A Scalable Sampling Method to High-Dimensional Uncertainties for Optimal and Reinforcement Learning-Based Controls

Junfei Xie, Yan Wan, Kevin Mills, James J. Filliben, and F. L. Lewis

***Abstract*—Modern dynamical systems often operate in environments of high-dimensional uncertainties that modulate system dynamics in a complicated fashion. These high-dimensional uncertainties, non-Gaussian in many realistic scenarios, complicate real-time system analysis, design, and control tasks. In this letter, we address the scalability of computation for systems of high-dimensional uncertainties by introducing new sampling methods, the multivariate probabilistic collocation method (M-PCM), and its extension called M-PCM-orthogonal fractional factorial design (OFFD) which integrates M-PCM with the OFFDs to break the curse of dimensionality. We explore the capabilities of M-PCM and M-PCM-OFFD-based optimal control and adaptive control using the reinforcement learning approach. The analyses and simulation studies illustrate the efficiency and effectiveness of these two approaches.**

***Index Terms*—Uncertain systems, optimal control, adaptive control.**

I. Introduction

MODERN dynamical systems, such as complex information systems, power networks, and air traffic systems, often operate in environments of high-dimensional uncertainties. These high-dimensional uncertainties typically have the following features: 1) they modulate system dynamics in a complicated fashion that cannot be captured by simple additive white noises, and hence significantly complicate real-time system analysis, design and control tasks, and 2) their statistical information can be obtained from environmental forecasting tools that are independent to system internal dynamics. For instance, strategic air traffic flow management (ATFM) plans traffic in the strategic time frame (with $2-15$ hours look-ahead time), during which a wide range of weather scenarios can possibly occur [1]. The uncertain weather information is available from probabilistic forecasting tools such as the *Very Short Range Ensemble Forecast System* (VSREF). The weather uncertainties modulate region capacities in the National Airspace System, posing uncertain nonlinear constraints to interweaving traffic flows [2]. Techniques are needed to quickly design optimal and adaptive control strategies that are robust and scalable to these high-dimensional uncertainties.

Stochastic optimal control has been widely studied in the literature (e.g., [3]–[7]). For systems of linear dynamics, additive noise and quadratic cost, linear quadratic control can easily find a solution [4]. However, for modern dynamical system applications, the dynamics often may only be captured by complicated simulators rather than clean mathematics, and uncertainties modulate system dynamics in a nonlinear fashion. Backward-in-time dynamic programming approaches (e.g., based on the *Bellman optimality equation*) and forward-in-time Monte Carlo (MC) reinforcement learning-based adaptive control [4], [6] have been used. These methods typically involve the discretization and sampling of the uncertainty space to estimate expected cost. As evaluating system dynamics over all discretized values of uncertain parameters is computationally expensive for large-scale dynamical systems [3], appropriate discretization and sampling schemes are pivotal to the accuracy and scalability of stochastic optimal control strategies [8].

The MC simulation method and its variants such as Markov Chain MC and Sequential MC have been used to sample the uncertainty space. However, they require a large number of simulations to converge to the mean cost estimates, making it computationally impractical for the real-time control of large-scale systems that typically consume considerable computation for a single simulation run. To reduce the computational cost, quadrature schemes such as Euler [9], Composite-Simpson [10], LGL-quadrature [11], and sparse-grid (SG) [12], have recently been applied to sample the uncertainty space. However, their accuracy and computational scalability are either unsatisfactory or unjustified for arbitrary

Manuscript received March 6, 2017; revised May 16, 2017; accepted May 19, 2017. Date of publication May 26, 2017; date of current version June 8, 2017. This work was supported in part by the National Institute of Standards and Technology under Grant 60NANB13D172, and in part by the National Science Foundation under Grant EAGER-1522458, Grant CAREER-1714519, and Grant CPS-1714826. Recommended by Senior Editor F. Blanchini. *(Corresponding author: Yan Wan.)*

J. Xie is with the Department of Computer Science, Texas A&M University at Corpus Christi, Corpus Christi, TX 78412 USA (e-mail: junfei.xie@tamucc.edu).

Y. Wan and F. L. Lewis are with the Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: yan.wan@uta.edu; lewis@uta.edu).

K. Mills and J. J. Filliben are with NIST, Gaithersburg, MD 20899 USA (e-mail: kmills@nist.gov).

Digital Object Identifier 10.1109/LCSYS.2017.2708598

forms of uncertainties. Another approach is to approximate the system dynamics modulated by uncertainties as a master Markov chain of the state space being the Cartesian product of the system state space and the uncertainty space, and solve the problem using Markov decision process approaches [13]. This approach is again not scalable. It works well for systems of simple dynamics, but is ineffective for systems of large state and uncertainty spaces.

To address these challenges, we develop in this letter an optimal control framework that is scalable to high-dimensional uncertainties. The framework builds on two multi-dimensional uncertainty evaluation approaches that we developed recently, the multivariate probabilistic collocation method (M-PCM) [14], and the scalable M-PCM-OFFD that integrates M-PCM with orthogonal fractional factorial designs (OFFDs) [15], [16].

In this letter, we explore the capabilities of M-PCM and M-PCM-OFFD in facilitating optimal control for systems of high-dimensional uncertainties, which are introduced in Section II. Three representative scenarios are considered: 1) finite-horizon optimal control with uncertain parameters changing independently across time (Section III-A), 2) finite-horizon optimal control with uncertain parameters evolving according to Markov chains (Section III-B), and 3) infinite-horizon forward-in-time optimal control using reinforcement learning (Section IV). For arbitrary system dynamics captured by a black-box simulator, we prove that the M-PCM and M-PCM-OFFD based stochastic optimal controls find accurate solutions with very limited computational costs, under simple assumptions that hold for broad realistic systems. These features are further validated through illustrative examples and comparative simulation studies with existing methods in Section V.

II. Uncertainty Sampling Methods: M-PCM and M-PCM-OFFD

M-PCM [14] and M-PCM-OFFD [15], [16] are newly developed computationally effective sampling methods to evaluate output statistics for system mappings of high-dimensional uncertainties.

A. M-PCM

The M-PCM smartly selects a small set of samples according to the statistics (e.g., pdfs) of uncertain parameters, and runs simulations at these samples to produce a reduced-order mapping that maintains precisely the output statistics of the original mapping (see [14, Sec. II.B] for detailed design procedures). Lemma 1 shows the key result on the performance of M-PCM in accurately estimating mapping statistics under uncertainty.

Lemma 1 [14, Th. 2]: Consider a system mapping modulated by m independent uncertain parameters:

$$G(a_1,\ldots,a_m)=\sum_{j_1=0}^{2n_1-1}\sum_{j_2=0}^{2n_2-1}\cdots\sum_{j_m=0}^{2n_m-1}\Psi_{j_1,\ldots,j_m}\prod_{i=1}^{m}a_i^{j_i} \quad (1)$$

where a_i is an uncertain parameter with the degree *up to* $2n_i-1$. n_i are positive integers for all $i\in\{1,2,\ldots,m\}$, and $\Psi_{j_1,\ldots,j_m}\in R$ are the coefficients. Each uncertain parameter a_i follows an independent pdf $f_{A_i}(a_i)$. The M-PCM approximates $G(a_1,\ldots,a_m)$ with the following low-order mapping

$$G'(a_1,\ldots,a_m)=\sum_{j_1=0}^{n_1-1}\sum_{j_2=0}^{n_2-1}\cdots\sum_{j_m=0}^{n_m-1}\Omega_{j_1,\ldots,j_m}\prod_{i=1}^{m}a_i^{j_i}, \quad (2)$$

with $E[G(a_1,\ldots,a_m)]=E[G'(a_1,\ldots,a_m)]$, where $\Omega_{j_1,\ldots,j_m}\in R$ are coefficients. The M-PCM reduces the number of simulations from $2^m\prod_{i=1}^{m}n_i$ to $\prod_{i=1}^{m}n_i$.

Remarks: The knowledge of uncertainty is typically available to realistic system studies in the form of probabilistic forecasts or historical data. We note that M-PCM is not limited by the knowledge of precise pdfs. When low-order moments or sample data of uncertain parameters are available, sample-moment based approaches can be used to select M-PCM simulation points (see [14, Sec. V.B]). We can also fit the sample data or low-order moments (e.g., mean and variance) with canonical pdfs. Beyond accurate output mean estimation, M-PCM also has other nice statistical properties such as accurate cross-statistics estimation and tight connection to minimum mean squares estimation (see [14, Sec. III.A]).

B. M-PCM-OFFD

M-PCM significantly reduces the number of simulations to estimate output mean, however its computation cost does not scale with the number of uncertain parameters. M-PCM-OFFD [15], [16] further breaks the curse of dimensionality through leveraging the systematic procedures and nice properties of 2-level OFFD such as *balance* and *orthogonality* [17], [18]. We show that M-PCM-OFFD has better performance in terms of accuracy and computational scalability for estimating output statistics, compared to existing uncertainty sampling approaches, such as stochastic response surface method and SG [12], [19]. Lemma 2 illustrates the key capability of M-PCM-OFFD. Please refer to [16, Sec. 5.2] for detailed design procedures.

Lemma 2 [16, Sec. 5.2]: Consider an m-parameter system mapping (Equation (1)) with each input parameter a_i of degree up to 3 (i.e., $n_i=2,\forall i\in\{1,2,\ldots,m\}$). Assume that its coefficients $\Psi_{j_1,\ldots,j_m}=0$ if more than τ of $j_1,\ldots,j_m$ are non-zero, where $1\le\tau\le m$. The integrated M-PCM and $2^{m-\gamma^*}$ OFFD approximates the original system mapping with the following low-order mapping.

$$G^*(a_1,\ldots,a_m)=\sum_{j_1=0}^{1}\sum_{j_2=0}^{1}\cdots\sum_{j_m=0}^{1}\Omega_{j_1,\ldots,j_m}\prod_{i=1}^{m}a_i^{j_i}, \quad (3)$$

such that $E[G(a_1,\ldots,a_m)]=E[G^*(a_1,\ldots,a_m)]$, where coefficients $\Omega_{j_1,\ldots,j_m}=0$ if more than τ of $j_1,\ldots,j_m$ are non-zero, $\gamma^*=max\{\gamma\mid 1\le\gamma\le m-\lceil log_2(\sum_{i=0}^{\tau}\binom{i}{m})\rceil$, *and* $2^{m-\gamma}_{\mathcal{R}}$ *OFFD exists, with* $\mathcal{R}\ge 2\tau+1\}$. The M-PCM-OFFD reduces the number of simulations from 2^{2m} to $2^{m-\gamma^*}$ in the range of $\left[2^{\lceil log_2(m+1)\rceil},2^{m-1}\right]$, making it scalable with the number of uncertain parameters.

Remarks: The assumption on $\Psi_{j_1,\ldots,j_m}$ reflects that the interacting effects that involve a large number of parameters are less important to the output statistics than those that involve a few parameters. This assumption typically holds for realistic modern complex systems, and has been widely used in the

field of statistical experimental designs [17], [18]. Other distinguishing features of M-PCM-OFFD that have been proved include the robustness to computational approximations in system simulators (see [16, Sec. 5.3]). We also note that in the lemma, the term of the highest degree is of degree $3m$. The lemma can be generalized to higher degrees through replacing the 2-level OFFD with higher-level OFFD, which we leave for the future work.

III. Optimal Control for Systems of High Dimensional Uncertainties

In this section, we study two optimal control problems for systems of high dimensional uncertainties.

A. Optimal Control for Systems of Uncertain Parameters Independent Across Time

Consider a generic dynamical system modulated by an m-dimensional time-varying uncertain vector $\mathbf{a}[k]$. Each element of $\mathbf{a}[k]$, $a_i[k]$, changes independently over time with pdf $f_{A_i[k]}(a_i[k])$. The system dynamic is

$$\mathbf{x}[k+1] = h_k(\mathbf{x}, \mathbf{u}, \mathbf{a}), \tag{4}$$

with state vector $\mathbf{x}[k] \in S$ and control input vector $\mathbf{u}[k] \in C$, where S and C are the known state space and control space, respectively. $h_k()$ is the system dynamic function of $\mathbf{x}[k]$, $\mathbf{u}[k]$ and $\mathbf{a}[k]$. The total expected cost equals [3]

$$J_N(\mathbf{x}[0]) = E_{\mathbf{a}[0]}\{\cdots E_{\mathbf{a}[N-1]}\{\sum_{k=0}^{N-1} \alpha^k g_k(\mathbf{x}, \mathbf{u}) + \alpha^N q_N(\mathbf{x})\}\cdots\} \tag{5}$$

where $E_{\mathbf{a}[k]}()$ is the expectation of function within () with respect to the uncertain vector $\mathbf{a}[k]$. $\alpha \in (0, 1]$ is a discount factor. $g_k()$ and $q_k()$ are the running and terminal cost at time step k, respectively. Consider the problem of finding an optimal *control policy* $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$, with $\mathbf{u}^*[k] = \mu_k^*(\mathbf{x}[k])$, such that the total expected cost is minimized, i.e.,

$$\pi^* = arg\min_{\pi}\{J_N(\mathbf{x}[0])\}. \tag{6}$$

μ_k is a control function that maps S into C.

This finite-horizon control problem can be solved using *backward-in-time* methods, e.g., dynamic programming. As the uncertain parameters are independent from the states, we define the *value function* as $V_k(\mathbf{x}[k]) = E_{\mathbf{a}[k]}\{\cdots E_{\mathbf{a}[N-1]}\{\sum_{i=k}^{N-1} \alpha^{i-k} g_i(\mathbf{x}, \mathbf{u}) + \alpha^{N-k} q_N(\mathbf{x})\}\cdots\} = E_{\mathbf{a}[k]}\{\cdots E_{\mathbf{a}[N-1]}\{g_k(\mathbf{x}, \mathbf{u}) + \alpha[\sum_{i=k+1}^{N-1} \alpha^{i-(k+1)} g_i(\mathbf{x}, \mathbf{u}) + \alpha^{N-(k+1)} q_N(\mathbf{x})]\}\cdots\} = E_{\mathbf{a}[k]}\{g_k(\mathbf{x}, \mathbf{u}) + \alpha E_{\mathbf{a}[k+1]}\{\cdots E_{\mathbf{a}[N-1]}\{\sum_{i=k+1}^{N-1} \alpha^{i-(k+1)} g_i(\mathbf{x}, \mathbf{u}) + \alpha^{N-(k+1)} q_N(\mathbf{x})\}\cdots\}\}$, and $V_N(\mathbf{x}[N]) = q_N(\mathbf{x})$. The following *Bellman's Equation* holds.

$$V_k(\mathbf{x}[k]) = E_{\mathbf{a}[k]}\big[g_k(\mathbf{x}, \mathbf{u}) + \alpha V_{k+1}(\mathbf{x}[k+1])\big], \tag{7}$$

The optimal cost $J_N^*(\mathbf{x}[0]) = V_0^*(\mathbf{x}[0])$ can then be derived by working backward in time using dynamic programming according to the *Bellman optimality equation* [20]:

$$V_k^*(\mathbf{x}[k]) = \min_{\pi} E_{\mathbf{a}[k]}\Big[g_k(\mathbf{x}, \mathbf{u}) + \alpha V_{k+1}^*(\mathbf{x}[k+1])\Big] \tag{8}$$

Defining $G_k(\mathbf{x}, \mathbf{u}, \mathbf{a}) = g_k(\mathbf{x}, \mathbf{u}) + \alpha V_{k+1}^*(\mathbf{x}[k+1])$, we then notice that given an admissible state $\mathbf{x}[k]$ and control value $\mathbf{u}[k]$, $E_{\mathbf{a}[k]}[G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})]$ can be approximated by the mean output of a system mapping $G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})$ using M-PCM or M-PCM-OFFD. When each uncertain parameter $a_i[k]$ has a degree up to $2n_i[k]-1$, $G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})$ has the following form

$$G_k(\mathbf{x}, \mathbf{u}, \mathbf{a}) = \sum_{j_1=0}^{2n_1[k]-1} \cdots \sum_{j_m=0}^{2n_m[k]-1} \Psi_{j_1,\ldots,j_m}(\mathbf{x}[k], \mathbf{u}[k]) \prod_{i=1}^{m} a_i^{j_i}[k], \tag{9}$$

Theorem 1 holds, where $\Psi_{j_1,\ldots,j_m}(\mathbf{x}[k], \mathbf{u}[k]) \in R$ are the coefficients determined by state $\mathbf{x}[k]$ and control input $\mathbf{u}[k]$.

Theorem 1: Consider a dynamical system described by Equation (4), with cost and value functions given by Equation (5) and Equations (7)–(9), respectively. By applying dynamic programming, and sampling the uncertainty space at each iteration using M-PCM, the true optimal control policy can be obtained with no error.

Proof: First, we introduce a set of notions for the optimal control obtained using M-PCM. According to Lemma 1, for a given admissible state $\mathbf{x}[k]$ and control input $\mathbf{u}[k]$, $G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})$ can be approximated using M-PCM as a low-order function of the form $G_k'(\mathbf{x}, \mathbf{u}, \mathbf{a}) = \sum_{j_1=0}^{n_1[k]-1} \cdots \sum_{j_m=0}^{n_m[k]-1} \Omega_{j_1,\ldots,j_m}(\mathbf{x}[k], \mathbf{u}[k]) \prod_{i=1}^{m} a_i^{j_i}[k]$, such that $E_{\mathbf{a}[k]}\big[G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})\big] = E_{\mathbf{a}[k]}\big[G_k'(\mathbf{x}, \mathbf{u}, \mathbf{a})\big]$. Denote $V_k'^*(\mathbf{x}[k]) = \min_{\mathbf{u}[k]} E_{\mathbf{a}[k]}[G_k'(\mathbf{x}, \mathbf{u}, \mathbf{a})]$ and $\pi'^* = \{\mu_0'^*, \mu_1'^*, \ldots, \mu_{N-1}'^*\}$. The optimal control policy can be found by evaluating $V_k'^*(\mathbf{x}[k])$ at only the M-PCM samples at each iteration. Denote $\mathbf{u}'^*[k] = \mu_k'^*(\mathbf{x}[k])$.

In order to prove this theorem, we need to show $\mathbf{u}^*[k] = \mathbf{u}'^*[k]$, i.e., for each $\mathbf{x}[k] \in S$, $arg\min_{\mathbf{u}[k]} E_{\mathbf{a}[k]}[G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})] = arg\min_{\mathbf{u}[k]} E_{\mathbf{a}[k]}\big[G_k'(\mathbf{x}, \mathbf{u}, \mathbf{a})\big]$. This is equivalent to the following two statements: 1) $\nexists \mathbf{u}'^*[k] \neq \mathbf{u}^*[k]$, such that $E_{\mathbf{a}[k]}\big[G_k'(\mathbf{x}, \mathbf{u}'^*, \mathbf{a})\big] < E_{\mathbf{a}[k]}\big[G_k(\mathbf{x}, \mathbf{u}^*, \mathbf{a})\big]$, and 2) $\nexists \mathbf{u}^*[k] \neq \mathbf{u}'^*[k]$, such that $E_{\mathbf{a}[k]}\big[G_k(\mathbf{x}, \mathbf{u}^*, \mathbf{a})\big] < E_{\mathbf{a}[k]}\big[G_k'(\mathbf{x}, \mathbf{u}'^*, \mathbf{a})\big]$. To prove statement 1 using contradiction, we assume that such a $\mathbf{u}'^*[k] \neq \mathbf{u}^*[k]$ exists. Lemma 1 leads to $E_{\mathbf{a}[k]}\big[G_k'(\mathbf{x}, \mathbf{u}'^*, \mathbf{a})\big] = E_{\mathbf{a}[k]}\big[G_k(\mathbf{x}, \mathbf{u}'^*, \mathbf{a})\big] < E_{\mathbf{a}[k]}\big[G_k(\mathbf{x}, \mathbf{u}^*, \mathbf{a})\big]$, which violates the fact that $\mathbf{u}^* = arg\min_{\mathbf{u}[k]} E_{\mathbf{a}[k]}[G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})]$. Similarly, to prove statement 2, we assume such a $\mathbf{u}^*[k] \neq \mathbf{u}'^*[k]$ exists. Lemma 1 leads to $E_{\mathbf{a}[k]}\big[G_k(\mathbf{x}, \mathbf{u}^*, \mathbf{a})\big] = E_{\mathbf{a}[k]}\big[G_k'(\mathbf{x}, \mathbf{u}^*, \mathbf{a})\big] < E_{\mathbf{a}[k]}\big[G_k'(\mathbf{x}, \mathbf{u}'^*, \mathbf{a})\big]$, which violates the fact that $\mathbf{u}'^* = arg\min_{\mathbf{u}[k]} E_{\mathbf{a}[k]}[G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})]$. The results that $\mathbf{u}^*[k] = \mathbf{u}'^*[k]$ and $\pi^* = \pi'^*$ naturally follow. ∎

We can also apply M-PCM-OFFD to estimate the mean cost $E_{\mathbf{a}[k]}[G_k(\mathbf{x}, \mathbf{u}, \mathbf{a})]$ using fewer samples. Following similar procedures in the above proof and through replacing $G_k'(\mathbf{x}, \mathbf{u}, \mathbf{a})$ according to Lemma 2 instead of Lemma 1 in the proof, we can derive Theorem 2. The proof is omitted here for the sake of space.

Theorem 2: Consider a dynamical system described by Equation (4), with cost and value functions given by Equation (5) and Equations (7)–(9), respectively. Then the control policy $\pi^* = \{\mu_0^*, \ldots, \mu_{N-1}^*\}$ optimal to the samples selected by M-PCM-OFFD at each time step, is also optimal to all possible values of the uncertain parameters, if $n_i[k] = 2, \forall i \in \{1, 2, \ldots, m\}$ and $k \in \{0, 1, \ldots, N-1\}$, and

$\Psi_{j_1,\ldots,j_m}(\mathbf{x}[k], \mathbf{u}[k]) = 0$ when more than τ of $j_1, \ldots, j_m$ are non-zero, where $1 \leq \tau \leq m$.

B. Optimal Control for Systems of Uncertain Parameters Evolving According to Markov Chains

Consider a generic dynamical system modulated by an m-dimensional time-varying uncertain vector $\mathbf{a}[k]$. Each element of $\mathbf{a}[k]$, $a_i[k]$, evolves according to a Markov chain of transition matrix $P_i \in R^{M_i \times M_i}$, where M_i is the total number of possible values (or states) of the uncertain parameter $a_i[k]$. The (j, l)th entry of P_i, $P_{i_{jl}}$, represents the probability for $a_i[k]$ to be at state l in the next step if the current state is j. The system dynamics and the total expected cost are shown in Equations (4) and (5). Similar to Equation (6), given the probability of $a_i[0]$ at each state for all $i \in \{1, 2, \ldots, m\}$, consider the problem of finding an optimal control policy $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$ that minimizes the total expected cost. This problem finds broad applications where dynamical systems operate in uncertain environments. For instance, in realistic systems such as the air traffic systems, uncertainties like convective weather have correlations across time, which can be captured by Markov chains.

Consider the i-th uncertain parameter, $a_i[k]$, in a system of m uncertain parameters. Suppose initially $a_i[0]$ is at state j, i.e., $a_i[0] = j$, then the conditional probability of $a_i[k] = l$ can be obtained using $p(a_i[k] = l \mid a_i[0] = j) = P_{i_{jl}}^k$. Suppose $\mathbf{r}_i[k] \in R^{1 \times M_i}$ is a vector of the possibilities of $a_i[k]$ at each state, with its j-th element, denoted as $r_{ij}[k]$, equal to the possibility of $a_i[k]$ at state j, i.e., $r_{ij}[k] = p(a_i[k] = j)$. Then we have $\mathbf{r}_i[k] = \mathbf{r}_i[0]P_i^k$. Since $\mathbf{r}_i[0]$ and the transition matrix P_i are given, we can calculate $\mathbf{r}_i[k]$ at each time step k. Let us now apply dynamic programming to solve this problem. Denoting the set $\{1, 2, \ldots, M_i\}$ as $\mathcal{M}_i$, the optimal value function is

$$
\begin{aligned}
V_k^*(\mathbf{x}[k]) &= \min_{\mathbf{u}[k]}\{E_{a_1[k]}\{\cdots E_{a_m[k]}\{g_k(\mathbf{x}, \mathbf{u}) \\
&\quad + \alpha V_{k+1}^*(\mathbf{x}[k+1])\}\cdots\}\} \\
&= \min_{\mathbf{u}[k]}\{E_{a_1[k]}\{\cdots E_{a_{m-1}[k]}\{\sum_{j \in \mathcal{M}_m} p(a_m[k] = j) \\
&\quad \{g_k(\mathbf{x}, \mathbf{u}) + \alpha V_{k+1}^*(\mathbf{x}[k+1])\}\}\cdots\}\} \\
&= \min_{\mathbf{u}[k]}\{E_{a_1[k]}\{\cdots E_{a_{m-1}[k]}\{\sum_{j \in \mathcal{M}_m} r_{mj}[k] \\
&\quad \{g_k(\mathbf{x}, \mathbf{u}) + \alpha V_{k+1}^*[h_k(\mathbf{x}, \mathbf{u}, a_1, \ldots, a_m = j)]\}\}\cdots\}\} \\
&= \min_{\mathbf{u}[k]}\{\sum_{l \in \mathcal{M}_1} r_{1j}[k]\{\cdots\{\sum_{j \in \mathcal{M}_m} r_{mj}[k]\{g_k(\mathbf{x}, \mathbf{u}) \\
&\quad + \alpha V_{k+1}^*[h_k(\mathbf{x}, \mathbf{u}, a_1 = l, \ldots, a_m = j)]\}\}\cdots\}\} \quad (10)
\end{aligned}
$$

This problem then transforms to the first problem discussed in Section III-A, with the pdfs of the uncertain parameters replaced by the probability mass functions (pmf) that can then be solved by M-PCM or M-PCM-OFFD based stochastic optimal controls. The sample points can be selected according to the pdfs approximated by pmfs, or directly using the sample-moment based approach [14]. The proof of Theorem 3 is omitted as it can be easily derived following the proof procedure of Theorem 1 with $V_k^*(\mathbf{x}[k])$ represented by Equation (10).

Theorem 3: Consider a dynamical system shown in Equation (4), with each element of $\mathbf{a}[k]$, $a_i[k]$, evolves according to a Markov chain and the total expected cost shown in Equation (5). The optimal control policy can be found by sampling the uncertainty space using M-PCM under the assumptions in Theorem 1, or using M-PCM-OFFD under the assumptions in Theorem 2.

IV. Reinforcement Learning-Based Infinite Horizon Control

The uncertainty sampling methods can also be integrated with reinforcement learning to bring the backward-in-time optimal control to forward-in-time adaptive control for systems of high-dimensional uncertainties [5]–[7]. For the infinite horizon case, the adaptive solution is also the optimal solution.

Consider a generic dynamical system described by Equation (4), but with time horizon $N \to \infty$ and $a_i[k]$ following a time-invariant pdf $f_{A_i}(a_i[k])$. Consider the problem of finding the optimal control policy $\pi^* = \{\mu, \mu, \ldots\}$ that minimizes the total expected cost:

$$J(\mathbf{x}[k]) = E_{\mathbf{a}[k]}\{E_{\mathbf{a}[k+1]}\{\cdots E_{\mathbf{a}[\infty]}\{\sum_{i=k}^{\infty} \alpha^{i-k} g_i(\mathbf{x}, \mathbf{u})\}\cdots\}\} \quad (11)$$

The value function $V(\mathbf{x}[k]) = J(\mathbf{x}[k])$, and the Bellman's equation becomes

$$V(\mathbf{x}[k]) = E_{\mathbf{a}[k]}\Big\{g_k(\mathbf{x}, \mathbf{u}) + \alpha V(\mathbf{x}[k+1])\Big\}, \quad (12)$$

where the same value function $V()$ appears on both sides. Therefore, given a control policy π, we can calculate the total expected cost by solving Equation (12). The optimal control policy

$$\pi^* = arg\min_{\pi} V(\mathbf{x}[k]) \quad (13)$$

can be found by repeatedly iterating two procedures *policy evaluation* and *policy improvement*. In the *policy evaluation* step, the value function $V(\mathbf{x}[k])$ is solved using Equation (12) for a set of admissible states $\mathbf{x}[k] \in S' \subseteq S$, given a current control policy π. A best control policy is then derived based on the value function $V(\mathbf{x}[k])$ determined in the previous step using Equation (13) in the *policy improvement* step. Various approaches can be used to implement these two procedures, and here we use *value iteration*. In particular, starting from an arbitrary initial control policy π_0, the following two steps are iterated until convergence [5], [6]:

Value Update: $\forall \mathbf{x}[k] \in S_j' \subseteq S$

$$V_{j+1}(\mathbf{x}[k]) = E_{\mathbf{a}[k]}\Big[g_k(\mathbf{x}, \mu_j(\mathbf{x})) + \alpha V_j(\mathbf{x}[k+1])\Big] \quad (14)$$

Policy Improvement: $\forall \mathbf{x}[k] \in S_j' \subseteq S$

$$\mu_{j+1}(\mathbf{x}[k]) = arg\min_{\mu(\cdot)} E_{\mathbf{a}[k]}\Big[g_k(\mathbf{x}, \mu(\mathbf{x})) + \alpha V_{j+1}(\mathbf{x}[k+1])\Big] \quad (15)$$

where j is the iteration step index. For systems of infinite state and control spaces, approximators (e.g., neural network and polynomial functions) can be used to estimate the value function by $V(\mathbf{x}) = W^T\Phi(\mathbf{x})$, where $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_L(\mathbf{x}))]$ is the basis vector, W is the weight

vector, and the estimation error approaches 0 when the number of terms in $\Phi(\mathbf{x})$, $L \to \infty$. The value update then becomes

$$W_{j+1}^T \Phi(\mathbf{x}[k]) = E_{\mathbf{a}[k]}\Big[g_k(\mathbf{x}, \mathbf{u}) + \alpha W_j^T \Phi(\mathbf{x}[k+1])\Big] \quad (16)$$

which can be solved using the least squares estimation. Similarly, the control function can be estimated by $\mu(\mathbf{x}) = U^T\sigma(\mathbf{x})$, where $\sigma(\mathbf{x})$ and U are the basis and weight vectors, respectively. For systems of $\mathbf{x}[k+1] = A(\mathbf{x}[k]) + B(\mathbf{x}[k])\mathbf{u}[k]$ and $g_k(\mathbf{x}, \mathbf{u}) = Q_1(\mathbf{x}[k]) + \mathbf{u}^T[k]Q_2\mathbf{u}[k]$, where $Q_1(\mathbf{x}[k]) > 0$ and $Q_2 > 0$, the policy improvement step can then be performed using the gradient descent method according to the following equation [5], [6]:

$$U_{j+1}^{i+1} = U_{j+1}^i - \beta\sigma(\mathbf{x}[k])\Big\{2Q_2(U_{j+1}^i)^T\sigma(\mathbf{x}[k]) + E_{\mathbf{a}[k]}\big[\alpha B(\mathbf{x}[k])^T \nabla\Phi^T(\mathbf{x}[k+1])W_{j+1}\big]\Big\}^T, \quad (17)$$

where $\nabla\Phi(\mathbf{x}) = \partial\Phi(\mathbf{x})/\partial\mathbf{x}$, and β is a tuning parameter and i is the tuning index [5], [6]. For systems of finite state and control spaces, value iteration can be achieved by storing and updating lookup tables. Please refer to [5] and [6] for detailed descriptions of different approaches.

We note that given an admissible state $\mathbf{x}[k]$ and control function $\mu_j(\mathbf{x})$, the value function $V_{j+1}(\mathbf{x}[k])$ in Equation (14) can be approximated by the mean output of a system mapping $G_k(\mathbf{x}, \mathbf{a}) = g_k(\mathbf{x}, \mu(\mathbf{x})) + \alpha V_j(\mathbf{x}[k+1])$, using M-PCM or M-PCM-OFFD. Similarly, to derive $\mu_{j+1}(\mathbf{x})$ in the policy improvement step, the expected cost of a candidate control function can also be approximated by M-PCM or M-PCM-OFFD. Theorem 4 describes the utilization of M-PCM and M-PCM-OFFD to find the optimal control policy, the proof of which can be easily derived based on above analysis and proof of Theorem 1 and is thus omitted.

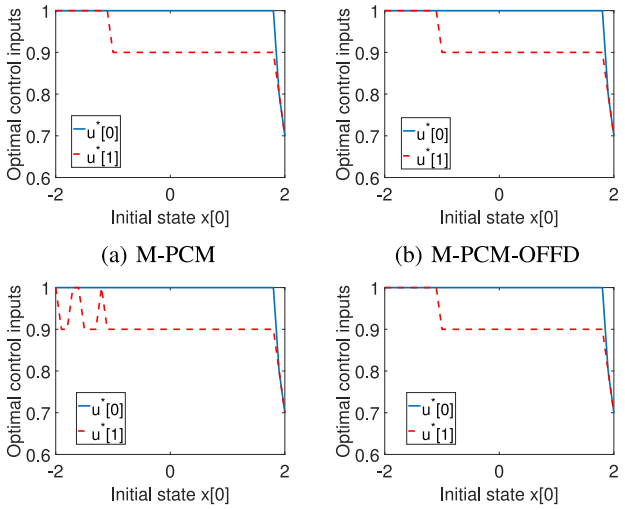
Theorem 4: Consider a dynamical system shown in Equation (4), with time horizon $N \to \infty$ and $a_i[k]$ following a time-invariant pdf $f_{A_i}(a_i[k])$. The optimal control policy can be found by applying value iteration of reinforcement learning, and approximating the value function using M-PCM under the assumptions in Theorem 1, or using M-PCM-OFFD under the assumptions in Theorem 2.

V. Illustrative Examples

In this section, we use three simple examples to illustrate and validate the M-PCM and M-PCM-OFFD based stochastic optimal controls. As the validation requires extensive Monte Carlo simulations, we limit the dimension of uncertainties in the first two examples. The third example illustrates the capability of the M-PCM-OFFD based approach in addressing high-dimensional uncertainties.

A. Dynamic Programming Based Finite-Horizon Control

Consider a system of the dynamics $x[k+1] = a_1[k]x[k] + a_2[k]u[k] + a_3[k]$, where $a_i[k]$ are uncertain parameters that follow independent non-Gaussian distributions: $f_{A_1}(a_1[k]) = \frac{5}{3}$ with $-0.1 \le a_1[k] \le 0.5$, $f_{A_2}(a_2[k]) = \frac{5}{3}$ with $-0.2 \le a_2[k] \le 0.4$ and $f_{A_3}(a_3[k]) = \frac{10}{7}$, $0 \le a_3[k] \le 0.7$. The optimal control policy $\pi^* = \{\mu_0^*, \ldots, \mu_{N-1}^*\}$ is sought to minimize the total expected cost given by Equation (5) with $q_N(x) = (x[N]-10)^2$, $g_k(x, u) = (x[k] - 10)^2 + u[k]^2$, and $\alpha = 1$. The state space and control space are $S = \{-2, -1.9, \ldots, 1.9, 2\}$ and $C = \{-1, -0.9, \ldots, 0.9, 1\}$, respectively.



(a) M-PCM (b) M-PCM-OFFD (c) SG (d) MC

Fig. 1. The optimal control policies found by applying the four approaches to sample the uncertainty space.

We apply M-PCM and M-PCM-OFFD (with $\tau = 1$) based stochastic optimal controls to find the optimal solution. Linear interpolation is used to approximate costs at intermediate points [7]. As a comparative study, we also apply the SG with Gauss-Legendre quadrature rules to sample the uncertainty space [21]. For a fair comparison, the accuracy level[1] of the SG is set to 2 such that a similar number of samples is produced. The MC method is also applied to estimate the true optimal solution for the validation purpose.

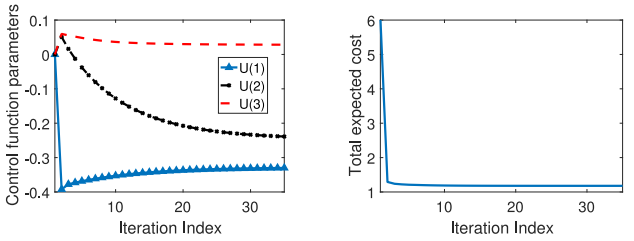
The optimal control policies found by the four methods are shown in Figure 1, with $N = 2$. The number of samples n selected by MPCM, MPCM-OFFD, SG and MC at each iteration are 8, 4, 7 and 100000, respectively. M-PCM-OFFD (or M-PCM) based method finds accurate control solutions, with significantly reduced number of simulations (4 or 8). The SG method is relatively efficient but is less accurate.

B. Reinforcement Learning Based Infinite-Horizon Control

Consider a system of the dynamic $\mathbf{x}[k+1] = A[k]\mathbf{x}[k] + Bu[k] + Ca_3[k]$, where $\mathbf{x}[k] = [x_1[k], x_2[k]]^T$, $A[k] = [a_1[k], 0; 0, a_2[k]]$, $B = [1, 0.5]^T$, and $C = [1, 1]^T$. $a_1[k]$, $a_2[k]$ and $a_3[k]$ are uncertain parameters following $f_{A_1}(a_1[k]) = \frac{5}{2}$ with $0.1 \le a_1[k] \le 0.5$, $f_{A_2}(a_2[k]) = 2$ with $-0.5 \le a_2[k] \le 0$ and $f_{A_3}(a_3[k]) = \frac{10}{3}$, $0.2 \le a_3[k] \le 0.5$, respectively. The optimal control policy π^* is sought to minimize the total expected cost given by Equation (11), with $\alpha = 0.8$, $g_k(\mathbf{x}, u) = 8x_1^2[k] + 2x_2^2[k] + u[k]^2$. Both state and control spaces are continuous and infinite.

We use value iteration, with cost and control functions approximated as polynomials, i.e., $J(\mathbf{x}) = W^T\Phi(\mathbf{x})$ and $\mu(\mathbf{x}) = U^T\sigma(\mathbf{x})$, where $\Phi(\mathbf{x}) = [1, x_1, x_2, x_1^2, x_2^2, x_1x_2]^T$,

[1]The accuracy level reflects the accuracy of the underlying quadrature rule [21]. Higher accuracy level requires more sample runs to reach convergence.



(a) U (b) J(**x**[0])

Fig. 2. Illustration of the a) convergence of U and b) trajectory of total expected cost $J(\mathbf{x}[0])$ with $\mathbf{x}[0] = [0.1, 0.1]^T$.

$\sigma(\mathbf{x}) = [1, x_1, x_2]^T$, $W \in R^{6\times 1}$ and $U \in R^{3\times 1}$ are weight vectors. We iteratively update W and U vectors according to Equations (16)–(17) ($\beta = 0.1$) until convergence, with mean values $E_{\mathbf{a}[k]}()$ in Equations (16) and (17) estimated using M-PCM, M-PCM-OFFD ($\tau = 1$), and MC. The number of samples n used by M-PCM, M-PCM-OFFD, and MC to estimate each mean value are 8, 4 and 10000, respectively. Figure 2 shows the convergence of U along with the associated total expected cost $J(\mathbf{x}[0])$ using M-PCM-OFFD. The weight vectors derived by M-PCM (or M-PCM-OFFD) based method are $W = [1.076, 0.118, -0.133, 8.192, 2.136, 0.104]^T$ and $U = [-0.329, -0.246, 0.028]^T$, which are close to $W = [1.076, 0.118, -0.130, 8.203, 2.130, 0.103]^T$ and $U = [-0.327, -0.249, 0.026]^T$ obtained by the MC based method. When $\mathbf{x}[0] = [0.1, 0.1]^T$, the optimal total expected cost found by M-PCM (or M-PCM-OFFD) and MC based methods are 1.175 and 1.180, respectively.

C. Control Under High-Dimensional Uncertainties

Consider a system modulated by 50 time-invariant uncertain parameters and 2 time-invariant control inputs. The system dynamics are described by $\mathbf{x}[k] = 0.2\begin{bmatrix}\sum_{i=1}^{10} a_i & -\sum_{i=11}^{20} a_i \\ 0 & \sum_{i=1}^{25} a_i^2\end{bmatrix}\mathbf{x}[k] + 0.1\begin{bmatrix}0 & 0 \\ 0 & \sum_{i=26}^{50} a_i\end{bmatrix}\mathbf{u}$, where $\mathbf{x}[k] = [x_1[k], x_2[k]]^T$ and $\mathbf{u} = [u_1, u_2]^T$ are state and control vectors, respectively. a_i, $i \in \{1, 2, \ldots, 50\}$, is an uncertain variable that follows a uniform distribution $f_{A_i}(a_i) = 1$, $0 \le a_i \le 1$. The state space is $\mathbf{x}[k] \in R^2$ and the control space is of size 121: $u_1 \in \{1, 1.1, \ldots, 1.9, 2\}$ and $u_2 \in \{-0.5, -0.4, \ldots, 0.4, 0.5\}$. The optimal control inputs $\mathbf{u}^*$ are sought to minimize the following total expected cost: $J_N(\mathbf{x}[0]) = E_{\mathbf{a}[0]}\{\cdots E_{\mathbf{a}[N-1]}[\sum_{k=0}^{N-1} \alpha^k(\mathbf{x}^T[k]Q_1\mathbf{x}[k] + \mathbf{u}^T Q_2 \mathbf{u})]\cdots\}$ where $\alpha = 0.8$, $N = 5$, $\mathbf{x}[0] = [0, 1]^T$, $Q_1 = \begin{bmatrix}2 & 0 \\ 0 & 3\end{bmatrix}$ and $Q_2 = \begin{bmatrix}0.5 & 0 \\ 0 & 1\end{bmatrix}$.

With the high number of uncertain parameters, MC simulations are impractical to use. M-PCM based control also requires 2^{50} simulations to estimate the total expected cost for each admissible control vector $\mathbf{u}$. Theorem 2 proves the feasibility of M-PCM-OFFD in handling this high dimensionality. Integrating M-PCM with 2^{50-44}_{III} OFFD ($\tau = 1$), the M-PCM-OFFD can further reduce the number of simulations from 2^{50} to $2^6 = 64$. The optimal control inputs and corresponding total expected cost obtained by M-PCM-OFFD based control are $\mathbf{u}^* = [1, -0.5]^T$ and $J_N^*(\mathbf{x}[0]) = 167.351$, respectively.

VI. Conclusion

This letter develops two multi-dimensional uncertainty evaluation based approaches to address the scalability issue of stochastic optimal control for systems of high-dimensional uncertainties. For three cases covering varying scenarios, we prove that the control solution optimal to the sampled uncertainty space produced by M-PCM or M-PCM-OFFD is also optimal to the original uncertainty space under simple assumptions on the forms of the cost functions and orders of uncertain parameters. Simulation and comparison studies demonstrate the accuracy and computational efficiency of these two approaches.

References

[1] C. Taylor, C. Wanke, Y. Wan, and S. Roy, "A framework for flow contingency management," in *Proc. AIAA Aviation Technol. Integr. Oper. Conf. (ATIO)*, 2011, pp. 6904–6926.

[2] Y. Wan and S. Roy, "A scalable methodology for evaluating and designing coordinated air-traffic flow management strategies under uncertainty," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 4, pp. 644–656, Dec. 2008.

[3] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, vol. 23. New York, NY, USA: Academic Press, 1978.

[4] H. J. Kappen, "An introduction to stochastic control theory, path integrals and reinforcement learning," in *Proc. Cooperative Behav. Neural Syst.*, vol. 887. 2007, pp. 149–181.

[5] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, 3rd Quart., 2009.

[6] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

[7] F. L. Lewis, L. Xie, and D. Popa, *Optimal and Robust Estimation: With an Introduction to Stochastic Control Theory*, vol. 29. Hoboken, NJ, USA: CRC Press, 2007.

[8] C. D. Phelps, "Computational optimal control of nonlinear systems with parameter uncertainty," Ph.D. dissertation, Dept. Appl. Math. Stat., Univ. California, Santa Cruz, CA, USA, 2014.

[9] H. Chung, E. Polak, J. O. Royset, and S. Sastry, "On the optimal detection of an underwater intruder in a channel using unmanned underwater vehicles," *Naval Res. Logistics*, vol. 58, no. 8, pp. 804–820, 2011.

[10] J. C. Foraker, "Optimal search for moving targets in continuous time and space using consistent approximations," Ph.D. dissertation, Naval Postgraduate School, Monterey, CA, USA, 2011.

[11] J. Ruths and S. Li, Jr., "Optimal control of inhomogeneous ensembles," *IEEE Trans. Autom. Control*, vol. 57, no. 8, pp. 2021–2032, Aug. 2012.

[12] Y. Matsuno, T. Tsuchiya, J. Wei, I. Hwang, and N. Matayoshi, "Stochastic optimal control for aircraft conflict resolution under wind uncertainty," *Aerosp. Sci. Technol.*, vol. 43, pp. 77–88, Jun. 2015.

[13] W. Liu and I. Hwang, "Probabilistic aircraft midair conflict resolution using stochastic optimal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 37–46, Feb. 2014.

[14] Y. Zhou *et al.*, "Multivariate probabilistic collocation method for effective uncertainty evaluation with application to air traffic flow management," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 10, pp. 1347–1363, Oct. 2014.

[15] J. Xie *et al.*, "Effective and scalable uncertainty evaluation for large-scale complex system applications," in *Proc. IEEE Winter Simulat. Conf. (WSC)*, Savannah, GA, USA, 2014, pp. 733–744.

[16] J. Xie *et al.*, "Effective uncertainty evaluation in large-scale systems," in *Principles of Cyber Physical Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[17] G. E. Box, J. S. Hunter, and W. G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery*, vol. 2. New York, NY, USA: Wiley, 2005.

[18] *Fractional Factorial Design Specifications and Design Resolution*. Accessed on May 15, 2017. [Online]. Available: http://www.itl.nist.gov/div898/handbook/pri/section3/pri3344.htm

[19] S. S. Isukapalli, "Uncertainty analysis of transport-transformation models," Ph.D. dissertation, Dept. Chem. Biochem. Eng., Rutgers Univ.-New Brunswick, New Brunswick, NJ, USA, 1999.

[20] R. E. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[21] F. Heiss and V. Winschel, "Likelihood approximation by numerical integration on sparse grids," *J. Econometrics*, vol. 144, no. 1, pp. 62–80, 2008.