The TD-error is

$$\delta_t = r_t + \gamma\,\frac{\pi(s_{t+1},a_{t+1})}{b(s_{t+1},a_{t+1})}\,Q_t(s_{t+1},a_{t+1}) - Q_t(s_t,a_t),$$

(3) The update rule for the action-value estimate is

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t(s,a)\,e_t(s,a)\,\delta_t \qquad \forall s,a$$

Convergence theorem

Theorem 1. *For any soft, stationary behavior policy b, and any $\lambda \in [0,1]$ that does not depend on the action a_t, the above algorithm with off-line updating converges w.p.1 to Q^π, under the usual step-size conditions on α_t.*

Proof. First, we consider the corrected truncated return of the off-line per-decision algorithm. After n steps the current estimate is used

$$R_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k r_{t+k} \prod_{i=t+1}^{t+k} \frac{\pi_i}{b_i} + \gamma^n Q(s_{t+n}, a_{t+n}) \prod_{i=t+1}^{t+n} \frac{\pi_i}{b_i} \ ,$$

where $r_t = r(s_t, a_t, s_{t+1})$, $b_i = b(s_i, a_i)$ and $\pi_i = \pi(s_i, a_i)$. As a first step, we show that $\| \mathbb{E}_b[R_t^{(n)} \mid s_t, a_t] - Q^\pi \|_\infty \leq \gamma^n \| Q - Q^\pi \|_\infty$.

Let $\Omega(s, a, j)$ denote the set of all possible trajectories of j state-action pairs starting with (s, a). For example, $\omega = \langle a_0, s_0, s_1, a_1, \ldots, s_{j-1}, a_{j-1} \rangle$.
Then the expected return of the corrected truncated return for (s, a) is

$$\mathbb{E}_b \left[R^{(n)} \mid s_0 = s, a_0 = a \right] =$$

$$= \sum_{k=0}^{n-1} \sum_{\omega \in \Omega(s,a,k+1)} \mathbb{P}(\, \omega \mid b, s_0 = s, a_0 = a \,)\, \gamma^k r_k \prod_{i=1}^{k} \frac{\pi_i}{b_i} \; +$$

$$+ \sum_{\omega \in \Omega(s,a,k+1)} \mathbb{P}(\, \omega \mid b, s_0 = s, a_0 = a \,)\, \gamma^n Q(s_n, a_n) \prod_{i=1}^{n} \frac{\pi_i}{b_i} = (*)$$

Now, we can apply the equality as follows (from the Markov property)

$$\mathbb{P}(\, \omega \mid b, s_0 = s, a_0 = a \,) = \prod_{i=1}^{k} \mathbb{P}(s_i \mid s_{i-1}, a_{i-1})\, b(s_i, a_i)$$

$$(*) = \sum_{k=0}^{n-1} \sum_{\omega \in \Omega(s,a,k+1)} \left(\prod_{i=1}^{k} \mathbb{P}(s_i \mid s_{i-1}, a_{i-1})\, b(s_i, a_i) \right) \gamma^k r_k \prod_{i=1}^{k} \frac{\pi_i}{b_i} \; +$$

$$+ \sum_{\omega \in \Omega(s,a,k+1)} \left(\prod_{i=1}^{n} \mathbb{P}(s_i \mid s_{i-1}, a_{i-1})\, b(s_i, a_i) \right) \gamma^n Q(s_n, a_n) \prod_{i=1}^{n} \frac{\pi_i}{b_i} \; =$$

$$= \sum_{k=0}^{n-1} \sum_{\omega \in \Omega(s,a,k+1)} \gamma^k r_k \prod_{i=1}^{k} \mathbb{P}(s_i \mid s_{i-1}, a_{i-1})\, \pi(s_i, a_i) +$$

$$+ \sum_{\omega \in \Omega(s,a,k+1)} \gamma^n Q(s_n, a_n) \prod_{i=1}^{n} \mathbb{P}(s_i \mid s_{i-1}, a_{i-1})\, \pi(s_i, a_i)\,.$$

On the other hand, by applying the Bellman equation for Q^π iteratively n times

$$Q^\pi(s,a) = \sum_{k=0}^{n-1} \sum_{\omega \in \Omega(s,a,k+1)} \gamma^k r_k \prod_{i=1}^{k} \mathbb{P}\big(s_i \mid s_{i-1}, a_{i-1}\big)\, \pi\big(s_i, a_i\big) +$$

$$+ \sum_{\omega \in \Omega(s,a,k+1)} \gamma^n Q^\pi(s_n, a_n) \prod_{i=1}^{n} \mathbb{P}\big(s_i \mid s_{i-1}, a_{i-1}\big)\, \pi\big(s_i, a_i\big)\,.$$

Therefore, we obtain

$$\max_{(s,a)} \Big| \mathbb{E}_b\big[R^{(n)} \mid s, a \big] - Q^\pi(s,a) \Big| \le \gamma^n \max_{(s,a)} \big| Q(s,a) - Q^\pi(s,a) \big|$$

From this, the convergence of the original per-decision algorithm follows from the results of Jaakola, Jordan and Singh (1994).

Now, we consider the case of eligibility-traces. Assume $\lambda = 1$

$$e_t(s,a) = \gamma^{t-t_m} \prod_{i=t_m+1}^{t} \frac{\pi_i}{b_i}$$

We have

$$\sum_{k=0}^{n-1} e_{t+k}(s,a)\delta_{t+k} =$$

$$= \sum_{k=0}^{n-1} \gamma^k \left(\prod_{i=t+1}^{t+k} \frac{\pi_i}{b_i} \right) \left(r_{t+k} + \gamma \frac{\pi_{t+k+1}}{b_{t+k+1}} Q(s_{t+k+1}, a_{t+k+1}) - Q(s_{t+k}, a_{t+k}) \right) =$$

$$= \sum_{k=0}^{n-1} \gamma^k r_{t+k} \left(\prod_{i=t+1}^{t+k} \frac{\pi_i}{b_i} \right) + \gamma^n Q(s_{t+n}, a_{t+n}) \left(\prod_{i=t+1}^{t+n} \frac{\pi_i}{b_i} \right) - Q(s_t, a_t).$$

$$\sum_{k=0}^{n-1} \gamma^k r_{t+k} \left(\prod_{i=t+1}^{t+k} \frac{\pi_i}{b_i} \right) + \gamma^n Q(s_{t+n}, a_{t+n}) \left(\prod_{i=t+1}^{t+n} \frac{\pi_i}{b_i} \right) - Q(s_t, a_t) =$$

$$= R_t^{(n)} - Q(s_t, a_t).$$

Since the algorithm is equivalent to applying a convex combination of n-step updates, and each update converges to correct Q-values, the algorithm will converge to correct Q-values, as well. *Q.E.D.*

Action-value function approximation

Now, we consider the case of linear function approximation

$$Q^\pi(s,a) \approx \tilde{Q}(s,a) = \sum_{i=1}^{m} \theta(i)\phi_{sa}(i) = \theta^T \phi_{sa}$$

In what follows we restrict ourselves to per-episode (off-line) updating. The increment of the conventional TD(λ) under per-episode updating is

$$\Delta\theta_t = \alpha_t(R_t^\lambda - \tilde{Q}_t)\nabla_\theta \tilde{Q}_t = \alpha_t(R_t^\lambda - \theta^T\phi_t)\phi_t,$$

where $\tilde{Q}_t = \tilde{Q}(s_t, a_t)$ and $\phi_t = \phi_{s_t a_t}$. And R_t^λ is

$$R_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} \left(\sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \theta^T \phi_{t+n} \right)$$

Per-decision importance sampling

The per-decision importance sampling with function approximation is

$$\Delta\theta_t = \alpha_t(\tilde{R}_t^\lambda - \theta^T\phi_t)\phi_t \prod_{i=1}^{t} \varrho_i,$$

where $\varrho_i = \pi(s_i, a_i)/b(s_i, a_i)$ and $\tilde{R}_t^\lambda$ is

$$\tilde{R}_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1}\tilde{R}_t^{(n)},$$

and the corrected n-step return is defined as follows

$$\tilde{R}_t^{(n)} = \sum_{i=0}^{n-1} \gamma^i\, r_{t+i} \prod_{j=1}^{i} \varrho_{t+j} + \gamma^n \theta^T \phi_{t+n} \prod_{j=1}^{n} \varrho_{t+j}$$

Convergence and error bounds

Theorem 2. *Let $\Delta\theta$ and $\Delta\tilde{\theta}$ be the sum of the parameter increments over an episode under on-policy TD(λ) and importance sampling TD(λ) respectively, assuming that the starting vector is θ in both cases. Then, for all s_0 and a_0*

$$\mathbb{E}_b[\Delta\tilde{\theta} \mid s_0, a_0] = \mathbb{E}_\pi \left[\Delta\theta \mid s_0, a_0\right].$$

Now, we investigate error bounds. Let $d \in \Delta(S \times A)$ be an arbitrary distribution of starting state-action pairs. Let P_π be the state-action pair transition probability matrix for policy π. $D_\pi = \sum_{t=0}^{\infty} P_\pi^t d$, $D_\pi(s,a)$ is the expected number of visits to state-action pair (s,a). Define the norm $\|\cdot\|_\pi$ over state-action vectors by $\|v\|_\pi^2 = \sum_{s,a} v(s,a)^2 D_\pi(s,a)$.

Convergence and error bounds

There are a number of (natural?) assumptions:

1. the state and action sets are finite;
2. all state-action pairs are visited under the behavior policy b;
3. both behavior and target polices, b and π, are proper;
4. the rewards are bounded;
5. the step-size sequence satisfies the stochastic approximation conditions: $\forall k : \alpha_k \geq 0\,, \quad \sum_{k=0}^{\infty} \alpha_k = \infty \quad$ and $\quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty\,;$
6. the variance of the product of correction factors can be bounded for any initial state: $\mathbb{E}_b \left[\varrho_1^2 \varrho_2^2 \ldots \varrho_T^2\right] < B$ for all $s_1 \in S$.

Convergence and error bounds

Theorem 3. *Under the assumptions 1–6 above, episodic importance sampled TD(λ) converges with probability one to some θ_∞ such that*

$$\left\|\tilde{Q}_{\theta_\infty} - Q^\pi\right\|_\pi \leq \min_\theta \left\|\tilde{Q}_\theta - Q^\pi\right\|_\pi \frac{1}{1-\beta},$$

where β is the contraction factor of the matrix

$$M = (1-\lambda)\sum_{k=0}^{\infty} \lambda^k (\gamma P_\pi)^{k+1}.$$

The proof follows immediately from the results of Bertsekas and Tsitsiklis, page 312 of Neurodynamic Programming (1996).

Restarting within an episode

Restarting within an episode makes sense, since:

1. Assumption 6 (bounded variance of the correction factor product) can be satisfied with bounded episode lengths;
2. The importance sampling correction product often decay rapidly, and when it becomes very small, a very little more is learned.

Let g_t be a non-negative random variable, which is allowed to depend only on events up to time t. Function $g : \Omega_t \to \mathbb{R}^+$ gives the expected value of g_t for any trajectory up through t. A generalized algorithm (with forward view) is

$$\Delta\theta_t = \alpha_t(\tilde{R}_t^\lambda - \theta^T\phi_t)\phi_t \sum_{k=0}^{t} g_k \prod_{i=k+1}^{t} \varrho_i,$$

Restarting within an episode

Theorem 4. *Let $\Delta\theta$ and $\Delta\tilde{\theta}$ be the sum of the parameter increments over an episode under the original importance sampled TD(λ) and the generalized version respectively, assuming that the starting vector is θ in both cases. Then, $\forall g$, there exists an alternate starting distribution d_g such that*

$$\mathbb{E}_b[\Delta\tilde{\theta} \mid s_0, a_0 \sim d] = \mathbb{E}_b[\Delta\theta \mid s_0, a_0 \sim d_g].$$

Therefore, restarting in a general way, at any point during an episode, is equivalent to a conventional at-the-beginning starting distribution. (The latter case the convergence is already proven.) However, the value converged to will depend on d_g and thus on b, rather then d and π alone.

Thank you for your attention!