Neural Networks 22 (2009) 1399–1410

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

# Adaptive importance sampling for value function approximation in off-policy reinforcement learning[☆,☆☆]

Hirotaka Hachiya[a,*], Takayuki Akiyama[a], Masashi Sugiayma[a], Jan Peters[b]

[a] Department of Computer Science, Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

[b] Max-Planck Institute for Biological Cybernetics, Dept. Schölkopf, Spemannstraße 38, 72076 Tübingen, Germany

ARTICLE INFO

*Article history:*
Received 22 May 2008
Received in revised form 23 December 2008
Accepted 16 January 2009

*Keywords:*
Off-policy reinforcement learning
Value function approximation
Policy iteration
Adaptive importance sampling
Importance-weighted cross-validation
Efficient sample reuse

ABSTRACT

Off-policy reinforcement learning is aimed at efficiently using data samples gathered from a policy that is different from the currently optimized policy. A common approach is to use importance sampling techniques for compensating for the bias of value function estimators caused by the difference between the data-sampling policy and the target policy. However, existing off-policy methods often do not take the variance of the value function estimators explicitly into account and therefore their performance tends to be unstable. To cope with this problem, we propose using an adaptive importance sampling technique which allows us to actively control the trade-off between bias and variance. We further provide a method for optimally determining the trade-off parameter based on a variant of cross-validation. We demonstrate the usefulness of the proposed approach through simulations.



## 1. Introduction

*Policy iteration* is a reinforcement learning setup where the optimal policy is obtained by iteratively performing policy evaluation and improvement steps (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998). When policies are updated, many popular policy iteration methods require the user to gather new samples following the updated policy, and the new samples are used for *value function approximation*. However, this approach is inefficient particularly when the sampling cost is high and it would be more cost-efficient if we could reuse the data collected in the past. A situation where the sampling policy (a policy used for gathering data samples) and the current policy are different is called *off-policy* reinforcement learning (Sutton & Barto, 1998).

In the off-policy setup, simply employing a standard policy iteration method such as *least-squares* policy iteration (Lagoudakis & Parr, 2003) does not lead to the optimal policy as the sampling policy can introduce bias into value function approximation. This distribution mismatch problem can be eased by the use of *importance sampling* techniques (Fishman, 1996), which cancel the bias asymptotically. However, the approximation error is not necessarily small when the bias is reduced by importance sampling; the variance of estimators also needs to be taken into account since the approximation error is the sum of squared bias and variance. Due to large variance, existing importance sampling techniques tend to be unstable (Precup, Sutton, & Singh, 2000; Sutton & Barto, 1998).

To overcome the instability problem, we propose using an *adaptive importance sampling* technique used in statistics (Shimodaira, 2000). The proposed adaptive method, which smoothly bridges the ordinary estimator and importance-weighted estimator, allows us to control the trade-off between bias and variance. Thus, given that the trade-off parameter is determined carefully, the optimal performance can be achieved in terms of both bias and variance. However, the optimal value of the trade-off parameter is heavily dependent on data samples and policies, and therefore using a pre-determined parameter value may not be always effective in practice.

For optimally choosing the value of the trade-off parameter, we propose using an automatic model selection method based on a variant of cross-validation (Sugiyama, Krauledat, & Müller, 2007). The method called *importance-weighted cross-validation* enables us to estimate the approximation error of value functions in an almost unbiased manner even under off-policy situations. Thus we can adaptively choose the trade-off parameter based on data samples at hand. We demonstrate the usefulness of the proposed approach through simulations.

☆ This research was supported in part by JSPS Global COE program "Computationism as a Foundation for the Sciences" and MEXT (20680007 and 18300057).

☆☆ Contributed article.

* Corresponding author. Tel.: +81 3 5734 2699; fax: +81 3 5734 2699.
*E-mail addresses:* hachiya@sg.cs.titech.ac.jp (H. Hachiya), akiyama@sg.cs.titech.ac.jp (T. Akiyama), sugi@cs.titech.ac.jp (M. Sugiayma), jan.peters@tuebingen.mpg.de (J. Peters).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.01.002

## 2. Background and notation

In this section, we review how Markov decision problems can be solved using policy iteration based on value functions.

### 2.1. Markov decision problems

Let us consider a Markov decision problem (MDP) specified by

$(\mathcal{S}, \mathcal{A}, P_T, R, \gamma)$,

where

- $\mathcal{S}$ is a set of states,
- $\mathcal{A}$ is a set of actions,
- $P_T(s'|s, a)$ $(\in [0, 1])$ is the transition probability-density from state $s$ to next state $s'$ when action $a$ is taken,
- $R(s, a, s')$ $(\in \mathbb{R})$ is a reward for transition from $s$ to $s'$ by taking action $a$,
- $\gamma \in (0, 1]$ is the discount factor for future rewards.

Let $\pi(a|s) \in [0, 1]$ be a stochastic policy which is the conditional probability density of taking action $a$ given state $s$. The state–action value function $Q^\pi(s, a) \in \mathbb{R}$ for policy $\pi$ is the expected discounted sum of rewards the agent will receive when taking action $a$ in state $s$ and following policy $\pi$ thereafter, i.e.,

$$Q^\pi(s, a) \equiv \mathop{\mathbb{E}}_{\pi, P_T}\left[\sum_{n=1}^{\infty} \gamma^{n-1} R(s_n, a_n, s_{n+1}) \,\middle|\, s_1 = s, a_1 = a\right],$$

where $\mathbb{E}_{\pi, P_T}$ denotes the expectation over $\{s_n, a_n\}_{n=1}^{\infty}$ following $\pi(a_n|s_n)$ and $P_T(s_{n+1}|s_n, a_n)$.

The goal of reinforcement learning is to obtain the policy which maximizes the sum of future rewards; the optimal policy can be expressed[1] as

$$\pi^*(a|s) \equiv \delta(a - \arg\max_{a'} Q^*(s, a')),$$

where $\delta(\cdot)$ is the Dirac delta function and $Q^*(s, a)$ is the *optimal* state–action value function defined by

$$Q^*(s, a) \equiv \max_{\pi} Q^\pi(s, a).$$

$Q^\pi(s, a)$ can be expressed as the following recurrent form called the *Bellman equation* (Sutton & Barto, 1998):

$$Q^\pi(s, a) = R(s, a) + \gamma \mathop{\mathbb{E}}_{P_T(s'|s,a)} \mathop{\mathbb{E}}_{\pi(a'|s')} \left[Q^\pi(s', a')\right],$$

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \tag{1}$$

where $R(s, a)$ is the expected reward function when the agent takes action $a$ in state $s$:

$$R(s, a) \equiv \mathop{\mathbb{E}}_{P_T(s'|s,a)} \left[R(s, a, s')\right].$$

$\mathbb{E}_{P_T(s'|s,a)}$ denotes the conditional expectation of $s'$ over $P_T(s'|s, a)$ given $s$ and $a$. $\mathbb{E}_{\pi(a'|s')}$ denotes the conditional expectation of $a'$ over $\pi(a'|s')$ given $s'$.

### 2.2. Policy iteration

The computation of the value function $Q^\pi(s, a)$ is called *policy evaluation*. Using $Q^\pi(s, a)$, we can find a better policy $\pi'(a|s)$ by

$$\pi'(a|s) = \delta(a - \arg\max_{a'} Q^\pi(s, a')).$$

This is called (greedy) *policy improvement*. It is known that repeating policy evaluation and policy improvement results in the optimal policy $\pi^*(a|s)$ (Sutton & Barto, 1998). This entire process is called *policy iteration*:

$$\pi_1 \xrightarrow{E} Q^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} Q^{\pi_2} \xrightarrow{I} \pi_3 \xrightarrow{E} \cdots \xrightarrow{I} \pi^*,$$

where $\pi_1$ is an initial policy, and $E$ and $I$ indicate policy *evaluation* and *improvement* steps respectively. For technical reasons, we assume that all policies are strictly positive (i.e., all actions have non-zero probability densities). In order to guarantee this, we use policy improvement which generates explorative policies such as the *Gibbs policy* and the *$\epsilon$-greedy policy*. In the case of the Gibbs policy,

$$\pi'(a|s) = \frac{\exp(Q^\pi(s, a)/\tau)}{\int_{\mathcal{A}} \exp(Q^\pi(s, a')/\tau)\, da'}, \tag{2}$$

where $\tau$ is a positive parameter which determines the randomness of the new policy $\pi'$. In the case of the $\epsilon$-greedy policy,

$$\pi'(a|s) = \begin{cases} 1 - \epsilon + \epsilon/|\mathcal{A}| & \text{if } a = a^*, \\ \epsilon/|\mathcal{A}| & \text{otherwise,} \end{cases} \tag{3}$$

where

$$a^* = \arg\max_a Q^\pi(s, a),$$

and $\epsilon \in (0, 1]$ determines how stochastic the new policy $\pi'$ is.

### 2.3. Value function approximation

Although policy iteration is guaranteed to produce the optimal policy, it is often computationally intractable since the number of state–action pairs $|\mathcal{S}| \times |\mathcal{A}|$ is very large; $|\mathcal{S}|$ or $|\mathcal{A}|$ becomes infinite when the state space or action space is continuous. To overcome this problem, the authors of the references Lagoudakis and Parr (2003), Precup, Sutton, and Dasgupta (2001) and Sutton and Barto (1998) proposed to approximate the state–action value function $Q^\pi(s, a)$ using the following linear model:

$$\widehat{Q}^\pi(s, a; \boldsymbol{\theta}) \equiv \sum_{b=1}^{B} \theta_b \phi_b(s, a) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(s, a),$$

where

$$\boldsymbol{\phi}(s, a) = (\phi_1(s, a), \phi_2(s, a), \ldots, \phi_B(s, a))^\top$$

are the fixed basis functions, $\top$ denotes the transpose, $B$ is the number of basis functions, and

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_B)^\top$$

are model parameters. Note that $B$ is usually chosen to be much smaller than $|\mathcal{S}| \times |\mathcal{A}|$. For $N$-step transitions, we ideally want to learn the parameters $\boldsymbol{\theta}$ so that the approximation error is minimized:

$$\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{P_1, \pi, P_T}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{Q}^\pi(s_n, a_n; \boldsymbol{\theta}) - Q^\pi(s_n, a_n)\right)^2\right],$$

where $\mathbb{E}_{P_1, \pi, P_T}$ denotes the expectation over $\{s_n, a_n\}_{n=1}^{N}$ following the initial-state probability density $P_1(s_1)$, the policy $\pi(a_n|s_n)$, and the transition probability density $P_T(s_{n+1}|s_n, a_n)$.

A fundamental problem of the above formulation is that the target function $Q^\pi(s, a)$ cannot be observed directly. To cope with this problem, we use the square of the Bellman residual (Lagoudakis & Parr, 2003; Schoknecht, 2003) as

$$\begin{aligned} \boldsymbol{\theta}^* &\equiv \arg\min_{\boldsymbol{\theta}} G, \\ G &\equiv \mathop{\mathbb{E}}_{P_1, \pi, P_T}\left[\frac{1}{N}\sum_{n=1}^{N} g(s_n, a_n; \boldsymbol{\theta})\right], \\ g(s, a; \boldsymbol{\theta}) &\equiv \Bigg(\widehat{Q}^\pi(s, a; \boldsymbol{\theta}) - R(s, a) \\ &\qquad - \gamma \mathop{\mathbb{E}}_{P_T(s'|s,a)} \mathop{\mathbb{E}}_{\pi(a'|s)} \left[\widehat{Q}^\pi(s', a'; \boldsymbol{\theta})\right]\Bigg)^2, \end{aligned} \tag{4}$$

[1] We assume that given state $s$ there is only one action maximizing the optimal value function $Q^*(s, a)$.