Revisiting Dynamic Programming for Exploration: Insights from a Simple Dual Control Problem

Ying Wang, Kévin Colin, Yue Ju, Mirko Pasquini, Håkan Hjalmarsson

Abstract— **The dual control problem, first introduced by Feldbaum in the 1960s, is recognized as encapsulating the "exploration versus exploitation" dilemma, central to online learning and control. Numerous heuristic-based exploration methods have been developed to facilitate active learning. However, the theoretically optimal solution provided by dynamic programming (DP) remains computationally intractable for most problems due to the curse of dimensionality. In this paper, we revisit the DP framework within the context of regret minimization for a simple real-time optimization problem, aiming to identify valuable insights and uncover new avenues for simplified DP-based exploration strategies. By deriving the two-horizon DP solution in our simple setting, we observe that the optimal input is obtained by solving an optimization problem composed of two distinct components representing exploration and exploitation separately, clearly highlighting their inherent trade-off. For longer horizons, receding horizon control based on the iterative application of the two-horizon DP provide possible approximations, reducing computational complexity while yielding useful suboptimal control policies. A key advantage of the DP-based exploration method is its ability to automatically adjust the exploration based on the current exploitation and system uncertainty. The proposed method is studied numerically through comparative evaluations against classical heuristic exploration methods from the literature.**

I. Introduction

Learning-based control, a powerful approach for addressing a wide range of scientific and industrial challenges [1], has a long history rooted in adaptive control [2]. A key challenge in this framework is the inherent system uncertainty, which requires control input to balance its dual effects: *actively* reducing uncertainty while ensuring control performance. This is the classic, yet still open, dual control problem, initially introduced by Feldbaum in the 1960s [3]. It has been framed as the fundamental trade-off between exploration and exploitation (E2): (i) exploitation, which aims to achieve the best possible short range control performance, and (ii) exploration, i.e., actively exciting the system to enhance system information helping to improve long range control performance, at the cost of compromising short-term control performance [4].

A theoretical framework based on DP is provided by Feldbaum to optimally balance E2 [3]. Although DP guarantees theoretical optimality, its practical applicability is severely limited by the well-known curse of dimensionality [2]. We focus on a simplified, scalar, static, Real-Time Optimization (RTO) problem [5], [6]. RTO addresses the challenge of optimizing plant operating conditions when the input-output steady-state relationship is dependent on unknown parameters that must be learned from noisy data. Our contributions are twofold: (i) We derive a two-horizon DP solution for a simple RTO problem, revealing that the optimal input is achieved by balancing two distinct components associated with E2, which clearly demonstrates their trade-off. (ii) To address longer horizons, we propose suboptimal online input policy through a receding horizon method, which iteratively applies the two-horizon DP solution.

II. A Simple Dual control Problem

We focus on the following simple dual control problem with a tractable DP horizon $T = 2$:

$$\Phi(u, \theta_0) = u^2 + 2\theta_0 u + 2\theta_0^2,$$
$$y_t = \theta_0 u_t + e_t,\ e_t \sim \mathcal{N}(0, 1).$$

The optimal input, assuming perfect knowledge of the true parameter θ_0, is simply found as $\mathcal{U}(\theta_0) = -\theta_0$. However, due to uncertainty in θ_0, this optimal input is unknown in practice. A common exploitation policy in data-driven control is the certainty equivalent (CE) controller, which relies on the estimate $\hat{\theta}_t$ and computes the input as $u_t^{CE} = -\hat{\theta}_t$. In the DP framework, θ_0 is treated as a realization of the prior conditional distribution $p(\theta \mid Z^0) = \mathcal{N}(\hat{\theta}_1, P_1)$, where $\hat{\theta}_1$ and P_1 denote the prior mean and variance, respectively. After applying u_1 and observing y_1, the posterior distribution is updated as $p(\theta \mid Z^1) = \mathcal{N}(\hat{\theta}_2, P_2)$, where $Z^1 = \{Z^0, u_1, y_1\}$ and $\hat{\theta}_2$, P_2 denote the posterior mean and variance, respectively. The Gaussian assumptions guarantee that Bayesian updates preserve the Gaussian form, leading to explicit expressions for the posterior mean and variance:

$$P_2 = \frac{P_1}{1 + P_1 u_1^2},\quad \hat{\theta}_2 = P_2\big(\frac{\hat{\theta}_1}{P_1} + u_1 y_1\big). \tag{1}$$

A. Dynamic Programming for $T = 2$ (DP-2)

(i) At step $T = 2$, the general cost-to-go function is

$$V_2(\hat{\theta}_2, P_2) = \min_{u_2} \mathbb{E}_\theta[\Phi(u_2, \theta) \mid Z^1]$$
$$= \min_{u_2} [u_2^2 + 2\hat{\theta}_2 u_2 + 2(\hat{\theta}_2^2 + P_2)] = \hat{\theta}_2^2 + 2P_2.$$

The corresponding optimal solution is $u_2^{DP} = -\hat{\theta}_2$, which coincides with the CE input.

(ii) At step $t = 1$, the Bellman equation, used to recursively obtain the earlier optimal input, becomes:

$$V_1(\hat{\theta}_1, P_1) = \min_{u_1} \{\mathbb{E}_\theta[\Phi(u_1, \theta) \mid Z^0] + \mathbb{E}_{y_1}[V_2(\hat{\theta}_2, P_2) | Z^0]\},$$

Expanding the first expectation term leads to

$$\mathbb{E}_\theta[\Phi(u_1, \theta) \mid Z^0] = u_1^2 + 2\hat{\theta}_1 u_1 + 2(\hat{\theta}_1^2 + P_1). \tag{2}$$

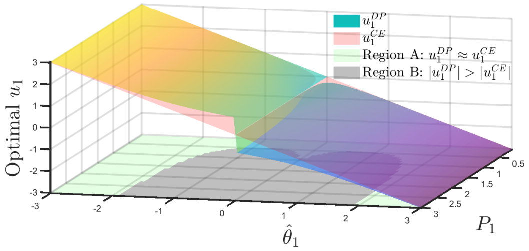

Fig. 1: The CE input u_1^{CE} and the DP-2 input u_1^{DP} under various initial distribution $\mathcal{N}(\hat{\theta}_1, P_1)$.

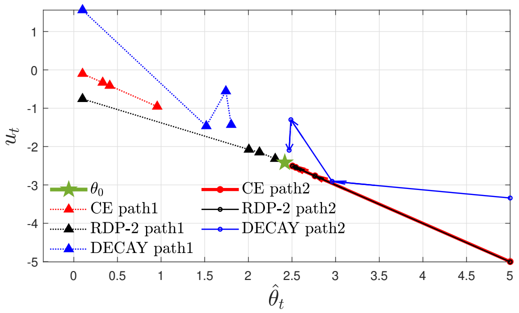

Fig. 2: Trajectories over horizon $T = 4$ generated by CE, RDP-2 and DECAY with the true parameter θ_0 sampled from two prior conditions $\mathcal{N}(0.1, 2)$ and $\mathcal{N}(5, 2)$.

The second expectation implicitly depends on u_1, since both its arguments, $\hat{\theta}_2$ and P_2, are functions of u_1, as indicated by (1). After applying a selected input u_1, the observation y_1 is stochastic due to the noise and parameter uncertainty. Thus the posterior mean $\hat{\theta}_2$, updated in (1), is a random variable, that should be averaged w.r.t. y_1 to obtain

$$\mathbb{E}_{y_1}[V_2(\hat{\theta}_2, P_2) \mid Z^0] = \mathbb{E}_{y_1}[\hat{\theta}_2^2 + 2P_2 \mid Z^0] \tag{3}$$
$$=(\mathbb{E}_{y_1}[\hat{\theta}_2 \mid Z^0])^2 + \mathrm{Var}_{y_1}(\hat{\theta}_2 \mid Z^0) + 2P_2 = \hat{\theta}_1^2 + P_1 + P_2,$$

where the last equality is based on the following equations

$$\mathbb{E}_{y_1}[\hat{\theta}_2 \mid Z^0] = \hat{\theta}_1, \quad \mathrm{Var}_{y_1}(\hat{\theta}_2 \mid Z^0) + P_2 = P_1,$$

Combining (2), (3), the cost-to-go function at $t = 1$ becomes,

$$V_1(\hat{\theta}_1, P_1) = \min_{u_1}[(u_1^2 + 2\hat{\theta}_1 u_1 + P_2) + 3(\hat{\theta}_1^2 + P_1)].$$

where $P_2 = \frac{P_1}{1+P_1 u_1^2}$ and the term $3(\hat{\theta}_1^2 + P_1)$ is a constant. This clearly sheds light on the involved trade-off: (i) u_1 should be close to the exploitation input $-\hat{\theta}_1$ to minimize the current cost, and (ii) a large magnitude of u_1 is desirable to reduce P_2, i.e., the uncertainty in the next step.

Given that the maximum tractable DP horizon is two (i.e. $N = 2$), we approximate longer-horizon DP solutions with a receding horizon control strategy, named RDP-2.

III. Numerical Results

In MC simulations, we first sample 500 true system parameters from the prior distribution $\mathcal{N}(\hat{\theta}_1, P_1)$. For each true system, we conduct 500 MC simulations by generating 500 noise realizations. We use expected regret as a criterion.

For the horizon $T = 2$, DP-2 is theoretically optimal. For a longer horizon $T = 10$, we evaluated the DP-based methods by comparing them with five established exploration strategies: CE input, decaying binary noise exploration (DECAY), binary immediate exploration (IMM), upper confidence bound (UCB), Thompson sampling (TS).

The RDP-2 method exhibits encouraging performance in this small study. However, it is important to note that no single suboptimal method is guaranteed to consistently outperform the others for long-horizon cases. That is because all these approaches are just approximations of the optimal but intractable DP solution. RDP-2 distinguishes itself by adaptively incorporating system feedback, enhancing its effectiveness in dynamically adjusting exploration. This insight highlights the potential for developing more sophisticated exploration strategies to further leverage adaptive learning and system feedback for decision-making under uncertainty.

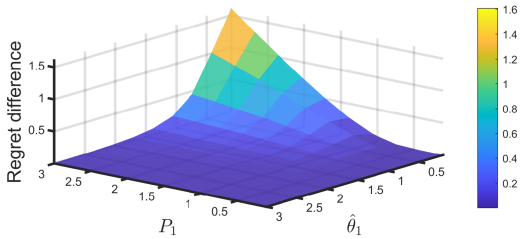

Fig. 3: Expected regret difference (CE minus DP-2) from MC simulations under various initial distribution $\mathcal{N}(\hat{\theta}_1, P_1)$.

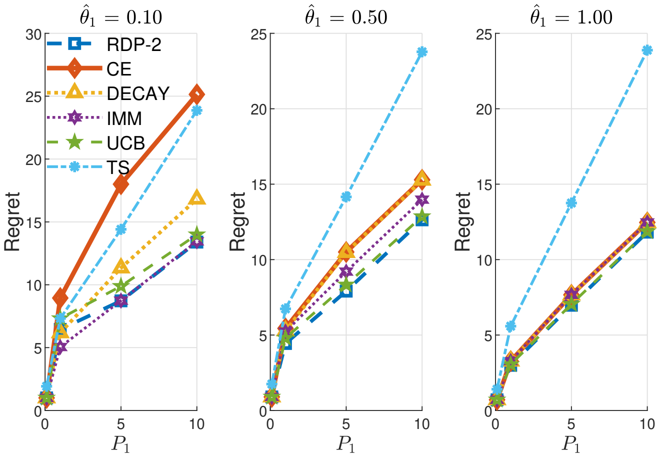

Fig. 4: Expected regret comparison using six methods across twelve different system configurations over horizon $T = 10$.

References

[1] Z.-P. Jiang, T. Bian, W. Gao *et al.*, "Learning-based control: A tutorial and some recent results," *Foundations and Trends® in Systems and Control*, vol. 8, no. 3, pp. 176–284, 2020.

[2] B. Wittenmark, "Adaptive dual control methods: An overview," *Adaptive Systems in Control and Signal Processing 1995*, pp. 67–72, 1995.

[3] A. Fel'dbaum, "Dual-control theory," *Automn. Remote Control*, vol. 21, pp. 874–880, 1960.

[4] A. Mesbah, "Stochastic model predictive control with active uncertainty learning: A survey on dual control," *Annu. Rev. Control*, vol. 45, pp. 107–117, 2018.

[5] A. G. Marchetti, G. François, T. Faulwasser, and D. Bonvin, "Modifier adaptation for real-time optimization—methods and applications," *Processes*, vol. 4, no. 4, p. 55, 2016.

[6] A. Ahmad, W. Gao, and S. Engell, "A study of model adaptation in iterative real-time optimization of processes with uncertainties," *Computers & Chemical Engineering*, vol. 122, pp. 218–227, 2019.