Proceedings of Machine Learning Research vol 120:1-10, 2020

Bayesian model predictive control: Efficient model exploration and regret bounds using posterior sampling

Kim P. Wabersich WKIM@ETHZ.CH
Institute for Dynamic Systems and Control, ETH Zurich, Switzerland

Melanie N. Zeilinger MZEILINGER@ETHZ.CH
Institute for Dynamic Systems and Control, ETH Zurich, Switzerland

Abstract

Tight performance specifications in combination with operational constraints make model predictive control (MPC) the method of choice in various industries. As the performance of an MPC controller depends on a sufficiently accurate objective and prediction model of the process, a significant effort in the MPC design procedure is dedicated to modeling and identification. Driven by the increasing amount of available system data and advances in the field of machine learning, data-driven MPC techniques have been developed to facilitate the MPC controller design. While these methods are able to leverage available data, they typically do not provide principled mechanisms to automatically trade off exploitation of available data and exploration to improve and update the objective and prediction model. To this end, we present a learning-based MPC formulation using posterior sampling techniques, which provides finite-time regret bounds on the learning performance while being simple to implement using off-the-shelf MPC software and algorithms. The performance analysis of the method is based on posterior sampling theory and its practical efficiency is illustrated using a numerical example of a highly nonlinear dynamical car-trailer system.

Keywords: Thompson sampling, posterior sampling , control, predictive control, regret bounds

1. Introduction

Many autonomous systems of practical relevance such as autonomous cars, delivery drones, or chemical synthesis processes need to be optimally controlled using limited input authority subject to safety specifications in terms of state constraints. To meet these requirements, model predictive control (MPC) techniques have been developed since the 1970's. A distinct property of MPC is the possibility to ensure state and input constraint satisfaction in a principled way while providing approximately optimal control performance. As a result, MPC has enabled the development of a wide range of high performance control applications as discussed, e.g., in Morari and Lee (1999); Qin and Badgwell (2000).

The central mechanism of MPC is based on solving an open-loop optimal control problem, the MPC problem, at discrete time instances based on the current system state. More precisely, the future system evolution starting from the currently measured system state is simultaneously predicted and optimized in real time using a prediction model of the plant. Due to uncertainties in the prediction model and external disturbances, however, only the first element of the resulting optimal input sequence is applied to the system. At each time instance, the procedure of prediction and optimization is repeated, which introduces state feedback and therefore allows for disturbance compensation. Nevertheless, the resulting control performance heavily relies on a sufficiently accurate prediction model of the underlying system dynamics, which typically results in time-consuming system modeling and identification procedures. In addition to the prediction model, the closed loop behavior is essentially determined through the objective function used in the MPC problem. While it is commonly assumed that the objective is given in closed form it may not be explicitly available, e.g., in case of complex or interactive applications. For example, the objective of a pick and place application could be determined by a person who provides feedback whether objects are placed correctly. The MPC objective function would then need to be infered from noisy samples.

To account for the modeling challenge, MPC approaches that are capable of leveraging learning-based prediction models have been investigated, see e.g. Hewing et al. (2018); Carron et al. (2019); Kamthe and Deisenroth (2018); Koller et al. (2018); Soloperto et al. (2018), which generally assume availability of sufficiently informative system data. However, by *passively* relying on available data, the resulting prediction model will only improve performance if the data is informative for the current task. Another option to generate sufficiently informative task-independent system data is to apply classical offline system identification procedures as described e.g. in Ljung (1998). While this enables sensible model estimation and even completely data-driven MPC controllers as proposed in Yang and Li (2015); Coulson et al. (2019); Berberich et al. (2019), the main limitation is the high cost of obtaining informative data regardless of the control objective. Therefore, such approaches can become impractical, especially for nonlinear or high dimensional systems.

To balance between passive knowledge exploitation and objective-independent exploration, effective exploration-exploitation strategies have been developed in so-called dual-MPC approaches, see e.g. Mesbah (2018) for an overview. Thereby the idea is to consider the potential advantage of obtaining relevant data in the future, e.g. through approximate stochastic dynamic programming as in Hanssen and Foss (2015); Klenske and Hennig (2016); Heirung et al. (2017); Arcari et al. (2020). While these techniques show promising results for simple tasks up to two state dimensions in combination with very short MPC prediction horizons, they are, so far, fundamentally limited to systems of low complexity. In addition, due to the rather crude approximation of the underlying stochastic dynamic programming problem, no theoretical performance guarantees have been reported so far.

The goal of this paper is to address these limitations for episodic learning tasks through a Bayesian learning-based MPC controller that automatically trades off exploration and exploitation while maintaining the computational complexity of conventional MPC. This is achieved by combining MPC with posterior sampling for reinforcement learning (RL) as originally proposed in Strens (2000) and theoretically investigated by Osband et al. (2013) and Osband and Van Roy (2014).

To trade off extraction of informative data and exploitation of already cumulated data, we propose a simple mechanism that samples an MPC controller at the beginning of each episode according to its posterior probability of being optimal with respect to the uncertain system dynamics and objective. Thereby, initial uncertainty about the optimal soft-constrained MPC controller leads to exploration of MPC controllers and generates explorative data collection in closed loop. As the posterior belief about the optimal soft-constrained MPC controller gets more certain through such explorative episodes, the MPC samples begin to aggregate around the optimal soft-constrained MPC controller for the plant, therefore automatically trading off exploration and exploitation. The resulting learning-based MPC controller yields a standard MPC problem and can be implemented using available algorithms and software packages, such as Wang and Boyd (2010); Houska et al. (2011); Domahidi et al. (2012); Zanelli et al. (2017). The presented MPC allows for a rigorous, finite-time performance analysis with respect to the a-priori unknown optimal soft-constrained MPC controller for a specific system at hand by applying results from model-based RL, relating the degree of sub-optimality of a model-based controller to the respective model discrepancy that vanishes at a provable learning rate.

In the remainder of the paper we begin by formalizing the considered class of system dynamics and objective functions, provide the necessary background on MPC and state the formal problem formulation using the notion of Bayesian expected regret. Afterwards, the learning-based MPC scheme is presented, analyzed, and demonstrated using a highly nonlinear learning task.

2. Model predictive control as an approximate optimal control policy

We consider discrete-time stochastic dynamical systems of the form

$$x(k+1) = f(x(k), u(k); \theta_f) + w(k), \quad k = 0, 1, 2, .., N-1 \tag{1}$$

with inputs $u(k) \in \mathbb{R}^m$, states $x(k) \in \mathbb{R}^n$, parameters $\theta_f \in \mathbb{R}^{n_{\theta_f}}$, zero mean σ_w-sub-Gaussian process noise $w(k) \sim \mathcal{Q}_w$, and random initial condition $x(0) \sim \mathcal{Q}_{x(0)}$. The system is subject to input constraints $u(k) \in \mathbb{U} \subseteq \mathbb{R}^m$ and state constraints $x(k) \in \mathbb{X} \subseteq \mathbb{R}^n$, which should be satisfied point-wise in time. The control objective is to minimize a time-varying stage cost function $\ell : \mathbb{N} \times \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ along system trajectories up to a finite time horizon N, i.e. minimizing

$$\mathbb{E}_W \left[\sum_{k=0}^{N-1} \ell(k, x(k), u(k); \theta_\ell) \right], \tag{2}$$

where $W \coloneqq [w(0), w(1), .., w(N-2)]$ and $\theta_\ell \in \mathbb{R}^{n_{\theta_\ell}}$ parametrizes the objective function.

While many relevant control problems can be stated in the above form, it is generally intractable to compute an optimal control policy that minimizes (2), which motivates the use of MPC techniques. A simple yet efficient approximate control strategy to minimize (2) is based on repeatedly solving a constrained optimal control problem initialized at the currently measured state $x(k)$ in a shrinking horizon fashion. While corresponding MPC formulations vary greatly in their complexity, the most simplistic formulation, sometimes referred to as nominal MPC, often provides sufficient practical properties in terms of performance and constraint satisfaction. Thereby we optimize over a control sequence $\{u_{i|k}\}$ subject to state and input constraints while neglecting zero mean additive disturbances. The resulting MPC problem is given by

$$J_k^\theta(x) \coloneqq \min_{u_{i|k}} \; c_\epsilon I(\epsilon) + \sum_{i=k}^{N-1} \ell(i, x_{i|k}, u_{i|k}; \theta_\ell) \tag{3a}$$

$$\text{s.t. } x_{k|k} = x, \;\; \epsilon \geq 0, \tag{3b}$$

$$x_{i+1|k} = f(x_{i|k}, u_{i|k}; \theta_f), \; i = k, .., N-2, \tag{3c}$$

$$x_{i|k} \in \mathbb{X}(\epsilon_i), \; i = k, .., N-1, \tag{3d}$$

$$u_{i|k} \in \mathbb{U}, \; i = k, .., N-1, \tag{3e}$$

where the additional cost term $c_\epsilon I(\epsilon)$ together with (3d) corresponds to a soft-constraint reformulation of the state constraints $x_{i|k} \in \mathbb{X}$, accounting for the fact that closed loop system trajectories might differ from nominal predictions and therefore ensuring feasibility of (3). For example, if $\mathbb{X} = \{x \in \mathbb{R}^n | g(x) \leq 0_{n_g}\}$ a soft-constraint reformulation can be obtained as $\mathbb{X}(\epsilon_i) = \{x \in \mathbb{R}^n | g(x) \leq \epsilon_i \mathbb{1}_{n_g}\}$ with $\epsilon_i \geq 0$, and $I(\epsilon) = c_{1,\epsilon}^\top \epsilon + c_{2,\epsilon} \epsilon^\top \epsilon$ (Kerrigan and Maciejowski, 2000).

In the following, we denote the expected closed loop cost-to-go at time k and state x as

$$V^{\theta}_{\tilde{\theta},k}(x) := \mathbb{E}_W \left[\sum_{j=k}^{N-1} \ell(j, x(j), u(j); \theta_\ell) \,\middle|\, \begin{array}{l} x(k) = x, \\ u(j) = \pi_{\tilde{\theta}}(j, x(j)), \\ x(j+1) = f(x(j), u(j); \theta_f) + w(j) \end{array} \right] \tag{4}$$

with $\theta := (\theta_\ell, \theta_f)$ and $\pi_{\tilde{\theta}}(j, x(j)) := u^*_{j|j}(x; \tilde{\theta})$, being the first element of the optimal input sequence of the MPC problem (3) at time step j with parameters $\tilde{\theta} := (\tilde{\theta}_\ell, \tilde{\theta}_f)$.

3. Problem formulation

We consider the case of unknown system dynamics and objective parametrization. More specifically, the learning problem is to improve the performance of the MPC controller π_θ through data-based refinements of a-priori unknown parameters θ. The efficient collection of system data with respect to the objective (2) is carried out through repeated episodic interactions with the system (1). During each episode $e = 0, 1, .., N_E - 1$, we need to provide a control policy that trades off information extraction and knowledge exploitation when applied to system (1) at each sampling time step $k = 0, 1, .., N - 1$. The corresponding data, obtained up to N_E episodes, is denoted by

$$\mathbb{D}_{N_E} := \left\{ (k, x_{k,e}, u_{k,e}, f(x_{k,e}, u_{k,e}; \theta_f) + w_{k,e}, \ell(k, x_{k,e}, u_{k,e}; \theta_\ell) + \epsilon_{k,e})_{k=0}^{N-1} \right\}_{e=0}^{N_E-1} \tag{5}$$

with zero mean σ_ϵ-sub-Gaussian measurement noise $\epsilon_{k,e}$ on objective observations. Prior knowledge about the system parameters θ such as production or sensor tolerances of the plant to be controlled is considered to be given as $\theta \sim \mathcal{Q}_\theta$. In the following, $\theta_e \sim \mathcal{Q}_{\theta|\mathbb{D}_e}$ denotes the posterior belief about θ during episode e after data (5) has been observed. The learning progress based on acquired data (5) after N_E episodes is measured w.r.t. the Bayesian cumulative expected regret

$$\mathrm{CR}(N_E) := \mathbb{E}_{\theta, \theta_e, \mathbb{D}_e} \left[\sum_{e=0}^{N_E-1} \Delta_e \right] \text{ with episodic regret } \Delta_e := \mathbb{E}_x \left[V^{\theta}_{\theta_e,0}(x) - V^{\theta}_{\theta,0}(x) \right]. \tag{6}$$

Here, the regret Δ_e during each learning episode is taken with respect to the optimal soft-constrained MPC in terms of the nominal model and objective accuracy, that is, the MPC controller based on (3) using the true system parameters θ.

4. Bayesian model predictive control

The proposed learning-based MPC controller is based on posterior sampling as first proposed in the general reinforcement learning (RL) setting by Strens (2000). The resulting procedure is given in the Bayesian MPC Algorithm and works as follows. Based on the prior information $\mathcal{Q}_\theta$ about the system (1) and objective (2), a parameter realization $\theta_e = (\theta_{\ell,e}, \theta_{f,e})$ is sampled from the posterior belief at the beginning of each episode. The sample parametrizes the MPC problem (3) during the e-th episode, which results in an MPC controller that is sampled according to its a-posteriori probability of being optimal.

Bayesian MPC Algorithm

Data: Parametric model f, ℓ; Prior $\mathcal{Q}_\theta$
Initialize $\mathbb{D}_0 = \emptyset$
for *episodes* $e = 0, 1, .., N_E$ **do**
  sample $\theta_e \sim \mathcal{Q}_{\theta|\mathbb{D}_e}$
  for *time steps* $k = 0, 1, .., N-1$ **do**
    apply $u(k) = \pi_{\theta_e}(k, x(k))$
    measure objective and state
  end
  extend data set to obtain $\mathbb{D}_{e+1}$
end

By applying the sampled controller $u(k) = \pi_{\theta_e}(k, x(k))$, we obtain measurements of the state evolution and the objective value, leading to an update of the data set to $\mathbb{D}_{e+1}$ after N time steps. The collected data then refines the posterior belief about θ and the process is repeated in the subsequent episode. This mechanism naturally causes exploration in case of large uncertainties in the posterior distribution $\mathcal{Q}_{\theta|\mathbb{D}_e}$ due to rich variation in system trajectories through diverse MPC controller samples. At the same time, this mechanism also exploits collected knowledge as the posterior belief starts to cumulate around a consistent model of the true system. As a consequence, the performance of the sampled MPC will converge to that of the nominally optimal soft-constrained MPC.

5. Bound on finite-time learning performance

We apply the analysis provided by Osband et al. (2013); Osband and Van Roy (2014) to bound the cumulative regret (6) w.r.t. the nominal soft-constrained MPC controller using the true system dynamics. First, we reformulate the regret (6) in terms of the optimal cost-to-go that corresponds to the sampled parameters θ_e in episode e. This allows us in a second step to express the regret in terms of the learning progress of the system dynamics and objective function. By enforcing a regularity assumption on the expected cost-to-go under sampled MPC controllers we finally bound the regret in terms of posterior mean estimation errors of f and ℓ, allowing us to state the desired regret bound.

For the instant regret in episode e we have that

$$\mathbb{E}_{\theta,\theta_e,x,\mathbb{D}_e}\left[\Delta_e\right] = \mathbb{E}_{\theta,x,\mathbb{D}_e}\left[\mathbb{E}_{\theta_e}\left[\underbrace{V^{\theta}_{\theta_e,0}(x)}_{\text{Measured}} - \underbrace{V^{\theta}_{\theta,0}(x)}_{\text{Unknown}} | \theta, x, \mathbb{D}_e\right]\right].$$

Since $V^{\theta}_{\theta,0}(x)$ is unknown, we instead consider the regret in terms of the sampled MPC controller applied to the corresponding sampled system, for which it is optimal:

$$\mathbb{E}_{\theta,\theta_e,x,\mathbb{D}_e}\left[\tilde{\Delta}_e\right] = \mathbb{E}_{\theta,x,\mathbb{D}_e}\left[\mathbb{E}_{\theta_e}\left[\underbrace{V^{\theta}_{\theta_e,0}(x)}_{\text{Measured}} - \underbrace{V^{\theta_e}_{\theta_e,0}(x)}_{\text{Known}} | \theta, x, \mathbb{D}_e\right]\right]. \tag{7}$$

Using standard Thompson sampling (posterior matching) arguments we can verify that

$$\mathbb{E}_{\theta,\theta_e,x,\mathbb{D}_e}\left[\Delta_e - \tilde{\Delta}_e\right] = \mathbb{E}_{\mathbb{D}_e,x}\left[\mathbb{E}_{\theta,\theta_e}\left[\Delta_e - \tilde{\Delta}_e|\mathbb{D}_e, x\right]\right] \Rightarrow \mathbb{E}_{\theta,\theta_e,x,\mathbb{D}_e}\left[\Delta_e\right] = \mathbb{E}_{\theta,\theta_e,x,\mathbb{D}_e}[\tilde{\Delta}_e]$$

holds, since $\theta|\mathbb{D}_e$ and θ_e are equally distributed, yielding equally distributed MPC controllers $\pi_{\theta|\mathbb{D}_e}$, and π_{θ_e}, i.e. equally distributed time/state-to-input mappings, as well as equally distributed regrets[1] Δ_e and $\tilde{\Delta}_e$ (Russo and Van Roy, 2014).

Next, the goal is to express the regret $\tilde{\Delta}_e$ explicitly in terms of the posterior estimation accuracy of the functions f and ℓ to be learned instead of the episodic cost difference (7). As introduced in Osband et al. (2013), we define the recursive operator

$$\mathcal{T}^{\theta}_{\tilde{\theta},k}V(x) := \ell(k, x, \pi_{\tilde{\theta}}(k, x); \theta) + \mathbb{E}_w\left[V(x^+)|x^+ = f(x, \pi_{\tilde{\theta}}(k, x); \theta) + w\right] \tag{8}$$

1. Formally this requires that Δ_e is measurable with respect to the σ-algebra generated by the observed data $\mathbb{D}_e$.

at time steps $k = 0, 1, .., N-1$ to express the cost-to-go for system parameters θ under an MPC controller using potentially different parameters $\tilde{\theta}$. The cost under the optimal soft-constrained MPC (4) can therefore be written as $V^{\theta}_{\theta,k}(x) = \mathcal{T}^{\theta}_{\theta,k} V^{\theta}_{\theta,k+1}(x)$, in which case $\mathcal{T}^{\theta}_{\theta,k}$ relates to the Bellman operator. Repeated application of this relation allows us to eliminate the term $V^{\theta}_{\theta_e,0}(x)$ in (7). We sketch the corresponding derivation for the special case of zero process noise, i.e. $w(k) = 0$, for which we expand $\tilde{\Delta}_e(x)$ recursively using (8):

$$
\begin{aligned}
&V^{\theta}_{\theta_e,0}(x) - V^{\theta_e}_{\theta_e,0}(x) = \ell(0, x, \pi_{\theta_e}(0,x); \theta) + V^{\theta}_{\theta_e,1}(f(x, \pi_{\theta_e}(0,x); \theta)) - \ell(0, x, \pi_{\theta_e}(0,x); \theta_e) \\
&\qquad - V^{\theta_e}_{\theta_e,1}(f(x, \pi_{\theta_e}(0,x); \theta_e)) + V^{\theta_e}_{\theta_e,1}(f(x, \pi_{\theta_e}(0,x); \theta)) - V^{\theta_e}_{\theta_e,1}(f(x, \pi_{\theta_e}(0,x); \theta)) \\
&= \left(\mathcal{T}^{\theta}_{\theta_e,0} - \mathcal{T}^{\theta_e}_{\theta_e,0}\right) V^{\theta_e}_{\theta_e,1}(x) + V^{\theta}_{\theta_e,1}(x(1)) - V^{\theta_e}_{\theta_e,1}(x(1)) \\
&= \left(\mathcal{T}^{\theta}_{\theta_e,0} - \mathcal{T}^{\theta_e}_{\theta_e,0}\right) V^{\theta_e}_{\theta_e,1}(x) + \left(\mathcal{T}^{\theta}_{\theta_e,1} - \mathcal{T}^{\theta_e}_{\theta_e,1}\right) V^{\theta_e}_{\theta_e,2}(x(1)) + V^{\theta}_{\theta_e,2}(x(2)) - V^{\theta_e}_{\theta_e,2}(x(2)) \\
&\vdots \\
&= \sum_{k=0}^{N-1} \left(\mathcal{T}^{\theta}_{\theta_e,k} - \mathcal{T}^{\theta_e}_{\theta_e,k}\right) V^{\theta_e}_{\theta_e,k+1}(x(k)) + \underbrace{V^{\theta}_{\theta_e,N}(x(N))}_{=0} - \underbrace{V^{\theta_e}_{\theta_e,N}(x(N))}_{=0},
\end{aligned}
$$

where $x(0) = x$ and $x(k+1) = f(x(k), \pi_{\theta_e}(k, x(k)); \theta)$. Including again the process noise $w(k)$, this result enables us to bound

$$
\begin{aligned}
\mathbb{E}[\tilde{\Delta}_e] \leq &\mathbb{E}\left[\sum_{k=0}^{N-1} \mathbb{E}_{w(k)}\left[|V^{\theta_e}_{\theta_e,k+1}(f(x(k), u(k); \theta) + w(k)) - V^{\theta_e}_{\theta_e,k+1}(f(x(k), u(k); \theta_e) + w(k))|\right]\right] + \\
&\mathbb{E}\left[\sum_{k=0}^{N-1} |\ell(k, x(k), u(k); \theta) - \ell(k, x(k), u(k); \theta_e)|\right], \qquad (9)
\end{aligned}
$$

where the outer expectation is taken w.r.t. $\theta, \theta_e, x, \mathbb{D}_e$. Consequently, we can bound the second term in (9) in terms of $\mathbb{E}_{\theta,x,\mathbb{D}_e}[\mathbb{E}_{\theta_e}[\sum_{k=0}^{N-1} |\ell(k, x(k), u(k); \theta) - \ell(k, x(k), u(k); \theta_e)| \mid \theta, x, \mathbb{D}_e]]$, that is, bounding the conditional posterior mean error of the cost $|\ell(.; \theta) - \ell(.; \theta_e)|$ based on the real problem parameter realization θ, initial condition x, and observed data $\mathbb{D}_e$ up to episode e. To derive a similar bound on the first term in (9) with respect to the conditional posterior mean error of the dynamics, we follow Osband and Van Roy (2014) and assume the following regularity property on the expected cost-to-go.

Assumption 1 *For all $\theta_e \in \mathbb{R}^{n_\theta}$ and $x^+, \tilde{x}^+ \in \mathbb{X}$ there exists a constant $L_V > 0$ such that*

$$
\mathbb{E}_{w(k)}\left[|V^{\theta_e}_{\theta_e,k+1}(x^+ + w(k)) - V^{\theta_e}_{\theta_e,k+1}(\tilde{x}^+ + w(k))|\right] \leq L_V \left\|x^+ - \tilde{x}^+\right\|_2.
$$

Note that Assumption 1 can, e.g., be satisfied for the standard case of linear dynamics and positive definite quadratic objectives. This allows us to bound the first term in (9) in a similar fashion by $\mathbb{E}_{\theta,x,\mathbb{D}_e}[\mathbb{E}_{\theta_e}[\sum_{k=0}^{N-1} L_V \|f(x(k), u(k); \theta) - f(x(k), u(k); \theta_e)\|_2 \mid \theta, x, \mathbb{D}_e]]$.

The previously outlined analysis steps provide a bound on the expected regret in terms of the deviation between the posterior mean estimates of f and ℓ, conditioned on observed data, and the true underlying system dynamics and objective function. As first proposed by Russo and Van Roy (2014) for the case of real-valued functions in the bandit optimization setting and later extended by

Osband and Van Roy (2014) to vector-valued functions in the context of RL, the magnitude of this deviation can be described using two distinct measures of complexity. The first measure is given by the classical Kolmogorov dimensions $\dim_K(\ell)$ and $\dim_K(f)$ describing the complexity of ℓ and f in the parameters θ, see, e.g., Russo and Van Roy (2014, Section 7.1). The other measure is called Eluder dimension, denoted by $\dim_E(\ell)$ and $\dim_E(f)$ and describes the complexity of the mean inference problem based on sequentially obtained measurements. More details and explicit bounds on $\dim_E(\ell)$ can be found in Osband and Van Roy (2014, Section 4.1). Using these measures of complexity, we get the following regret bound as an immediate consequence of Osband and Van Roy (2014, Theorem 1) with $\tilde{\mathcal{O}}$ neglecting terms that are logarithmic in N_E.

Corollary 1 *Let Assumption 1 hold. If there exist constants c_ℓ and c_f such that for all admissible $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $\theta \in \mathbb{R}^{n_{\theta_\ell}}$, and $k = 0, 1, .., N$ it holds $\ell(k, x, u; \theta_\ell) \leq c_\ell$, and $f(x, u; \theta_f) \leq c_f$, then it follows that*

$$\mathrm{CR}(N_E) \leq \tilde{\mathcal{O}}\left(\sigma_\epsilon \sqrt{\dim_K(\ell)\dim_E(\ell)N_E N} + L_V \sigma_w \sqrt{\dim_K(f)\dim_E(f)N_E N}\right).$$

As a direct consequence of Osband and Van Roy (2014, Proposition 2) we obtain the following bound for the important special case of linear Bayesian regression.

Corollary 2 *Let the assumptions of Corollary 1 hold. If $f(x, u; \theta_f) = \theta_f^\top \Phi_f(x, u)$ and $l(x, u; \theta_f) = \theta_\ell^\top \Phi_\ell(x, u)$ with $\theta_\ell \in \mathbb{R}^{n_\ell}$ and $\theta_f \in \mathbb{R}^{n_f \times n}$, then*

$$\mathrm{CR}(N_E) \leq \tilde{\mathcal{O}}\left(\sigma_\epsilon \sqrt{n_\ell N_E N} + L_V \sigma_w n \sqrt{n n_f N_E N}\right).$$

While Corollary 1 provides a general regret bound in terms of $\dim_E$ and $\dim_K$, Corollary 2 ensures a finite-time learning progress through a sub-linear bound on the cumulated regret in case of linear Bayesian regression. However, also in the general case of Corollary 1, the regret bound scales naturally with the process and measurement noise, as well as with the regularity property of the expected cost-to-go according to Assumption 1.

Note that the regret bounds are valid for the objective function (3a) including the slack variables that indicate constraint violations. Consequently, the cumulative regret in this case also bounds the cumulated amount of expected constraint violation during different learning episodes.

6. Numerical results

We consider the problem of learning how to drive a car-trailer system with partially known dynamics backwards. Starting from a random initial system configuration, the goal is to reach an uncertain goal position as depicted in Figure 1 (left) with the car and trailer being horizontally aligned. The system dynamics according to Figure 1 (middle) are obtained through a Euler-forward discretization with sampling time $T_s = 0.1[\mathrm{s}]$ of the model dynamics presented in Rouchon et al. (1993). By denoting $x^+ := x(k+1)$, $x := x(k)$, and $u := u(k)$ we get a prediction model of the form $x^+ = f(x, u; \theta_f) + w$ with states $x = [y_c, \phi, \delta, \kappa, x_c, v_c]^\top$, inputs $u = [\omega_\delta, a_c]$, and dynamics

$$\begin{array}{lll} x_c^+ = x_c + T_s v_c & y_c^+ = y_c + T_s v_c \sin(\phi) & v_c^+ = v_c + T_s a_c \\ \phi^+ = \phi + T_s a^{-1} v \tan(\delta) & \delta^+ = \delta + \theta_1^\top \Phi_1(x) & \kappa^+ = \kappa + \theta_2^\top \Phi_2(x). \end{array}$$

The process noise $w \sim \mathcal{Q}_w$ accounts for model mismatch and is given as $\mathcal{Q}_w = \mathbb{N}(0, \Sigma_w)$ with $\Sigma_w = \mathrm{diag}(T_s[0.03, 0.017, 0.1, .01, 0.01, 0.01])$. State and input constraints are $|\kappa - \phi| \leq 0.7$ [rad],

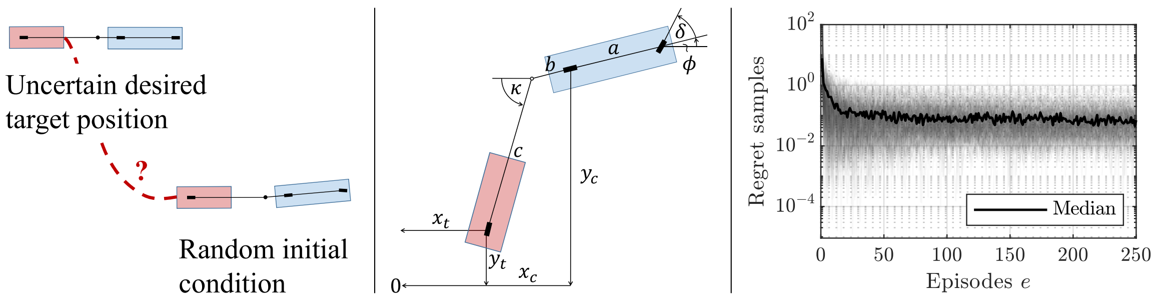

Figure 1: **Left:** Problem setting of steering a partially known trailer system from a random initial condition to an uncertain target position. **Middle:** Car-trailer system. **Right:** Sampled regret Δ_e and median of 200 different systems over 250 learning episodes.

$x > 1$ [m], $|\delta| > 0.7$ [rad], $|\omega_\delta| \leq 1.22$ [rad], and $|a| \leq 2[\mathrm{m/s^2}]$. The unknown terms in δ^+ and κ^+ are parametrized by features $\Phi_1(x) = \omega_\delta$ and $\Phi_2(x) = [v\sin(\kappa - \phi), v\tan(\delta)\cos(\kappa - \phi)]^\top$ and parameters θ_1 and θ_2, describing the steering dynamics and the trailer geometry. The objective of reaching the goal configuration can be encoded into a terminal cost $\ell(N-1, x, u) := \phi^2 + \kappa^2 + v^2 + y_t^2 + x_t^2 + \theta_\ell \Phi_3(x)$, with $\ell(i, x, u) = 0$ for $i = 0, 1, ..N-2$, where $\Phi_3(x) = [x_c^2, x_c, y_c^2, y_c]^\top$ describes the desired target position. After each episode, we obtain very noisy feedback from, e.g., a vision system or a person that is modeled by zero mean normally distributed measurement noise $\epsilon_\ell \sim \mathbb{N}(0, 0.5^2)$.

For learning, we consider a prior distribution $\mathcal{Q}_\theta$ that corresponds to a standard deviation of 0.45 [m] in the trailer length, 10 [s/deg] in steering dynamics, and 0.5 [m] in the desired position $[x_d, y_d]^\top$. Note that the theoretical results from Corollary 1 only hold, if the states and objectives are bounded within one episode, which is practically fulfilled in this example due to the MPC controller that ensures bounded input signals.

In Figure 1 (right), we plot the measured difference between the optimal and sampled MPC (=sampled regret), simulated with 200 different system and objective realizations that are sampled according to their prior distribution $\mathcal{Q}_\theta$. During the first 15 episodes the median drops quickly to a slowly degrading regret, depending on the process noise and measurement noise magnitude.

7. Conclusion

In this paper, we considered episodic learning tasks for unknown dynamical systems and objective functions subject to state and input constraints. To enable efficient, easily implementable learning-based control, we combined Bayesian posterior sampling theory with model predictive control techniques. The learning performance of the proposed approach can formally be bounded in terms of the regret w.r.t. the optimal model predictive controller. The efficiency of the algorithm was demonstrated in simulation using a reverse driving task with a nonlinear car-trailer system.

Acknowledgments

We thank Benjamin Van Roy for the fruitful discussion at ETH Zurich in spring 2017. This work was supported by the Swiss National Science Foundation under grant no. PP00P2_157601/1.

References

Elena Arcari, Lukas Hewing, and Melanie N Zeilinger. An approximate dynamic programming approach for dual stochastic model predictive control. *IFAC World Congress (accepted)*, 2020.

Julian Berberich, Johannes Köhler, Matthias A. Müller, and Frank Allgöwer. Data-driven model predictive control with stability and robustness guarantees. *arXiv:1906.04679*, 2019.

A. Carron, E. Arcari, M. Wermelinger, L. Hewing, M. Hutter, and M. N. Zeilinger. Data-driven model predictive control for trajectory tracking with a robotic arm. *IEEE Robotics and Automation Letters*, 4(4):3758–3765, Oct 2019.

J. Coulson, J. Lygeros, and F. Dörfler. Data-enabled predictive control: In the shallows of the DeePC. In *2019 European Control Conference (ECC)*, pages 307–312, 2019.

A. Domahidi, A. U. Zgraggen, M. N. Zeilinger, M. Morari, and C. N. Jones. Efficient interior point methods for multistage problems arising in receding horizon control. In *2012 IEEE Conference on Decision and Control (CDC)*, pages 668–674, 2012.

Kristian G. Hanssen and Bjarne Foss. Scenario based implicit dual model predictive control. Proc. 5th IFAC Conf. Nonlinear Model Predictive Control, pages 416–421, 2015.

Tor Aksel N. Heirung, B. Erik Ydstie, and Bjarne Foss. Dual adaptive model predictive control. *Automatica*, 80:340–348, 2017.

Lukas Hewing, Juraj Kabzan, and Melanie N. Zeilinger. Cautious model predictive control using Gaussian process regression. *arXiv*, page 1711.06586v3, 2018.

B. Houska, H.J. Ferreau, and M. Diehl. ACADO Toolkit – An Open Source Framework for Automatic Control and Dynamic Optimization. *Optimal Control Applications and Methods*, 32(3): 298–312, 2011.

Sanket Kamthe and Marc Deisenroth. Data-efficient reinforcement learning with probabilistic model predictive control. In *Proc. of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1701–1710. PMLR, 2018.

Eric C Kerrigan and Jan M Maciejowski. Soft constraints and exact penalty functions in model predictive control. In *Proc. UKACC International Conference (Control 2000)*, 2000.

Edgar D. Klenske and Philipp Hennig. Dual control for approximate Bayesian reinforcement learning. *J. Mach. Learn. Res.*, 17(127):1–30, 2016.

T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause. Learning-based model predictive control for safe exploration. 2018 IEEE Conference on Decision and Control (CDC), pages 6059–6066, 2018.

Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

Ali Mesbah. Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annu. Rev. in Control*, 45:107–117, 2018. ISSN 1367-5788. doi: 10.1016/j.arcontrol.2017.11.001.

Manfred Morari and Jay H. Lee. Model predictive control: Past, present and future. *Comput. & Chem. Eng.*, 23(4):667–682, 1999.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the Eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

S. Joe Qin and Thomas A. Badgwell. An overview of nonlinear model predictive control applications. Nonlinear Model Predictive Control, pages 369–392. Birkhäuser Basel, 2000.

Pierre Rouchon, Michel Fliess, Jean Lévine, and Philippe Martin. Flatness, motion planning and trailer systems. In *IEEE Conference on Decision and Control*, volume 3, pages 2700–2700, 1993.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Raffaele Soloperto, Matthias A. Müller, Sebastian Trimpe, and Frank Allgöwer. Learning-based robust model predictive control with state-dependent uncertainty. Proc. 6th IFAC Conf. Nonlinear Model Predictive Control, pages 442–d447, 2018.

Malcolm Strens. A bayesian framework for reinforcement learning. In *Proc. of the Seventeenth International Conference on Machine Learning*, pages 943–950, 2000.

Y. Wang and S. Boyd. Fast model predictive control using online optimization. *IEEE Transactions on Control Systems Technology*, 18(2):267–278, 2010.

Hua Yang and Shaoyuan Li. A data-driven predictive controller design based on reduced hankel matrix. Proc. 10th Asian Control Conf., pages 1–7, 2015.

A. Zanelli, A. Domahidi, J. Jerez, and M. Morari. Forces nlp: an efficient implementation of interior-point methods for multistage nonlinear nonconvex programs. *International Journal of Control*, pages 1–17, 2017.