Random-Sampling Monte-Carlo Tree Search Methods for Cost Approximation in Long-Horizon Optimal Control

Shankarachary Ragi, *IEEE Senior Member*, and Hans D. Mittelmann

***Abstract*—In this paper, we develop a Monte-Carlo based heuristic approach to approximate the objective function in long horizon optimal control problems. In this approach, we evolve the system state over multiple trajectories into the future while sampling the noise disturbances at each time-step, and find the weighted average of the costs along all the trajectories. We call these methods *random sampling - multipath hypothesis propagation* or RS-MHP. These methods (or variants) exist in the literature; however, the literature lacks convergence results for a generic class of nonlinear systems. This paper fills this knowledge gap to a certain extent. We derive convergence results for the cost approximation error from the MHP methods and discuss their convergence (in probability) as the sample size increases. As a case study, we apply RS-MHP to approximate the cost function in a linear quadratic control problem and demonstrate the benefits of our approach against an existing and closely related approximation approach called *nominal belief-state optimization*.**

***Index Terms*—Long horizon optimal control, cost approximation, approximate dynamic programming, multipath hypothesis propagation.**

I. Introduction

Long-horizon optimal control problems appear naturally in robotics, advanced manufacturing, and economics, especially in applications requiring decision making in stochastic environments. Often these problems are solved via dynamic programming (DP) formulation [1]. DP problems are notorious for their computational complexity, and require approximation approaches to make them tractable. A plethora of approximation techniques called *approximate dynamic programs* (ADPs) exist in the literature to solve these problems approximately. Some of the commonly used ADPs include *policy rollout* [2], *hindsight optimization* [3], [4], etc. A survey of the ADP approaches can found in [1]. Feature-based techniques and deep learning methods are gaining importance in the development of ADP approaches as discussed in [5]. These approximation techniques have been successfully adopted to solve real-time problems such as a UAV guidance control problem in [6]–[8]. Certain ADP approaches, especially the methods based on approximation in value space, require numerical approximation of the expectation in the objective function [6]. In this study, our objective is to develop Monte-Carlo-based approaches to approximate the expectation in the objective function in the long (but finite) horizon optimal control problems, and study their convergence.

This work was supported in part by Air Force Office of Scientific Research under grant FA9550-19-1-0070.

Shankarachary Ragi is with Department of Electrical Engineering, South Dakota School of Mines and Technology, Rapid City, SD 57701, USA shankarachary.ragi@sdsmt.edu

Hans D. Mittelmann is with the School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85281, USA mittelmann@asu.edu

A. Preliminaries

A long horizon optimal control problem is described as follows. Let x_k be the state vector for a system at time k, which evolves according to a discrete stochastic process as follows:

$$x_{k+1} = f(x_k, u_k, w_k) \tag{1}$$

where $f(\cdot)$ represents the state-transition mapping, u_k is the control vector, and w_k random disturbance. Let $g(x_k, u_k)$ represent the cost (a real value) of being in state x_k and performing action u_k. The functions f and g are independent of k in our study, but can generally depend on k. The goal is to optimize the control vectors $u_k, k = 0, \ldots, H-1$ such that the expected cumulative cost is minimized, i.e., the goal leads to solving the following optimization problem

$$\min_{u_k, k=0,\ldots,H-1} \mathrm{E}\left[\sum_{k=0}^{H-1} g(x_k, u_k)\right], \tag{2}$$

where H is the length of the planning horizon. Let x_0 be the initial state and according to the dynamic programming formulation the optimal cost function is given by

$$J_0^*(x_0) = \min_{u_0} \mathrm{E}\left[g(x_0, u_0) + J_1^*(x_1)\right], \tag{3}$$

where J_1^* represents the optimal cost-to-go from time $k = 1$, and $x_1 = f(x_0, u_0, w_0)$. In this study, *long horizon* refers to the condition that H is sufficiently large that the optimal policy is approximately *stationary* (independent of k). Solving the above optimization problem is not tractable mainly due to two reasons: the expectation $E[\cdot]$ and the optimal cost-to-go J_1^* are hard to evaluate, which are usually approximated by numerical methods or ADP approaches.

An ADP approach called *nominal belief-state optimization* (NBO) [6], [9] was developed primarily to approximate the above expectation. In NBO, the expectation is replaced by a sample state trajectory generated

with an assumption that the future noise variables in the system take so called nominal or mean values, thus making the above objective function deterministic. The NBO method was developed to solve a UAV path optimization problem, which was posed as a *partially observable Markov decision process* (POMDP). POMDP generalizes the long horizon optimal control problem described in Eq. 2 in that the system state is assumed to be "partially" observable, which is inferred via using noisy observations and Bayes rules. Although the performance of the NBO approach was satisfactory, in that it allowed to obtain reasonably optimal control commands for the UAVs, it ignored the uncertainty due to noise disturbances thus leading to inaccurate evaluation of the objective function. To address this challenge, several methods exist in the literature usually referred to as Monte-Carlo Tree Search (MCTS) methods as surveyed in [10].

Inspired from the NBO method and MCTS methods, we develop a new MCTS method called *random sampling - multipath hypothesis propagation* (RS-MHP) and derive convergence results. In this study, we mainly use the NBO approach as a benchmark for performance assessment since RS-MHP builds on the NBO approach.

II. Random Sampling Multipath Hypothesis Propagation (RS-MHP)

In the NBO method, the expectation is replaced by a sample trajectory of the states (as opposed to random states) generated by

$$\tilde{x}_{k+1} = f(\tilde{x}_k, u_k, \bar{w}_k),\ k = 0, \ldots \tag{4}$$

where $\tilde{x}_0 = x_0$ (initial state or current state), and $\bar{w}_k$ is the mean of the random variable w_k. Thus, the long horizon optimal control problem, with NBO approximation, reduces to

$$\min_{u_k} \sum_{k=0}^{H-1} g(\tilde{x}_k, u_k). \tag{5}$$

The above reduced problem, without the need for evaluating the expectation, can significantly reduce the computational burden in solving the long horizon control problems. However, the downside with this approach is it completely ignores the uncertainty in the state evolution, and may generate severely sub-optimal controls. To overcome this trivialization, we develop a Monte-Carlo approach to approximate the expectation described as follows. For time step $k = 1$, we sample the probability distribution of the noise disturbance N times to generate the samples w_0^i with corresponding probability p_0^i, $i = 1, \ldots, N$. Using these, we generate N sample states at $k = 1$ generated according to

$$x_1^i = f(x_0, u_0, w_0^i),\ \forall i. \tag{6}$$

We repeat this sampling approach for time $k = 2$, i.e., we generate N noise samples w_1^i with corresponding probability p_1^i, $i = 1, \ldots, N$. Using these noise samples and the sample states from the previous time step, we generate N^2 sample states at $k = 2$ according to

$$x_2^{i,j} = f(x_1^i, u_1, w_1^j),\ \forall i, j. \tag{7}$$

We repeat the above sampling procedure until the last time step $k = H - 1$ to generate N^{H-1} possible state evolution trajectories using N noise samples generated in each time step.

One can now replace the expectation in Eq. 2 with the weighted average of the cumulative cost corresponding to each state evolution trajectory, where the weights are the probabilities or likeliness of the trajectories. Clearly, the number of possible state trajectories grow exponentially with the horizon length H. Although this approach is not novel as many such methods exist in the literature often classified as Monte-Carlo Tree Search methods, our study is focused on deriving convergence results of RS-MHP approaches.

To avoid the exponential growth in our RS-MHP approach, at each time step we retain only M sample states and prune the remaining states, and if the number of sample states at a given time instance is less than or equal to M, we do not perform pruning. For pruning, at each time k, we rank the state trajectories up to time k according to their likeliness (obtained by multiplying the probabilities of all the noise samples that generated the trajectory) and retain the top M trajectories with highest likeliness and prune the rest. With this procedure, at $k = H - 1$, there would be only M state trajectories. With pruning, the number of trajectories remains a constant irrespective of the time horizon length. An illustration of the above RS-MHP approach is shown in Figure 1 along with the NBO approach. Here, we consider pruning based on likeliness of the state trajectories as the costs from these trajectories have higher contribution in the cost function in Eq. 1 than the less likely trajectories. We will consider other pruning strategies to further improve the approximation error in our future study.

Let $i = 1, \ldots, M$ represent the indices of the M distinct state trajectories with $q_1, q_2, \ldots$ being their probabilities (likeliness). The probability q_i is evaluated by simply multiplying the probabilities of the noise samples that generate the trajectory i over time. These probabilities are normalized, i.e., $\sum_{i=1}^{M} q_i = 1$. Let J represent the actual objective function as described below

$$J = \mathrm{E}\left[\sum_{k=0}^{H-1} g(x_k, u_k)\right]. \tag{8}$$

We can now approximate the objective function J in four possible ways as described below (assuming $N >$

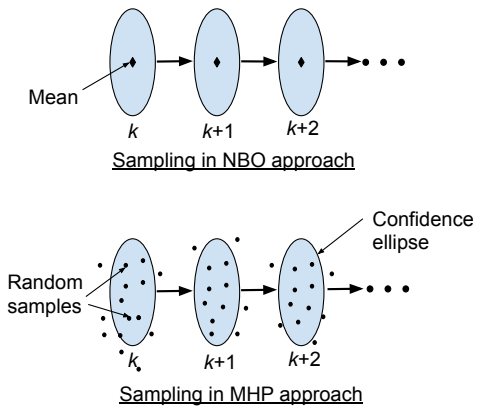

Fig. 1. Sampling probability distributions of noise variables: NBO vs. MHP.

M). Let x_k^i represent the state at time k in the ith state trajectory.

(I) *Sample Averaging*. We can simply approximate the expectation with an average over all possible trajectories as follows:

$$\text{No pruning: } J \approx \tilde{J}_{NP} = \frac{1}{N^{H-1}} \sum_{i=1}^{N^{H-1}} \left(\sum_{k=0}^{H-1} g(x_k^i, u_k) \right)$$
$$\text{With pruning: } J \approx \tilde{J}_P = \frac{1}{M} \sum_{i=1}^{M} \left(\sum_{k=0}^{H-1} g(x_k^i, u_k) \right) \tag{9}$$

(II) *Weighted Sample Averaging*. We can also approximate the expectation with a weighted average with weights being the normalized likeliness indices of the state trajectories given by $q_i, i = 1, \ldots$ (and $\bar{q}_i$ in the pruned case) as follows:

$$\text{No pruning: } J \approx \bar{J}_{NP} = \sum_{i=1}^{N^{H-1}} q_i \left(\sum_{k=0}^{H-1} g(x_k^i, u_k) \right)$$
$$\text{With pruning: } J \approx \bar{J}_P = \sum_{i=1}^{M} \bar{q}_i \left(\sum_{k=0}^{H-1} g(x_k^i, u_k) \right). \tag{10}$$

For a given sequence of control decisions $u_0, u_1, \ldots$, let g_i denote the cost of the ith trajectory given by

$$g_i = \sum_{k=0}^{H-1} g(x_k^i, u_k). \tag{11}$$

Clearly, $g_1, g_2, \ldots$ are independent and identically distributed or i.i.d. random variables, where $\mathrm{E}[g_i] = J, \forall i$. In dynamic programming formulations, we do not typically optimize the decision variables $u_0, u_1, \ldots$ together, except in certain ADP schemes such as the NBO, where the decision variables over a finite horizon are indeed optimized together.

The below result suggests that with sufficient number of sample state trajectories (large N), the approximation error in $\tilde{J}_{NP}$ becomes small enough to ignore.

Lemma 2.1: For any given sequence of actions $u_0, u_l, \ldots$, if the random variables $g_1, g_2, \ldots$ have finite variances, $\tilde{J}_{NP}$ converges to J almost surely.

We can verify the above result using the strong law of large numbers as stated below

$$\tilde{J}_{NP} = \frac{1}{N^{H-1}} \sum_{i=1}^{N^{H-1}} g_i \xrightarrow{\text{a.s.}} \mathrm{E}[g_i] = J, \tag{12}$$

where $\xrightarrow{\text{a.s.}}$ represents almost sure convergence.

In most applications, normal distributions capture the system or model uncertainties and noise characteristics well, as can be seen in our previous studies [6], [11]. Suppose, for a given sequence of actions $u_0, \ldots, u_{H-1}$, the trajectory cost variables $g_1, g_2, \ldots$ (i.i.d.) too follow normal distribution with $\mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are the mean and the variance respectively. Of course, if μ is known, then we do not need an approximation strategy as $J = \mu$. However, if $g_1, g_2, \ldots$ are known to follow a normal distribution with unknown mean (μ) and variance (σ) with possibly known bounds, i.e., $\mu_{\min} \leq \mu \leq \mu_{\max}$ and $\sigma_{\min} \leq \sigma \leq \sigma_{\max}$, the following result holds significance.

Lemma 2.2: For a given sequence of actions $u_0, \ldots, u_{H-1}$

$$\bar{J}_{NP} \xrightarrow{\text{a.s.}} \frac{J}{\sqrt{2\pi\sigma^2}} \tag{13}$$

Proof: The likeliness probability of g_i is q_i. We also know that $q_i g_1, q_2 g_2, \ldots$ are i.i.d., where the expectation of this sequence is evaluated below

$$\mathrm{E}[q_i g_i] = \int_{-\infty}^{\infty} P(g_i) g_i P(g_i) dg_i. \tag{14}$$

Since $g_i \sim \mathcal{N}(\mu, \sigma^2)$, the following holds:

$$\begin{aligned}
\mathrm{E}[q_i g_i)] &= \int_{-\infty}^{\infty} g_i P(g_i)^2 dg_i \\
&= \int_{-\infty}^{\infty} g_i \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(g_i-\mu)^2/2\sigma^2} \right)^2 dg_i \\
&= \frac{1}{\sqrt{4\pi\sigma^2}} \int_{-\infty}^{\infty} g_i \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(g_i-\sqrt{2}\mu)^2/2\sigma^2} \right)^2 dg_i \\
&= \frac{\sqrt{2}\mu}{\sqrt{4\pi\sigma^2}} = \frac{\mu}{\sqrt{2\pi\sigma^2}} = \frac{J}{\sqrt{2\pi\sigma^2}}.
\end{aligned} \tag{15}$$

Therefore, due to the strong law of large numbers

$$\bar{J}_{NP} = \sum_{i=1}^{N^{H-1}} q_i g_i \xrightarrow{\text{a.s.}} \mathrm{E}[q_i g_i] = \frac{J}{\sqrt{2\pi\sigma^2}}. \tag{16}$$

Although the above result does not guarantee that the approximation error for $\bar{J}_{NP}$ converges to zero, we know that the ratio $\bar{J}_{NP}/J$ converges (in probability) to a limit bounded above by the constant $1/\sqrt{2\pi\sigma^2_{\min}}$.

III. Case Study

We implement the above-discussed MHP methods in the context of a linear quadratic Gaussian control (LQG) problem as discussed below. Although there are closed-form solutions for LQG problems, this example allows us to quantify the benefits of using RS-MHP methods over existing similar methods, particularly NBO.

A. Linear Quadratic Problem

Let the system state evolve according to the following linear equation:

$$x_{k+1} = (1-a)x_k + au_k + w_k, \quad w_k \sim \mathcal{N}(0,\sigma^2), \quad (17)$$

where $0 < a < 1$ is a constant, and w_k is a random disturbance modeled by a zero-mean Gaussian distribution with variance σ^2. The cost function over the time-horizon H is defined as follows:

$$J = \mathrm{E}\left[r(x_H - T)^2 + \sum_{k=0}^{H-1} u_k^2 \right], \quad (18)$$

where r and T are constants. This is a simplified oven temperature control example borrowed from [12].

If we apply the traditional NBO method, assuming $H = 2$, the cost function J is approximated (assuming nominal values or zeros for w_0 and w_1) as

$$J_{\mathrm{NBO}} = r\left((1-a)^2 x_0 + a(1-a)u_0 + au_1 - T\right)^2 + u_0^2 + u_1^2 \quad (19)$$

and the exact cost function J can be evaluated analytically as

$$J = r\left((1-a)^2 x_0 + a(1-a)u_0 + au_1 - T\right)^2 + u_0^2 + u_1^2 + r\sigma^2\left((1-a)^2 + 1\right). \quad (20)$$

We notice the approximation error due to the NBO method is

$$|J_{\mathrm{NBO}} - J| = r\sigma^2\left((1-a)^2 + 1\right). \quad (21)$$

This approximation error for a generic time-horizon H (the above error term is derived for $H = 2$) is given by

$$|J_{\mathrm{NBO}} - J| = r\sigma^2 \sum_{n=0}^{H-1} (1-a)^{2n}. \quad (22)$$

The above expression suggests that the NBO approximation error can be significantly high depending on the parameters a, σ, and r. With MHP approximation, the cost function reduces to

$$J_{\mathrm{MHP}} = \frac{1}{P}\left(\sum_{i=1}^{P} r(x_H^i - T)^2 \right) + \sum_{k=0}^{H-1} u_k^2, \quad (23)$$

where P is the number of state-trajectories generated using the MHP approach, and x_H^i is the final state in the ith trajectory. Lemma 2.1 suggests that the approximation error due to the above MHP method converges (in probability) to zero. We verify this result with a numerical simulation, where we implement the NBO and the MHP methods with the following assumptions: $x_0 = 0, r = 10, T = 1, H = 2, u_0 = 0.55, u_1 = 0.17, \sigma = 1$. We vary P from 100 to 10000 with increments of 100. Figure 2 shows the cost function approximated using MHP and NBO methods. The figure clearly demonstrates that the error due to NBO approximation can be significantly high, while MHP performs better in cost approximation.

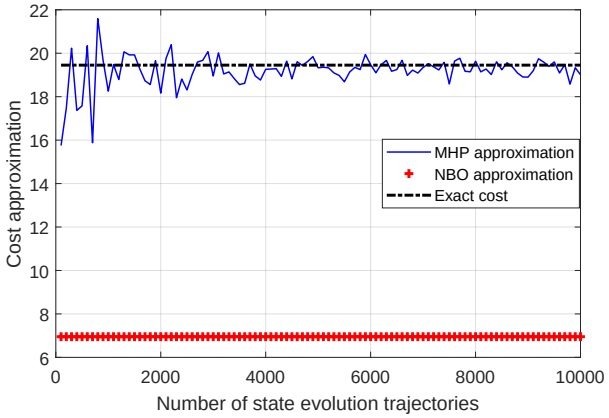

Fig. 2. MHP vs. NBO

RS-MHP has better capability in approximating the expectation operator in Eq. 1 than the NBO approach as we consider multiple hypotheses of state trajectories in RS-MHP as opposed to a single hypothesis in NBO. This is demonstrated in the above case study. In our future study, we will derive quantitative performance guarantees of RS-MHP over NBO for generic long horizon optimal control problems. The impact of parameters M and N on the approximation error will also be considered in our future study.

IV. Conclusions

In this paper, we developed two *approximate dynamic programming* or ADP methods to approximate the cost function in long horizon optimal control problems. Specifically, our methods called *random sampling - multipath hypothesis propagation* or RS-MHP methods are inspired from Monte-carlo Tree Search methods. The basic theme of these methods is to evolve the system state over multiple trajectories into the future while sampling the noise disturbances at each time-step. We derive convergence results that show that the cost approximation error from our RS-MHP methods

converges (in probability) toward zero as the sample size increases. As a case study, we applied our methods to approximate the cost function in a linear quadratic control problem, where we demonstrated the benefits of our approach against an existing approach called *nominal belief-state optimization* or NBO. In our future study, we will apply the above methods to more complex control problems such as the UAV motion control problem we studied in the past [6], where we applied the NBO method to approximate the cost function. Additionally, we will derive convergence results for a general classs of nonlinear systems.

V. Acknowledgment

The authors would like to thank Nicolas Lanchier, Arizona State University, for his valuable inputs and feedback on the convergence results discussed in this paper.

References

[1] E. K. P. Chong, C. M. Kreucher, and A. O. Hero, "Partially observable Markov decision process approximations for adaptive sensing," *Discrete Event Dynamic Systems*, vol. 19, no. 3, pp. 377–422, Sep 2009.

[2] D. P. Bertsekas and D. A. Castanon, "Rollout algorithms for stochastic scheduling problems," *J. Heuristics*, vol. 5, pp. 89–108, 1999.

[3] E. K. P. Chong, R. L. Givan, and H. S. Chang, "A framework for simulation-based network control via hindsight optimization," in *Proc. 39th IEEE Conf. Decision and Control*, Sydney, Australia, 2000, pp. 1433–1438.

[4] G. Wu, E. K. P. Chong, and R. Givan, "Burst-level congestion control using hindsight optimization," *IEEE Trans. Autom. Control*, vol. 47, pp. 979–991, 2002.

[5] D. Bertsekas, "Feature-based aggregation and deep reinforcement learning: A survey and some new implementations," *IEEE/CAA Journal of Automatica Sinica*, no. 1, 2019.

[6] S. Ragi and E. K. P. Chong, "UAV path planning in a dynamic environment via partially observable Markov decision process," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, pp. 2397–2412, 2013.

[7] ——, "Dynamic UAV path planning for multitarget tracking," in *Proc. American Control Conf.*, Montreal, Canada, 2012, pp. 3845–3850.

[8] S. Ragi and H. D. Mittelmann, "Mixed-integer nonlinear programming formulation of a UAV path optimization problem," in *Proc. American Control Conf.*, Seattle, WA, 2017, pp. 406–411.

[9] S. Miller, Z. Harris, and E. K. P. Chong, "A POMDP framework for coordinated guidance of autonomous UAVs for multitarget tracking," *EURASIP Journal on Advances in Signal Processing*, 2009.

[10] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of Monte Carlo tree search methods," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, March 2012.

[11] S. Ragi and E. K. P. Chong, "Decentralized guidance control of UAVs with explicit optimization of communication," *J. Intelligent & Robotic Systems*, vol. 73, no. 1, pp. 811–822, 2014.

[12] D. P. Bertsekas. Lecture on reinforcement learning and optimal control. [Online]. Available: http://www.mit.edu/~dimitrib/Slides_Lecture2_RLOC.pdf