Contents lists available at ScienceDirect

Automatica

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

Coupling based estimation approaches for the average reward performance potential in Markov chains[☆]

Yanjie Li[a,*], Xinyu Wu[b], Yunjiang Lou[a], Haoyao Chen[a], Jiangang Li[a]

[a] *Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, China*
[b] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*

ARTICLE INFO

Article history:
Received 28 August 2016
Received in revised form 27 September 2017
Accepted 28 December 2017

Keywords:
Value function
Performance potential
Perturbation realization factor
Estimation with geometric variance reduction (EGVR)
Coupling techniques

ABSTRACT

Performance potential is an important concept in the sensitivity analysis of Markov chains. The estimation of performance potential provides the basis for the simulation-based optimization and sensitivity analysis of Markov chains. In this study, we present novel estimation approaches for the average reward (or cost) performance potential by combining perturbation realization factors and coupling techniques for Markov chains with finite state space. These approaches can effectively implement estimation with geometric variance reduction for average reward performance potential. Meanwhile, a number of coupling methods, including two optimal coupling methods, can be applied to further reduce estimation variance or simulation time. The numerical tests show that our approaches can significantly enhance the simulation efficiency.

1. Introduction

Performance potential (Cao, 2007) is referred to by different terms in various fields. For the average reward (cost) case, performance potential is also called "bias" (Puterman, 1994), "relative cost function" (Bertsekas, 1995), or "value function" (Munos, 2006). Performance potential plays an important role in solving the Markov decision process (Cao, 2007; Jia, 2011; Li & Wu, 2016; Xia & Jia, 2015). The estimation of performance potential provides the basis for the simulation-based optimization (Chang, Fu, Hu, & Marcus, 2007; Cooper, Henderson, & Lewis, 2003; Marbach & Tsitsiklis, 2001) and sensitivity analysis of Markov chains (Cao, 2007). Monte Carlo simulation is a powerful tool for estimating performance potential (Cooper et al., 2003). Regular Monte Carlo estimation can provide variance reduction with order $O(1/N)$, where N is the number of sample paths of a Markov chain. In general, variance reduction simulation techniques include importance sampling, correlated sampling, control variate, stratified sampling and so on Hammerley and Handscomb (1964). Estimations with less variance can be obtained by combining the aforementioned techniques. Estimations with geometric variance reduction (EGVR), the variances of which have a geometric reduction rate, have been investigated by processing the aforementioned methods iteratively (Halton, 1994; Kollman, Baggerly, Cox, & Picard, 1999) and have been applied to continuous-time Markov processes (Gobet & Maire, 2006), discrete-time Markov processes, and gradient estimation (Munos, 2006). As described in Munos (2006), the EGVR of average reward performance potential is subtle and deserves intensive treatment. An estimation approach with geometric variance reduction was proposed in Munos (2006) for average reward performance potential. However, additional variables, such as average reward (also called average expected gain in Munos (2006)) and steady state probability distribution, should be estimated and the sample path functional should be truncated artificially. These problems eliminate the benefit of EGVR and make reducing variance with a geometric rate difficult. A novel generalized fundamental matrix was proposed in Xia and Glynn (2016) to compute average reward performance potential. This fundamental matrix may provide new computation approaches for average reward performance potential; however, it seems to be difficult to find the corresponding sample path functional for EGVR.

[☆] The authors were partially supported by the National Natural Science Foundation of China (61004036, 61673131), Shenzhen basic research programs (JCYJ20150731105106111, JCYJ20160427183958817, JCYJ20150731105134064), and the National Natural Science Foundation-Shenzhen Joint Research Program (U1613219). The material in this paper was presented at the 31st Chinese Control Conference, July 20-27, 2012, Hefei, China. This paper was recommended for publication in revised form by Associate Editor Hyeong Soo Chang under the direction of Editor Ian R. Petersen.

[*] Corresponding author.
E-mail addresses: autolyj@hit.edu.cn (Y. Li), xy.wu@siat.ac.cn (X. Wu), louyj@hit.edu.cn (Y. Lou), hychen5@hit.edu.cn (H. Chen), jiangang_lee@hit.edu.cn (J. Li).

https://doi.org/10.1016/j.automatica.2018.03.011
0005-1098/© 2018 Elsevier Ltd. All rights reserved.

The perturbation realization factor (PRF) proposed in Cao (2007) and Cao and Chen (1997) provides a relative value of performance potentials and can be used to estimate performance potential based on a single sample path (Cao, 2007) and to optimize the performance of queueing systems (Xia & Cao, 2012). In this study, we apply PRF to implement EGVR for average reward performance potential. Coupling methods have been widely used in statistics (Griffeath, 1975; Lindvall, 1992) and simulations. We do not intend to provide a comprehensive survey of coupling methods, and instead, we limit our review to studies that are most relevant to our problem. As a coupling method, common random number (CRN) (Glynn, 1985; Pflug, 1996) has been widely used in simulations to reduce variance in the estimation of the mean difference of two different random variables. In Cao (2007), the application of coupling methods in average reward performance potential was presented from the perspective of numerical solutions. This study found that coupling methods cannot improve the convergence rate of the numerical algorithms. Four coupling methods were presented in perturbation analysis to estimate the performance derivative in Dai (2000). This work pointed out that finding a better coupling scheme is desirable. The maximal coupling in Griffeath (1975) is maximal in the sense that merging is attained as efficiently as possible but leads to non-Markov coupling, and thus, the linear equation required by EGVR cannot be obtained. The basic idea regarding the use of coupling to estimate performance potential was briefly introduced in Li (2012) but details were not provided.

In this study, we present novel approaches for the EGVR of average reward performance potential based on the sample paths of Markov chains, which avoid the estimations of additional variables, such as average reward and steady state probability distribution, and thus implement EGVR in the true sense. On the basis of estimation algorithms, we introduce the correlation of the simulations of Markov chains by using coupling methods. Various coupling methods can be used to further reduce estimation variance or simulation time. Although certain coupling methods require information of transition probabilities, these methods can be applied to states where transition probabilities are known, to introduce the correlation of simulations or to combine them with physical simulations. In particular, we propose two optimal coupling methods that can minimize simulation time and estimation variance, respectively. To our knowledge, the simulation of Markov chains is frequently time-consuming and costly. The reduction in simulation time may significantly reduce estimation time. The coupling-based estimation methods used in this study can reduce estimation variance and simulation time and improve simulation efficiency.

2. Performance potential and its estimations

Consider an ergodic (irreducible, aperiodic, and positive recurrent) Markov chain $\boldsymbol{X} = \{X_l, l = 0, 1, \ldots\}$ on a finite state space $S = \{1, 2, \ldots, M\}$ with a transition probability matrix $P = [p(i,j)]_{i,j=1}^{M}$. Let $\pi = (\pi(1), \pi(2), \ldots, \pi(M))$ be a row vector that represents its steady state probability. Then, we derive the following balance equations:

$$\pi P = \pi, \pi e_M = 1, \tag{1}$$

where $e_M = (1, 1, \ldots, 1)^T$ is an M-dimensional column vector whose components are all equal to 1, and superscript "T" denotes the transpose. Let $f : S \to R$ be a reward function and occasionally a (column) reward vector, i.e., $f = (f(1), f(2), \ldots, f(M))^T$. We consider the long-run average reward (or simply the average reward) as a performance measure, which is defined as

$$\eta = \lim_{L\to\infty} \frac{1}{L} E\left\{\sum_{l=0}^{L-1} f(X_l)\right\} = \sum_{i=1}^{M} \pi(i)f(i) = \pi f.$$

For the aforementioned Markov chain, the following Poisson equation holds (Bertsekas, 1995; Cao, 2007; Puterman, 1994):

$$(I_M - P)g + \eta e_M = f, \tag{2}$$

where I_M denotes an $M \times M$ identity matrix. Its solution $g = (g(1), g(2), \ldots, g(M))^T$ is called average reward performance potential (Cao, 2007) (it is equivalent to the "relative cost function" (Bertsekas, 1995), "bias" (Puterman, 1994) or "value function" (Munos, 2006)). The solution to (2) can be obtained only up to an additive constant, i.e., if g is a solution to (2), then so is $g + ce_M$, where c is a constant. It can be proven that

$$g(i) = E\left\{\sum_{l=0}^{\infty}[f(X_l) - \eta]\middle| X_0 = i\right\}, i \in S \tag{3}$$

is a specific average-reward performance potential since

$$\begin{aligned} g(i) &= E\left\{\sum_{l=0}^{\infty}[f(X_l) - \eta]\middle| X_0 = i\right\} \\ &= f(i) - \eta + \sum_{j\in S} p(i,j)E\left\{\sum_{l=1}^{\infty}[f(X_l) - \eta]\middle| X_1 = j\right\} \\ &= f(i) - \eta + \sum_{j\in S} p(i,j)g(j), \end{aligned}$$

whose matrix form is the same as Eq. (2). To clearly describe EGVR, we also introduce the discounted reward performance potential, which is defined as follows:

$$g_\alpha(i) = E\left\{\sum_{l=0}^{\infty}\alpha^l f(X_l)\middle| X_0 = i\right\}, \tag{4}$$

where $0 < \alpha < 1$ denotes a discount factor. Similarly, the discounted reward performance potential satisfies the following Poisson equation (Puterman, 1994):

$$(I_M - \alpha P)g_\alpha = f, \tag{5}$$

where $g_\alpha = (g_\alpha(1), g_\alpha(2), \ldots, g_\alpha(M))^T$.

When the transition probabilities in P are known, performance potential g may be obtained by solving linear equation (2) or using its value iteration. However, certain transition probabilities in P may be generally unknown. In such case, the linear equation or its value iteration methods will not works. Estimation approaches can be applied for such cases. Monte Carlo algorithms are generally used to estimate performance potential and implement simulation-based optimization (Cooper et al., 2003; Marbach & Tsitsiklis, 2001). Assume that N samples of performance potential are obtained via simulation. Then, we have N estimates $\hat{g}^n(i)$, $n = 1, 2, \ldots, N$ of performance potential $g(i)$, $i \in S$. The final estimate is obtained by $\sum_{n=1}^{N}\hat{g}^n(i)/N$. The variance of this estimate is $\frac{1}{N^2}\sum_{n=1}^{N} var\hat{g}^n(i)$, which is reduced with order $1/N$. Compared with Monte Carlo algorithms, the EGVR has a geometric variance reduction rate. The related results about EGVR (Munos, 2006) are reviewed in the succeeding paragraphs.

Consider a function vector $g = (g(1), \ldots, g(M))^T$, where $g(i) = E[\Psi_i(f)]$ and $\Psi_i(f)$ is a linear sample path functional of reward function f that depends on the sample path of a Markov chain with initial state i. For example, discounted reward performance potential (4) can be described as $g_\alpha(i) = E[\Psi_i(f)]$ and

$$\Psi_i(f) = \sum_{l=0}^{\infty} \alpha^l f(X_l),$$