In: Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), Paris, July 2-7, 2006, pp. 66-73.

Balanced Importance Sampling Estimation

Paweł Wawrzyński
P.Wawrzynski@elka.pw.edu.pl

Andrzej Pacut
A.Pacut@ia.pw.edu.pl

Institute of Control and Computation Engineering
Warsaw University of Technology
00–665 Warsaw, Poland

Abstract

In this paper we analyze a particular issue of estimation, namely the estimation of the expected value of an unknown function for a given distribution, with the samples drawn from other distributions. A motivation of this problem comes from machine learning. In reinforcement learning, an intelligent agent that learns to make decisions in an unknown environment encounters the problem of judging an arbitrary decision policy (the given distribution) on the basis of previous decisions and their outcomes suggested by previous policies (other distributions).

The problem can be solved with the use of well established *importance sampling* estimators. To overcome a potential problem of excessive variance of such estimators, we introduce the family of *balanced importance sampling* estimators, prove their consistency and demonstrate empirically their superiority over the classical counterparts.

Keywords: Estimation, Importance Sampling, Machine Learning, Reinforcement Learning.

1 Introduction

Reinforcement Learning (RL) algorithms may be viewed as computational processes that transform observations of states, actions, and rewards, into policy parameters. Popular RL algorithms, like Q-Learning [17], and Actor-Critic methods, [1, 3, 4] process the data sequentially. A single observation is utilized for an adjustment the algorithms' parameters and becomes unavailable for further use. This sequential processing approach follows a common understanding that applications of RL to real life learning control problems require large amounts of data which cannot be kept in the limited memory assigned to the algorithm.

Recently, this reason for the algorithms to process data sequentially has lost its practical justification. The faster computers are and the larger their memory capacity is, the more practical the alternative assumption becomes, namely, that the RL algorithm can put all its input data in a database and utilize them in exhaustive processing. This idea is not new, and probably the first RL method based on this alternative assumption was DYNA was proposed by Sutton in [15]. Recently, algorithms of this kind have been gaining increasing attention [2, 5, 7–9, 13].

In this paper we analyze a core problem in the design of a RL algorithm that shape an agent policy on the basis of the entire available experience. The problem consists in inferring a quality of an arbitrary policy on the basis of observations of the previously applied policies. Usually such an inference is based on *importance sampling* estimators [12]. We present estimators that form a sound alternative to the existing ones.

The paper is organized as follows. In Sec. 2 we formulate the problem and introduce some notational conventions. In Sec. 3, we discuss the classical importance sampling (IS) estimators. Section 4 contains the main contribution of this paper. We introduce there the novel family of balanced IS estimators. We derive upper bound of the mean-squared error of these estimators which allows us to form conditions of their mean-squared consistency. Also, we derive the form of the optimal estimator within the family as the one that minimizes the upper bound of the mean-squared error. In the next section we demonstrate an example for which the error bound is tight. In Sec. 6, we report an experimental study that compares the novel estimators and their classical equivalents. Section 7 summarizes the paper.

2 Problem formulation

We consider an abstract problem that can be understood as a core issue that emerges in the design of RL algorithms.

Let $\varphi(\cdot\,;\theta)$ be a density of random variables with values in $\mathcal{A}$. The density is parametrized by $\theta \in \Theta$. Let $t \in \{1, 2, \dots\}$ be a discrete time. At each moment t, the parameter is equal to θ_t and an *action*, a_t, is drawn from $\varphi(\cdot\,;\theta_t)$. We denote it by

$$a_t \sim \varphi(\cdot\,;\theta_t).$$

The action yields a *payment*, $d(a_t)$, where $d : \mathcal{A} \mapsto \mathfrak{R}^{n_d}$ is an unknown function. The parameter at each moment t is generated on the basis of previous parameters and actions, namely

$$\theta_t = \widetilde{\theta}_t(\theta_1, a_1, d(a_1), \dots, \theta_{t-1}, a_{t-1}, d(a_{t-1}))$$

where $\widetilde{\theta}_t$ is a certain known function, and θ_1 is a given constant.

Our problem is to estimate, on the basis of events that have taken place up to the moment t, the integral

$$D(\theta) = \int_{\mathcal{A}} d(\alpha)\varphi(\alpha;\theta)\mathrm{d}\alpha.$$

Note that the above quantity is equal to the expected value of $d(a)$ when a is drawn with the use of the density $\varphi(\cdot\,;\theta)$.

The analyzed problem can be easily interpreted in the context of reinforcement learning. Namely, treating t as a time index or a trial index, we can interpret a_t as the action performed by the agent at moment t or the sequence of states visited and actions taken by the agent at trial t. Furthermore, d may represent a certain quality measure of an action/trial or an improvement direction. In such a context, D may be a measure of quality of the policy represented by θ or an averaged improvement direction. Finally, $\widetilde{\theta}_t$ may define a mechanism of adaptation of θ that employs the previous data.

3 Basic Importance Sampling Estimators

In order to construct an estimator of $D(\theta)$, we may apply an estimation technique called *importance sampling* [12] frequently applied in RL [8–10, 13]. Namely, given θ, θ_i, and a_i drawn from $\varphi(\cdot\,;\theta_i)$, the statistic

$$d(a_i)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} \tag{1}$$

is an unbiased estimator of $D(\theta)$, since

$$\begin{aligned}\mathcal{E}\left(d(a_i)\,\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}\right) &= \int d(\alpha)\frac{\varphi(\alpha;\theta)}{\varphi(\alpha;\theta_i)}\,\varphi(\alpha;\theta_i)\,\mathrm{d}\alpha \\ &= \int d(\alpha)\,\varphi(\alpha;\theta)\,\mathrm{d}\alpha \qquad (2)\\ &= D(\theta).\end{aligned}$$

Let $c \in \mathfrak{R}^{n_d}$ be a constant *baseline*. A derivation similar to the one above shows that for given c, also

$$c + (d(a_i) - c)\,\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} \tag{3}$$

is an unbiased estimator of $D(\theta)$ for given θ. We can manipulate c to confine the variance of (3). Because we want to estimate $D(\theta)$, it is reasonable to set c to a certain non-random assessment of this quantity. Such c would reduce the difference $d(a_i) - c$ in (3) for large values of the fraction $\varphi(a_i;\theta)\,\varphi(a_i;\theta_i)^{-1}$.

The standard estimator of $D(\theta)$ can be obtained by averaging (3) over all i-s, namely

$$c + \frac{1}{t}\sum_{i=1}^{t}(d(a_i) - c)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}. \qquad (4)$$

As an ordinary average value of unbiased estimators, (4) is unbiased. Under certain additional regularity conditions it is also consistent [12]. It yet has two important disadvantages: a possibility of excessive variance and equal treatment of samples regardless of their quality. The first disadvantage can be removed by a normalization, which we discuss below. The second one can be removed by balancing, which we introduce in the next section.

Normalization

Estimator (4) is impractical: for some combinations of θ and θ_i, the density in the denominator is likely to be very small. In this case variance of the estimator may be arbitrarily large. Because of these difficulty, the *normalized estimator* was introduced [8,10]. In order to derive it, we first note that for any random vector a in $\mathcal{A}$ and a function $d : \mathcal{A} \mapsto \mathfrak{R}^n$, a minimum of the index $\mathcal{E}\|d(a) - x\|^2$ is attained for $x = \mathcal{E}d(a)$. Let us first estimate $\mathcal{E}\|d(a) - x\|^2$ for θ and x given. Consider the estimator of the form

$$\|d(a_i) - x\|^2 \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}. \qquad (5)$$

It is, for each i, unbiased, since

$$\begin{aligned}\mathcal{E}\|d(a_i) - x\|^2 \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} &= \int_{\mathcal{A}} \|d(\alpha) - x\|^2 \varphi(\alpha;\theta)\,\mathrm{d}\alpha \\ &= \mathcal{E}\|d(a) - x\|^2\end{aligned}$$

The average of (5), namely

$$\frac{1}{t}\sum_{i=1}^{t}\|d(a_i) - x\|^2 \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}, \qquad (6)$$

is obviously also an unbiased estimator of $\mathcal{E}\|d(a) - x\|^2$. We now apply the property of the expected value mentioned above, and minimize (6) with respect to x, to obtain another estimator of $D(\theta)$, which was used in [8, 10]. Because (6) is a quadratic function of x, its minimum can be derived analytically to obtain the estimator of $D(\theta)$, namely

$$\frac{\sum_{i=1}^{t} d(a_i)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}}{\sum_{i=1}^{t}\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}}. \qquad (7)$$

If d is bounded, then the above estimator is also bounded as a weighted average of $d(a_i)$-s and, consequently, has bounded variance. Under the same regularity conditions as before, it is also consistent, even though it is biased for each t.

4 Balanced Estimators

Both classic estimators, (4) and (7), have important disadvantage: They weight the component estimators equally (by $1/t$), hence treating all the samples equally, regardless of their "quality". We introduce a family of *balanced estimators* particularly useful in solving the problem analyzed here. The estimators have the form

$$\widehat{D}_t^q(\theta) = c + \frac{\sum_{i=1}^{t}(d(a_i) - c)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}\, q(\theta,\theta_i)}{\sum_{i=1}^{t} q(\theta,\theta_i)} \qquad (8)$$

where c is constant and the *balance function* $q : \Theta \times \Theta \mapsto (0, 1]$ controls a relative impact of each sample on the aggregated estimator. Each estimator (8) is a weighted average of unbiased estimators (3), with the weights defined by q. Below we show how to obtain certain desired properties of $\widehat{D}_t^q(\theta)$ by an appropriate shaping of q. Because θ_i-s are random for $i > 1$, estimator $\widehat{D}_t^q(\theta)$ is in general biased. However, we will see that it is still consistent.

In order to analyze properties of estimator $\widehat{D}_t^q(\theta)$ (8), we introduce *the expected squared density ratio*

$$\kappa(\theta,\theta_i) = \mathcal{E}\left(\varphi(a_i;\theta)\,\varphi(a_i;\theta_i)^{-1}\right)^2. \qquad (9)$$

Certainly

$$\kappa(\theta,\theta_i) \geq \left(\mathcal{E}\varphi(a_i;\theta)\,\varphi(a_i;\theta_i)^{-1}\right)^2 = 1.$$

The following proposition is the key one in analysis of the properties of estimator $\widehat{D}_t^q(\theta)$.

Proposition 1 *If d is bounded, then the inequality*

$$\mathcal{E}\big\|\widehat{D}_t^q(\theta) - D(\theta)\big\|^2 \leq (\|c\| + \|c_0\|)^2 \mathcal{E}\left(\frac{\sum_{i=1}^t \kappa(\theta,\theta_i) q(\theta,\theta_i)^2}{(\sum_{i=1}^t q(\theta,\theta_i))^2}\right) \tag{10}$$

holds where $c_0 = \sup_a \|d(a)\|$.

Proof: See the Appendix.

From theoretical point of view, Proposition 1 allows us to define conditions for mean-squared convergence of $\widehat{D}_t^q(\theta)$ to $D(\theta)$. For instance, if the product $\kappa(\theta,\theta)q(\theta,\theta_i)^2$ is bounded and the sum $\sum_{i=1}^t q(\theta,\theta_i)$ grows faster than $m\sqrt{t}$ for each m, then the convergence takes place.

From practical point of view, Proposition 1 allows us to shape q in order to obtain favorable properties of $\widehat{D}_t^q(\theta)$. The most obvious objective is to minimize $\mathcal{E}\big(\widehat{D}_t^q(\theta) - D(\theta)\big)^2$. In order to do so, we will minimize the fraction that is averaged on the right hand side of inequality (10).

Differentiation of

$$\frac{\sum_{i=1}^t \kappa(\theta,\theta_i) q(\theta,\theta_i)^2}{(\sum_{i=1}^t q(\theta,\theta_i))^2}$$

with respect to $q(\theta,\theta_i)$ for an arbitrary i reveals that this fraction is minimized for

$$q(\theta,\theta_i) \propto 1/\kappa(\theta,\theta_i).$$

which makes (8) equal to the *optimized balanced estimator* in the form

$$\widehat{D}_t^\kappa(\theta) = c + \frac{\sum_{i=1}^t (d(a_i) - c)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}\kappa(\theta,\theta_i)^{-1}}{\sum_{i=1}^t \kappa(\theta,\theta_i)^{-1}}. \tag{11}$$

This estimator has a simple interpretation: $\kappa(\theta,\theta_i)$ defines a certain discrepancy measure for distributions $\varphi(\cdot\,;\theta_i)$ and $\varphi(\cdot\,;\theta)$. The larger the measure, the less "reliable" is i-th sample, and the smaller becomes its weight in (11).

Let us analyze convergence of the optimized balanced estimator (11). Applying Proposition 1 we obtain

$$\mathcal{E}\big(\widehat{D}_t^\kappa(\theta) - D(\theta)\big)^2 \leq (\|c\| + \|c_0\|)^2 \mathcal{E}\left(\frac{1}{\sum_{i=1}^t \kappa(\theta,\theta_i)^{-1}}\right).$$

Consequently, mean-squared convergence of $\widehat{D}_t^\kappa(\theta)$ to $D(\theta)$ takes place if the sum $\sum_{i=1}^t \kappa(\theta,\theta_i)^{-1}$ diverges to infinity with probability one. It is a rather weak condition. It is obviously satisfied if κ is bounded from above in $\Theta \times \Theta$. Otherwise, it is enough that with probability 1 there is an infinite subsequence of indexes i for which $\kappa(\theta,\theta_i)$ does not diverge too fast (or does not diverge at all).

Normalization

We may also employ the balance function q to derive an estimator similar to (7). In this order, consider estimators of $\mathcal{E}\|d(a) - x\|^2$ in the form

$$\frac{\sum_{i=1}^t \|d(a_i) - x\|^2 \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}\, q(\theta,\theta_i)}{\sum_{i=1}^t q(\theta,\theta_i)}. \tag{12}$$

Minimization with respect to x leads to the *normalized balanced* estimator of $D(\theta)$ in the form

$$\frac{\sum_{i=1}^t d(a_i)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}\, q(\theta,\theta_i)}{\sum_{i=1}^t \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)}\, q(\theta,\theta_i)}. \tag{13}$$

Since it is a weighted average of $d(a_i)$-s, this estimator is also bounded and has a bounded variance, provided d is bounded. Applying $q(\theta,\theta_i) = \kappa(\theta,\theta_i)^{-1}$, we further decrease its variance. Notice that (13) corresponds to (7). It is biased, it is also a weighted average, yet here the weights take into account a "reliability" of samples.

5 An Example

In this section we analyze a certain simple problem to illustrate the introduced solutions. The example also shows that the bound defined by Proposition 1 is tight in the following sense: The fraction of the true value of

$\mathcal{E}\|\widehat{D}_t^q(\theta) - D(\theta)\|^2$ and the bound can be arbitrarily close to 1. Let $\mathcal{A} = \Theta = \mathfrak{R}$ and $\varphi(\cdot\,;\theta)$ be a density of the normal distribution $N(\theta, 1)$, namely

$$\varphi(a;\theta) = (2\pi)^{-1/2} \exp\left(-0.5(a-\theta)^2\right).$$

We consider the payment function d of the form

$$d(a) = \begin{cases} 0 & \text{iff } a \leq 0 \\ 1 & \text{iff } a > 0. \end{cases}$$

φ and d define D of the form

$$D(\theta) = \int_0^{+\infty} \varphi(\alpha;\theta)\mathrm{d}\alpha = \Phi(\theta)$$

where Φ is the cumulative distribution function of the normal distribution $N(0,1)$. Let $\{\theta_1, \ldots, \theta_t\}$ be a set of non-random parameters in $\mathfrak{R}$; their non-randomness will greatly simplify an analysis of estimators of $D(\theta)$.

We analyze the variance of the estimator $\widehat{D}_t^q$ (8) for $c = 0$ in the above setting. An auxiliary function

$$\kappa_0(\theta,\theta_i) = \int_0^{+\infty} \left(\frac{\varphi(\alpha;\theta)}{\varphi(\alpha;\theta_i)}\right)^2 \varphi(\alpha;\theta_i)\mathrm{d}\alpha$$

allows us to derive a lower bound of $\mathcal{E}(\widehat{D}_t^q(\theta) - D(\theta))^2$, namely

$$\begin{aligned}
&\mathcal{E}(\widehat{D}_t^q(\theta) - D(\theta))^2 \\
&= \mathcal{E}\left(\frac{\sum_{i=1}^t \left(d(a_i)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} - D(\theta)\right) q(\theta,\theta_i)}{\sum_{i=1}^t q(\theta,\theta_i)}\right)^2 \\
&= \frac{\sum_{i=1}^t \mathcal{E}\left(d(a_i)\frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} - D(\theta)\right)^2 q(\theta,\theta_i)^2}{(\sum_{i=1}^t q(\theta,\theta_i))^2} \\
&= \frac{\sum_{i=1}^t \left(\kappa_0(\theta,\theta_i) - D(\theta)^2\right) q(\theta,\theta_i)^2}{(\sum_{i=1}^t q(\theta,\theta_i))^2}.
\end{aligned}$$

Let $\theta_i < 0 < \theta$ for all i. We then have

$$\begin{aligned}
&\kappa(\theta,\theta_i) - \kappa_0(\theta,\theta_i) \\
&\quad = \int_{-\infty}^0 \frac{\varphi(\alpha;\theta)}{\varphi(\alpha;\theta_i)}\varphi(\alpha;\theta)\mathrm{d}\alpha < 1.
\end{aligned}$$

because $\varphi(\alpha;\theta) < \varphi(\alpha;\theta_i)$ for $\alpha < 0$. Because obviously $D(\theta) \leq 1$, we obtain a lower bound of $\mathcal{E}(\widehat{D}_t^q(\theta) - D(\theta))^2$ in the form

$$\frac{\sum_{i=1}^t \left(\kappa(\theta,\theta_i) - 2\right) q(\theta,\theta_i)^2}{(\sum_{i=1}^t q(\theta,\theta_i))^2} \tag{14}$$

For the analyzed φ we have

$$\kappa(\theta,\theta_i) = \exp\left((\theta-\theta_i)^2\right).$$

If we divide (14) by the upper bound implied by Proposition 1, we will obtain the lower-upper bound ratio that in this case is of the form

$$\frac{\sum_{i=1}^t \left(e^{(\theta-\theta_i)^2} - 2\right) q(\theta,\theta_i)^2}{\sum_{i=1}^t e^{(\theta-\theta_i)^2} q(\theta,\theta_i)^2}.$$

We can see that it can be arbitrarily close to 1 if only the difference $\theta - \theta_i$ is large enough for all i.

The above analysis can be easily extended to multidimensional normal distributions. Namely, let φ be normal densities of mean θ and constant nonsingular covariance matrix C. In this case

$$\varphi(a;\theta) = \frac{\exp\left(-0.5\,(a-\theta)^T C^{-1}(a-\theta)\right)}{\sqrt{(2\pi)^{n_\theta}|C|}}$$

and we have

$$\kappa(\theta,\theta_i) = \exp\left((\theta-\theta_i)^T C^{-1}(\theta-\theta_i)\right).$$

This value grows very fast with $\|\theta - \theta_i\|$. Suppose we want to infer the value $D(\theta)$ on the basis of the classical estimator (4). Its variance can be bounded by (10) with $q \equiv 1$. It depends mainly on those θ_i-s which are very far from θ, no matter how many samples are drawn that use parameters close to θ. On the contrary, the balanced estimator, especially in its optimized form (11) can be understood as based on samples drawn with the use of θ_i-s closest to θ and is not disturbed if some other θ_i-s are very far from θ.

6 Experimental Study

We check behavior of all the discussed estimators in a simple estimation problem.

Let the payment function $d : \mathfrak{R} \mapsto \mathfrak{R}$ be defined as

$$d(a) = \text{sign}(\sin(a)).$$

Let $\varphi(\cdot\,;\theta)$ be a normal density with mean θ and variance 1. It can be derived that in this case

$$\begin{aligned} D(\theta) &= \int_{\mathcal{A}} d(\alpha)\varphi(\alpha;\theta)\mathrm{d}\alpha \\ &= \sum_{k\geq 0} \frac{4}{(2k+1)\pi} \frac{\sin\left((2k+1)\theta\right)}{\exp\left(0.5(2k+1)^2\right)}. \end{aligned}$$

Even though it is an infinite sum, only few of its components are larger than a numeric error. Fig. 1 presents both d and D. Notice that D is a local average of d.

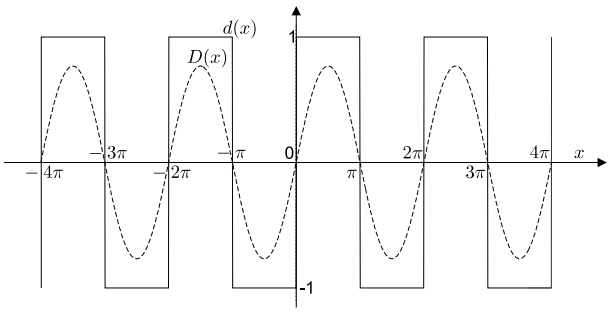

Figure 1: Functions d and D.

Let $\{\theta_t, t = 1, 2, \dots\}$ be a sequence of parameters drawn independently from the uniform distribution $U(-10\pi, 10\pi)$. We will investigate how fast the estimators of $D(\theta)$ converge to this quantity for all $\theta \in \Theta$. In this order, we will analyze how fast the discrepancy measure

$$\widehat{e}_t = \sqrt{\frac{1}{20\pi}\int_{\Theta} (D(\theta) - \widehat{D}_t(\theta))^2 \mathrm{d}\theta} \qquad (15)$$

converges to 0 for various estimators $\widehat{D}_t(\theta)$. In (15) $\Theta = [-10\pi, 10\pi]$. We have analyzed 4 estimators that can be applied: the standard one (4), the normalized one (7), the balanced one (8) (here we will apply its optimized form (11)) and the normalized balanced one (13). In all performed experiments the baseline c is equal to 0 and the balance function is equal to $\kappa(\theta, \theta_i)^{-1} = \exp(-(\theta - \theta_i)^2)$.

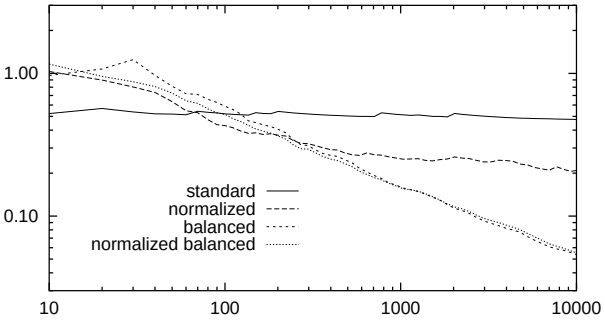

Figure 2: Average $\widehat{e}_t$ vs. time index for various estimators. Each above curve averages 10 runs.

Fig. 2 demonstrates the behavior of the four discussed estimators. The standard estimator behaves very poorly, hardly improving with time. The normalized estimator behaves much better, but is definitely outperformed by the balanced and the normalized balanced estimator.

7 Summary

In this paper we developed importance sampling estimation for its applications in reinforcement learning. To overcome a variance problem, a family of balanced estimators has been introduced. Conditions for their mean-squared consistency have been analyzed. The experimental study shows that the proposed estimators converge faster to the appropriate values that their traditional equivalents.

References

[1] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike Adaptive Elements That Can Learn Difficult Learning Control Problems," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-13, pp. 834-846, Sept.-Oct. 1983.

[2] M. Kearns, Y. Mansour, and A. Y. Ng, "Approximate Planning In Large POMDPs Via Reusable Trajectories," *Advances in Neural Information Processing Systems,* pp. 1001-1007, 1999.

[3] H. Kimura and S. Kobayashi, "An Analysis of Actor/Critic Algorithm Using Eligibility

Traces: reinforcement learning with imperfect value functions," *Proceedings of the Fifteenth International Conference on Machine Learning,* 1998.

[4] V. R. Konda and J. N. Tsitsiklis, "Actor-Critic Algorithms," *SIAM Journal on Control and Optimization,* Vol. 42, No. 4, pp. 1143-1166, 2003.

[5] Y. Mansour, "Reinforcement Learning And Mistake Bounded Algorithms," *Proceedings of the 12th Annual Conference on Computational Learning Theory,* pp. 183-192, ACM Press, New York, 1999.

[6] N. Meuleau, L. Peshkin, and K.-E. Kim, "Exploration in Gradient-Based Reinforcement Learning," Technical Report AI Memo 2001-003, MIT, 2001.

[7] L. Peshkin, N. Meuleau, and L. P. Kaelbling, "Learning Policies With External Memory," *Proceedings of the Sixteenth International Conference on Machine Learning,* Morgan Kaufmann, 1999.

[8] L. Peshkin and S. Mukherjee, "Bounds on Sample Size For Policy Evaluation in Markov Environments," *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory,* pp. 608-615, 2001.

[9] L. Peshkin and Ch. R. Shelton, "Learning from Scarce Experience," *Proceedings of the Nineteenth International Conference on Machine Learning,* pp. 498-505, 2002.

[10] D. Precup, R. S. Sutton, and S. Singh, "Eligibility Traces for Off-Policy Policy Evaluation," *Proceedings of the Seventeenth International Conference on Machine Learning,* Morgan Kaufmann, 2000.

[11] D. Precup, R. S. Sutton, and S. Dasgupta, "Off-Policy Temporal-Difference Learning with Function Approximation," *Proceedings of the Eighteenth International Conference on Machine Learning,* 2001.

[12] R. Rubinstein, *Simulation and The Monte Carlo Method*. New York, Wiley, 1981.

[13] Ch. R. Shelton, "Policy Improvement for POMDPs Using Normalized Importance Sampling," *Proceedings of the Seventeenth International Conference on Uncertainty in Artificial Intelligence,* pp. 496-503, 2001.

[14] C. R. Shelton, "Importance Sampling for Reinforcement Learning with Multiple Objectives," PhD Thesis, MIT, August 2001.

[15] R. S. Sutton, "Integrated Architectures For Learning, Planning, and Reacting Based on Approximating Dynamic Programming," *Proceedings of the Seventh Int. Conf. on Machine Learning,* pp. 216-224, Morgan Kaufmann, 1990.

[16] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[17] C. Watkins and P. Dayan, "Q-Learning," *Machine Learning*, vol. 8, pp. 279-292, 1992.

Proof of Proposition 1:
We fix θ and denote
1. $\widehat{b}_t = \left\| \widehat{D}_t^q(\theta) - D(\theta) \right\|^2$.
2. $d_i^c = d(a_i) - c$,
3. $D^c = D(\theta) - c$,
4. $Q_t = \sum_{i=1}^t q(\theta, \theta_i)$,
5. $\mathcal{F}_i$ – σ-algebra generated by the events up to moment i, $\mathcal{F}_i = \sigma(\theta_1, a_1, \ldots, \theta_i, a_i, \theta_{i+1})$.
6. $M_t = (\|c_0\| + \|c\|)^2 \kappa(\theta, \theta_t) q(\theta, \theta_t)^2$.

Step 1. We express $\widehat{b}_t$ as a function of $\widehat{b}_{t-1}$. First, note that

$$\begin{aligned}
\widehat{b}_t &= \left\| c + \frac{\sum_{i=1}^t (d(a_i) - c) \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} q(\theta, \theta_i)}{\sum_{i=1}^t q(\theta, \theta_i)} - D(\theta) \right\|^2 \\
&= \left\| \frac{\sum_{i=1}^t d_i^c \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} q(\theta, \theta_i)}{\sum_{i=1}^t q(\theta, \theta_i)} - D^c \right\|^2 \\
&= \left\| \frac{\sum_{i=1}^t \left(d_i^c \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} - D^c \right) q(\theta, \theta_i)}{Q_t} \right\|^2 .
\end{aligned}$$

We split the sum in the numerator into the first $t-1$ elements and the last one to obtain

$$\widehat{b}_t = \frac{\left\|\sum_{i=1}^{t-1}\left(d_i^c \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} - D^c\right) q(\theta,\theta_i)\right\|^2}{Q_t^2} \quad (16)$$

$$+ 2\frac{\left(\sum_{i=1}^{t-1}\left(d_i^c \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} - D^c\right) q(\theta,\theta_i)\right)^T}{Q_t^2} \times$$

$$\times \left(d_t^c \frac{\varphi(a_t;\theta)}{\varphi(a_t;\theta_t)} - D^c\right) q(\theta,\theta_t) \quad (17)$$

$$+ \frac{\left\|d_t^c \frac{\varphi(a_t;\theta)}{\varphi(a_t;\theta_t)} - D^c\right\|^2 q(\theta,\theta_t)^2}{Q_t^2}.$$

$\widehat{b}_{t-1}$ is hidden in (16) since

$$\frac{\left\|\sum_{i=1}^{t-1}\left(d_i^c \frac{\varphi(a_i;\theta)}{\varphi(a_i;\theta_i)} - D^c\right) q(\theta,\theta_i)\right\|^2}{Q_t^2} = \frac{Q_{t-1}^2}{Q_t^2}\widehat{b}_{t-1}.$$

Step 2. We derive a bound for $\mathcal{E}(\widehat{b}_t|\mathcal{F}_{t-1})$. Under $\mathcal{F}_{t-1}$, θ_t is constant and the only random variable is a_t. The expected value of (17) vanishes and we have:

$$\mathcal{E}(\widehat{b}_t|\mathcal{F}_{t-1}) = \frac{Q_{t-1}^2}{Q_t^2}\widehat{b}_{t-1}$$

$$+ \frac{1}{Q_t^2}\mathcal{E}\left(\left(d_t^c \frac{\varphi(a_t;\theta)}{\varphi(a_t;\theta_t)} - D^c\right)^2 q(\theta,\theta_t)^2 \Big|\theta_t\right)$$

Because $\|d_t^c\|$ is bounded by $\|c_0\| + \|c\|$, we have

$$\mathcal{E}\left(\left(d_t^c \frac{\varphi(a_t;\theta)}{\varphi(a_t;\theta_t)} - D^c\right)^2 \Big|\theta_t\right) = \mathcal{V}\left(d_t^c \frac{\varphi(a_t;\theta)}{\varphi(a_t;\theta_t)}\Big|\theta_t\right)$$
$$\leq (\|c_0\| + \|c\|)^2 \kappa(\theta,\theta_t),$$

which leads to

$$\mathcal{E}(\widehat{b}_t|\mathcal{F}_{t-1}) \leq \frac{Q_{t-1}^2}{Q_t^2}\widehat{b}_{t-1}$$
$$+ \frac{1}{Q_t^2}(\|c_0\| + \|c\|)^2 \kappa(\theta,\theta_t) q(\theta,\theta_t)^2$$
$$\leq \frac{Q_{t-1}^2}{Q_t^2}\widehat{b}_{t-1} + \frac{1}{Q_t^2}M_t.$$

Step 3. We derive recursively a bound for $\mathcal{E}(\widehat{b}_t|\mathcal{F}_0)$. We have

$$\mathcal{E}(\widehat{b}_t|\mathcal{F}_{t-2}) = \mathcal{E}(\mathcal{E}(\widehat{b}_t|\mathcal{F}_{t-1})|\mathcal{F}_{t-2})$$
$$\leq \mathcal{E}\left(\frac{Q_{t-1}^2}{Q_t^2}\widehat{b}_{t-1} + \frac{1}{Q_t^2}M_t\big|\mathcal{F}_{t-2}\right).$$

We can decompose $\widehat{b}_{t-1}$ in a similar manner

$$\mathcal{E}(\widehat{b}_t|\mathcal{F}_{t-2}) \leq \mathcal{E}\left(\frac{Q_{t-2}^2}{Q_t^2}\widehat{b}_{t-2} + \frac{M_{t-1}}{Q_t^2} + \frac{M_t}{Q_t^2}\big|\mathcal{F}_{t-2}\right).$$

The decomposition can be repeated for $\mathcal{F}_{t-3}, \mathcal{F}_{t-4}, \ldots$ to obtain

$$\mathcal{E}\widehat{b}_t = \mathcal{E}(\widehat{b}_t|\mathcal{F}_0) \leq \mathcal{E}\left(\frac{\sum_{i=1}^t M_i}{Q_t^2}\big|\mathcal{F}_0\right)$$
$$\leq (\|c_0\| + \|c\|)^2 \mathcal{E}\left(\frac{\sum_{i=1}^t \kappa(\theta,\theta_i) q(\theta,\theta_i)^2}{(\sum_{i=1}^t q(\theta,\theta_i))^2}\right).$$

■