http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
*Articles in Advance*, pp. 1–3
ISSN 1946-5238 (online)

Special Section: Open Problems in Applied Probability

# Open Problem—Convergence and Asymptotic Optimality of the Relative Value Iteration in Ergodic Control

**Ari Arapostathis**[a]

[a] Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712
**Contact:** ari@utexas.edu, https://orcid.org/0000-0003-2207-357X

**Published Online in Articles in Advance:** September 17, 2019

https://doi.org/10.1287/stsy.2019.0040



**History:** This paper was accepted for the *Stochastic Systems* Special Section on Open Problems in Applied Probability, presented at the 2018 INFORMS Annual Meeting in Phoenix, Arizona, November 4–7, 2018.



**Funding:** The work of A. Arapostathis was supported in part by the National Science Foundation through Grant DMS-1715210 and in part by the Army Research Office through Grant W911NF-17-1-001.

**Keywords:** stochastic networks • control • optimization

The relative value iteration scheme (RVI) for Markov decision processes (MDP) dates back to White (1963), a seminal work, which introduced an algorithm for solving the ergodic dynamic programming equation for the finite state, finite action case. Its ramifications have given rise to popular learning algorithms ($Q$-learning). More recently, this algorithm gained prominence because of its implications for model predictive control (MPC). For stochastic control problems on an infinite time horizon, especially for problems that seek to optimize the average performance (ergodic control), obtaining the optimal policy in explicit form is only possible for a few classes of well-structured models. What is often used in practice is a heuristic method called the rolling horizon, or receding horizon, or MPC. This works as follows: one solves the finite horizon problem for a given number of steps $N$, or for an interval $[0, T]$ in the case of a continuous time problem. The result is a nonstationary Markov policy, which is optimal for the finite horizon problem. We fix the initial action (this is the action determined at the $N$th step of the *value iteration* (VI) algorithm) and apply it as a stationary Markov control. We refer to this Markov control as the rolling horizon control. This of course depends on the length of the horizon $N$. One expects that for well-structured problems, if $N$ is sufficiently large, then the rolling horizon control is near optimal. Of course, this is a heuristic. The rolling horizon control might not even be stable. For a good discussion on this problem, we refer the reader to Della Vecchia et al. (2012). Obtaining such solutions is further complicated by the fact that the value of the ergodic cost required in the successive iteration scheme is not known. This is the reason for the RVI.

Naturally, for the rolling horizon control to have the desirable properties, the value function, suitably normalized, must converge. For nonfinite state space models, the literature contains several results on convergence for problems enjoying uniform stability properties (Montes-de Oca and Hernández-Lerma 1996, Cavazos-Cadena 1998, Hernández-Lerma and Lasserre 1990), even in fairly abstract Borel models (Yu 2015), and more recently in continuous time, even for stochastic differential games (Arapostathis et al. 2013) and mean-field games (Arapostathis et al. 2017). But the real challenge is the study of models that do not exhibit uniform ergodicity under all Markov controls, but are only stabilizable. This is of course the paradigm of linear systems with a quadratic running penalty, which is very well understood. For nonlinear problems, there are few results. For discrete time problems, we refer the reader to Cavazos-Cadena (1996), Chen and Meyn (1999), and Arapostathis and Borkar (2019), and for continuous time, to Arapostathis et al. (2014). There has also been some work on models where the drift of the equation is the control, and the objective is to minimize a running penalty, which is the sum of a potential function and the control effort (Ichihara 2012).

The work cited above concerns the convergence of the VI or RVI. Stability and asymptotic optimality of the rolling horizon control is a much harder problem, and results for nonlinear models are scarce. Let us review such a result for an MDP on a countable state space. Consider a model with state space $\mathcal{S} := \{0, 1, 2, \ldots\}$, action

space $\mathbb{U}$, a compact metric space, transition probability kernel $P_u(i,j)$, with $i,j \in \mathcal{S}$ and $u \in \mathbb{U}$, and running cost $r(i,u)$, which is assumed nonnegative. Denote the state process as $X_n$, $n = 0, 1, \ldots$. The optimization criterion is the infinite horizon average cost (or ergodic cost)

$$J(\bar{v}) := \limsup_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}^{\bar{v}}[r(X_k, v_k(X_k))],$$

where $\bar{v} = (v_0, v_1, \ldots)$ denotes a Markov strategy. Let $\varrho$ denote the infimum of $J(\bar{v})$ over all Markov strategies. We assume that $u \mapsto P_u(i,j)$ and $u \mapsto r(i,u)$ are continuous, that the process is irreducible and aperiodic, and that $\liminf_{i\to\infty} \min_{u\in\mathbb{U}} r(i,u) = \infty$. The last hypothesis is a structural assumption on the running cost and has the effect of discouraging unstable behavior. As a result, any stationary Markov control that results in a finite ergodic cost renders the process Harris recurrent. Recall that the Bellman equation, or average cost optimality equation for this problem takes the form $\min_{u\in\mathbb{U}} [P_u V(i) + r(i,u)] = \varrho + V(i)$.

The VI is given by

$$\varphi_{n+1}(i) = \min_{u\in\mathbb{U}} \big[P_u \varphi_n(i) + r(i,u)\big] - \varrho\,, \tag{1}$$

with $\varphi_0$ a given function, which we set here equal to zero for simplicity. Since $\varrho$ is not known, we replace it with $\varphi_n(0)$ in the above equation to obtain the RVI. Consider the following structural hypothesis: There exist positive constants $\theta_1$ and $\theta_2$ such that

$$\min_{u\in\mathbb{U}} r(i,u) \geq \theta_1 V(i) - \theta_2 \quad \forall\, i \in \mathcal{S}. \tag{2}$$

Equation (2) implies that under an optimal Markov control, the resulting process is geometrically ergodic. It can be viewed as a hypothesis on the rate of "stabilizability" of the process. Note that if (2) is verified for some Markov control, with $V$ replaced by the solution of a Lyapunov equation corresponding to that specific control, it also holds for the solution of the Bellman equation. For an example of a class of problems that satisfy this hypothesis in a generic manner, see Arapostathis and Borkar (2019). As shown in Arapostathis and Borkar (2019), which addresses MDPs living in a Euclidean space, this hypothesis implies that

$$(1-\beta^n)\Big(V(i) - \tfrac{\varrho+\theta_2}{\theta_1}\Big) \leq \varphi_n(i) \leq V(i) \quad \text{for } n \in \mathbb{N}. \tag{3}$$

Thus, the VI converges to a neighborhood of the solution of the Bellman equation at a geometric rate, and using a separate argument, we can show that it indeed converges to the function $V$ up to an additive constant. The same applies to the RVI. Moreover, one can show that the rolling horizon policy $\mathfrak{v}_n$ is "stabilizing" for all $n$ such that $(1+\theta_1)(1-\beta^{n-1}) > 1$, and the process is geometrically ergodic under any such $\mathfrak{v}_n$. Here, $\mathfrak{v}_n$ is a selector from the minimizer of (1), and it is important to note that the VI and the RVI have the same set of minimizers. In addition, we can derive bounds on the performance of $\mathfrak{v}_n$, or, in other words, for $J(\mathfrak{v}_n)$, and show that this quantity converges to $\varrho$. Analogous results can be obtained for continuous Markov processes (see theorem 3.20 in Arapostathis et al. 2014).

The result above shows that the convergence of the RVI may depend on the rate of ergodicity of the process under the optimal control, or, more generally, under a stabilizing control. This of course was demonstrated in the analysis above only for a system that exhibits a geometric rate of convergence to the stationary distribution in total variation, and for which the running cost serves as the storage function of a solution to an associated Lyapunov equation. The rate of convergence to the invariant distribution for Markov processes has been a topic of intense study lately (see for example Douc et al. 2009 and Hairer 2016), and various results have been obtained via the Foster–Lyapunov theory, or coupling techniques. This might be a promising direction for the study of the problem that we described above and can be summarized as follows:

(A) Determine suitable structural properties for the system model that result in the rolling horizon control to become stabilizing after running the RVI a finite number of steps. Also determine the rate of convergence.

(B) Determine conditions for the rolling horizon control to be asymptotically optimal and bounds on its performance.

## References

Arapostathis A, Borkar VS (2019) Average cost optimal control under weak hypotheses: Relative value iterations. *arXiv*:1902.01048.

Arapostathis A, Biswas A, Carroll J (2017) On solutions of mean field games with ergodic cost. *J. Math. Pures Appl. (9)* 107(2):205–251.

Arapostathis A, Borkar VS, Kumar KS (2013) Relative value iteration for stochastic differential games. Křivan V, Zaccour G, eds. *Advances in Dynamic Games*, Annals of the International Society of Dynamic Games, vol. 13 (Birkhäuser/Springer, Cham), 3–27.

Arapostathis A, Borkar VS, Kumar KS (2014) Convergence of the relative value iteration for the ergodic control problem of nondegenerate diffusions under near-monotone costs. *SIAM J. Control Optim.* 52(1):1–31.

Cavazos-Cadena R (1996) Value iteration in a class of communicating Markov decision chains with the average cost criterion. *SIAM J. Control Optim.* 34(6):1848–1873.

Cavazos-Cadena R (1998) A note on the convergence rate of the value iteration scheme in controlled Markov chains. *Systems Control Lett.* 33(4): 221–230.

Chen RR, Meyn S (1999) Value iteration and optimization of multiclass queueing networks. *Queueing Systems Theory Appl.* 32(1-3):65–97.

Della Vecchia E, Di Marco S, Jean-Marie A (2012) Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs. *Ann. Oper. Res.* 199:193–214.

Douc R, Fort G, Guillin A (2009) Subgeometric rates of convergence of f-ergodic strong Markov processes. *Stochastic Process. Appl.* 119(3): 897–923.

Hairer M (2016) Convergence of Markov processes. Lecture notes, University of Warwick. Accessed September 1, 2019, http://www.hairer.org/notes/Convergence.pdf.

Hernández-Lerma O, Lasserre JB (1990) Error bounds for rolling horizon policies in discrete-time Markov control processes. *IEEE Trans. Automat. Control* 35(10):1118–1124.

Ichihara N (2012) Large time asymptotic problems for optimal stochastic control with superlinear cost. *Stochastic Process. Appl.* 122(4):1248–1275.

Montes-de Oca R, Hernández-Lerma O (1996) Value iteration in average cost Markov control processes on Borel spaces. *Acta Appl. Math.* 42(2): 203–222.

White DJ (1963) Dynamic programming, Markov chains, and the method of successive approximations. *J. Math. Anal. Appl.* 6:373–376.

Yu H (2015) On convergence of value iteration for a class of total cost Markov decision processes. *SIAM* J. *Control Optim.* 53(4):1982–2016.