2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

# On the Convergence of Reinforcement Learning in Nonlinear Continuous State Space Problems

Raman Goyal, Suman Chakravorty, Ran Wang, Mohamed Naveed Gul Mohamed

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9682829

***Abstract*— We consider the problem of Reinforcement Learning for nonlinear stochastic dynamical systems. We show that in the RL setting, there is an inherent "Curse of Variance" in addition to Bellman's infamous "Curse of Dimensionality", in particular, we show that the variance in the solution grows factorial-exponentially in the order of the approximation. A fundamental consequence is that this precludes the search for anything other than "local" feedback solutions in RL, in order to control the explosive variance growth, and thus, ensure accuracy. We further show that the deterministic optimal control has a perturbation structure, in that the higher order terms do not affect the calculation of lower order terms, which can be utilized in RL to get accurate local solutions.**

***Index Terms*— RL, Optimal control, Nonlinear systems**

## I. Introduction

A large class of decision making problems under uncertainty can be posed as a nonlinear stochastic optimal control problem that requires the solution of an associated Dynamic Programming (DP) problem, however, as the state dimension increases, the computational complexity goes up exponentially in the state dimension [1]: the manifestation of Bellman's infamous "curse of dimensionality (CoD)" [2]. To understand the CoD better, consider the simpler problem of estimating the cost-to-go function of a feedback policy $\mu_t(\cdot)$. Let us further assume that the cost-to-go function can be "linearly parametrized" as: $J_t^\mu(x) = \sum_{i=1}^M \alpha_t^i \phi_i(x)$, where the $\phi_i(x)$'s are some *a priori* basis functions. Then the problem of estimating $J_t^\mu(x)$ becomes that of estimating the parameters $\bar{\alpha}_t = [\alpha_t^1, \cdots, \alpha_t^M]$. This can be shown to be the recursive solution of the linear equations $\bar{\alpha}_t = \bar{c}_t + L_t \bar{\alpha}_{t+1}$, where $\bar{c}_t = [c_t^i]$, with $c_t^i = \int c(x, \mu_t(x))\phi_i(x)\, dx$, and $L_t^{ij} = \int\int p^{\mu_t}(x'|x)\phi_i(x')\phi_j(x)\, dx'\, dx, \quad i, j \in \{1, \ldots, M\}$, where $p^{\mu_t}(./.)$ is the transition density of the Markov chain under policy $\mu_t$. This can be done using numerical quadratures given knowledge of the model $p^\mu(x'/x)$, termed Approximate DP (ADP), or alternatively, in Reinforcement Learning (RL), simulations of the process under the policy $\mu_t$, $x_t \xrightarrow{\mu_t(x_t)} x_{t+1} \to \cdots$, is used to get an approximation of the $L_t^{ij}$ by sampling, and solve the equation above either batchwise or recursively [3], [1]. But, as the dimension $d$ increases, the number of basis functions and the number of evaluations required to evaluate the integrals go up exponentially. There has been recent success using the Deep RL paradigm where deep neural networks are used as nonlinear function approximators to keep the parametrization tractable [4], [5], [6], [7], [8], however, the training times required for these approaches is still prohibitive. Hence, the primary problem with ADP/ RL techniques is the CoD inherent in the complex representation of the cost-to-go function, and the exponentially large number of evaluations required for its estimation. In this paper, we show that there is an additional "Curse of Variance" that afflicts the RL solution, the fact that the variance grows at a factorial-exponential rate in the order of the approximation, that precludes us from solving for higher order approximations of the feedback law. Prior research has focused in some details on the sample complexity of RL for finite control problems [9], [10], [11], and the case of optimal Linear Quadratic Control (LQR) in continuous state and control spaces [12], [13]. We study the general nonlinear problem, and show the scale of the variance inherent in an RL estimate. Albeit anecdotal and empirical evidence of the variance phenomenon has always existed in the RL literature [14], we believe we are the first to exactly enumerate the factorial-exponential growth and its consequences: it is necessary that we look for local solutions in order to find accurate solutions, that stochastic control problems are fundamentally intractable, and the best we can hope for is a suitably accurate deterministic approximation (see points 1-5 of contributions below). However, this does not mean we need to give up on global optimality. In [15], we established the local optimality of the deterministic feedback law, in that the nominal (zero noise) action, and the linear feedback action, of the optimal stochastic and deterministic policies are close to fourth order in a small noise parameter, starting at any given state, which, when allied with replanning, recovers a near-optimal solution. Thus, a local solution allied with replanning is an efficient and near-optimal way to solve nonlinear stochastic control problem rather than solve for a global (higher order) solution. In particular, one should look for local deterministic solutions to ensure accuracy (which are locally optimal due to the results of [15]), and re-plan when necessary as in Model Predictive Control (MPC), to recover global optimality.

We summarize our contributions as follows.

1. It is fundamentally intractable to solve for a high order approximation of a feedback law for optimal control via RL (global/ nonlocal), since the variance of the solution grows factorial-exponentially in the order of the approximation.
2. The deterministic problem has a perturbation structure, in that higher order terms do not affect the calculation of lower order terms, and thus, when a model is known, the calculations can be closed at any order without affecting the accuracy of the lower order terms.

The authors are with the Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843 USA. {ramaniitrgoyal92, schakrav, rwang0417, naveed}@tamu.edu

3. If the deterministic problem is solved in an RL fashion, then an accurate solution can be found, if and only if we concentrate on a suitably local solution, enforced via constraining the random exploration around a nominal trajectory.
4. The perturbation structure and locality of the solution are key to an accurate RL implementation.

**Outline of Paper.** The rest of the document is organized as follows: Section II outlines the Problem Formulation, Section III studies the convergence of Policy evaluation in a finite time RL setting and the resulting variance in the solution. Section IV derives a perturbation structure inherent to the deterministic policy evaluation problem, and shows how to leverage this for accurate local RL solutions. Section V gives empirical results in a simple example to validate the theoretical development.

## II. Problem Formulation

The problem of control under uncertainty can be formulated as a stochastic optimal control problem in the space of feedback policies. We assume here that the uncertainty in the problem lies in the system's process model.

***System Model:*** For a dynamic system, we denote the state and control vectors by $x_t \in \mathbb{X} \subset \mathbb{R}^{n_x}$ and $u_t \in \mathbb{U} \subset \mathbb{R}^{n_u}$ respectively at time $t$. The motion model $h : \mathbb{X} \times \mathbb{U} \times \mathbb{R}^{n_u} \to \mathbb{X}$ is given by the equation

$$x_{t+1} = h(x_t, u_t, w_t);\ w_t \sim \mathcal{N}(0, \Sigma_{w_t}), \tag{1}$$

where $\{w_t\}$ are zero mean independent, identically distributed (i.i.d) random sequences with variance $\Sigma_{w_t}$.

***Stochastic optimal control problem:*** The stochastic optimal control problem for a dynamic system with initial state $x_0$ is defined as:

$$J_{\pi^*}(x_0) = \min_{\pi}\ E\left[\sum_{t=0}^{T-1} c(x_t, \pi_t(x_t)) + g(x_T)\right], \tag{2}$$

$s.t.\ x_{t+1} = h(x_t, \pi_t(x_t), w_t)$, where: the optimization is over feedback policies $\pi := \{\pi_0, \pi_1, \ldots, \pi_{T-1}\}$ and $\pi_t(\cdot)$: $\mathbb{X} \to \mathbb{U}$ specifies an action given the state, $u_t = \pi_t(x_t)$; $J_{\pi^*}(\cdot) : \mathbb{X} \to \mathbb{R}$ is the cost function on executing the optimal policy $\pi^*$; $c(\cdot,\cdot) : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is the one-step cost function; $g(\cdot) : \mathbb{X} \to \mathbb{R}$ is the terminal cost function; $T$ is the horizon of the problem. The solution to the above problem is given by the Dynamic Programming equation:

$$J_t(x) = \min_u [c(x,u) + E[J_{t+1}(x')]], \tag{3}$$

where $x' \sim p(./x, u)$, and $p(./x, u)$ denotes the transition density of the state at the next time step arising from the system dynamics, given the control $u$ is taken at state $x$, solved with the terminal condition $J_T(x) = g(x)$. The DP equation can be solved using the so-called Policy Iteration method, where given a time varying feedback policy $\pi_t^{(k)}(\cdot)$, one first solves for the cost function corresponding to it:

$$J_t^{(k)}(x) = c(x, \pi_t^{(k)}(x)) + E[J_{t+1}^{(k)}(x')], \tag{4}$$

where $x' \sim p(./x, \pi_t^{(k)}(x))$, and the above equation is solved with the terminal condition $J_T^{(k)}(x) = g(x)$, which is followed by a policy improvement step:

$$\pi_t^{(k+1)}(x) = \arg\min_u [c(x,u) + E[J_{t+1}^{(k)}(x')]], \tag{5}$$

where $x' \sim p(./x, u)$. This process is followed till convergence, starting with some initial time varying policy $\pi_t^{(0)}(x)$ [1].

## III. Convergence of Policy Evaluation

In the following, we shall concentrate on the Policy Evaluation (PE) part of Policy iteration, in particular, a single Policy Evaluation step, to show the convergence issues inherent, and at the end of this section, outline the issues arising from the dynamic recursion. We shall consider a synchronous model of computing, i.e., where all the experiments are done first, and the cost functions at any step updated using all the experiments. We here consider the deterministic scalar state case for simplicity. Let us rewrite the policy evaluation equation from above as:

$$J_t(x_t) = c_t(x_t) + J_{t+1}(f(x_t)), \tag{6}$$

where the deterministic dynamics are $x_{t+1} = f(x_t)$, and given that the terminal cost $J_T(x_T) = g(x_T)$. In the context of Policy Iteration, the dynamics corresponds to the closed loop under some feedback policy $\pi(x)$, i.e., $f(x) = h(x, \pi(x), 0)$. Note that the policy is, in general, time varying, and thus, the dynamics should also be time varying. But we consider time invariant dynamics for simplicity, and all the results obtained below generalize to the time varying case in a straightforward fashion.

*a) Computing Model:* Suppose that we have basis functions $\{\phi^1(x), \cdots \phi^N(x)\}$ such that any $J_t(x) = \sum_i \alpha_t^i \phi^i(x)$ for suitably chosen coefficients $\alpha_t^i$, and such that $c_t(x) = \sum_i c_t^i \phi^i(x)$. Suppose now that we are given $R$ samples from the dynamical system, say $\{x_1^{(k)}, \cdots x_t^{(k)}\}$, for $k = 1, \cdots R$, that are sampled from some time varying density $p_t(\cdot)$. We leave the question of what this density ought to be to later on in our development. Given the samples, we write:

$$J_t(x_t^{(k)}) = c(x_t^{(k)}) + J_{t+1}(f(x_t^{(k)})) + v_t^{(k)}, \tag{7}$$

where $v_t^{(k)}$ is an independent identically distributed (i.i.d.) noise sequence for all time steps $t$. Representing the cost functions in terms of the basis functions, we obtain:

$$\bar{\alpha}_t \phi_t^{(k)} = \bar{c}_t \phi_t^{(k)} + \bar{\alpha}_{t+1} \phi_{t+1}^{(k)} + v_t^{(k)}, \tag{8}$$

where $\bar{\alpha}_t = \begin{bmatrix}\alpha_t^1 \cdots \alpha_T^N\end{bmatrix}$, $\bar{c}_t = \begin{bmatrix}c_t^1 \cdots c_t^N\end{bmatrix}$, $\phi_t^{(k)} = \begin{bmatrix}\phi^1(x_t^{(k)}) \\ \vdots \\ \phi^N(x_t^{(k)})\end{bmatrix}$, and $\phi_{t+1}^{(k)} = \begin{bmatrix}\phi^1(x_{t+1}^{(k)}) \\ \vdots \\ \phi^N(x_{t+1}^{(k)})\end{bmatrix}$, where note that $x_{t+1}^{(k)} = f(x_t^{(k)})$.

*b) RL as Least Squares:* Then, we may view the above as the following least squares problem:

$$\bar{\alpha}_t^R = \arg\min_{\bar{\alpha}_t} ||\bar{\alpha}_t \Phi_t^R - \bar{c}_t \Phi_t^R - \bar{\alpha}_{t+1}^R \Phi_{t+1}^R||^2, \tag{9}$$

where $\Phi_t^R = \begin{bmatrix}\phi_t^{(1)}, \cdots, \phi_t^{(R)}\end{bmatrix}$, and the supercase $R$ is used to denote the solution after $R$ samples. In this case, we are

sweeping back in time starting at the final time $T$, and thus, it is assumed above that we have solved for $\alpha_{t+1}^R$ already. The solution to this problem is standard and given by:

$$\bar{\alpha}_t^R = \bar{c}_t + \bar{\alpha}_{t+1}^R \Phi_{t+1}^R \Phi_t^{R'} (\Phi_t^R \Phi_t^{R'})^{-1}, \tag{10}$$

where $A'$ denotes the transpose of a matrix $A$.

Now, we shall establish some properties of the least squares (LS) solution above in the context of RL. First, let us find the "true" solution of the Policy Evaluation equation (6). First, we make the following assumption.

*Assumption 1:* There exist a set of constants $\beta_{ij}$, $i = 1, 2, \cdots N$, and $j = 1, \cdots N'$, where $N' > N$, such that for any $\phi^i(f(x)) = \sum_{j=1}^{N'} \beta^{ij} \phi^j(x)$.

The reason $N' > N$ is that, in general, unless $f(\cdot)$ is linear, it will require more basis functions to represent $J_t(x)$ than $J_{t+1}(x)$. The reason is that if $\phi^i$ is an $i$ degree polynomial, then $\phi^i(f(x))$ will be a $ki$ degree polynomial if $f(\cdot)$ is a $k$ degree polynomial. Thus, in general, we will need an expanding basis to represent the functions $J_t(\cdot)$ as we sweep back in time from $T$ to 0. Then, we can characterize the "true" solution to the Policy Evaluation equation (6) as follows.

*Proposition 1:* The true solution to the policy evaluation equation (6) is given by: $\bar{\alpha}_t^* = \bar{c}_t + \bar{\alpha}_{t+1}^* B_t$, where $B_t = \begin{bmatrix} \beta^{11}, \cdots, \beta^{1N_t} \\ \ddots \\ \beta^{N_{t+1}1}, \cdots, \beta^{N_{t+1}N_t} \end{bmatrix}$, where $N_{t+1}$ is the number of basis functions required to represent $J_{t+1}(\cdot)$ and $N_t$ is the number of basis functions required to represent $J_t(\cdot)$.

*Proof: arxiv report* [16].

Next, we show that the RL least squares solution (10) converges to the above true solution in the mean square sense as the number of samples $R$ becomes large.

*Proposition 2:* ***PE convergence.*** Let Assumption 1 hold. Further, let the number of basis functions at time $t$ required be $N_t$. Given that all the required basis functions at time $t$ are considered, the RL least square estimate (10) converges to the true solution in the mean square sense.

*Proof: arxiv report* [16].

The above result shows that the RL least squares procedure is a randomized approximation of the PE equation. However, the above result follows under an idealized situation when all necessary basis functions are considered, and $R$ becomes very large. Thus, in the following, we characterize the bias and the variance of the estimate, which in turn will allow us to find the sample complexity of the estimates.

*Corollary 1: Bias in RL estimate* (10). Suppose that the number of basis functions required at time $t$ is $N_t$ and only $N < N_t$ are used. Then, the RL least squares estimate (10) is biased for all $\tau < t$.

*Proof: arxiv report* [16].

Next, we consider the variance of the estimate. The key role here is played by the Gram matrix $\mathcal{G}_t = [< \phi_i, \phi_j >_t]$, $i, j = 1 \cdots N_t$, where $< ., . >_t$ denotes the inner product with respect to the sampling distribution $p_t(.)$. In general, the variance of the solution is determined by the variance in the R-sample empirical Gram matrix estimate $\mathcal{G}_t^R = \frac{1}{R} \Phi_t^R \Phi_t^{R'}$. In the following, we characterize the variance of the empirical Gram matrix $\mathcal{G}_t^R$ for suitable choice of basis functions and sampling distribution.

*Assumption 2:* Let the basis functions used at time $t$ be $\{\phi^1, \cdots \phi^{N_t}\}$. We assume that there exists a constant matrix $H_t$ such that $\begin{bmatrix} \phi^1 \\ \vdots \\ \phi^{N_t} \end{bmatrix} = H_t \begin{bmatrix} 1 \\ \vdots \\ x^{M_t} \end{bmatrix}$, i.e., the basis functions at time $t$ can be represented as $M_t$ degree polynomials.

Define $H_t = [H_1, \cdots H_{M_t}] = \begin{bmatrix} H^1 \\ \vdots \\ H^{N_t} \end{bmatrix}$, i.e., we define the rows and columns of the matrix $H_t$. The covariance of the error in the LS estimate is given by $P_t^R = \sigma_v^2 (G_t^R)^{-1}$, where $G_t^R = R\mathcal{G}_t^R$, and $\mathcal{G}_t^R = \frac{1}{R} \Phi_t^R \Phi_t^{R'}$ (see Proof of Proposition 2). Now, we can characterize the size of the error in the LS estimate as a function of number of samples $R$ required.

*Theorem 1:* **Variance of RL least squares estimate.** Let $x_t \sim \mathcal{N}(0, \sigma_X^2)$, i.e. the sampling distribution is zero mean Gaussian with variance $\sigma_X^2$. Let $\beta < 1, \delta > 0$, and $n < \infty$ be given. Then, to probabilistically bound the norm of the error covariance:

$$Prob(||P_t^R|| \leq \delta) > 1 - 2e^{-n^2/2}, \tag{11}$$

the number of samples required are:

$$R > \max[(\frac{n}{\beta} CC')^2 \sigma_{2M_t}^2, \frac{\sigma_v^2 C}{\delta(1-\beta)}], \tag{12}$$

where $\sigma_{2M_t}^2 = [(4M_t - 1)!! - (2M_t - 1)!!^2] \sigma_X^{4M_t}$,
and $C$ and $C'$ are constants such that $||\mathcal{G}_t^{-1}|| \leq C$, and $||H_{M_t}|| ||H^{N_t}|| \leq C'$.

*Proof: arxiv report* [16].

The above result establishes the number of samples required to get an accurate RL least squares estimate. Next, we shall see the implications of the above result for particular choices of basis functions.

### A. Sample Complexity

*a) Monomial Basis:* For a monomial basis, $C' = 1$ since $H$ is the identity matrix and $N_t = M_t$. Thus, the number of samples has to satisfy:

$$R \sim O([(4N_t - 1)!! - (2N_t - 1)!!^2] \sigma_X^{4N_t}), \tag{13}$$

for the LS error to be small enough.

*b) Hermite (Orthonormal) Basis:* It is well known that the Hermite polynomials form the orthonormal basis for Gaussian sampling distributions [17]. Noting that $M_t = N_t$, in the case of Hermite polynomials, owing to their orthonormality, one can show that $C' = \frac{1}{N_t!^2 \sigma_X^{4N_t}}$, and this, in turn, implies that for the Hermite basis, the number of samples $R$ need to satisfy:

$$R \sim O\left( \frac{(4N_t - 1)!! - (2N_t - 1)!!^2}{N_t!^2} \right), \tag{14}$$

for the LS error to be small enough.

*c) Nonlinear Basis:* In this case, we mean that the cost function $J_t(x) = h(x, \theta_t)$, where $h(x, \theta)$ is a suitable nonlinear approximation architecture, such as a (deep) neural net, parametrized nonlinearly by the (vector) parameter $\theta$. The PE equation in this case becomes:

$$h(x, \theta_t) = c(x) + h(f(x), \theta_{t+1}), \tag{15}$$

where $\theta_t$ parametrizes the cost function at time $t$, and the same holds for $\theta_{t+1}$. It is reasonable to assume, given the change in the parameter $\theta$ between consecutive steps is small enough, that:

$$h(x, \theta_t) \approx h(x, \theta_{t+1}) + \sum_{i=1}^{N_t} H_{t+1}^i(x)\delta\theta_i, \tag{16}$$

where $H_{t+1}^i(x) = \frac{\partial h(x,\theta)}{\partial \theta_i}|_{\theta_{t+1}}$, where $\theta_i$ are the components of the vector parameter $\theta$. Rewriting the policy evaluation equation, one obtains:

$$\begin{aligned} &H_{t+1}^1(x)\delta\theta_1 + \cdots + H_{t+1}^{N_t}(x)\delta\theta_{N_t} \\ &\quad = c(x) + \underbrace{[h(f(x), \theta_{t+1}] - h(x, \theta_{t+1})]}_{\delta h_{t+1}(x)} + v, \end{aligned} \tag{17}$$

where $v$ is a noise term, which may be written in matrix form for $R$ samples as: $[\delta\theta_1 \cdots \delta\theta_{N_t}] \underbrace{\begin{bmatrix} H_{t+1}^1(x_t^{(1)}) & \cdots & H_{t+1}^1(x^{(R)_t}) \\ \vdots & \vdots & \vdots \\ H_{t+1}^{N_t}(x_t^{(1)}) & \cdots & H_{t+1}^{N_t}(x_t^{(R)}) \end{bmatrix}}_{\mathcal{H}_t^R}$

$= \underbrace{[c(x^{(1)}) \ \cdots \ c(x^{(R)})]}_{\mathcal{C}_t^R} + \underbrace{[\delta h_{t+1}(x^{(1)}) \ \cdots \ \delta h_{t+1}(x^{(R)})]}_{\delta\mathcal{H}_t^R}$

$+ \underbrace{\begin{bmatrix} v_t^{(1)} & \cdots & v_t^{(R)} \end{bmatrix}}_{V_t^R}$, where we assume that the noise terms $v_t^{(i)}$ are i.i.d with variance $\sigma_v^2$ as before. The least squares solution to the above equation is, as usual, given by:

$$\delta\theta_t^R = \mathcal{C}_t^R\mathcal{H}_t^{R'}(\mathcal{H}_t^R\mathcal{H}_t^{R'})^{-1} + \delta H_t^R\mathcal{H}_t^{R'}(\mathcal{H}_t^R\mathcal{H}_t^{R'})^{-1}. \tag{18}$$

Denote the empirical Gram matrix for the instantaneous basis functions, $H_{t+1}^i(x)$, at time $t+1$ as: $\mathcal{G}_t^R = \frac{1}{R}\mathcal{H}_t^R\mathcal{H}^{R'}$, and thus, the covariance of the error in the LS solution is given by $\frac{\sigma_v^2}{R}(\mathcal{G}_t^R)^{-1}$. Note that this situation is no different from the linear case considered previously, in that we are approximating the change in the cost function via the change in the parameter $\theta$ at time $t+1$, and the instantaneous basis functions $H_{t+1}^i(x)$. The primary difference with the linear case is that our basis functions $H_{t+1}^i(x)$ change with time unlike the fixed basis in the linear case. Suppose that we have: $\begin{bmatrix} H_{t+1}^1(x) \\ \vdots \\ H_{t+1}^{N_t}(x) \end{bmatrix} = \mathcal{D}_{t+1}\begin{bmatrix} 1 \\ x \\ \vdots \\ x^{M_t} \end{bmatrix}$, then we are back to the conditions enumerated in Theorem 1. Therefore, it follows that the number of samples $R$ required such that we get a small enough error in the solution should be:

$$R \sim O([(4M_t - 1)!! - (2M_t - 1)!!^2]\sigma_X^{4M_t}). \tag{19}$$

**Discussion.** It can be seen clearly from above that choosing orthonormal (o.n.) polynomials to the sampling distribution greatly reduces the variance of the RL least squares estimate (10). Further, the variance of the estimate if we use unnormalized polynomials is very high. The situation is no different even when using a nonlinear basis. Typically in RL, one uses rollouts and the basis functions are almost never chosen to satisfy orthonormality. Further, in general, since the rollout sampling distributions $p_t(\cdot)$ need not be Gaussian, finding such o.n. basis functions is a challenge in itself since one does not have an idea about these sampling distributions in advance, or they are never known explicitly. However, in our opinion, the sampling distributions can, and should, be chosen at our convenience and need not arise from rollouts if viewed in the context of the Galerkin interpretation of Proposition 1. In fact, the above analysis suggests that if we are to reduce the variance of the estimates, the sampling distributions should not be chosen from rollouts.

A significantly more intractable problem arises due to the exponential growth of the basis functions required at every time step. Suppose that the dynamics could be represented by a $k$ degree polynomial, and the terminal cost function was degree $N_T$, then the number of basis functions required at time $t$ is $k^{T-t}N_T$. When coupled with the variance estimate above, one can see that, even with o.n. basis, this leads to an explosive variance growth: a factorial-exponential rate of growth.

## IV. A Perturbation Structure for PE

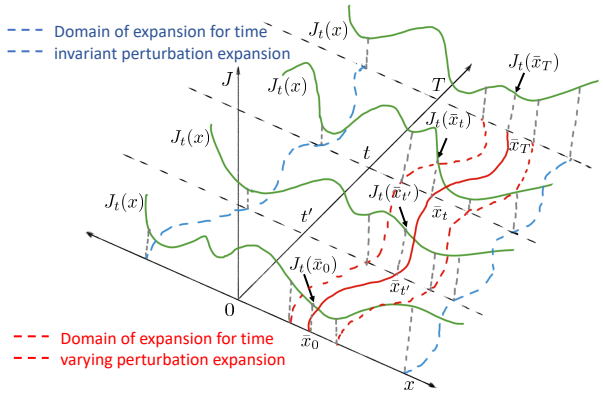


Fig. 1: Time-varying versus time-invariant perturbation expansion.

In this section, we present a construct that ideally allows a perturbation structure to the Policy Evaluation equations, i.e., a structure where the higher order terms do not affect the evolution of the lower order terms which implies that we can close our computations at any desired order without incurring an error.

Recall the Policy evaluation equation (6). Now, consider that we are given a nominal trajectory under the dynamics,

say $\bar{x}_t = f(\bar{x}_{t-1})$, $t = 0, 1, \cdots T$ given some initial condition $\bar{x}_0 = x_0$. Next, expand the dynamics about this nominal trajectory as: $f(x_t) = f(\bar{x}_t + \delta x_t) = f(\bar{x}_t) + F_t^1 \delta x_t + F_t^2 \delta x_t^2 + \cdots$, where $F_t^i$ denotes the $i^{th}$ term in the expansion of the dynamics around the nominal trajectory. Similarly: $J_{t+1}(f(x_t)) = J_{t+1}(f(\bar{x}_t + \delta x_t)) = J_{t+1}(f(\bar{x}_t)) + K_{t+1}^1(F_t^1 \delta x_t + \frac{1}{2} F_t^2 \delta x_t^2 + \cdots) + \frac{1}{2} K_{t+1}^2 (F_t^1 \delta x_t + F_t^2 \delta x_t^2 + \cdots)^2 + \cdots$, where $K_{t+1}^i$ denotes the $i^{th}$ term in the Taylor expansion of $J_{t+1}(f(x_t))$ around the nominal. Similarly the incremental cost function $c_t(x_t) = c_t(\bar{x}_t) + C_t^1 \delta x_t + \frac{1}{2} C_t^2 \delta x_t^2 + \cdots$, and the optimal cost function at time $t$, $J_t(x_t) = J_t(\bar{x}_t) + K_t^1 \delta x_t + \frac{1}{2} K_t^2 \delta x_t^2 + \cdots$. Substituting the above expressions into the policy evaluation equation (6), we obtain:

$$\begin{aligned} J_t(x_t) &= \bar{J}_t + K_t^1 \delta x_t + \frac{1}{2} K_t^2 \delta x_t^2 + \cdots \\ &= (\bar{c}_t + C_t^1 \delta x_t + \frac{1}{2} C_t^2 \delta x_t^2 + \cdots) + \bar{J}_{t+1} \\ &\quad + K_{t+1}^1 (F_t^1 \delta x_t + \frac{1}{2} F_t^2 \delta x_t^2 + \cdots) \\ &\quad + \frac{1}{2} K_{t+1}^2 (F_t^1 \delta x_t + F_t^2 \delta x_t^2 + \cdots)^2 + \cdots, \end{aligned}$$

where $\bar{J}_t = J(\bar{x}_t)$, $c(\bar{x}_t) = \bar{c}_t$ and $\bar{J}_{t+1} = J_{t+1}(\bar{x}_{t+1}) = J_{t+1}(f(\bar{x}_t))$. Now, grouping the different terms of $\delta x_t^i$ on both sides of the equation allows for writing the following vector-matrix form equation as:

$$\begin{aligned} [K_t^1 \, K_t^2 \cdots] &= [C_t^1 \, C_t^2 \cdots] \\ &\quad + [K_{t+1}^1 \, K_{t+1}^2, \cdots] \underbrace{\begin{bmatrix} F_t^1 & F_t^2 & \cdots \\ 0 & (F_t^1)^2 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \vdots \\ 0 & 0 & \cdots \end{bmatrix}}_{\mathcal{B}_t}, \end{aligned} \quad (20)$$

note the equations have a beautiful perturbation/ upper triangular structure. Practically, this means that we can close our computations at any order we desire without worrying about the effect of the higher order terms on the lower order terms, given we have knowledge of the dynamics, and hence, the Taylor coefficients $F_t^i$.

A special case of the above equation is when $f(0) = 0$, and the nominal trajectory is simply $\bar{x}_t = 0$. In such a case the expansion is about a nominal trajectory that stays at the origin. In such a case, the Taylor coefficients, rather than being time varying, will be time invariant, i.e., $F^1, F^2, \cdots$ etc. Typically, we are given problems where the initial state $x_0 \neq 0$, and can be far from the origin. In such a case, we may see that the number of terms required for a static expansion, i.e., about $\bar{x}_t = 0$, will require far more terms than would an expansion that was centered on a nominal trajectory starting at $x_0$. The situation is illustrated in Fig. 1. Thus, it is much more efficient to seek the time varying expansion above. In particular, we shall explore the implication further when we solve the perturbed policy evaluation (PPE) equation (20).

*Remark 1:* Note that given the Taylor coefficients $F_t^i$, the second and higher rows of the "dynamics" matrix encoding the structure of the PPE equation are perfectly known. This knowledge can be used to solve the PPE equation in a "model based" fashion, as opposed to a model-free approach, where this structure is actually teased out of the data from the system.

### A. RL type solution to the PPE

First, we show a model-based solution to the PPE equation (20), i.e., one where we explicitly estimate the first $M$ Taylor coefficients of the dynamics $\mathcal{F}_t \equiv [F_t^1, F_t^2, \cdots F_t^M]$ which are then substituted into (20) to solve the PE equation. Next, we show how this can be extended to the model-free case: one where we do not solve for $\mathcal{F}_t$, and instead directly solve (20) and infer the matrix $\mathcal{B}_t$ from the system data.

We can write the following approximation, after neglecting the higher order terms beyond $\delta x_t^M$:

$\delta x_{t+1}^{(i)} \approx \mathcal{F}_t \begin{bmatrix} \delta x_t^{(i)} \\ (\delta x_t^{(i)})^2 \\ \vdots \\ (\delta x_t^{(i)})^M \end{bmatrix} + v_t^{(i)}$, where as before $v_t^{(i)}$ is an i.i.d. noise sequence and $i = 1, 2 \cdots R$. A least squares estimate of $\mathcal{F}_t$ is quite straightforward and may be written as:

$$\mathcal{F}_t^R = \delta X_{t+1}^R \delta \chi_t^{R'} (\delta \chi_t^R \delta \chi_t^{R'})^{-1}, \quad (21)$$

where $\delta \chi_t^R = \begin{bmatrix} \delta x_t^{(1)} & \cdots & \delta x_t^{(R)} \\ \vdots & \vdots & \vdots \\ (\delta x_t^{(1)})^M & \cdots & (\delta x_t^{(R)})^M \end{bmatrix}$, and $\delta X_{t+1}^R = \begin{bmatrix} \delta x_{t+1}^{(1)} & \cdots & \delta x_{t+1}^{(R)} \end{bmatrix}$, where $\delta x_t^{(i)} = x_t^{(i)} - \bar{x}_t$, and $\delta x_{t+1}^{(i)} = f(x_t^{(i)}) - \bar{x}_{t+1}$, where $\bar{x}_{t+1} = f(\bar{x}_t)$. Further, we assume that $\delta x_t^{(i)} \sim \mathcal{N}(0, \sigma_X^2)$. Thus, the data is obtained by perturbing the system from the nominal trajectory.

The following development characterizes the error in the LS solution (21) incurred from neglecting the higher order terms of the dynamics (beyond $\delta x_t^M$) and shows that it can be made arbitrarily small by choosing the perturbation $\delta x_t$ to be suitably small. In the following, unlike in Section III, the Gram matrix $\mathcal{G}$ size will not change, since we are looking at order $M$ approximation throughout time.

*Lemma 1:* Let $\Delta_t^R = [\Delta_t^{R,1}, \Delta_t^{R,2} \cdots \Delta_t^{R,M}]$, where $\Delta_t^{R,l} = \frac{1}{R} \sum_{k>M} \sum_{i=1}^R F_t^k (\delta x_t^{(i)})^k (\delta x_t^{(i)})^l$. Let the empirical Gram matrix $\mathcal{G}^R = [\mathcal{G}_{ij}^R]$, where $i, j = 1, 2, \cdots M$, and $\mathcal{G}_{ij}^R = \frac{1}{R} \sum_{k=1}^R (\delta x_t^{(k)})^i (\delta x_t^{(k)})^j$. Then, $\mathcal{F}_t^R = \mathcal{F}_t + \Delta_t^R (\mathcal{G}^R)^{-1} + \frac{1}{R} V_t^R \delta \chi_t^{R'} (\mathcal{G}^R)^{-1}$.
*Proof: arxiv report* [16].

The above result makes it clear that our estimates are biased for any finite $R$ (in fact, even for the limit), but if $\Delta_t^R$ is small enough, then this bias can be made small. In the following, we show precisely such a result.

*Proposition 3:* Let $\Delta_t = \lim_R \Delta_t^R$ and let $\mathcal{G} = \lim_R \mathcal{G}^R$. Given any $M$, any $\epsilon > 0$, there exists a variance $\sigma_X^2 < \infty$ such that $|\Delta_t^j| \leq \epsilon |\tilde{f}_t^j|$, where $\mathcal{F}_t \mathcal{G} = [\tilde{f}_t^1, \cdots, \tilde{f}_t^j, \cdots \tilde{f}_t^M]$.

*Proof: arxiv report* [16].

Next, we have the following consequence.

*Corollary 2:* The least squares estimate (21), $\mathcal{F}_t^R \to \mathcal{F}_t + \Delta_t \mathcal{G}^{-1}$ as $R \to \infty$ in mean square sense.

*Proof: arxiv report* [16].

***Accuracy of the Solution.*** To understand the result above, let us rewrite the limiting solution as $\mathcal{F}_t^R = (\mathcal{F}_t\mathcal{G} + \Delta_t)\mathcal{G}^{-1}$. Thus, the limiting solution can be understood as the solution to the linear equation $\mathcal{F}_t\mathcal{G} = \tilde{\mathcal{F}}_t + \Delta_t$ where $\tilde{\mathcal{F}}_t = \mathcal{F}_t\mathcal{G}$. We have shown in Proposition 3 that for small enough variance, $||\Delta_t|| \leq \epsilon||\tilde{\mathcal{F}}_t||$, and thus we can expect a small error on the right hand side of the equation above. However, albeit the error $\Delta_t$ may be small compared to the "signal $\tilde{\mathcal{F}}_t$", the actual error in the solution is affected by the conditioning of the Gram matrix $\mathcal{G}$. In fact, one can show that:

$$\frac{||\mathcal{F}_t^\infty - \mathcal{F}_t||}{||\mathcal{F}_t||} \leq \kappa(\mathcal{G})\epsilon, \tag{22}$$

where $\mathcal{F}_t^\infty = \lim_R \mathcal{F}_t^R$, and $\kappa(\mathcal{G})$ denotes the condition number of the Gram matrix $\mathcal{G}$. Thus, the true pacing item in the accuracy of the solution is the conditioning of the matrix $\mathcal{G}$. In fact, if the input is Gaussian, then the conditioning of the matrix rapidly deteriorates as $M$ increases since the higher moments of a Gaussian increase as $(2M-1)!!\sigma_X^{2M}$, i.e, in factorial-exponential fashion. Thus, in practice, one cannot make $M$ large as the solution becomes highly sensitive due to the ill conditioning of the Gram matrix.

***Variance of Solution.*** As shown previously, the variance of the solution is directly proportional to the variance in the empirical Gram matrix $\mathcal{G}^R$, and thus, the number of samples required is $R \sim \mathcal{O}([(4M-1)!!-(2M-1)!!^2]\sigma_X^{4M})$ as in (13). Hence, the central role in the accuracy and in the variance of the least squares estimate is played by the empirical Gram matrix $\mathcal{G}^R$.

**Model-based or Model-free?** The number of computations in the model-free method is $\approx 2\times$ the computation in the model-based method, since $R$ is typically large. Thus, the model-free approach will result in higher computational efforts or $\approx 2\times$ the variance as compared to model-based methods. *(Detailed Discussion: arxiv report [16].)*

***Orthonormal and Nonlinear Basis Functions:*** As we saw previously, the variance of the RL-LS solution (10) is significantly reduced when using o.n. basis functions such as the Hermite polynomials, see (13) vs (14). Thus, it is of interest to see if any added advantage can be gained by using such o.n. functions in the PPE case.

We may write the Hermite polynomials in terms of the monomials as: $\begin{bmatrix} h_1(\delta x) \\ \vdots \\ h_N(\delta x) \end{bmatrix} = H \begin{bmatrix} \delta x \\ \vdots \\ \delta x^N \end{bmatrix}$, where $h_i(\cdot)$ represent the Hermite polynomials, and $H$ is a lower triangular matrix that encodes the linear transformation from the monomials to the Hermite polynomials. Similarly, let $H^{-1}$ represent the inverse transformation from the Hermite polynomials to the monomials. Then, the PPE equation (20) may be written in the Hermite basis as:

$$\mathcal{K}_t^H = \mathcal{C}_t^H + \mathcal{K}_{t+1}^H \mathcal{B}_t^H, \tag{23}$$

where $\mathcal{K}_t^H$, and $\mathcal{C}_t^H$ consist of the Hermite coefficients (rather than the Taylor coefficients) and $\mathcal{B}_t^H = H\mathcal{F}_t H^{-1}$. Note that $H$ and $H^{-1}$ are lower triangular, while $\mathcal{F}_t$ is upper triangular, and therefore, $\mathcal{B}_t^H$ is fully populated, i.e., the PPE equations in the Hermite basis lose their perturbation/ upper triangular structure. This is the primary shortcoming of the o.n. representation, in particular, we shall show the necessity of the perturbation structure to an accurate solution.

Further, in the nonlinear basis case, nothing changes from the LS problem in (18), since the formulation cannot distinguish deviations from a trajectory. Thus, the equations for the parameters, in general, will be fully coupled, and not have a perturbation structure.

## V. Empirical Results

Thus far in this paper, we have established theoretical results that show that finding a global (higher order) solution for the policy evaluation problem is subject to very high error and variance. In this section, we provide empirical evidence of this explosive growth of errors. In order to accomplish this, we need a system for which we know the true solution to the PE problem. To this end, we assume a discretized system of the form: $x_{t+1} = x_t + \delta(-x_t + \epsilon x_t^3)$, where $\delta$ is the time discretization, let us define the terminal cost to be of the quadratic form: $J_T(x_T) = \alpha x_T^2$ and the incremental cost to be: $c_t(x_t) = cx_t^2$, so that the DP equation can be written as: $J_t(x_t) = cx_t^2 + J_{t+1}(x_{t+1})$. Since this system has a polynomial nonlinearity, it is straightforward to find the cost functions backward in time as polynomials. Given that we have this true answer, we can now find the error in our solution as we sweep back in time, as a function of the number of samples, as well as the exploration parameter $\sigma_X$ which determines whether we explore locally or globally. We show results for polynomial approximations of order $N = M = \{6, 12, 18\}$ as we back propagate 3 steps in time. The required number of basis functions at time $t = \{T, T-1, T-2, T-3\}$ are $M_t = N_t = \{2, 6, 18, 54\}$, respectively. The parameter values assumed are: $\epsilon = 1, \delta = 0.1, c = 10$, and $\alpha = 10$.

For the large exploration case with $\sigma_X = 1$, the mean error and its variance become very high (Fig. 3). Also, notice that even with a large number of samples, the error values do not decrease as in the small noise case of $\sigma_X = 0.1$. Another important observation here is that albeit the performance at $T-1$ is very good, as we sweep back in time, the errors and their variance show explosive growth and become unacceptable.

Thus, albeit preliminary, this empirical evidence suggests that solution accuracy, measured via the mean error and its variance, is indeed adversely affected by higher order approximations and/ or large exploration. Conversely, the only way to ensure accuracy is to have a suitably low order approximation which needs to be enforced by a suitably local exploration. For a "small" exploration noise of $\sigma_X = 0.1$, as

we increase the number of basis functions $M_t$, better results in the mean error are obtained as we propagate back in time due to the increase in number of required basis functions (1st column of Fig. 2). Although the mean error improves till M= 12, there is a marked increase in the variance of the error with increased basis function (2nd column of Fig. 2) from 12 to 18. Also, notice the large number of samples required to reduce the variance of the error even for a one-dimensional problem.

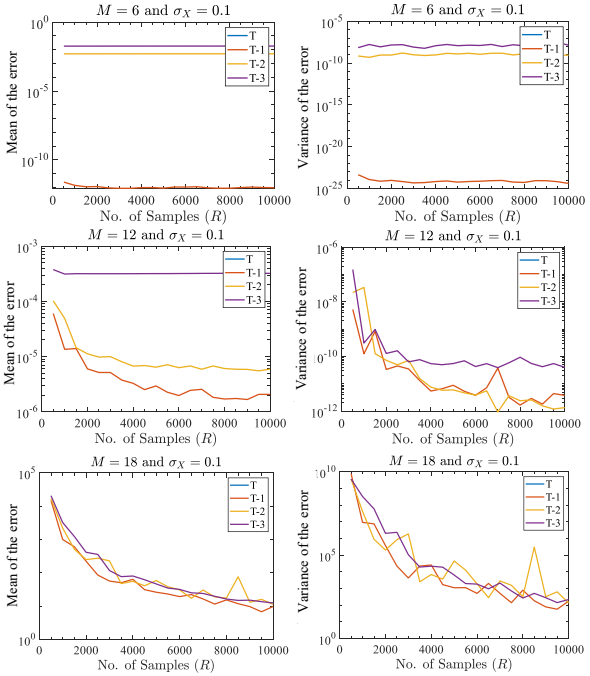


Fig. 2: Results for varying number of samples for exploration parameter $\sigma_X = 0.1$ and different orders of approximation, $M = 6, 12, 18$, for the backward sweep in time till $T - 3$.

## VI. Conclusions and Implications

In this paper, we have studied the inherent structure of the Reinforcement Learning problem. Concentrating on the policy evaluation problem, we have shown that unless we seek local solutions, the answers found are bound to suffer from high variance, and thus, be inaccurate. In particular, the deterministic problem has a perturbation structure that can be exploited to obtain arbitrary accurate local solutions. The primary issue one has to worry about now: "what now for stochastic control?" It is intractable so the best seems to be the deterministic approximation. However, the deterministic solution is optimal locally, and thus, when allied with replanning of the nominal trajectory, we can recover at least near-optimal solution for the stochastic case. We also conjecture that this local replanning based approach is "fundamentally" the best that one can hope to achieve via computation/ RL. Our future research will concentrate on doing an extensive version of the experiments that we started in this paper, and provide further evidence that local RL methods are the ones to pursue.

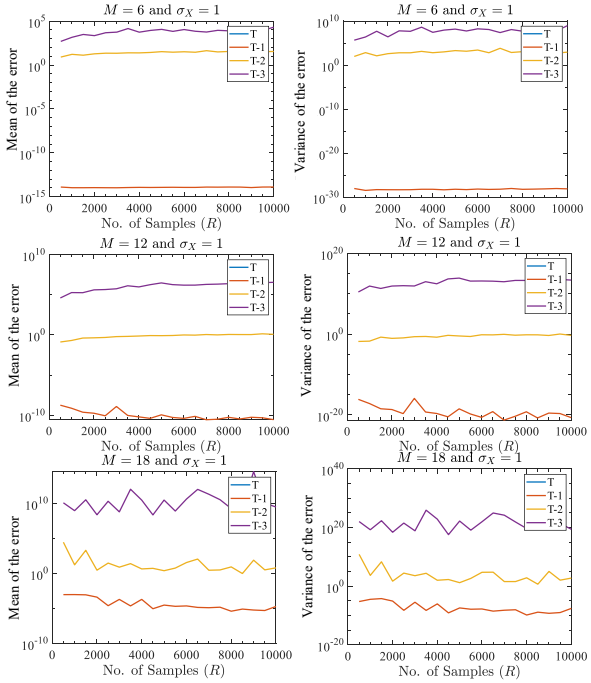


Fig. 3: Results for varying number of samples for $\sigma_X = 1$.

## References


[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control, vols I and II*. Cambridge, MA: Athena Scientific, 2012.

[2] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[3] M. Lagoudakis and R. Parr, "Least squares policy iteration," *Journal of Machine Learning Research*, vol. 4, pp. 1107–1149, 2003.

[4] R. Akrour, A. Abdolmaleki, H. Abdulsamad, and G. Neumann, "Model free trajectory optimization for reinforcement learning," in *Proc. of the ICML*, 2016.

[5] E. Todorov and Y. Tassa, "Iterative local dynamic programming," in *Proc. of the IEEE Int. Symposium on ADP and RL.*, 2009.

[6] E. Theodorou, Y. Tassa, and E. Todorov, "Stochastic differential dynamic programming," in *Proc. of the ACC*, 2010.

[7] S. Levine and P. Abbeel, "Learning neural network policies with guided search under unknown dynamics," in *Advances in NIPS*, 2014.

[8] S. Levine and K. Vladlen, "Learning complex neural network policies with trajectory optimization," in *Proc. of the ICML*, 2014.

[9] M. G. Azar, R. Munos, and B. Kappen, "On the sample complexity of reinforcement learning with a generative model," *arXiv preprint arXiv:1206.6461*, 2012.

[10] S. M. Kakade, "On the sample complexity of reinforcement learning," Ph.D. dissertation, UCL (University College London), 2003.

[11] R. Munos and C. Szepesvári, "Finite-time bounds for fitted value iteration." *Journal of Machine Learning Research*, vol. 9, no. 5, 2008.

[12] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, pp. 1–47, 2019.

[13] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[14] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[15] M. N. G. Mohamed, S. Chakravorty, R. Goyal, and R. Wang, "On the optimal feedback law in stochastic optimal nonlinear control," *arXiv preprint arXiv:2004.01041*, 2020.

[16] R. Goyal, S. Chakravorty, R. Wang, and M. N. G. Mohamed, "On the convergence of reinforcement learning in nonlinear continuous state space problems," *arXiv preprint arXiv:2011.10829*, 2020.

[17] R. Courant and D. Hilbert, *Methods of Mathematical Physics, vol. II*. New York: Interscience publishers, 1953, vol. 336.