Interval Estimation for Reinforcement-Learning Algorithms in Continuous-State Domains: Supplementary Material

Martha White
Department of Computing Science
University of Alberta
whitem@cs.ualberta.ca

Adam White
Department of Computing Science
University of Alberta
awhite@cs.ualberta.ca

Applicability Proofs for Block Bootstrapping in Reinforcement Learning

We prove the consistency and coverage error for the bootstrapped studentized interval around our the sample mean of the sequence of parameters for global function approximation. Each parameter vector $\boldsymbol{\theta}_t$ on time step t corresponds to the action-value function Q_t on that time step, with $Q(s,a) = f(\boldsymbol{\theta}, s, a)$ for some bounded function f. A common example of f is a linear function $f(\boldsymbol{\theta}, s, a) = \boldsymbol{\theta}^T \boldsymbol{\phi}(s,a)$ with features given by the function $\boldsymbol{\phi} : S \times A \rightarrow \mathbb{R}^d$.

Let $\{\boldsymbol{\theta}_t\}$ be the sequence of weight vectors, drawn from probability distributions $P_a(< \boldsymbol{\Theta}_t, s_t > | < \boldsymbol{\theta}_{t-1}, s_{t-1} >, \ldots, < \theta_{t-k-1}, s_{t-k-1})$ with means μ_t. For a given state-action pair $(s,a) \in S \times A$, with $g(\boldsymbol{\theta}) = f(\boldsymbol{\theta}, s, a)$, we are estimating

$$\bar{g}_n = n^{-1} \sum_{i=1}^{n} g_i$$

where $g_i = E[(g(\boldsymbol{\Theta}_i)]$.

In order to prove a coverage error of $o(n^{-1/2})$ for the studentized interval on $f(\mu_n, s, a)$ for any given $(s,a) \in S \times A$, we will need the following assumptions, simplified from Lahiri's Theorem 4.1 [5] for our scenario. The proof will be for any (s,a), so we fix an $(s,a) \in S \times A$ and let $g(\theta) = f(\theta, s, a)$.

Let $Y_n =< S_n, A_n, R_n >$, the triplet obtained from acting in the given MDP, $G = (S, A, P, R)$. The triplets are drawn from the implicit from the probability distribution $P_R(s,a,r)$ computed using $P(s,a,s')$ and $R(s,a,s')$, giving the probability of receiving reward r after taking action a in state s. Let $D_j = \sigma(Y_j)$, the σ-fields of the random variables Y_n.

Assumption 1 *For any $(s,a) \in S \times A$, the density function, $P_R(s,a,\dot{}) : R \rightarrow \mathbb{R}$ is continuous and bounded.*

This assumption is required so that we can approximate this (continuous) density function with infinitely many samples from Y_n. This in turn enables us to define the σ-fields in terms of the Y_n, placing the strong mixing assumptions instead on Y_n, rather than on our sequence, $\{\boldsymbol{\theta}_t\}$. The next two assumptions are the typical assumptions placed on the sequence of σ-fields for bootstrapping proofs. Essentially, Assumptions 2 and 3 both place a strong mixing assumptions on Y_n. Mixing assumptions are a common restriction in reinforcement learning, related to ergodicity of an MDP [2, 7]. Any MDP satisfying the above mixing assumptions is ergodic [3]. Any stationary positive recurrent Markov chain (essentially ergodic) with trivial tail field is strongly mixing [6](Page 553).

Assumption 2 *There exists $d > 0$ such that for all $m, n \in \mathbb{N}$, $A \in D_{-\infty}^{n}$ and $B \in D_{n+m}^{\infty}$,*

$$|P(A \cap B) - P(A)P(B)| \leq d^{-1} e^{-dm}$$

Assumption 3 *There exists $d > 0$ such that for all $m, n, q \in \mathbb{N}$, $A \in D_{m-q}^{m+q}$,*

$$E|P(A|D_j : j \neq n) - P(A|D_j : 0 < |n - j| \leq m + q)| \leq d^{-1}e^{-dm}$$

The remaining assumptions are on the sequence of function values, $\{g(\boldsymbol{\theta}_t)\}$ (implicitly assumptions on the sequence of weights $\{\boldsymbol{\theta}_t\}$). The assumptions include boundedness assumptions on certain moments of the sequence, a smoothness condition and an m-dependence requirement.

Assumption 4 *g is a continuos bounded function and $\sup_{j\geq 1} E||g(\boldsymbol{\theta}_j)||^{12} < \infty$*

Assumption 5 *(Conditional Cramer condition) There exists $d > 0$ such that for all $m, n \in \mathbb{N}$, $d^{-1} < m < n$ and all $t \in \mathbb{R}$ with $t \geq d$,*

$$E|E[e^{it(g(\boldsymbol{\theta}_{n-m})+\ldots+g(\boldsymbol{\theta}_{n+m})}|D_j : j \neq n]| \leq e^{-d}$$

Assumption 6 *For $|s - t| > m$, $\mathrm{Cov}(g(\boldsymbol{\theta}_s), g(\boldsymbol{\theta}_t)) = 0$*

We expect our Q-values to change smoothly and to remain reasonably bounded, so Assumptions 4 and 5 are unrestrictive assumptions. The m-dependence assumption is slightly stronger, though m can be quite large, so we can still have quite long-range dependence.

Finally, we want to say something about the properties of the algorithm, essentially restricting to non-divergent algorithms.

Assumption 7 *$M_n = \mathrm{Var}(n^{-1/2}(g(\boldsymbol{\theta}_1)+\ldots+g(\boldsymbol{\theta}_n)))$ is non-singular $\forall\, n$ and $M = \lim_{n\to\infty} M_n$ exists and is non-singular.*

This assumptions is necessary to ensure that we reach a normal distribution in the limit and can therefore use an Edgeworth expansion to approximate the true distribution. Notice therefore that for non-convergent (but not divergent) sequences of $\boldsymbol{\theta}_n$, we can still apply the bootstrap, as long as the conditions on M_n are met. For convergent sequences, $M_n \to 0$, because the weights stop changing; however, in practice, the weights will always oscillate around the true $\boldsymbol{\theta}_0$. Therefore, assuming the above does not practically restrict convergence. We expect that we can drop this condition and in fact expect many of the conditions to simplify assuming $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_0$, but we leave this to future work.

With just these assumptions, we can proof the main result.

Theorem 1 *Given that Assumption 1-7 are satisfied and there exists constants $C_1, C_2 > 0$, $0 < \alpha \leq \beta < 1/4$ such that $C_1 n^\alpha < l < C_2 n^\beta$ (i.e. l increases with n), then the moving block bootstrap produces a one-sided confidence interval that is consistent and has a coverage error of $o(n^{-1/2})$ for the studentization of the mean of the process $\{Q_1(s, a) = f(\boldsymbol{\theta}_1, s, a), Q_2(s, a) = f(\boldsymbol{\theta}_2, s, a), \ldots\}$.*

Proof: For the proof, we need to satisfy the seven assumptions in Lahiri's Theorem 4.1 [5]. We will call these assumptions *requirements* to distinguish them from our assumptions. The proof will be organized based on these requirements (which will be stated below). The statements of the requirements will be italicized with justification of how that requirement is satisfied following the italicized statement.

Requirement 1 (C.1'): $\sup_{j\geq 1} E||g(\boldsymbol{\theta}_j)||^4 < \infty$ *and* $M = \lim_{n->\infty} M_n$ *exists and is non-singular* Satisfied by Assumptions 4 and 7

Requirement 2 (C.2): *There exists a $d > 0$ such that for $n, m \in \mathbb{N}$ with $m > d^{-1}$, there exists a D_{n-m}^{n+m} measurable p-variate random vector $\bar{Y}_{n,m}$ for which*

$$E||g(\boldsymbol{\theta}_n) - \bar{Y}_{n,m}|| \leq d^{-1}e^{-dm}$$

C.2 is satisfied by Assumption 1 and by the construction of Y_n, which we justify in the following. Essentially, this requirement says that for infinitely many Y_n, we can accurately approximate $g(\boldsymbol{\theta}_n)$. In Theorem 1 [4], an exponential decrease rate is proved for a kernel density estimator. Choose a kernel K satisfying:

1. The kernel function K is a probability density of bounded variation such that $\int K^2(u)du < \infty$; further, the derivative K' exists and is integrable.

2. For the parameters h_n, p_n defined in [4], $nh_n/p_n \to \infty$.

Now, by the theorem, because P_R is bounded and continuous and Y_n is strongly mixing, the error in the approximation of P_R using samples from Y_n decreases exponentially with n for some constants c_1, c_2 (based on norms on K, some constants, etc.), at a rate of

$$err \leq c_1 e^{-c_2 n} \tag{1}$$

With an accurate P_R distribution for a given (s, a), we can exactly compute the mean for the distribution $P_a(< \boldsymbol{\Theta}_t, s_t > \mid < \boldsymbol{\theta}_{t-1}, s_{t-1} >, \ldots, < \boldsymbol{\theta}_{t-k}, s_{t-k} >)$. Let $\bar{Y}_{n,m}$ be the function g applied to the approximated mean of this distribution using the current approximation of P_R with the $2m$ samples, $Y_{n-m}, \ldots, Y_{n+m}$. The construction of $\bar{Y}_{n,m}$ is possible because we have a measurable function (Borel function) between the σ-fields and $g(\boldsymbol{\Theta})$ drawn from $P_a(< g(\boldsymbol{\Theta}_t), s_t > \mid < \boldsymbol{\theta}_{t-1}, s_{t-1} >, \ldots, < \boldsymbol{\theta}_{t-k}, s_{t-k} >)$: the kernel estimator is continuous and g is continuous, between Borel sets. Therefore, the error between this approximated mean and the L_2 norm of the random variable will decrease with the some C times the above rate (Equation 1), for some large enough C (which exists because P_R and g are bounded).

Setting $d = \min\{(C \times c_1)^{-1}, c_2\}$, we obtain the desired result.

Requirement 2 (C.2 (ii)): $\sup_{j \geq 1} E||g(\boldsymbol{\theta}_j)||^{12} < \infty$
The twelfth moment is bounded by Assumption 4.

Requirements 3 and **5** (C.3 and C.5) correspond exactly to our Assumptions 2 and 3.

Requirement 4 (C.4) corresponds exactly to our smoothness Cramer condition, Assumption 5.

Finally, there are two more requirements that restrict the heterogeneity of the means μ_t asymptotically. Let

$$\begin{aligned} m_t &= EU_t = l^{-1} \sum_{j=1}^{l} g_{t+j-1} \\ M_{nt} &= \mathrm{Var}(\sqrt{l} U_t) \\ U_t &= l^{-1} \sum_{j=1}^{l} g(\boldsymbol{\Theta}_{t+j-1}) \end{aligned}$$

where l is the block length and $1 \leq t \leq b$. Recall $M_n = \mathrm{Var}(n^{-1/2}(g(\boldsymbol{\Theta}_1) + \ldots + g(\boldsymbol{\Theta}_n)))$ and $\bar{g}(\mu_n) = n^{-1}(g_1, \ldots, g_n)$, where $g_i = E[g(\boldsymbol{\Theta}_i]$.

Since the sequence of μ_t eventually reaches it's limiting distribution (stationary) with mean μ and variance σ, for $n_0 \in \mathbb{N}$, $\mu = \mu_t = \mu_{t+1} = \ldots$ for all $t > n_0$. Note that this Markov chain becomes stationary because the $\boldsymbol{\theta}_t$ are drawn from a time-homogenous Markov chain, and time-homogenous Markov chains always reach a limiting distribution [1]. This fact enables us to satisfy the asymptotic heterogeneity conditions.

Requirement 6 (C.6):

$$\lim_{n \to \infty} \max\{l^2 ||m_t - \bar{\mu}_n|| : 1 \leq t \leq b\} = 0$$

Requirement 7 (C.7):[1]

$$\lim_{n \to \infty} \max\{l^2 ||M_{nt} - M_n|| : 1 \leq t \leq b\} = 0$$

Note that as $n \to \infty$, then by the assumption that $C_1 n^\alpha < l < C_2 n^\beta$, l also goes to ∞. As $l \to \infty$, the finite initial number of samples with means different from μ (the stationary distribution) will be dominated by the infinite tail of samples in the stationary distribution with mean μ. Therefore, clearly both $m_t \to g(\mu)$ and $\bar{\mu}_n \to g(\mu)$as $n \to \infty$. Similarly, the tail has the same variances; therefore, the difference between $M_{n,t}$ and M_n goes to zero.

[1]Note that C.7 is slightly different in the theorem, but Lahiri mentions that it can be simplified to what we have here because of requirement C.1

Therefore, because our data meets the assumptions in [5], we know that the bootstrap is consistent and the coverage error for the studentized confidence interval on $\{Q_n\}$ is $o(n^{-1/2})$.

■

References

[1] D. Aldous. Random walks on finite groups and rapidly mixing Markov chains. *Séminaire de Probabilités XVII 1981/82*, pages 243–297, 1983.

[2] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, 19:49, 2007.

[3] R.M. Gray. *Probability, random processes, and ergodic properties*. Springer Verlag, 2009.

[4] C Henriques and P Oliveira. Exponential rates for kernel density estimation under association. *Statistica Neerlandica*, Jan 2005.

[5] SN Lahiri. Edgeworth correction by moving blockbootstrap for stationary and nonstationary data. *Exploring the Limits of Bootstrap*, pages 183–214, 1992.

[6] T. Orchard, MA Woodbury, L.M. Le Cam, J. Neyman, and E.L. Scott. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1972.

[7] C. Szepesvári. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010.