Combining Model-based Controller and ML Advice via Convex Reparameterization

Junxuan Shen JSHEN@CALTECH.EDU
Adam Wierman ADAMW@CALTECH.EDU
Caltech

Guannan Qu GQU@ANDREW.CMU.EDU
CMU

Editors: A. Abate, K. Margellos, A. Papachristodoulou

Abstract

Machine Learning (ML) based control, particularly Reinforcement Learning (RL), has achieved impressive advancements but is often black-box and lacks worst-case guarantees in safety-critical systems. In contrast, classical model-based control offers stability guarantees but usually underperforms the machine-learned black-box controller. This motivates us to combine machine-learned black-box and model-based controllers. Due to the nonconvexity of the space of stable controllers, a simple convex combination of the two controllers can lead to instability. We propose using Disturbance Response Control (DRC) to reparameterize the two controllers, ensuring the convexity of the stable controller space. We then propose λ-CLEAC, which adaptively combines the machine-learned black-box controller and the model-based controller in the DRC parameterization. We prove that our approach achieves the best of both worlds: stability as in model-based control and similar regret bounds as the machine-learned controller.

Keywords: Learning-augmented control, stability, black-box policy

1. Introduction

Machine-learned controllers, such as those learned by Reinforcement Learning (RL), have demonstrated remarkable success in various domains, including game playing (Kiran et al. (2021); Silver et al. (2018)), fine-tuning of large language models (Chang et al. (2023); Lee et al. (2023)), online advertising (Zhao et al. (2021)). However, when it comes to safety-critical systems such as robotics, energy management, and autonomous driving, machine-learned controllers are often black-box controllers that fall short in providing worst-case guarantees. This is in sharp contrast with model-based control, where tools such as Linear Quadratic Regulators (LQR), Model Predictive Control (MPC), and robust control (Anderson and Moore (2007); García et al. (1989); Dorato (1987)) can help provide stability guarantees (Dean et al. (2017); Doyle (1996)). This discrepancy motivates the need for learning-augmented control, which is to combine a machine-learned black-box controller with a model-based controller to obtain a controller with good performance when the machine-learned black-box controller is effective while maintaining stability guarantees of the model-based controller.

The motivation described above has driven significant progress in the design of learning-augmented algorithms for a wide range of problems. This line of work aims to leverage machine-learned black-box advice to counterbalance the conservatism of traditional algorithms designed to optimize worst-

case performance bounds, thus achieving near-optimal performance when the machine-learned advice is reliable while ensuring performance bounds when the machine-learned advice is not dependable. For example, learning-augmented algorithms have been designed for various online problems, such as ski rental (Shah and Rajkumar (2021); Wang et al. (2020)), caching (Lykouris and Vassilvtiskii (2018); Rohatgi (2020); Wei (2020)), convex body/function chasing (Christianson et al. (2022); Sellke (2023); Rutten et al. (2023)), and more generally, metrical task systems (Antoniadis et al. (2023); Christianson et al. (2023)). The underlying idea of these algorithms is usually via switching between the untrusted machine-learned and the model-based decisions or, more generally speaking, combining the decisions in a convex way.

When it comes to designing learning-augmented algorithms for dynamical systems, there have been several recent works (Nagabandi et al. (2018); Rosolia and Borrelli (2018); Pong et al. (2018); Qu et al. (2021); Li et al. (2022a,b), where the idea of a convex combination between machine-learned and model-based controllers is also employed. However, there is a critical issue with directly combining two controllers via a convex combination. It is widely known that the space of stable controllers is not convex (Fazel et al. (2019)), so combining two stable controllers through a naive convex combination could lead to instability. To see this, in the most basic linear dynamical system setting, it is easy to construct examples where some convex combination of two stable linear controllers K and K' can be unstable Fazel et al. (2019). This shows that the convex combination idea is not suitable for designing learning-augmented algorithms for dynamical systems. In previous work (Li et al. (2022a)), to counteract the above problem, the convex combination parameter for the machine-learned black-box controller must be set to close to 0 after a finite number of steps, meaning that it does not fully utilize the potential benefits of the machine-learned black-box controller. This leads us to ask the question: *Rather than combining the controllers directly using a convex combination, is there another way to combine the controllers to fully realize the potential benefits of the machine-learned black-box controller?*

Contributions. In this paper, we propose a novel approach to address the above question. Rather than simply using a convex combination of the two controllers, our high-level idea is to reparameterize the controllers to make the space of the stable controllers convex. This allows us to combine these reparameterized controllers while ensuring stability. Specifically, we use the Disturbance Response Control (DRC) parameterization proposed by (Simchowitz et al. (2020)). Although DRC was originally used for linear controllers in LTI systems, we generalize the scheme to nonlinear controllers and prove the convexity of the space of stable controllers in LTI systems. With the DRC parameterization of the machine-learned black-box controller and the model-based controller, we propose λ-CLEAC, which adaptively selects the confidence parameters that combine the two controllers. We show that in the LTI setting, λ-CLEAC achieves a bounded state norm (stability) as well as similar regret as the machine-learned black-box controller. This implies that when the machine-learned black-box controller performs well, our algorithm also performs well while maintaining stability guarantees. Additionally, our work extends beyond LTI systems, generalizing to LTV systems and LTI systems with model mismatch. In both settings, we provide adaptive policies to combine the two controllers and show similar guarantees: the bounded state norm (stability) and similar regret bounds as the machine-learned black-box controller. Finally, we perform experiments to validate the effectiveness of our approach.

Related Work. Our research is broadly related to reinforcement learning literature (Oh et al. (2020); Canese et al. (2021); Moerland et al. (2023); Szepesvári (2022); Kaelbling et al. (1996)) and the learning-based control literature (Levine (2018); Wabersich and Zeilinger (2018); Hewing et al.

(2020); Fisac et al. (2019); Buşoniu et al. (2018)). More specifically, work is connected to a range of works aiming to bridge machine-learned black-box and model-based approaches.

Combination of model-based controllers with model-free controllers. Our study contributes to the latest research endeavors that aim to combine model-free and model-based controllers for online control. We include some of the prominent works in this field below: Pong et al. (2018) establishes a connection between Q-learning and model predictive control (MPC). Rosolia and Borrelli (2018) explores MPC methods incorporating penalty terms acquired through model-free algorithms. Nagabandi et al. (2018) employs deep neural network dynamics models to initialize a model-free learner, enhancing sample efficiency while preserving high task-specific performance. In Qu et al. (2021), the authors examine a specific dynamical system described by $x_{t+1} = Ax_t + Bu_t + r(x_t)$, where f represents residual dynamics, and demonstrate that initializing a model-free policy with a model-based approach is guaranteed to converge toward a nearly optimal linear controller. Our work is mostly related to Li et al. (2022a), where the authors provide a policy that adaptively chooses the confidence parameters to combine the machine-learned black-box controller with a model-based one, ensuring stability and a competitive ratio. However, a downside of this approach is its tendency to overlook the machine-learned black-box controller after a finite number of steps. In contrast, our algorithm addresses this limitation by consistently upholding a positive confidence parameter for the machine-learned black-box controller.

Learning-augmented online problems. The idea of harnessing the machine-learned black-box advice with traditional robust online algorithms, known as *learning-augmented algorithms*, has motivated significant research advancement in various online algorithm settings, such as ski rental (Shah and Rajkumar (2021); Wang et al. (2020)), caching (Lykouris and Vassilvtiskii (2018); Rohatgi (2020); Wei (2020)), convex body/function chasing (Sellke (2023); Rutten et al. (2023)), and metrical task systems (Antoniadis et al. (2023); Christianson et al. (2023)). Typically, these algorithms make decisions by dynamically switching between the machine-learned black-box decision and the model-based decision or more broadly speaking, by combining them in a convex manner. However, when dealing with dynamical systems, directly combining two stable controllers in a convex way can lead to instability. Our contribution lies in addressing this challenge by reparameterizing the controller space so that the space of stable controllers becomes convex.

2. Problem Formulation and Preliminaries

2.1. Problem Formulation

Consider a general dynamical system

$$x_{t+1} = f_t(x_t, u_t, w_t), \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the system state, $u_t \in \mathbb{R}^m$ is the control action at time t, $f_t : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ is the system dynamic, and $w_t \in \mathbb{R}^n$ are the disturbance. We study the following control problem:

$$\min_{u_t : t \geq 0} \sum_{t=0}^{T} b_t(x_t) + c_t(u_t), \text{ subject to (1)},$$

where $b_t : \mathbb{R}^n \to \mathbb{R}$ and $c_t : \mathbb{R}^m \to \mathbb{R}$ are convex and differentiable for all $t \geq 0$, with a fixed initial state $x_0 \in \mathbb{R}^n$.

Suppose we are provided with a machine-learned black-box controller g_{ml} and a model-based controller g_{mb}. The model-based controller is known to be stable, whereas the machine-learned black-box controller has potentially superior control performance, but may sometimes be unstable. Our goal is to design a controller that achieves similar stability guarantees as the model-based controller, and similar performance as the machine-learned black-box controller when it performs well.[1] More formally, by stability, we use the following definitions.

Definition 1 *A system (1) is **input-to-state stable (ISS)** under a controller if, for any bounded disturbance sequence $\|w_t\| \leq W$, there exists some $C_1, C_2 > 0$, such that for all $t \geq 0$ and $x_0 \in \mathbb{R}^n$, $\|x_t\| \leq C_1\|x_0\| + C_2 W$. A system (1) is **exponential input-to-state stable (Exp-ISS)** under a controller if, there exists some $C_1, C_2 > 0$, $\beta \in (0,1)$, such that for all $t \geq 0$, $\|x_t\| \leq C_1\beta^t\|x_0\| + C_2 W$. Further, a system (1) is **incrementally exponential input-to-state stable (Inc-Exp-ISS)** under a controller, if there exists $C_1, C_2 > 0$, $\beta \in (0,1)$, such that for all $t \geq 0$,*

$$\|x_t^1 - x_t^2\| \leq C_1\beta^t\|x_0^1 - x_0^2\| + C_2\|\mathcal{W}_1 - \mathcal{W}_2\|_\infty, \tag{2}$$

where $\{x_t^1\}$ and $\{x_t^2\}$ are two trajectories under the controller starting at x_0^1, x_0^2 and under disturbances $w_0^1, w_1^1 \ldots$ and $w_0^2, w_1^2, \ldots$ respectively. Notations $\mathcal{W}_1 = [w_0^1 \mid w_1^1 \mid \cdots \mid w_t^1]^T, \mathcal{W}_2 = [w_0^2 \mid w_1^2 \mid \cdots \mid w_t^2]^T$ are the disturbances stacked in matrix form for the two trajectories respectively.

Further, the performance is formally defined as regret.

Definition 2 *Let ALG be an online algorithm that chooses control action u_t at each time t. Define its performance cost as $\mathsf{Cost}(ALG) = \sum_{t=0}^{T} b_t(x_t) + c_t(u_t)$. Similarly, let $\mathsf{Cost}(OPT)$ be the cost of the offline optimal controller; let $\mathsf{Cost}(ALG)$ be the cost of the algorithm ALG. Then we define regret of ALG as $\mathsf{Regret}(\mathsf{ALG}) = \mathsf{Cost}(\mathsf{ALG}) - \mathsf{Cost}(\mathsf{OPT})$.*

With the above definition, our goal can be formally stated as *How to combine the machine-learned black-box controller and the model-based controller to get a controller that is ISS and further achieves similar regret bound as the machine-learned black-box controller?*

In answering the above question, we will first consider the LTI setting (in Section 3),

$$f_t(x_t, u_t, w_t) = Ax_t + Bu_t + w_t, \tag{3}$$

as the LTI setting streamlines the presentation and best illustrates our ideas. Here, we assume that A satisfies $\|A^t\| \leq C_A\rho^t$ for some $C_A > 0$ and $\rho \in (0,1)$. Our result easily generalizes beyond LTI, and in Section 4, we consider the Linear Time-Varying (LTV) system and LTI system with model mismatch, which are general enough to capture many realistic dynamical systems.

Finally, in the LTI setting, we assume the system matrices A, B are known, which is reasonable if one has a simulator for the underlying dynamics. In addition, this assumption will be relaxed in the LTV and the LTI with a model mismatch setting. In the LTV setting, only the systems matrices up to the current time step need to be known, which is realistic as the LTV system can be thought as a linearization of the nonlinear system around the past trajectory; in the LTI with nonlinear model mismatch setting, the nonlinear mismatch is unknown, which is also realistic as oftentimes, one has a good linear approximation of the unknown nonlinear system, and the mismatch can be thought as the linearization error.

1. In the paper, we label variables related to the machine-learned black-box controller as ml and the model-based controller as mb.

2.2. DRC parameterization

As mentioned earlier, if we naively combine two controllers g_{ml} and g_{mb} with convex parameter $\lambda \in [0,1]$ to get $\lambda g_{\mathrm{ml}} + (1-\lambda) g_{\mathrm{mb}}$, the combined controller can be unstable even if both g_{ml} and g_{mb} are stable. This issue motivates us to reparameterize the controllers using Disturbance Response Control (DRC), which is commonly used in online control.

We now introduce DRC parameterization for *linear* controllers in LTI systems (3), which was proposed by Simchowitz et al. (2020). We first define the natural state.

Definition 3 *Given a sequence of disturbances w_t and initial state x_0, the **natural states** for the LTI system (3) are*

$$x_t^n = A^t x_0 + \sum_{i=0}^{t-1} A^{t-1-i} w_i. \tag{4}$$

DRC parameterization means that instead of writing the controller as a function of the current observed state $g(x_t)$, it writes the controller as a function of the *natural states*, which according to (4) are independent of the past control actions u_t. As an example, we show that the linear controller $u_t = Kx_t$ can be written in the DRC form. Note that we have $w_t = x_{t+1}^n - Ax_t^n$, so we can then reparameterize the state x_t as a linear function of the natural states:

$$x_t = (A+BK)x_{t-1} + w_{t-1} = (A+BK)^t x_0^n + \sum_{i=0}^{t-1} (A+BK)^{t-1-i}(x_{i+1}^n - Ax_i^n).$$

Then we can rewrite $u_t = Kx_t$ as $u_t = \sum_{i=0}^{t} M_i x_{t-i}^n \approx \sum_{i=0}^{h} M_i x_{t-i}^n$, where $M_i = KBK(A+BK)^{i-1}$. Here, the controller is parameterized by $(M_0, \ldots, M_h)$ for some constant $h \ll t$ that acts upon the *natural states* x_i^n. The approximation is valid up to error $O(\exp(-h))$ if K is a stable controller, i.e. the spectral radius $\rho(A+BK) < 1$.

The benefit of DRC is that the control input u_t can be written as a *convex* function of $(M_0, \ldots, M_h)$, which also means the space of stable controllers is convex. Note that the original definition of DRC is only for LTI systems and linear controllers. In the next subsections, we will generalize DRC to nonlinear controllers, which will be vital to developing our approach.

3. Main Results

As mentioned in the previous section, throughout the section, we focus on the LTI setting (3) as this streamlines the presentation and best illustrates our core idea. We generalize beyond LTI in Section 4.

To present our result, we first generalize the DRC parameterization to nonlinear controllers and show that the space of stable nonlinear controllers is convex after DRC parameterization (Section 3.1). Then, utilizing this result, we propose to combine the machine-learned black-box controller and the model-based controller in DRC parameterization, and we propose λ-CLEAC that adaptively selects the confidence parameter (Section 3.2). Finally, we formally prove the stability and regret bounds for λ-CLEAC (Section 3.3). Due to space limits, all proofs of the paper can be found in the online appendix, accessible at Shen et al. (2023).

3.1. DRC for nonlinear controllers

As Section 2.2 gives the DRC parameterization for linear controllers in LTI systems, we now generalize the DRC for nonlinear controllers in LTI systems. Consider a nonlinear controller $g : \mathbb{R}^n \to \mathbb{R}^m$, which gives the control input $u_t = g(x_t)$ and generates the state sequence

$$x_t = Ax_{t-1} + Bg(x_{t-1}) + w_{t-1}, \tag{5}$$

where A satisfies $\|A^t\| \leq C_A\rho^t$ for some $C_A > 0$ and $\rho \in (0, 1)$ and g is L_0-Lipschitz, i.e. $\|g(x) - g(y)\| \leq L_0\|x - y\|$ for all $x, y \in \mathbb{R}^n$.

Define the natural states the same way as in (4), i.e. $x_t^n = A^t x_0 + \sum_{i=0}^{t-1} A^{t-1-i} w_i$, so that it only depends on the initial state x_0 and the noise inputs w_i. We show that we can rewrite any Inc-Exp-ISS controller as a function of the natural states with an exponentially small approximation error:

Theorem 4 *Consider the LTI system (3) with bounded disturbance sequence $\|w_t\| \leq W$, with the norm satisfying $\|v\|_1 \leq C_0\|v\|$ for some $C_0 > 0$ for all v. Let $u_t = g(x_t)$ be an Inc-Exp-ISS controller with parameters C_1, C_2, β. Then for any $h \leq t$, $h \in \mathbb{Z}_+$, there is a function $\mathcal{G} : \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{h+1 \text{ times}} \to \mathbb{R}^m$ s.t. $\|g(x_t) - \mathcal{G}(x_{t-h}^n, \cdots, x_t^n)\| \leq L_0 C_0 C_1 C_2 \beta^{h+1} W$.*

The proof of Theorem 4 can be found in Appendix A.1 Shen et al. (2023). The above theorem shows that any stable nonlinear controller can be converted to DRC parameterization as given by the function $\mathcal{G}$, which shows the generality of DRC. This theorem also justifies our following algorithm which combines two controllers in their DRC form.

In what follows, we show an important benefit of DRC parameterization, which is when combining two stable controllers, the resulting controller is stable. The proof of Theorem 5 can be found in Appendix A.2 in Shen et al. (2023).

Theorem 5 *Let $\mathcal{G}_1, \mathcal{G}_2$ be two DRC controllers that are Exp-ISS. Then for any $\lambda \in [0, 1]$, the combined controller $\mathcal{G} = \lambda\mathcal{G}_1 + (1 - \lambda)\mathcal{G}_2$ is also Exp-ISS.*

3.2. Proposed Algorithm

The previous section shows the set of stabilizing controllers is convex in DRC parameterization. We utilize this fact and propose to combine the model-based and machine-learned black-box controller in the DRC parameterization. Our approach features adaptively selecting confidence coefficients, $\lambda_t \in [0, 1]$ to combine the machine-learned and the model-based controllers in the DRC parameterization, which ensures that the resulting controller is stable and performs well when the machine-learned black-box controller performs well.

Consider the LTI system (3). Suppose we are given a machine-learned black-box controller $u_t^{\text{ml}} = \mathcal{G}^{\text{ml}}(x_{t-h}^n, \cdots, x_t^n)$ and a model-based controller $u_t^{\text{mb}} = \mathcal{G}^{\text{mb}}(x_{t-h}^n, \cdots, x_t^n)$, yielding two state sequences given by

$$x_t^{\text{ml}} = Ax_{t-1}^{\text{ml}} + Bu_{t-1}^{\text{ml}} + w_{t-1} \text{ and } x_t^{\text{mb}} = Ax_{t-1}^{\text{mb}} + Bu_{t-1}^{\text{mb}} + w_{t-1}, \tag{6}$$

with initial states $x_0^{\text{ml}} = x_0^{\text{mb}} = x_0$. The machine-learned black-box controller can potentially be unstable, but the model-based controller is stable, indicated in Assumption 6(a). Otherwise, we do

not place other assumptions, other than the control action should have bounded norms (Assumption 6(b)), which is a reasonable assumption, as in many problem settings, there is a saturation limit for the control actions.

Assumption 6 *We place the following two assumptions on the two controllers*

(a) The model-based policy $\mathcal{G}_{\text{mb}}$ stabilizes the system (3), i.e. for any fixed initial state $x_0 \in \mathbb{R}^n$ and norm $\|\cdot\|$, there exists $R_{\text{mb}} > 0$ such that for all $t \geq 0$, $\|x_t^{\text{mb}}\| \leq R_{\text{mb}}$, where x_t^{mb} is defined in (6).

(b) For both policies $\mathcal{G}_{\text{ml}}$ and $\mathcal{G}_{\text{mb}}$, the control actions have bounded norms, i.e. for any fixed initial state $x_0 \in \mathbb{R}^n$ and norm $\|\cdot\|$, there exists $U > 0$ such that for all $t \geq 0$, $\|u_t^{\text{ml}}\|, \|u_t^{\text{mb}}\| \leq U$.

Algorithm 1: λ CONVEX LEARNING ASSISTED CONTROL (λ-CLEAC)

Data: System parameters A, B; DRC controllers $\mathcal{G}_{\text{ml}}, \mathcal{G}_{\text{mb}}$

1 $R \leftarrow R_{\text{mb}} + L$
2 $x_0^{\text{ml}} = x_0^{\text{mb}} = x_0$
3 **for** $t \geq 0$ **do**
4 Observe x_t, calculate $w_{t-1} = x_t - Ax_{t-1} - Bu_{t-1}$
5 $x_t^n = Ax_{t-1}^n + w_{t-1}$
6 $u_t^{\text{ml}} = \mathcal{G}_{\text{ml}}(x_{t-h}^n, \cdots, x_t^n)$
7 $u_t^{\text{mb}} = \mathcal{G}_{\text{mb}}(x_{t-h}^n, \cdots, x_t^n)$
8 **if** $\|x_t^{\text{ml}}\| \leq R$ **then**
9 $\lambda_t \leftarrow 1$
10 **else**
11 $\lambda_t = \frac{R - R_{\text{mb}}}{\|x_t^{\text{ml}}\| - R_{\text{mb}}}$
12 **end**
13 Apply control action $u_t = \lambda_t u_t^{\text{ml}} + (1 - \lambda_t) u_t^{\text{mb}}$
14 **end**

Given the above two controllers, the pseudocode of the proposed algorithm is provided in λ-CLEAC, which works intuitively as follows. Initially, the user can adjust the threshold L (line 1). At each time step t, we first compute the natural state x_t^n (line 5), then compute the control inputs of both machine-learned black-box and model-based controllers via DRC parameterization (lines 6-7). Then we can generate the machine-learned black-box state input and compute its state norm (line 8). We then compare the state norm with the threshold and set the confidence parameter accordingly (lines 9 - 11).

One key step in our algorithm is the selection of λ_t, which is designed so that each time the state norm of the machine-learned black-box policy exceeds the threshold, we set the confidence in the machine-learned black-box controller λ_t to be $\frac{R - R_{\text{mb}}}{\|x_t^{\text{ml}}\| - R_{\text{mb}}} < 1$, and set it as 1 otherwise. The rationale behind this approach lies in the inherent property of the DRC parameterization, where $x_t = \lambda x_t^{\text{ml}} + (1 - \lambda) x_t^{\text{mb}}$ as long as $u_t = \lambda u_t^{\text{ml}} + (1 - \lambda) u_t^{\text{mb}}$, with a fixed λ.

We can show that, with a varying λ_t, we can still approximately have $x_t \approx \lambda_t x_t^{\text{ml}} + (1 - \lambda_t) x_t^{\text{mb}}$, resulting in the inequality $\|x_t\| \lesssim \lambda_t \|x_t^{\text{ml}}\| + (1 - \lambda_t)\|x_t^{\text{mb}}\| \leq \lambda_t \|x_t^{\text{ml}}\| + (1 - \lambda_t) R_{\text{mb}}$. Our selection of λ_t exactly makes the right-hand side of the above inequality to be bounded by R.

What sets our approach apart from the previous learning-augmented control policies is the fact that we combine two controllers in the DRC parameterization. This allows us to use non-monotonic confidence parameter λ_t. Compared to the state-of-art algorithm (Li et al. (2022a)), instead of diminishing to close to zero, our algorithm produces λ_t's that fluctuate in response to the machine-learned black-box controller's performance. This ensures that the system consistently considers the

machine-learned black-box control actions, avoiding complete disregard after a finite number of time steps. In the next Section, we show the stability and regret bounds for the proposed approach.

3.3. Stability and Regret Guarantees

The following theorem provides the stability guarantee of λ-CLEAC, with the notation introduced in Assumption 6. The proof of this theorem can be found in Appendix A.3 in Shen et al. (2023).

Theorem 7 *λ-CLEAC is ISS-stable. That is, for all $t \geq 0$, with $\|w_t\| \leq W$, we have*

$$\|x_t\| \leq R_{\mathrm{mb}} + L + \frac{2C_A U\|B\|}{1-\rho^2}. \tag{7}$$

We then provide the regret bound for λ-CLEAC with respect to the machine-learned black-box controller. To state the regret, we define the notion of *Stability Violation (SV)*, which characterizes how much the machine-learned black-box controller violates the stability constraint.

Definition 8 *Let λ_t be as defined in λ-CLEAC. The **stability violation** of the machine-learned black-box controller at time step t is*

$$\mathrm{SV}_t := 1 - \lambda_t = \begin{cases} 0, & \text{if } \|x_t^{\mathrm{ml}}\| \leq R \\ \frac{\|x_t^{\mathrm{ml}}\| - R}{\|x_t^{\mathrm{ml}}\| - R_{\mathrm{mb}}}, & \text{otherwise} \end{cases} \tag{8}$$

The intuition behind this definition is that: when the machine-learned black-box state norm violates the threshold, λ_t will be smaller, so $\mathrm{SV}_t := 1 - \lambda_t$ is large, which is why we call it *stability violation*. Given the above definition, the theorem below provides the regret bound of λ-CLEAC using the regret of the machine-learned black-box controller and the stability violations. The proof can be found in Appendix A.4 Shen et al. (2023).

Theorem 9 *We have*

$$\textit{Regret}(\lambda\textit{-CLEAC}) \leq \textit{Regret}(\mathcal{G}^{\mathrm{ml}}) + \frac{4C_L C_A U\|B\|}{1-\rho} \cdot \sum_{t=1}^{T} \mathrm{SV}_t, \tag{9}$$

where $C_L = \max_{t\in[T]} \|\nabla h_t(x_t)\| \in O(1)$. In particular, if we have that both $\textit{Regret}(\mathcal{G}^{\mathrm{ml}}) \in o(T)$ and $\sum_{t=1}^{T} \mathrm{SV}_t \in o(T)$, then $\textit{Regret}(\lambda\textit{-CLEAC}) \in o(T)$.

Note that since we do not make any assumptions on the machine-learned black-box controller, $\mathsf{Regret}(\mathcal{G}^{\mathrm{ml}}) \in o(T)$ happens when the machine-learned black-box controller performs particularly well, e.g. when it is adaptive to the data or when it is exactly the optimal policy.

The insights from Theorem 7 and Theorem 9 reveal an important balance between stability and regret, controlled by the adjustable parameter L. When we pick a higher L, it increases the stability bound. However, this choice makes the threshold $R = R_{\mathrm{mb}} + L$ more lenient for the machine-learned black-box controller, causing λ_t to be often large. This results in a small value of $\sum_{t=1}^{T} \mathrm{SV}_t$ and then leads to small regret bounds. Importantly, when both $\mathsf{Regret}(\mathcal{G}^{\mathrm{ml}})$ and $\sum_{i=1}^{T} \mathrm{SV}_t$ are on the order of $o(T)$, we achieve $\mathsf{Regret}(\lambda\text{-CLEAC}) \in o(T)$ as well. These results not only ensure

the stability of our combined controller but also ensure that it performs well when the machine-learned black-box controller performs well. More importantly, when the machine-learned black-box controller performs very well, we will have $\lambda_t = 1$ and $\mathrm{SV}_t = 0$ for all t, so then both $\mathsf{Regret}(\mathcal{G}^{\mathrm{ml}})$ and $\sum_{t=1}^{T} \mathrm{SV}_t$ will be close to 0, meaning that our algorithm achieves close to 0 regret.

Compared to the previous algorithm (Algorithm 1, Section 4.1 of Li et al. (2022a)) which is also a λ-confident policy that combines a machine-learned black-box controller and a model-based controller, our algorithm has the following benefits:

Firstly, our adaptive policy ensures that it never rules out the machine-learned black-box controller, in the sense that even when it performs badly at a time step where we allocate a very small confidence parameter λ_t to it, we may increase the confidence when it performs well in the future, utilizing the future potential of the machine-learned black-box controller; whereas for their algorithm, the confidence parameter for the machine-learned black-box controller monotonically decreases to as small value, so the machine-learned black-box controller will be permanently ignored even if it performs well in future time steps.

Secondly, as the result of the above, when the machine-learned black-box controller performs well, our algorithm will fully trust it and achieve similar regret as the black-box controller. In contrast, their algorithm does not necessarily achieve 1-competitiveness [2] even when the black-box controller is the optimal controller because their algorithm never fully trusts the machine-learned black box controller.

4. Beyond LTI

In this section, we seek to generalize the results in Section 3 to LTV systems and LTI systems with nonlinear model mismatch. In particular, in both settings, we define the DRC parameterization for the nonlinear controllers, show that the space of stable nonlinear controllers is convex after DRC parameterization, provide adaptive algorithms, and prove stability and regret bounds. Due to space limits, we only provide an overview of our results here. The details can be found in Appendix B and Appendix C in Shen et al. (2023).

LTV. We consider the LTV system of the form

$$f_t(x_t, u_t, w_t) = A_t x_t + B_t u_t + w_t, \tag{10}$$

The DRC parameterization generalizes to this setting (Definition 24 in Appendix B Shen et al. (2023)) and we also show stable controllers can be converted to DRC format (15 in Appendix B Shen et al. (2023)). The algorithm naturally generalizes to the LTV setting, and we can achieve the following stability and regret result.

Theorem 10 (Informal version of Theorem 17 in Appendix B, Shen et al. (2023)) *The adaptive policy (λ-CLEAC) is ISS-stable for system (23), with the regret bound*

$$\mathsf{Regret}(\lambda\text{-CLEAC}) \le \mathsf{Regret}(\mathcal{G}^{\mathrm{ml}}) + O(1) \sum_{t=1}^{T} \mathrm{SV}_t. \tag{11}$$

2. We say ALG is c-**competitive** if $\mathsf{Cost}(\mathsf{ALG}) \le c \cdot \mathsf{Cost}(\mathsf{OPT}) + b$ on any problem instance, where $b \in \mathbb{R}$ is some constant independent of the problem instance.

LTI with model mismatch. Another generalization we consider is the LTI system with an unknown model mismatch, which is of the form:

$$f_t(x_t, u_t, w_t) = Ax_t + Bu_t + w_t + r(x_t), \tag{12}$$

where $r : \mathbb{R}^n \to \mathbb{R}^n$ is ϵ-Lipschitz with $\epsilon > 0$. We show that the DRC parameterization generalizes to this setting (Definition 19 in Appendix C Shen et al. (2023)) and we also show stable controllers can be converted to DRC format (Equation (36) in Appendix C Shen et al. (2023)). With some careful modifications, λ-CLEAC can be generalized for LTI systems with an unknown model mismatch (λ-CLEAC-M, Algorithm 3 in Appendix C Shen et al. (2023)), and we can achieve the following stability and regret result.

Theorem 11 (Informal version of Theorem 22 and 24 in Appendix C, Shen et al. (2023)) *The adaptive λ-confident policy (λ-CLEAC-M) is ISS-stable for system (34). The regret bound is*

$$\textit{Regret}(\lambda\text{-CLEAC-M}) \leq (1 + O(\epsilon))\textit{Regret}(\mathcal{G}^{\mathrm{ml}}) + O(1)\sum_{t=1}^{T} \mathrm{SV}_t + O(\epsilon T). \tag{13}$$

5. Simulations

To verify the efficacy of λ-CLEAC, we apply it to a synthetic dataset for the LTI systems. We also conduct more simulations beyond the LTI setting, which due to space limit is postponed to Appendix D (Shen et al., 2023).

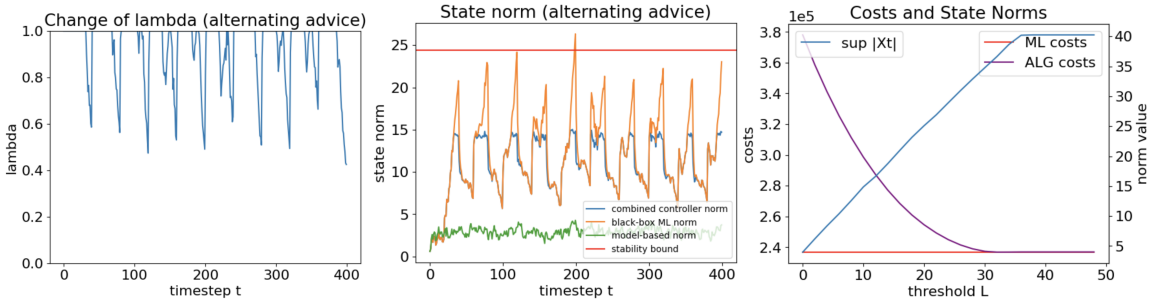
We use synthetic data to show that our algorithm can stabilize an LTI system when given a machine-learned black-box and a model-based controller. We consider a machine-learned black-box controller that is periodically unstable, alongside a model-based controller that is always stable. We choose the threshold parameter $L = 10$ and use λ-CLEAC to obtain a combined controller.

The outcomes are provided in Figure 5: The leftmost plot of Figure 5 demonstrates the variation of the confidence parameter λ_t, reflecting fluctuations based on the performance of the machine-learned black-box controller. In the middle plot, changes in the state norm of the combined controller, the machine-learned black-box controller, and the model-based controller are presented; the red line represents the theoretical stability bound, confirming its validity as a uniform upper limit for the state norm of the combined controller. Lastly, the rightmost plot illustrates the tradeoff between stability and regret of the combined controller. We choose the quadratic cost functions with some shifts. As shown in the plot, with an increase in the threshold parameter L, the state norm of the combined controller also increases, while the cost of the combined controller decreases and ultimately converges to that of the machine-learned black-box controller. This aligns with our earlier discussion on the tradeoff in Section 3.3.

6. Concluding Remarks

This work proposes a novel method to combine model-based and machine-learned black-box controllers by using DRC parameterization. Adaptive policies are proposed for settings including LTI, LTV, and LTI with model mismatch along with theoretical guarantees for stability and regret. The effectiveness of the adaptive policies is validated through experiments. Future directions include exploring theoretical results for more general dynamical systems and implementing our algorithms for other practical settings.

Figure 1: LTI Simulation Results



Acknowledgments

Junxuan Shen was supported by the Summer Undergraduate Research Fellowship from California Institute of Technology. Adam Wierman was supported by NSF grants CNS-2146814, CPS-2136197, CNS-2106403, and NGSDI-2105648. Guannan Qu was supported by NSF Grants 2154171, 2339112, CMU CyLab Seed Funding, C3 AI Institute.

References

Brian D. O. Anderson and John B. Moore. *Optimal Control: Linear Quadratic Methods*. Courier Corporation, February 2007. ISBN 9780486457666. Google-Books-ID: fW6TAwAAQBAJ.

Antonios Antoniadis, Christian Coester, Marek Eliáš, Adam Polak, and Bertrand Simon. Online Metric Algorithms with Untrusted Predictions. *ACM Transactions on Algorithms*, 19(2):19:1–19:34, April 2023. ISSN 1549-6325. doi: 10.1145/3582689. URL https://doi.org/10.1145/3582689.

Lucian Buşoniu, Tim de Bruin, Domagoj Tolić, Jens Kober, and Ivana Palunko. Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, 46:8–28, January 2018. ISSN 1367-5788. doi: 10.1016/j.arcontrol.2018.09.005. URL https://www.sciencedirect.com/science/article/pii/S1367578818301184.

Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. *Applied Sciences*, 11(11):4948, January 2021. ISSN 2076-3417. doi: 10.3390/app11114948. URL https://www.mdpi.com/2076-3417/11/11/4948.

Jonathan D. Chang, Kiante Brantley, Rajkumar Ramamurthy, Dipendra Misra, and Wen Sun. Learning to Generate Better Than Your LLM, June 2023. URL http://arxiv.org/abs/2306.11816. arXiv:2306.11816 [cs].

Nicolas Christianson, Tinashe Handina, and Adam Wierman. Chasing Convex Bodies and Functions with Black-Box Advice. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 867–908. PMLR, June 2022. URL https://proceedings.mlr.press/v178/christianson22a.html.

Nicolas Christianson, Junxuan Shen, and Adam Wierman. Optimal robustness-consistency tradeoffs for learning-augmented metrical task systems. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 9377–9399. PMLR, April 2023. URL https://proceedings.mlr.press/v206/christianson23a.html.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the Sample Complexity of the Linear Quadratic Regulator. 2017. doi: 10.48550/ARXIV.1710.01688. URL https://arxiv.org/abs/1710.01688. Publisher: arXiv Version Number: 3.

P. Dorato. A historical review of robust control. *IEEE Control Systems Magazine*, 7(2):44–47, April 1987. ISSN 2374-9385. doi: 10.1109/MCS.1987.1105273.

J. Doyle. Robust and optimal control. In *Proceedings of 35th IEEE Conference on Decision and Control*, volume 2, pages 1595–1598, Kobe, Japan, 1996. IEEE. ISBN 978-0-7803-3590-5. doi: 10.1109/CDC.1996.572756. URL http://ieeexplore.ieee.org/document/572756/.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator, March 2019. URL http://arxiv.org/abs/1801.05039. arXiv:1801.05039 [cs, stat].

Jaime F. Fisac, Anayo K. Akametalu, Melanie N. Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J. Tomlin. A General Safety Framework for Learning-Based Control in Uncertain Robotic Systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, July 2019. ISSN 1558-2523. doi: 10.1109/TAC.2018.2876389. URL https://ieeexplore.ieee.org/abstract/document/8493361?casa_token=JUBobBJruCwAAAAA:QsrHEXwlYOHZmPlQvTyZYP9IojWKG3cXl56r5myDHAw1l7kNHLNOokp8OMkt9VMUzMT-o_xjodA.

Carlos E. García, David M. Prett, and Manfred Morari. Model predictive control: Theory and practice—A survey. *Automatica*, 25(3):335–348, May 1989. ISSN 0005-1098. doi: 10.1016/0005-1098(89)90002-2. URL https://www.sciencedirect.com/science/article/pii/0005109889900022.

Lukas Hewing, Kim P. Wabersich, Marcel Menner, and Melanie N. Zeilinger. Learning-Based Model Predictive Control: Toward Safe Learning in Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):269–296, May 2020. ISSN 2573-5144, 2573-5144. doi: 10.1146/annurev-control-090419-075625. URL https://www.annualreviews.org/doi/10.1146/annurev-control-090419-075625.

Patricia Hidalgo-Gonzalez, Rodrigo Henriquez-Auba, Duncan S. Callaway, and Claire J. Tomlin. Frequency Regulation using Data-Driven Controllers in Power Grids with Variable Inertia due to Renewable Energy. In *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5, August 2019. doi: 10.1109/PESGM40551.2019.8973437. URL https://ieeexplore.ieee.org/document/8973437. ISSN: 1944-9933.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, May 1996. ISSN 1076-9757. doi: 10.1613/jair.301. URL https://www.jair.org/index.php/jair/article/view/10166.

B. Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep Reinforcement Learning for Autonomous Driving: A Survey, January 2021. URL http://arxiv.org/abs/2002.00444. arXiv:2002.00444 [cs].

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, September 2023. URL http://arxiv.org/abs/2309.00267. arXiv:2309.00267 [cs].

Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review, May 2018. URL https://arxiv.org/abs/1805.00909v3.

Tongxin Li, Ruixiao Yang, Guannan Qu, Yiheng Lin, Steven Low, and Adam Wierman. Equipping Black-Box Policies with Model-Based Advice for Stable Nonlinear Control. 2022a. doi: 10.48550/ARXIV.2206.01341. URL https://arxiv.org/abs/2206.01341.

Tongxin Li, Ruixiao Yang, Guannan Qu, Guanya Shi, Chenkai Yu, Adam Wierman, and Steven Low. Robustness and Consistency in Linear Quadratic Control with Untrusted Predictions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1):1–35, February 2022b. ISSN 2476-1249. doi: 10.1145/3508038. URL https://dl.acm.org/doi/10.1145/3508038.

Thodoris Lykouris and Sergei Vassilvtiskii. Competitive Caching with Machine Learned Advice. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3296–3305. PMLR, July 2018. URL https://proceedings.mlr.press/v80/lykouris18a.html.

Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based Reinforcement Learning: A Survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, January 2023. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000086. URL https://www.nowpublishers.com/article/Details/MAL-086.

Anusha Nagabandi, Gregory Kahn, Ronald S. Fearing, and Sergey Levine. Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566, May 2018. doi: 10.1109/ICRA.2018.8463189. URL https://ieeexplore.ieee.org/document/8463189. ISSN: 2577-087X.

Junhyuk Oh, Matteo Hessel, Wojciech M. Czarnecki, Zhongwen Xu, Hado P van Hasselt, Satinder Singh, and David Silver. Discovering Reinforcement Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pages 1060–1070. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0b96d81f0494fde5428c7aea243c9157-Abstract.html.

Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal Difference Models: Model-Free Deep RL for Model-Based Control. February 2018. URL https://openreview.net/forum?id=Skw0n-W0Z.

Guannan Qu, Chenkai Yu, Steven Low, and Adam Wierman. Exploiting Linear Models for Model-Free Nonlinear Control: A Provably Convergent Policy Gradient Approach. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6539–6546, December 2021. doi: 10.1109/CDC45484.2021.9683735. URL https://ieeexplore.ieee.org/document/9683735. ISSN: 2576-2370.

Dhruv Rohatgi. Near-Optimal Bounds for Online Caching with Machine Learned Advice. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, January 2020. doi: 10.1137/1.9781611975994. URL https://epubs.siam.org/doi/book/10.1137/1.9781611975994.

Ugo Rosolia and Francesco Borrelli. Learning Model Predictive Control for Iterative Tasks. A Data-Driven Control Framework. *IEEE Transactions on Automatic Control*, 63(7):1883–1896, July 2018. ISSN 1558-2523. doi: 10.1109/TAC.2017.2753460. URL https://ieeexplore.ieee.org/abstract/document/8039204.

Daan Rutten, Nicolas Christianson, Debankur Mukherjee, and Adam Wierman. Smoothed Online Optimization with Unreliable Predictions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–36, February 2023. ISSN 2476-1249. doi: 10.1145/3579442. URL https://dl.acm.org/doi/10.1145/3579442.

Mark Sellke. Chasing Convex Bodies Optimally. In Ronen Eldan, Bo'az Klartag, Alexander Litvak, and Emanuel Milman, editors, *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2020-2022*, Lecture Notes in Mathematics, pages 313–335. Springer International Publishing, Cham, 2023. ISBN 9783031263002. doi: 10.1007/978-3-031-26300-2_12. URL https://doi.org/10.1007/978-3-031-26300-2_12.

Anant Shah and Arun Rajkumar. Sequential Ski Rental Problem, April 2021. URL http://arxiv.org/abs/2104.06050. arXiv:2104.06050 [cs].

Junxuan Shen, Adam Wierman, and Guannan Qu. Combining model-based controller and ml advice via convex reparameterization, 2023. URL https://drive.google.com/file/d/1Gl0FYkbBu-BCQcQ16sOUg1ioZwVpxrU-/view?usp=sharing.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, December 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/10.1126/science.aar6404.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper Learning for Non-Stochastic Control. 2020. doi: 10.48550/ARXIV.2001.09254. URL https://arxiv.org/abs/2001.09254.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Springer Nature, May 2022. ISBN 9783031015519. Google-Books-ID: g4RyEAAAQBAJ.

Kim P. Wabersich and Melanie N. Zeilinger. Linear Model Predictive Safety Certification for Learning-Based Control. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 7130–7135, December 2018. doi: 10.1109/CDC.2018.8619829. URL https://ieeexplore.ieee.org/abstract/document/8619829?casa_token=KqJOqawDFYYAAAAA:LttfBp1WkpQ9a9im-dkpc6Ty613tTnzP_Drzl97PJgJORg8IM8659USnT8-fyWi6WmJ1qbhYfhk. ISSN: 2576-2370.

Shufan Wang, Jian Li, and Shiqiang Wang. Online Algorithms for Multi-shop Ski Rental with Machine Learned Advice. In *Advances in Neural Information Processing Systems*, volume 33, pages 8150–8160. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/5cc4bb753030a3d804351b2dfec0d8b5-Abstract.html.

Alexander Wei. Better and Simpler Learning-Augmented Online Caching, May 2020. URL http://arxiv.org/abs/2005.13716. arXiv:2005.13716 [cs].

Xiangyu Zhao, Changsheng Gu, Haoshenglun Zhang, Xiwang Yang, Xiaobing Liu, Jiliang Tang, and Hui Liu. DEAR: Deep Reinforcement Learning for Online Advertising Impression in Recommender Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1): 750–758, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i1.16156. URL https://ojs.aaai.org/index.php/AAAI/article/view/16156.