On the Feedback Law in Stochastic Optimal Nonlinear Control

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

***Abstract*— We consider the problem of nonlinear stochastic optimal control. This problem is thought to be fundamentally intractable owing to Bellman's "curse of dimensionality". We present a result that shows that repeatedly solving an open-loop deterministic problem from the current state, similar to Model Predictive Control (MPC), results in a feedback policy that is $O(\epsilon^4)$ near to the true global stochastic optimal policy. Furthermore, empirical results show that solving the Stochastic Dynamic Programming (DP) problem is highly susceptible to noise, even when tractable, and in practice, the MPC-type feedback law offers superior performance even for stochastic systems.**

***Index Terms*— Stochastic Optimal Control, Nonlinear Systems, Model Predictive Control.**

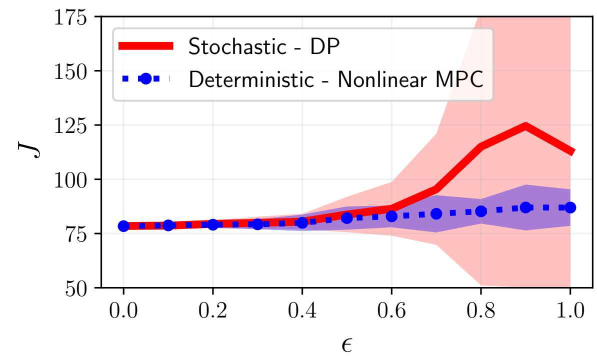


Fig. 1: Practical optimality of the deterministic nonlinear feedback law i.e. MPC on stochastic problems. The data shown are results of solving the stochastic optimal control problem on a nonlinear 1-D system shown in Section V-A (System 1) using the two methods. The lines in the plot denote the mean value and the shade denotes the standard deviation of the corresponding metric. J represents the cost incurred and ϵ is a parameter used to modulate the noise level. It is easy to infer from the figure that in practice, the deterministic feedback law is not only better at handling higher noises, it is also much more reliable, as seen from the low variance in the performance. While DP, in addition to the lack of scalability is also susceptible to noise.

I. INTRODUCTION

In this paper, we consider the problem of finite time nonlinear stochastic optimal control. We present a fundamental result that establishes that repeatedly solving a deterministic optimal control, or open-loop problem, from the current state, results in a feedback policy that is $O(\epsilon^4)$ near-optimal to the optimal stochastic feedback policy, in terms of a small noise parameter ϵ. Although near-optimal, empirical evidence shows that this Model Predictive Control (MPC)-type policy is the best we can do in practice, in the sense that albeit the optimal stochastic law should, in theory, have better performance, solving these problems is highly susceptible to noise, and in reality, the MPC law gives better performance. Thus, this result cuts the Gordian knot of the trade-off between tractability and optimality in stochastic feedback control problems, showing that, in practice, "what is tractable is also optimal". In this paper, we consider the case where a model is available for the control synthesis, we consider the case of data-based control in another paper [26].

A large majority of sequential 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 [4]: the manifestation of the so called Bellman's infamous "curse of dimensionality (CoD)" [3]. 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 done using numerical quadratures given knowledge of the model, termed Approximate DP (ADP), or alternatively, in Reinforcement Learning (RL), simulations of the process under the policy μ_t, $x_t \xrightarrow{\mu_t(x_t)} x_{t+1} \rightarrow \cdots$, is used to get an approximation of the parameters by sampling, and solve the equation above [4], [10]. But, as the dimension d increases, the number of basis functions, and more importantly, the number of evaluations required 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 [2], [11], [12], [23], [24], however, the training times required for these approaches, and the variance of the solutions, 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 resulting in high solution variance which makes them unreliable and inaccurate.

In the case of continuous state, control and observation space problems, the Model Predictive Control [15], [20] approach has been used with a lot of success in the control system and robotics community. For deterministic systems, the process results in solving the original DP problem in a recursive online fashion. However, stochastic control problems, and the control of uncertain systems in general, is still an unresolved problem in MPC. As succinctly noted in [15], the problem arises due to the fact that in stochastic control problems, the MPC optimization at every time step

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

cannot be over deterministic control sequences, but rather has to be over feedback policies, which is, in general, difficult to accomplish since a tractable parameterization of such policies to perform the optimization over, is, in general, unavailable. Thus, the tube-based MPC approach, and its stochastic counterparts, typically consider linear systems [6], [16], [21] for which a linear parametrization of the feedback policy suffices but the methods become intractable when dealing with nonlinear systems [14]. In more recent work, event-triggered MPC [9], [13] keeps the online planning computationally efficient by triggering replanning in an event driven fashion rather than at every time step. We note that event-triggered MPC inherits the same issues mentioned above with respect to the stochastic control problem, and consequently, the techniques are intractable for nonlinear systems.

The basic issue at work above is that, albeit solving the open-loop problem via the Minimum Principle [5] is much easier, solving for the optimal feedback control under uncertainty requires the solution of the DP equation, which is intractable. Moreover, this also begs the question, since all systems are subject to uncertainty, what is the utility of deterministic optimal control?

Contributions: In this work, we establish that the basic MPC approach of solving the deterministic open-loop problem at every time step results in a near-optimal policy, to $O(\epsilon^4)$, for a nonlinear stochastic system. The result uses a perturbation expansion of the cost-to-go function in terms of a perturbation parameter ϵ. We show the global optimality of the open-loop solution obtained by satisfying the Minimum Principle using the classical Method of Characteristics [7] thereby establishing that the MPC feedback law is indeed the optimal deterministic feedback law. We also obtain the true linear feedback gain equations of the optimal deterministic policy as a by-product, which shows it to be very different from the Riccati equation governing a typical LQR perturbation feedback design [5]. Finally, albeit the MPC law is only "near-optimum", our empirical evidence shows that this deterministic law has better performance than the optimal stochastic law, even for stochastic systems where the DP problem can be solved numerically, showing the susceptibility of the DP problem to noise. Thus, in practice, the MPC law is optimal. In contrast to [18], we show fourth order near-optimality to the optimal stochastic solution, the global optimality of the open-loop solution without which MPC is simply a heuristic, and empirical evidence regarding the superiority of MPC to stochastic DP even when DP is feasible.

The rest of the document is organized as follows: Section II states the problem, Section III presents three fundamental results that represent the three legs of the stool that supports the fact that the MPC feedback law is near-optimal, which is established in Section IV. We illustrate our results empirically in Section V using simple 1-dimensional examples for which the stochastic DP problem can be solved, and more practical examples from nonlinear robotic planning.

II. PROBLEM FORMULATION

The following outlines the finite time optimal control problem formulation that we study in this work.

a) 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 $h : \mathbb{X} \times \mathbb{U} \times \mathbb{R}^{n_x} \to \mathbb{X}$ is given by the equation

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

where $\{w_t\}$ are zero mean independent, identically distributed (i.i.d) random sequences with variance Σ_{w_t}, and ϵ is a parameter modulating the noise input to the system.

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

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

$s.t.\ x_{t+1} = h(x_t, \pi_t(x_t), \epsilon w_t)$, 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 π^*; $c_t(\cdot,\cdot) : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is the one-step cost function; $c_T(\cdot) : \mathbb{X} \to \mathbb{R}$ is the terminal cost function; T is the horizon of the problem.

III. A PERTURBATION ANALYSIS OF OPTIMAL FEEDBACK CONTROL

In order to derive the results in this section, we need some additional structure on the dynamics. *In essence, the results in this section require that the time discretization of the dynamics be small enough.* Thus, let the dynamics given in Eq.(1) be written in the form:

$$x_{t+1} = x_t + (f(x_t) + g(x_t)u_t)\Delta t + \epsilon w_t \sqrt{\Delta t}, \tag{3}$$

where $\epsilon < 1$ is a perturbation parameter, ω_t is a white noise sequence, and the sampling time Δt is small enough that the $O(\Delta t^\alpha)$ terms are negligible for $\alpha > 1$. The noise term above stems from Brownian motion, and hence the $\sqrt{\Delta t}$ factor. We also assume that the instantaneous cost $c(\cdot,\cdot)$ has the following simple form, $c(x,u) = (l(x) + \frac{1}{2}u'Ru)\Delta t$, where R is symmetric and $R \succ 0$. The main reason to use the above assumptions is to simplify the Dynamic Programming (DP) equation governing the optimal cost-to-go function of the system developed in section III-B.

In the following three subsections, we establish three basic results that we shall use to establish the near optimality of the MPC law in Section IV. First, we characterize the performance of any given feedback policy as a perturbation (series) expansion in the parameter ϵ. We establish that the $O(\epsilon^0)$ term depends only on the nominal action, while the $O(\epsilon^2)$ depends only on the linear part of the feedback law. Next, we find the differential equations satisfied by these different perturbation costs using DP and show that the stochastic and deterministic optimal feedback laws share the same nominal and first order costs. In the subsequent section, we analyze the nominal/ open-loop problem using the classical Method of Characteristics and show that the open-loop optimal control has a unique global minimum. As a by-product, we also obtain the equations governing the optimal linear feedback term in the nonlinear problem, which is shown to be very different from a traditional LQR [5].

A. Characterizing the Performance of a Feedback Policy

Let us consider a noiseless version of the system dynamics given by (3), obtained by setting $w_t = 0$ for all t: $\bar{x}_{t+1} = \bar{x}_t + (f(\bar{x}_t) + g(\bar{x}_t)\bar{u}_t)\Delta t$, where we denote the "nominal" state trajectory as $\bar{x}_t$ and the "nominal" control as $\bar{u}_t$, with $\bar{u}_t = \pi_t(\bar{x}_t)$, and $\Pi = \{\pi_t\}_{t=1}^{T-1}$ is a given control policy.

Assuming that $f(\cdot)$ and $\pi_t(\cdot)$ are sufficiently smooth, we can expand the dynamics about the nominal trajectory using a Taylor series. Denoting $\delta x_t = x_t - \bar{x}_t, \delta u_t = u_t - \bar{u}_t$, we can express,

$$\delta x_{t+1} = A_t \delta x_t + B_t \delta u_t + S_t(\delta x_t) + \epsilon w_t \sqrt{\Delta t}, \quad (4)$$

$$\delta u_t = K_t \delta x_t + \tilde{S}_t(\delta x_t), \quad (5)$$

where $A_t = I_{n_x \times n_x} + \frac{\partial (f+gu)\Delta t}{\partial x}|_{\bar{x}_t,\bar{u}_t}$, $B_t = \frac{\partial (f+gu)\Delta t}{\partial u}|_{\bar{x}_t,\bar{u}_t} = g(\bar{x}_t)\Delta t$, $K_t = \frac{\partial \pi_t}{\partial x}|_{\bar{x}_t}$, and $S_t(\cdot), \tilde{S}_t(\cdot)$ are second and higher order terms in the respective expansions. Similarly, we can expand the instantaneous cost $c(x_t, u_t)$ about the nominal values $(\bar{x}_t, \bar{u}_t)$ as,

$$c(x_t, u_t) = \Big(l(\bar{x}_t) + L_t \delta x_t + H_t(\delta x_t) + \frac{1}{2}\bar{u}_t' R \bar{u}_t + \delta u_t' R \bar{u}_t + \frac{1}{2}\delta u_t' R \delta u_t\Big)\Delta t, \quad (6)$$

$$c_T(x_T) = c_T(\bar{x}_T) + C_T \delta x_T + H_T(\delta x_T), \quad (7)$$

where $L_t = \frac{\partial l}{\partial x}|_{\bar{x}_t}$, $C_T = \frac{\partial c_T}{\partial x}|_{\bar{x}_t}$, and $H_t(\cdot)$ and $H_T(\cdot)$ are second and higher order terms in the respective expansions.

Using (4) and (5), we can write the closed-loop dynamics of the trajectory $(\delta x_t)_{t=1}^T$ as,

$$\delta x_{t+1} = \underbrace{(A_t + B_t K_t)}_{\bar{A}_t} \delta x_t + \underbrace{B_t \tilde{S}_t(\delta x_t) + S_t(\delta x_t)}_{\bar{S}_t(\delta x_t)} + \epsilon w_t \sqrt{\Delta t}, \quad (8)$$

where $\bar{A}_t$ represents the linear part of the closed-loop systems and the term $\bar{S}_t(\cdot)$ represents the second and higher order terms in the closed-loop system. Similarly, the closed-loop incremental cost given in (6) can be expressed as

$$c(x_t, u_t) = \underbrace{\{l(\bar{x}_t) + \frac{1}{2}\bar{u}_t' R \bar{u}_t\}\Delta t}_{\bar{c}_t} + \underbrace{[L_t + \bar{u}_t' R K_t]\Delta t}_{\bar{C}_t} \delta x_t + \bar{H}_t(\delta x_t).$$

Therefore, the cumulative cost of any given closed-loop trajectory $(x_t, u_t)_{t=1}^T$ can be expressed as, $\mathcal{J}^\pi = \sum_{t=1}^{T-1} c(x_t, u_t = \pi_t(x_t)) + c_T(x_T)$, which can be written in the following form:

$$\mathcal{J}^\pi = \sum_{t=1}^{T} \bar{c}_t + \sum_{t=1}^{T} \bar{C}_t \delta x_t + \sum_{t=1}^{T} \bar{H}_t(\delta x_t), \quad (9)$$

where $\bar{c}_T = c_T(\bar{x}_T), \bar{C}_T = C_T$.

We first show the following critical result. *Note:* Due to paucity of space, the proofs for the results shown here are given in the extended version's appendix [17].

Lemma 1: Given any sample path, the state perturbation equation given in (8) can be equivalently characterized as

$$\delta x_t = \delta x_t^l + e_t, \ \delta x_{t+1}^l = \bar{A}_t \delta x_t^l + \epsilon w_t \sqrt{\Delta t} \quad (10)$$

where e_t is an $O(\epsilon^2)$ function that depends on the entire noise history $\{w_0, w_1, \cdots w_t\}$ and δx_t^l evolves according to the linear closed-loop system. Furthermore, $e_t = e_t^{(2)} + O(\epsilon^3)$, where $e_t^{(2)} = \bar{A}_{t-1} e_{t-1}^{(2)} + \delta x_t^{l'} \bar{S}_{t-1}^{(2)} \delta x_t^l$, $e_0^{(2)} = 0$, and $\bar{S}_t^{(2)}$ represents the Hessian corresponding to the Taylor series expansion of the function $\bar{S}_t(\cdot)$.

Next, we have the following result for the expansion of the cost-to-go function J^π.

Lemma 2: Given any sample path, the cost-to-go under a policy can be expanded about the nominal as:

$$\mathcal{J}^\pi = \underbrace{\sum_t \bar{c}_t}_{\bar{J}^\pi} + \underbrace{\sum_t \bar{C}_t \delta x_t^l}_{\delta J_1^\pi} + \underbrace{\sum_t \delta x_t^{l'} \bar{H}_t^{(2)} \delta x_t^l + \bar{C}_t e_t^{(2)}}_{\delta J_2^\pi} + O(\epsilon^3),$$

where $\bar{H}_t^{(2)}$ denotes the second order coefficient of the Taylor expansion of $\bar{H}_t(\cdot)$.

Now, we show the following important result.

Proposition 1: The mean of the cost-to-go function J^π obeys: $E[\mathcal{J}^\pi] = J^{\pi,0} + \epsilon^2 J^{\pi,1} + \epsilon^4 J^{\pi,2} + \mathcal{R}$, for some constants $J^{\pi,k}$, $k = 0, 1, 2$, where $\mathcal{R}$ is $o(\epsilon^4)$, i.e., $\lim_{\epsilon \to 0} \epsilon^{-4}\mathcal{R} = 0$. Furthermore, the term $J^{\pi,0}$ arises solely from the nominal control sequence while $J^{\pi,1}$ is solely dependent on the nominal control and the linear part of the perturbation closed-loop.

Remark 1: The physical interpretation of the result above is as follows: it shows that the ϵ^0 term, $J^{\pi,0}$, in the cost, stems from the nominal action of the control policy, the ϵ^2 term, $J^{\pi,1}$, stems from the linear feedback action of the closed-loop, while the higher order terms stem from the higher order terms in the feedback law. In the next section, we use Dynamic Programming (DP), to find the equations satisfied by these terms.

B. A Closeness Result for Optimal Stochastic and Deterministic Control

The DP equation for the optimal control problem on the system in Eq.(3) is given by:

$$J_t(x) = \min_{u_t}\{c(x, u_t) + E[J_{t+1}(x')]\}, \quad (11)$$

where $x' = x + f(x)\Delta t + g(x)u_t\Delta t + \epsilon\omega_t\sqrt{\Delta t}$ and $J_t(x)$ denotes the cost-to-go of the system given that it is at state x at time t. The above equation is marched back in time with terminal condition $J_T(x) = c_T(x)$, and $c_T(\cdot)$ is the terminal cost function. Let $u_t(\cdot)$ denote the corresponding optimal policy. Then, it follows that the optimal control u_t satisfies (since the argument to be minimized is quadratic in u_t)

$$u_t = -R^{-1} g' J_{t+1}^x, \quad (12)$$

where $J_{t+1}^x = \frac{\partial J_{t+1}}{\partial x}$. Further, let $u_t^d(\cdot)$ be the optimal control policy for the deterministic system, i.e., Eq. (3) with $\epsilon = 0$. The optimal cost-to-go of the deterministic system, $\phi_t(\cdot)$ satisfies the deterministic DP equation:

$$\phi_t(x) = \min_{u_t}[c(x, u_t) + \phi_{t+1}(x')], \tag{13}$$

where $x' = x + (f(x) + g(x)u_t)\Delta t$. Then, identical to the stochastic case, $u_t^d = -R^{-1}g'\phi_t^x$. Next, let $\varphi_t(\cdot)$ denote the cost-to-go of the deterministic policy when applied to the stochastic system, i.e., u_t^d applied to Eq. (3) with $\epsilon > 0$. The cost-to-go $\varphi_t(\cdot)$ satisfies the policy evaluation equation:

$$\varphi_t(x) = c(x, u_t^d(x)) + E[\varphi_{t+1}(x')], \tag{14}$$

where now $x' = x + (f(x) + g(x)u_t^d(x))\Delta t + \epsilon\omega_t\sqrt{\Delta t}$. Note the difference between the equations (13) and (14). Then, we have the following key result. An analogous version of the following result was originally proved in a seminal paper [8] for first passage problems. We provide a simple derivation of the result for a finite time final value problem below.

Proposition 2: The cost function of the optimal stochastic policy, J_t, and the cost function of the "deterministic policy applied to the stochastic system", φ_t, satisfy: $J_t(x) = J_t^0(x) + \epsilon^2 J_t^1(x) + \epsilon^4 J_t^2(x) + \cdots$, and $\varphi_t(x) = \varphi_t^0(x) + \epsilon^2\varphi_t^1(x) + \epsilon^4\varphi_t^2(x) + \cdots$. Furthermore, $J_t^0(x) = \varphi_t^0(x)$, and $J_t^1 = \varphi_t^1(x)$, for all t, x.

Proof: We show a sketch here for the case of a scalar state, please refer to [17] for the complete proof.
Due to Proposition 1, the optimal cost function satisfies: $J_t(x) = J_t^0 + \epsilon^2 J_t^1 + \epsilon^4 J_t^2 + \cdots$. Next, we substitute the above equation into the DP equation (11), along with the minimizing control (12) to obtain a perturbation expansion of the optimal cost function as a power series in ϵ^2. Equating the $O(\epsilon^0)$ and $O(\epsilon^2)$ terms on both sides results in governing equations for the J_t^0 and J_t^1 terms:
$J_t^0 = l\Delta t + \frac{1}{2}\frac{g^2}{r}(J_{t+1}^{0,x})^2\Delta t + \underbrace{(f + g\frac{-g}{r}J_{t+1}^{0,x})}_{\bar{f}^0} J_{t+1}^{0,x}\Delta t + J_{t+1}^0$, with the terminal condition $J_T^0 = c_T$, and
$J_t^1 = \underbrace{(f + g\frac{-g}{r}J_{t+1}^{0,x})}_{=\bar{f}^0} J_{t+1}^{1,x}\Delta t + \frac{1}{2}J_{t+1}^{0,xx}\Delta t + J_{t+1}^1$, with terminal condition $J_T^1 = 0$.
We also know that the cost function of the deterministic policy when applied to the stochastic system satisfies $\varphi_t = \varphi_t^0 + \epsilon^2\varphi_t^1 + \cdots$. Similar to above, we substitute this expression into the policy evaluation equation (14), along with the deterministic optimal control expression $u_t^d = -R^{-1}g'\phi_{t+1}^x$, to obtain the governing equations for φ_t^0 and φ_t^1. These equations, when compared with those for J_t^0 and J_t^1, are seen to be identical with the same terminal conditions thereby proving the result. ■

C. A Perturbation Expansion of Deterministic Optimal Feedback Control: the Method of Characteristics (MOC)

In this section, we will use the classical Method of Characteristics to derive some results regarding the deterministic optimal control problem, and in particular, regarding the open-loop solution [7]. In particular, we will show that satisfying the Minimum Principle is sufficient to assure us of a global optimum to the open-loop problem. We shall also do a perturbation expansion of the DP equation around the Characteristic curves to obtain the equations governing the linear feedback term, and show that this gain is entirely different from an LQR design. **Since the classical MOC is derived in continuous-time, we derive the following results in continuous-time, the extension to the discrete-time case is given in Remark 3. Also, for simplicity, we derive the following for the case of a scalar state and control, please see [17] for the vector case.**

Let us recall the Hamilton-Jacobi-Bellman (HJB) equation in continuous-time under the same assumptions as above, i.e., quadratic in control cost $c(x, u) = l(x) + \frac{1}{2}ru^2$, and affine in control dynamics $\dot{x} = f(x) + g(x)u$, (the discrete-time case for a sufficiently small discretization time then follows, please see Remark 3) [5]:

$$\frac{\partial J}{\partial t} + l - \frac{1}{2}\frac{g^2}{r}J_x^2 + fJ_x = 0, \tag{15}$$

where $J = J_t(x_t),\ J_x = \frac{\partial J}{\partial x_t}$, and the equation is integrated back in time with terminal condition $J_T(x_T) = c_T(x_T)$. Define $\frac{\partial J}{\partial t} = p,\ J_x = q$, then the HJB can be written as $F(t, x, J, p, q) = 0$, where $F(t, x, J, p, q) = p + l - \frac{1}{2}\frac{g^2}{r}q^2 + fq$. One can now write the Lagrange-Charpit equations [7] for the HJB as:

$$\dot{x} = F_q = f - \frac{g^2}{r}q, \tag{16}$$

$$\dot{q} = -F_x - qF_J = -l^x + \frac{gg^x}{r}q^2 - f^x q, \tag{17}$$

with the terminal conditions $x(T) = x_T,\ q(T) = c_T^x(x_T)$, where $F_x = \frac{\partial F}{\partial x},\ F_q = \frac{\partial F}{\partial q},\ g^x = \frac{\partial g}{\partial x},\ l^x = \frac{\partial l}{\partial x},\ f^x = \frac{\partial f}{\partial x}$ and $c_T^x = \frac{\partial c_T}{\partial x}$.
Given a terminal condition x_T, the equations above can be integrated back in time to yield a characteristic curve of the HJB PDE. Now, we show how one can use these equations to get a local solution of the HJB, and consequently, the feedback gain K_t.
Suppose now that one is given an optimal nominal trajectory $\bar{x}_t, t \in [0, T]$ for a given initial condition x_0, from solving the open-loop optimal control problem. Let the nominal terminal state be $\bar{x}_T$. We now expand the HJB solution around this nominal optimal solution. To this purpose, let $x_t = \bar{x}_t + \delta x_t$, for $t \in [0, T]$. Then, expanding the optimal cost function around the nominal yields: $J(x_t) = \bar{J}_t + G_t\delta x_t + \frac{1}{2}P_t\delta x_t^2 + \cdots$, where $\bar{J}_t = J_t(\bar{x}_t), G_t = \frac{\partial J}{\partial x_t}|_{\bar{x}_t}, P_t = \frac{\partial^2 J}{\partial x_t^2}|_{\bar{x}_t}$. Then, the co-state $q = \frac{\partial J}{\partial x_t} = G_t + P_t\delta x_t + \cdots$.
For simplicity, we assume that $g^x = 0$ (this is relaxed but at the expense of a rather tedious derivation shown in the

appendix of [17]). Hence,

$$\underbrace{\frac{d}{dt}(\bar{x}_t+\delta x_t)}_{\dot{\bar{x}}_t+\delta\dot{x}_t} = \underbrace{f(\bar{x}_t+\delta x_t)}_{(\bar{f}_t+\bar{f}^x_t\delta x_t+\cdots)} - \frac{g^2}{r}(G_t+P_t\delta x_t+\cdots),$$

where $\bar{f}_t = f(\bar{x}_t), \bar{f}^x_t = \frac{\partial f}{\partial x_t}|_{\bar{x}_t}$. Expanding in powers of the perturbation variable δx_t, the equation above can be written as (after noting that $\dot{\bar{x}}_t = \bar{f}_t - \frac{g^2}{r}G_t$ due to the nominal trajectory $\bar{x}_t$ satisfying the characteristic equation):

$$\delta\dot{x}_t = (\bar{f}^x_t - \frac{g^2}{r}P_t)\delta x_t + O(\delta x_t^2). \tag{18}$$

Next, we have: $\frac{dq}{dt} = -l_x - f_x q$

$$\frac{d}{dt}(G_t+P_t\delta x_t+\cdots) = -(\bar{l}^x_t+\bar{l}^{xx}_t\delta x_t+\cdots)$$
$$-(\bar{f}^x_t+\bar{f}^{xx}_t\delta x_t+\cdots)(G_t+P_t\delta x_t+\cdots), \tag{19}$$

where $\bar{f}^{xx}_t = \frac{\partial^2 f}{\partial x^2}|_{\bar{x}_t}, \bar{l}^x_t = \frac{\partial l}{\partial x}|_{\bar{x}_t}, \bar{l}^{xx}_t = \frac{\partial^2 l}{\partial x^2}|_{\bar{x}_t}$. Using $\frac{d}{dt}P_t\delta x_t = \dot{P}_t\delta x_t + P_t\delta\dot{x}_t$, substituting for $\delta\dot{x}_t$ from (18), and expanding the two sides above in powers of δx_t yields: $\dot{G}_t+(\dot{P}_t+P_t(\bar{f}^x_t-\frac{g^2}{r}P_t))\delta x_t+\cdots = -(\bar{l}^x_t+\bar{f}^x_t G_t)-(\bar{l}^{xx}_t+\bar{f}^x_t P_t+\bar{f}^{xx}_t G_t)\delta x_t+\cdots$.
Equating the first two powers of δx_t yields:

$$\dot{G}_t + \bar{l}^x_t + \bar{f}^x_t G_t = 0, \tag{20}$$

$$\dot{P}_t + l^{xx}_t + P_t\bar{f}^x_t + \bar{f}^x_t P_t - P_t\frac{g^2}{r}P_t + \bar{f}^{xx}_t G_t = 0. \tag{21}$$

The optimal feedback law is given by: $u_t(x_t) = \bar{u}_t + K_t\delta x_t + O(\delta x_t^2)$, where $K_t = -\frac{g}{r}P_t$.

Now, we provide the final result for the general vector case, with a state dependent control influence matrix (please see [17] for details). Let the control influence matrix be gives as: $\mathcal{G} = \begin{bmatrix} g^1_1(x)\cdots g^p_1(x) \\ \ddots \\ g^1_n(x)\cdots g^p_n(x)\end{bmatrix} = \begin{bmatrix}\Gamma^1(x)\cdots\Gamma^p(x)\end{bmatrix}$, i.e., Γ^j represents the control influence vector corresponding to the j^{th} input. Let $\bar{\mathcal{G}}_t = \mathcal{G}(\bar{x}_t)$, where $\{\bar{x}_t\}$ represents the optimal nominal trajectory. Further, let $\mathcal{F} = \begin{bmatrix}f_1(x)\cdots f_n(x)\end{bmatrix}^\intercal$ denote the drift/ dynamics of the system. Let $G_t = [G^1_t\cdots G^n_t]^\intercal$, and $R^{-1}\bar{\mathcal{G}}^\intercal_t G_t = -[\bar{u}^1_t\cdots\bar{u}^p_t]^\intercal$, denote the optimal nominal co-state and control vectors respectively. Let $\bar{\mathcal{F}}^x_t = \begin{bmatrix}\frac{\partial f_1}{\partial x_1}\cdots\frac{\partial f_1}{\partial x_n} \\ \ddots \\ \frac{\partial f_n}{\partial x_1}\cdots\frac{\partial f_n}{\partial x_n}\end{bmatrix}|_{\bar{x}_t}$, and similarly $\bar{\Gamma}^{j,x}_t = \nabla_x\Gamma^j|_{\bar{x}_t}$. Further, define: $\bar{\mathcal{F}}^{xx,i}_t = \begin{bmatrix}\frac{\partial^2 f_1}{\partial x_1\partial x_i}\cdots\frac{\partial^2 f_1}{\partial x_n\partial x_i} \\ \ddots \\ \frac{\partial^2 f_n}{\partial x_1\partial x_i}\cdots\frac{\partial^2 f_n}{\partial x_n\partial x_i}\end{bmatrix}|_{\bar{x}_t}$, and $\bar{\Gamma}^{j,xx,i}_t$ similarly for the vector function Γ^j, and $\bar{\mathcal{G}}^{x,i}_t = \begin{bmatrix}\frac{\partial g^1_1}{\partial x_i}\cdots\frac{\partial g^p_1}{\partial x_i} \\ \ddots \\ \frac{\partial g^1_n}{\partial x_i}\cdots\frac{\partial g^p_n}{\partial x_i}\end{bmatrix}|_{\bar{x}_t}$. Finally, define $\mathcal{A}_t = \bar{\mathcal{F}}^x_t + \sum_{j=1}^p\bar{\Gamma}^{j,x}_t\bar{u}^j_t$, $\bar{L}^x_t = \nabla_x l|_{\bar{x}_t}$, and $\bar{L}^{xx}_t = \nabla^2_{xx}l|_{\bar{x}_t}$.

Proposition 3: Given the above definitions, the following result holds for the evolution of the co-state/ gradient vector G_t, and the Hessian matrix P_t, of the optimal cost function $J_t(x_t)$, evaluated on the optimal nominal trajectory $\bar{x}_t, t\in[0,T]$:

$$\dot{G}_t + \bar{L}^x_t + \mathcal{A}^\intercal_t G_t = 0, \tag{22}$$

$$\dot{P}_t + \mathcal{A}^\intercal_t P_t + P_t\mathcal{A}_t + \bar{L}^{xx}_t + \sum_{i=1}^n[\bar{\mathcal{F}}^{xx,i}_t + \sum_{j=1}^p\bar{\Gamma}^{j,xx,i}_t\bar{u}^j_t]G^i_t - K^\intercal_t R K_t = 0, \tag{23}$$

$$K_t = -R^{-1}[\sum_{i=1}^n\bar{\mathcal{G}}^{x,i,\intercal}_t G^i_t + \bar{\mathcal{G}}^\intercal_t P_t]. \tag{24}$$

with terminal conditions $G_T = \nabla_x c_T|_{\bar{x}_T}$, and $P_T = \nabla^2_{xx}c_T|_{\bar{x}_T}$ and the control input with the optimal linear feedback is given by $u_t = \bar{u}_t + K_t\delta x_t$.

Remark 2: **Not LQR.** The co-state equation (22) above is identical to the co-state equation in the Minimum Principle [5], [19]. However, the Hessian P_t equation (23) is Riccati-like with some important differences: note the extra second order terms due to $\bar{\mathcal{F}}^{xx,i}_t$ and $\bar{\Gamma}^{xx,i}_t$ in the second line stemming from the nonlinear drift and input influence vectors and an extra term in the gain equation (24) coming from the state dependent influence matrix. These terms are not present in the LQR Riccati equation, and thus, it is clear that this cannot be an LQR, or perturbation feedback design (Ch. 6, [5]). If the input influence matrix is independent of the state, the first term in the second line remains, and hence, it is still different from the LQR case.

Remark 3: **Discrete-time case.** For the discrete-time case with small discretization time Δt, one would discretize the noiseless model with a forward Euler approximation as $x_{t+1} = x_t + (\mathcal{F}(x_t)+\mathcal{G}(x_t)u_t)\Delta t$ and the above equations as:

$$G_t = \bar{L}^x_t + A^\intercal_t G_{t+1}, \tag{25}$$

$$P_t = A^\intercal_t P_{t+1}A_t + \bar{L}^{xx}_t + \sum_{i=1}^n[\bar{\mathcal{F}}d^{xx,i}_t + \sum_{j=1}^p\bar{\Gamma}d^{j,xx,i}_t\bar{u}^j_t]G^i_{t+1} - K^\intercal_t(R_t + B^\intercal_t P_{t+1}B_t)K_t, \tag{26}$$

$$K_t = -(R_t + B^\intercal_t P_{t+1}B_t)^{-1}[\sum_{i=1}^n\bar{\mathcal{G}}d^{x,i,\intercal}_t G^i_{t+1} + B^\intercal_t P_{t+1}A_t]. \tag{27}$$

where, $A_t = I + (\bar{\mathcal{F}}^x_t + \sum_{j=1}^p\bar{\Gamma}^{j,x}_t\bar{u}^j_t)\Delta t$, $B_t = \bar{\mathcal{G}}\Delta t$, $\bar{\mathcal{F}}d^{xx,i}_t = \bar{\mathcal{F}}^{xx,i}_t\Delta t$, $\bar{\Gamma}d^{j,xx,i}_t = \bar{\Gamma}^{j,xx,i}_t\Delta t$, $\bar{\mathcal{G}}d^{x,i}_t = \bar{\mathcal{G}}^{x,i}_t\Delta t$.

Remark 4: **Convexity and Global Minimum.** Recall the Lagrange-Charpit equations for solving the HJB (16), (17). Given an unconstrained control, from the theory of the MOC (under standard smoothness assumptions on the involved functions), the characteristic curves are unique, and do not intersect. Therefore, the open-loop optimal trajectory, found by satisfying the Minimum Principle is also the unique global minimum even though the open-loop problem is non-convex. This observation is formalized in the following result.

Proposition 4: Global Optimality of open-loop solution. Let the cost functions $l(\cdot)$, $c_T(\cdot)$, the drift $f(\cdot)$ and the input influence function $g(\cdot)$ be $\mathcal{C}^2$, i.e., twice continuously differentiable. Then, an optimal trajectory that satisfies the Minimum Principle from a given initial state x_0, is the unique global minimum of the open-loop problem starting at the initial state x_0.

IV. THE NEAR-OPTIMALITY OF MODEL PREDICTIVE CONTROL

Consider now a Model Predictive approach to solving the stochastic control problem. We outline the algorithmic procedure below to highlight that our advocated procedure is slightly different from the traditional MPC approach studied in the literature [15], [20].

Algorithm 1: Shrinking Horizon MPC

Given: initial state x_0, time horizon T, cost $c(x,u) = l(x) + \frac{1}{2}ru^2$, and terminal cost $c_T(x)$.
Set $H = T$, $x_i = x_0$.
while $H > 0$ **do**
1) Solve the open-loop (noise free) optimal control problem (Eq. 2) for initial state x_i and horizon H. Let optimal sequence $U^* = \{u_0, u_1, \cdots, u_{H-1}\}$.
2) Apply the first control u_0 to the stochastic system, and observe the next state x_n.
3) set $H = H - 1$, $x_i = x_n$.

end

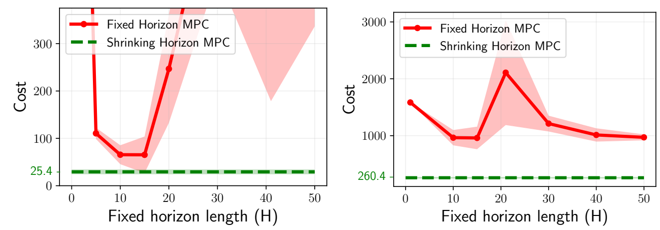
Remark 5: In traditional MPC [15], [20], the horizon H to solve the open-loop problem over is fixed. The setting is deterministic, and the necessity of replanning for the problem stems from the assumption that the actual problem horizon is infinite. In lieu, our problem horizon is finite, the repeated replanning takes place over progressively shorter horizons, and the setting is stochastic.

Theorem 1: Near-Optimality of MPC. The MPC feedback policy obtained from the recursive application of the MPC algorithm is near-optimal to $O(\epsilon^4)$ to the optimal stochastic feedback policy for the stochastic system (3).

Proof: We know that $J_t^0(x) = \varphi_t^0(x)$, and $J_t^1(x) = \varphi_t^1(x)$ from Proposition 2, for all (t, x). Owing to the uniqueness and global optimality of the open-loop from Proposition 4, it follows that the nominal control sequence found by the MPC procedure coincides with the nominal action of the optimal deterministic feedback law for any state x and any time t. Therefore, the result follows. ■

A. Summary of the Near-Optimality Result and its Implications

The result above establishes that repeatedly solving the deterministic optimal control problem from the current state results in a near-optimal stochastic policy. We examine two particularly important consequences in the following.



(a) Without state constraints (b) With state constraints

Fig. 2: The figure shows the sub-optimal behavior of the Fixed horizon MPC when applied to a state transition problem for a car-like robot with noise of $\epsilon = 0.5$. The horizon length H in fixed horizon MPC has to be carefully chosen to achieve a reasonable performance - albeit still worse - when compared to shrinking horizon MPC calculated for $T = 50$. The constraints in the state are imposed using penalty functions. The shrinking horizon MPC algorithm is shown in Algorithm 1.

a) Stochastic MPC: The MPC procedure we propose (shrinking horizon MPC) is slightly different from the traditional MPC (fixed horizon MPC) because of our shrinking horizon whereas typically MPC considers as a fixed horizon H that is small compared to the actual horizon T, i.e., $H \ll T$. Albeit the stability of the traditional MPC approach can be established, the actual domain of attraction of such a policy is limited by the fact that the policy needs to get the system into a control invariant subset containing the origin in at most H steps [15]. In practice, this shows up in the fact that such policies tend to be sluggish and almost never succeed in getting to the goal (please see Figure 2). Thus, in our opinion, in practice, as well as in theory, we should consider a problem with a finite horizon and with uncertainty, since that is a better approximation of the real problem and results in better performance.

A major bottleneck with MPC under uncertainty, in general, and stochastic MPC in particular [15], is that the MPC search needs to be over (time-varying) feedback policies rather than control sequences owing to the stochasticity of the problem, which leads to an intractable optimization for nonlinear systems. However, as our result demonstrates, the MPC feedback law we propose is near-optimal to fourth order. Further, as will be seen from our empirical results, in practice, solving the stochastic DP problem is highly sensitive to noise, and MPC still gives the best performance. A further important practical consequence of Theorem 1 is that we can get performance comparable to MPC, by wrapping the optimal linear feedback law around the nominal control sequence ($u_t = \bar{u}_t + K_t \delta x_t$), and replanning the nominal sequence only when the deviation is large enough. This is similar to the event driven MPC philosophy [9], [13]. This event driven replanning approach is also demonstrated in the next section.

b) Reinforcement Learning: The problem of Reinforcement learning can be construed as finding the optimal feedback policy for a stochastic nonlinear dynamical system [4]. Typically, this is done via simulations or rollouts of the dynamical system of interest, which allied with a suitable function approximator such as a (deep) neural net, yields a nonlinear feedback policy. However, these methods tend to be highly data intensive, slow to converge, and suffer from extremely high variance in the solution since they try to solve the DP equation. This is a manifestation of the inherent curse of dimensionality in trying to solve the DP problem. Thus, in our opinion, albeit the DP equation is an excellent analytical tool to study the structure of the feedback problem, nonetheless, it is not the correct synthesis tool. In fact, it is much easier to repeatedly solve the open-loop problem as prescribed by MPC. Of course, there remains the problem of whether we can solve the open-loop problem online. In our opinion, this is feasible today, when allied with efficient computational algorithms like iLQR [22] that exploit the beautiful causal structure of optimal control problems, suitable high performance computing (HPC) modifications, and suitable randomization of the computations that can help us very efficiently estimate the system parameters involved. In fact, this is the subject of the second part of this paper on data based control [26].

V. Empirical Results

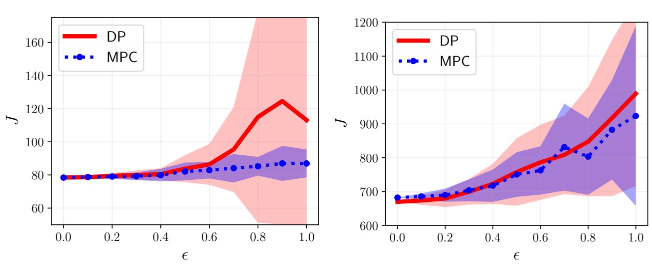
This section is divided into two subsections. Subsection V-A shows the practical optimality of MPC and the unreliable nature of the DP solution using simple 1-D problems. Subsection V-B shows the near-optimality of the linear feedback law and the effect of replanning to maintain near-optimality on stochastic problems using robotic planning problems. It also shows the suboptimal nature of applying MPC in its traditional form.

To all the systems considered below, process noise ω_t modeled as an additive white Gaussian noise with mean zero and a standard deviation of $\bar{u}_{\text{avg}}$ is added. ϵ is a scaling parameter that is varied to analyze the influence of the magnitude of the noise. Numerical optimization is performed using the `Casadi` [1] framework employing the `Ipopt` [25] solver.

A. Comparison of Stochastic DP and MPC: 1-D problems

The following two 1-D systems are considered to test the optimality of MPC on stochastic systems by comparing it to the DP solution.

System 1: $x_{t+1} = x_t + (-cos(x_t) + u_t)\Delta t + \epsilon\omega_t\sqrt{\Delta t}$,
System 2: $x_{t+1} = x_t+(-x_t-2x_t^2-0.5x_t^3+u_t)\Delta t+\epsilon\omega_t\sqrt{\Delta t}$.

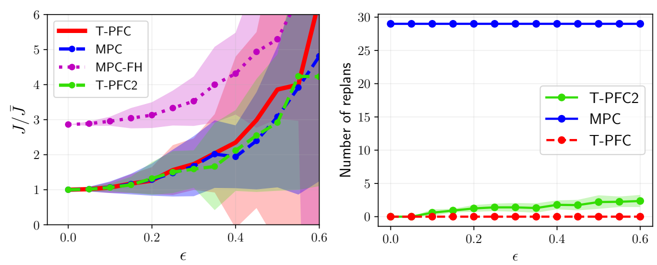


(a) System 1 (b) System 2

Fig. 3: Performance comparison between MPC and DP on the systems discussed. The data labeled as DP (red) corresponds to the exact stochastic DP, where it is solved for a particular ϵ value and tested out on the same value of ϵ. The lines in the plot denote the mean value and the shade denotes the standard deviation of the corresponding metric obtained from 100 Monte Carlo simulations.

The time horizon for both the problems is 50 steps with $\Delta t = 0.02s$. $x_0 = 1$ and $x_T = 4.8$. The cost function considered is $c(x,u) = 1/2(x'Qx + u'Ru)\Delta t$, $c_T(x) = (1/2)x'Q_T x$. MPC is solved as shown in Algorithm 1. DP is numerically solved by discretizing the space domain - $[0,5]$ into 200 states and solving Bellman's equation (11) for every state and at every time-step. The critical part in solving the DP problem is evaluating the expectation of the cost-to-go function which is done by taking samples and finding their mean. The optimal control is given by Eq. (12). One can infer from Fig. 3 that MPC, equivalently deterministic DP (DP with $\epsilon = 0.0$) actually performs better than its stochastic counterpart even for non-zero noise. Thus, there are no significant gains (in some cases makes it worse) when solving the stochastic DP problem, in practice even for simple cases such as these. The closeness between the DP solution and MPC also adds empirical evidence to the result in Proposition 4, that there is only a unique global optimum for the open-loop when working with cost functions and dynamics that are quadratic and affine in control, respectively.

B. Robotic Planning problems

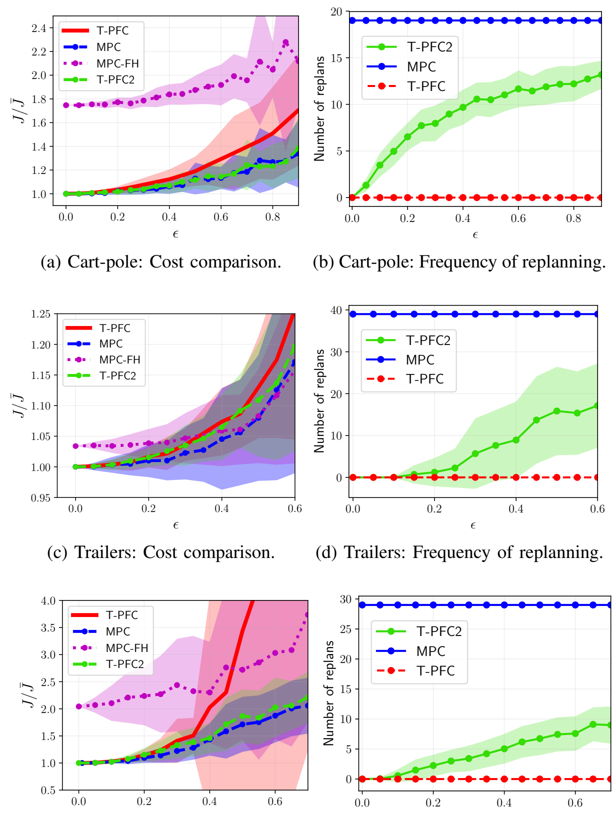


(a) Cost comparison. (b) Frequency of replanning.

Fig. 4: Cost evolution over a feasible range of ϵ for a car-like robot system, where ϵ is a measure of the noise in the system. Note that the performance of T-PFC is close to MPC for a wide range of noise levels ($\epsilon < 0.4$) but the cost and more importantly the standard deviation of the cost is seen to be larger than MPC as noise increases. T-PFC2 performs very similar to MPC, i.e. the mean and the standard deviation of the cost of T-PFC2 matches that of MPC, achieving it by replanning efficiently as seen in the subfigure (b). The key takeaways are: 1) the optimal policy for finite horizon stochastic optimal control problem is to use MPC as opposed to MPC-FH which is catastrophically off, 2) Significant gain in computation is achieved by using the linear feedback policy T-PFC/T-PFC2 without much loss in performance.

This section shows empirical results obtained by designing the feedback policy as discussed in section III-C and IV for a car-like robot, cart-pole, car with two trailers and a quadrotor tasked to move from an initial state to a goal state within a finite time. The linear feedback policy - called the Trajectory optimized Perturbation Feedback Controller (T-PFC) [18] here - involves 2 steps: 1) Solving the deterministic optimal control problem to obtain the nominal trajectory, 2) Calculating the feedback gains using Eqs. (25)-(27) for the discrete-time case. Note that there is no online computation involved in T-PFC. We also show the performance of our MPC and compare it with the traditional MPC, dubbed MPC-Fixed Horizon (MPC-FH). MPC-FH, unlike MPC, plans for a short horizon repeatedly rather than the full time horizon (as outlined in Section IV). In addition to that, we also show the performance of T-PFC2 which is simply T-PFC with cost triggered replanning, i.e. if the run time cost deviates beyond a threshold from the nominal cost, a new nominal is generated from the current state for the remainder of the horizon.

The car-like robot considered has the motion model described by $\dot{x}_t = v_t \cos(\theta_t)$, $\dot{y}_t = v_t \sin(\theta_t)$, $\dot{\theta}_t = \frac{v_t}{L} \tan(\phi_t)$, $\dot{\phi}_t = \omega_t$ and is discretized using forward Euler. The cost function used is $c(x,u) = 1/2(x'Qx + u'Ru)\Delta t$, $c_T(x) = (1/2)x'Q_T x$, $\Delta t = 0.01s$, Horizon = 30, Planning Horizon for MPC-FH = 5, Replanning threshold for T-PFC2 = 20%. Similar cost functions and parameters are chosen for the other nonlinear systems. It is evident from Figs. 4, and 5 that solving MPC for the entirety of the horizon gives the best possible solution and is significantly better than MPC-FH. It also shows that significant computational savings can be achieved without losing optimality if the linear feedback policy (T-PFC/T-PFC2) is used especially in low noise cases.



(a) Cart-pole: Cost comparison. (b) Cart-pole: Frequency of replanning.

(c) Trailers: Cost comparison. (d) Trailers: Frequency of replanning.

(e) Quadrotor: Cost comparison. (f) Quad: Frequency of replanning.

Fig. 5: Cost evolution over a feasible range of ϵ for different dynamical systems.

VI. Conclusion

In this paper, we have considered the problem of stochastic nonlinear control. We have shown that recursively solving the deterministic optimal control problem from the current state, a la MPC, results in a near-optimum policy to fourth order in a small noise parameter, and in practice, empirical evidence shows that the MPC law performs better than the law obtained by computationally solving the stochastic DP problem. An important limitation of the method is the smoothness of the nominal trajectory such that suitable Taylor expansions are possible, this breaks down when trajectories are non-smooth such as in hybrid systems like legged robots, or maneuvers have kinks for car-like robots such as in a tight parking application. It remains to be seen as to if, and how, one may extend the result to such applications that are piecewise smooth in the dynamics.

References

[1] J. A E Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl. CasADi – A software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, In Press, 2018.

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

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

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

[5] A. E. Bryson and Y. C. Ho. *Applied Optimal control*. Allied Publishers, NY, 1967.

[6] L. Chisci, J. A. Rossiter, and G. Zappa. Systems with persistent disturbances: predictive control with restricted contraints. *Automatica*, 37:1019–1028, 2001.

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

[8] W. H. Fleming. Stochastic control for small noise intensities. *SIAM Journal on Control*, 9(3):473–517, 1971.

[9] W. Heemels, K. Johansson, and P. Tabuada. An introduction to event triggered and self triggered control. In *Proc. IEEE CDC*, 2012.

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

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

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

[13] H. Li, Y. She, W. Yan, and K. Johansson. Periodic event-triggered distributed receding horizon control of dynamically decoupled linear systems. In *Proc. IFAC World Congress*, 2014.

[14] D. Mayne. Robust and stochastic mpc: Are we going in the right direction? *IFAC-PapersOnLine*, 48(23):1 – 8, 2015. 5th IFAC Conference on Nonlinear Model Predictive Control NMPC 2015.

[15] D. Q. Mayne. Model predictive control: Recent developments and future promise. *Automatica*, 50:2967–2986, 2014.

[16] D. Q. Mayne, E. C. Kerrigan, E. J. van Wyk, and P. Falugi. Tube based robust nonlinear model predictive control. *International journal of robust and nonlinear control*, 21:1341–1353, 2011.

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

[18] K. S. Parunandi and S. Chakravorty. T-pfc: A trajectory-optimized perturbation feedback control approach. *IEEE Robotics and Automation Letters*, 4(4):3457–3464, 2019.

[19] L. Pontryagin, V. Boltayanskii, R. Gamkrelidze, and E. Mishchenko. *The mathematical theory of optimal processes*. Wiley, New York, 1962.

[20] J. B. Rawlings and D. Q. Mayne. *Model Predictive Control: Theory and Design*. Nob Hill, Madison, WI, 2015.

[21] J. A. Rossiter, B. Kouvaritakis, and M. J. Rice. A numerically stable state space approach to stable predictive control strategies. *Automatica*, 34:65–73, 1998.

[22] Y. Tassa, T. Erez, and E. Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4906–4913. IEEE, 2012.

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

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

[25] A. Wächter and L. T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 2006.

[26] R. Wang, K. S. Parunandi, A. Sharma, R. Goyal, and S. Chakravorty. On the search for feedback in reinforcement learning. *Accepted to appear in Proc. of IEEE CDC*, 2021.