THE INSTITUTE FOR SYSTEMS RESEARCH

ISR TECHNICAL REPORT 2009-14

A Short Note on Combining Multiple Policies in Risk-Sensitive Exponential Average Reward Markov Decision Processes

Hyeong Soo Chang

A Short Note on Combining Multiple Policies in Risk-Sensitive Exponential Average Reward Markov Decision Processes

Hyeong Soo Chang*

Abstract

This short note presents a method of combining multiple policies in a given policy set such that the resulting policy improves all policies in the set for risk-sensitive exponential average reward Markov decision processes (MDPs), extending the work of Howard and Matheson for the singleton policy set case. Some applications of the method in solving risk-sensitive MDPs are also discussed.

Keywords: Risk-sensitive Markov decision process, policy improvement, policy iteration, controlled Markov chain

1 Introduction

Consider a risk-sensitive Markov decision process [3] (MDP) $M = (X, A, P, R)$ with a finite state set $X = \{1, 2, ..., N\}$, finite admissible action sets $A(x), x \in X$, a reward function $R : K \to \mathbb{R}$, and a transition function P that maps K to the set of probability distributions over X, where $K = \{(x, a)|x \in X, a \in A(x)\}$. We denote the probability of making a transition to state $y \in X$ when taking action $a \in A(x)$ at state $x \in X$ by p^a_{xy} and let q^a_{xy} be the "disutility contribution" of the transition from state x to state y by taking action $a \in A(x)$. The disutility contribution is given such that

$$q^a_{xy} := p^a_{xy} e^{-\gamma R(x,a)}, x, y \in X, a \in A(x).$$

*H. S. Chang is with the Department of Computer Science and Engineering at Sogang University, Seoul, Korea, and can be reached by e-mail at hschang@sogang.ac.kr, by phone +82-2-705-8925, or by fax +82-2-704-8273. This work was done while he was a visiting associate professor at ISR, University of Maryland, College Park.

The constant $\gamma \in \mathbb{R}$ is called the risk-sensitivity coefficient.

Let Π be the set of all Markovian (history-independent) stationary deterministic policies $\pi : X \to A$ with $\pi(x) \in A(x), x \in X$. We define *the certain equivalent gain* of a policy $\pi \in \Pi$ over a finite horizon $H < \infty$ with an initial state $x \in X$ as

$$J_H^\pi(x) = -\frac{1}{\gamma} \ln \left\{ E \left[\exp \left(-\gamma \sum_{t=0}^{H-1} R(x_t, \pi(x_t)) \right) \middle| x_0 = x \right] \right\},$$

where x_t denotes the random variable representing the state at time t by following π.

We denote the transition probability matrix with the xyth entry $p_{xy}^{\pi(x)}, x, y \in X$ as P^π and the disutility contribution matrix with the xyth entry $q_{xy}^{\pi(x)}, x, y \in X$ as Q^π. Throughout the present note, we assume that the risk-sensitivity coefficient γ is fixed with $\gamma > 0$. (We consider the risk-aversion case only—the risk-seeking case, $\gamma < 0$, can be treated with similar arguments.) Furthermore, we make the following assumption:

Assumption 1.1 *For any policy $\pi \in \Pi$, the probability transition matrix P^π is primitive.*

For an irreducible nonnegative matrix C, the matrix is called primitive if some power of C has all positive elements.

Howard and Matheson [3] showed that under Assumption 1.1, the sequence $J_H^\pi(x)/H$ converges independently of the initial state x, i.e.,

$$\lim_{H \to \infty} \frac{J_H^\pi(x)}{H} = -\frac{1}{\gamma} \ln \lambda^\pi =: \tilde{g}^\pi,$$

where λ^π is the dominant or “maximal” *positive* eigenvalue of the matrix Q^π. An irreducible nonnegative matrix C always has a positive eigenvalue λ and the moduli of all the other eigenvalues are less than λ. We use the notation $\tilde{g}^\pi$ (following [3]) for the certain equivalent gain of π (over infinite horizon), omitting x.

They also showed that if u^π is the unique eigenvector of Q^π with $u^\pi(N) = -(\operatorname{sgn} \gamma)$ corresponding to the maximal eigenvalue λ^π with the xth entry as $u^\pi(x), x \in X$, then the following is satisfied:

$$\lambda^\pi u^\pi(x) = \sum_{y \in X} q_{xy}^{\pi(x)} u^\pi(y), x \in X. \tag{1}$$

We refer to obtaining such u^π for a given π by solving (1) as *policy evaluation*.

This note's goal is to generalize the result of Howard and Matheson's single-policy improvement method. We provides a method of designing a policy π^m that improves all policies in a given nonempty subset $\Delta \subseteq \Pi$ by combining only the policies in Δ such that

$$\tilde{g}^{\pi^m} \geq \max_{\pi \in \Delta} \tilde{g}^{\pi}.$$

Howard and Matheson considered the case of $|\Delta| = 1$ and showed that the *policy iteration* (PI) algorithm that iteratively applies the policy evaluation for the current policy and the single-policy improvement method converges in a finite number of iterations to an optimal policy in Π that achieves $\max_{\pi \in \Pi} \tilde{g}^{\pi}$.

The present work is mainly motivated by problems for which we already have some heuristic policies available. For example, for the multiclass-scheduling problem with stochastically arriving prioritized tasks with deadlines [1], the "earliest-deadline-first" and "static-priority" heuristics are available candidate policies in hand for scheduling. It may even be the case that our heuristic policies are such that each policy is near-optimal over some part of the state space. If so, the decision maker may well wish to combine these policies into a single policy that somehow improves all of the heuristic policies. The result presented in this note is directly relevant to this goal.

2 Multi-Policy Improvement Method

Theorem 2.1 *Assume that Assumption 1.1 holds. For a given nonempty subset $\Delta \subseteq \Pi$, assume that u^π with $u^\pi(N) = -(\operatorname{sgn} \gamma)$ is obtained by policy evaluation in (1) for all $\pi \in \Delta$. Define a policy $\pi^m \in \Pi$ as*

$$\pi^m(x) \in \arg\max_{a \in A(x)} \left\{ \sum_{y \in X} q^a_{xy} \left(\max_{\pi \in \Delta} u^\pi(y) \right) \right\}, x \in X. \tag{2}$$

Then we have that

$$\max_{\pi \in \Delta} \tilde{g}^{\pi} \leq \tilde{g}^{\pi^m}.$$

Proof: Let $\Phi(x) = \max_{\pi \in \Delta} u^\pi(x), \forall x \in X$. We have that for any $\pi \in \Delta$ and $x \in X$,

$$\begin{aligned} \lambda^\pi u^\pi(x) &= \sum_{y \in X} q^{\pi(x)}_{xy} u^\pi(y) \text{ by (1) from the policy evaluation assumption} && (3) \\ &\leq \sum_{y \in X} q^{\pi(x)}_{xy} \max_{\pi \in \Delta} u^\pi(y) && (4) \\ &\leq \sum_{y \in X} q^{\pi^m(x)}_{xy} \max_{\pi \in \Delta} u^\pi(y) \text{ by the definition of } \pi^m. && (5) \end{aligned}$$

Therefore, for any $x \in X$,

$$\max_{\pi \in \Delta}\{\lambda^{\pi} u^{\pi}(x)\} \leq \sum_{y \in X} q_{xy}^{\pi^m(x)} \Phi(y).$$

Furthermore, because for any $\pi \in \Delta$, $\lambda^{\pi} > 0$ from Assumption 1.1 and $u^{\pi}(x) < 0, \forall x \in X$, we have that

$$\max_{\pi \in \Delta}\{\lambda^{\pi} u^{\pi}(x)\} \geq \max_{\pi \in \Delta}\{\lambda^{\pi}\} \max_{\pi \in \Delta} u^{\pi}(x) \geq \min_{\pi \in \Delta}\{\lambda^{\pi}\} \max_{\pi \in \Delta} u^{\pi}(x), \forall x \in X.$$

It follows that

$$\min_{\pi \in \Delta} \lambda^{\pi} \Phi(x) \leq \sum_{y \in X} q_{xy}^{\pi^m(x)} \Phi(y), \forall x \in X. \tag{6}$$

Now we consider the following two cases: Denote the maximal eigenvalue of Q^{π^m} as λ^{π^m}. From Assumption 1.1, such $\lambda^{\pi^m} > 0$ exists.

Case 1: Φ is an eigenvector of Q^{π^m} for λ^{π^m}. Then for all $x \in X$,

$$\sum_{y \in X} q_{xy}^{\pi^m(x)} \Phi(y) = \lambda^{\pi^m} \Phi(x) \geq \min_{\pi \in \Delta} \lambda^{\pi} \Phi(x),$$

where the last inequality comes from (6). Because $\Phi(x) < 0, \forall x \in X$, we have that $\min_{\pi \in \Delta} \lambda^{\pi} \geq \lambda^{\pi^m}$.

Case 2: Φ is not an eigenvector of Q^{π^m} for λ^{π^m}. Then because Q^{π^m} is an irreducible matrix (from Assumption 1.1) with nonnegative entries and $|\Phi(x)| > 0, \forall x \in X$ with not being an eigenvector of Q^{π^m} for λ^{π^m}, we have that (cf. A3 in Appendix A in [3])

$$\lambda^{\pi^m} < \max_{x \in X} \frac{\sum_{y \in X} q_{xy}^{\pi^m} |\Phi(y)|}{|\Phi(x)|}. \tag{7}$$

Furthermore, from $\Phi(x) < 0, \forall x \in X$ and (6), we have that

$$\max_{x \in X} \frac{\sum_{y \in X} q_{xy}^{\pi^m} |\Phi(y)|}{|\Phi(x)|} \leq \min_{\pi \in \Delta} \lambda^{\pi}. \tag{8}$$

Combining (7) and (8) implies $\min_{\pi \in \Delta} \lambda^{\pi} \geq \lambda^{\pi^m}$.

Therefore, by using the fact $\tilde{g}^{\pi'} = -\frac{1}{\gamma} \ln \lambda^{\pi'}, \pi' \in \Pi$, we finally have that

$$\max_{\pi \in \Delta} \tilde{g}^{\pi} \leq \tilde{g}^{\pi^m}$$

∎

3 Some Remarks

First, the multi-policy improvement method here is parallel to "parallel rollout" [1] [2] for risk-neutral MDPs. As such, the result in Theorem 2.1 is also useful for a *model-based* on-line control of risk-sensitive MDPs via simulation.

Second, the method can be used for designing a tractable PI-type algorithm for solving risk-sensitive MDPs with *large action spaces*, mitigating the curse of dimensionality as in Hu *et al.*'s work [4] for risk-neutral MDPs. Multi-policy improvement was exploited in designing the ERPS (Evolutionary Random Policy Search) algorithm in [4] (also for continuous action space) within the framework of evolutionary computational algorithm. In particular, it has been shown empirically that ERPS speeds up PI by an order of magnitude for a risk-neutral MDP with a large action space queueing problem. A similar idea can be applied to the risk-sensitive case here by using the result in Theorem 2.1.

References

[1] H. S. Chang, R. Givan, and E. K. P. Chong, "Parallel rollout for on-line solution of partially observable Markov decision processes," *Discrete Event Dynamic Systems: Theory and Application*, vol. 14, no. 3, 2004, pp. 309–341.

[2] H. S. Chang, M. C. Fu, J. Hu, and S. I. Marcus, *Simulation-Based Algorithms for Markov Decision Processes*. Springer, London, 2007.

[3] R. Howard and J. E. Matheson, "Risk-sensitive Markov decision processes," Management Science, vol. 18, no. 7, 1072, pp. 356–369.

[4] J. Hu, M.C. Fu, V. Ramezani, and S.I. Marcus, "An evolutionary random policy search algorithm for solving Markov decision processes," *INFORMS Journal on Computing*, vol. 19, no. 2, pp. 161–174, 2007.