Safe online nonstochastic control

Anonymous Author(s)
Affiliation
Address
email

Abstract

Online nonstochastic control has emerged as a promising strategy for online convex optimization of control policies for linear systems subject to adversarial disturbances and time-varying cost functions. However, ensuring safety in these systems remains a significant open problem, especially when the system parameters are unknown. Practical nonstochastic control algorithms for real-world systems must adhere to safety constraints without becoming overly conservative or relying on exact models. We address this challenge by presenting a safe nonstochastic control algorithm for systems with unknown parameters subject to state and input constraints. Given data of a single disturbed input-state trajectory, we design non-conservative constraint sets for the policy parameters and develop a robust strongly stabilizing controller. By drawing a connection to model predictive control, we propose a new analysis perspective and show how a slight change in the nonstochastic control algorithm can drastically improve performance if disturbances are constant or slowly time-varying.

1 Introduction

In reinforcement learning, gradient-based policy optimization has shown great success in practice Schulman et al. [2017]. For learning-based control, the paradigm of online convex optimization offers a powerful framework for iteratively updating control policies based on gradients and observed data. Nonstochastic control is such a gradient-based control method that has been proven effective for the control of linear dynamical systems in the face of deterministic, possibly adversarial, bounded disturbances and adversarially chosen cost functions [Agarwal et al., 2019, Hazan et al., 2020, Simchowitz, 2020]. At each time step, a convex cost function is revealed to the learner and the policy gradient is approximated by applying the cost function to the terminal state and action of a model-based rollout (simulation). Since optimizing over the function space of state or output feedback policies is computationally intractable, see for example [Goulart et al., 2006], disturbance feedback policies are employed. Nonstochastic control algorithms have been adapted or extended for different settings such as partial observability [Simchowitz et al., 2020], changing dynamics [Minasyan et al., 2021], bandit loss [Sun et al., 2023] or fully unknown linear systems [Chen and Hazan, 2021]. However, one critical challenge is the inclusion of a safety guarantee in the sense of adherence to state and input constraints, particularly in the presence of model uncertainty.

Little research on nonstochastic control so far has considered the addition of input and state constraints. In the related literature, this problem setting has only been considered with access to an exact model Li et al. [2021], Nonhoff et al. [2024], Liu et al. [2023], Zhou and Tzoumas [2023], Martin et al. [2023] or achieved results in high probability with i.i.d. disturbances and conservative fixed parameter constraints with careful transitions in between updates Li et al. [2024]. For systems with unknown parameters, most works propose a sequential approach of system identification via least squares estimation (LSE) and control. Recent advances in statistical learning theory and the related discovery of high-probability finite time guarantees for models obtained via LSE [Wagenmaker and Jamieson,

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

2020, Sarkar and Rakhlin, 2019, Foster and Simchowitz, 2020, Simchowitz et al., 2018] was leveraged by a multitude of works to obtain high-probability regret guarantees in the nonstochastic control setting [Hazan et al., 2020, Chen and Hazan, 2021, Simchowitz, 2020].

In this work, we change perspective from statistical learning LSE to data-driven robust control based on a set of unfalsified models Berberich et al. [2020], Van Waarde et al. [2023], Teutsch et al. [2024]. In the nonstochastic control setting of linear systems subject to bounded disturbances, such a set may be constructed by set membership identification (SMI). Informally, SMI begins by considering the whole space of model parameters and continually discards those that could not have reproduced the seen data. By leveraging the disturbance bounds, the resulting sets can be much smaller than LSE confidence regions Li et al. [2023], and always contain the system's true parameters, which allows for the design of robust controllers with certainty instead of high probability.

Contribution: This work presents a safe online optimal control algorithm for unknown linear systems subject to nonstochastic disturbances. Given an input-state data trajectory, we bridge the gap between low-regret nonstochastic control and safe data-driven robust control by designing safety constraints for online policy updates that hold for all models that may have produced the data. By drawing from concepts in model predictive control Rawlings et al. [2017], Lorenzen et al. [2019], we establish recursive feasibility of the safety constraints for all models and propose a subtle but effective change to the initial state of the rollouts used for the policy gradient, which leads to the elimination of steady-state errors in the case of constant or slowly time-varying disturbances. We show the practical potential of the approach in a small simulation example.

2 Preliminaries and problem setting

In the nonstochastic control setting, the learner is presented with a linear time-invariant dynamical system

$$x_{t+1} = Ax_t + Bu_t + w_t \tag{1}$$

where $x \in \mathbb{R}^{n_x}$ is the state of the system and $u \in \mathbb{R}^{n_u}$ is the input or action taken by the learner. The disturbance $w \in \mathbb{R}^{n_x}$ represents uncertainty and is not subject to any assumed stochastic properties, but may be chosen from a known compact set $\mathcal{W}$ by an adversary at each time step and remains unknown to the learner. In this work, we assume $\mathcal{W}$ is a convex polytope $\mathcal{W} = \{w \in \mathbb{R}^{n_x} \mid G_w w \leq g_w\}$. At each time step t, the learner measures the current state x_t and a cost function $c_t : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$ is revealed. The goal is to learn a policy that chooses inputs which minimize the cumulative costs $\sum_t c_t(x_t, u_t)$.

2.1 Disturbance-action policies and the Gradient perturbation controller

The considered policies are from the class of *disturbance-action policies* Agarwal et al. [2019], also called *affine disturbance feedback*, see for example [Goulart et al., 2006]. Instead of basing decisions on the current state directly, these policies compute the input based on estimates of past disturbances $\underline{\hat{w}}_t$. These estimates are based on a system model $(\hat{A}, \hat{B}) \approx (A, B)$. At time step t, the disturbance estimate $\hat{w}_{t-1}$ is computed as the prediction error

$$\hat{w}_{t-1} = x_t - (\hat{A}x_{t-1} + \hat{B}u_{t-1}). \tag{2}$$

Definition 1. *A disturbance-action policy (DAP) $\pi_{\mathrm{DAP}}(\underline{M})$ chooses inputs based on parameter matrices M_i via*

$$v_t = m_0 + \sum_{i=1}^{L} M_i \hat{w}_{t-i} = \underline{M}\,\underline{\hat{w}}_t \tag{3}$$

where L is the memory length and $\underline{M} = [m_0, M_1 \ldots, M_L]$, $\underline{\hat{w}}_t = [1, \hat{w}_{t-1}^{\mathrm{T}}, \ldots, \hat{w}_{t-L}^{\mathrm{T}}]^{\mathrm{T}}$ allow for shorter notation.

In order to guarantee stability, DAPs are often used together with a fixed stabilizing state feedback controller $u_t = Kx_t + v_t$. We will do the same and abbreviate $(A + BK) = A_{\mathrm{K}}$ in the following.

A *gradient perturbation controller* (GPC) iteratively updates the policy (3) based on gradients computed via a model rollout. Online, at each time step t, the state x_t is measured, the last disturbance $\hat{w}_{t-1}$ is estimated and the control parameters $\underline{M}_t$ are updated by taking a gradient-step as

$$M_{t+1,i} = M_{t,i} - \eta_t \nabla l_t(M_{t,i}) \tag{4}$$

where $\eta_t > 0$ is the learning rate. The loss l_t approximates the that would have been obtained under the fixed policy $\pi_{\mathrm{DAP}}(\underline{M}_t)$ and is defined based on the terminal state and input of an H-step simulation of the current with a given model $(\hat{A}, \hat{B})$ and the most recent disturbance estimates $\hat{\underline{w}}_{t-L-H:t}$. That is, $l_t(\underline{M}_t) = c_t(x_{H|t}(M_t), u_{H|t}(M_t))$ where $(x_{k|t}, u_{k|t})$ denotes the simulated states and inputs running from $k = 0, \ldots, L$ computed at time step t via

$$x_{0|t} = 0, \quad w_{k|t} = \hat{w}_{t-H+k}, \quad \underline{w}_{k|t} = [1, w_{k-1|t}^{\mathrm{T}}, \ldots, w_{k-L|t}^{\mathrm{T}}]^{\mathrm{T}}, \quad u_{k|t} = Kx_{k|t} + \underline{M}_t \underline{w}_{k|t}, \tag{5}$$
$$x_{k+1|t} = \hat{A}x_{k|t} + \hat{B}u_{k|t} + w_{k|t},\ k = 0, \ldots, H-1.$$

The justification is here that the actual state may be well approximated by such a simulation, since it evolves as $x_t = A_{\mathrm{K}}^t x_0 + \sum_{k=0}^{t-1} A_{\mathrm{K}}^k B v_{t-1-k} + A_{\mathrm{K}}^k w_{t-1-k}$ whereas a simulation with horizon H, initial state zero, and the latest H disturbances reads $\tilde{x}_t = \sum_{k=0}^{H-1} A_{\mathrm{K}}^k B v_{t-1-k} + A_{\mathrm{K}}^k w_{t-1-k} \approx x_t$. The resulting approximation error reads

$$x_t - \tilde{x}_t = \sum_{k=H}^{t-1} A_{\mathrm{K}}^k (B v_{t-1-k} + w_{t-1-k}) = A_{\mathrm{K}}^H \sum_{k=0}^{t-1-H} A_{\mathrm{K}}^k (B v_{t-1-k} + w_{t-1-k}) = A_{\mathrm{K}}^H x_{t-H} \tag{6}$$

and is small for stable A_{K} and large memory H. How small is captured by the following quantitative notion of stability introduced in Cohen et al. [2018].

Definition 2. *K is a (κ, γ)-strongly stable controller for (A, B) if $\|A_{\mathrm{K}}^t\| \le \kappa\gamma^t$ for all $t \ge 0$.*

Equipped with convergence bounds for the dynamics A_{K}, the presented gradient pertubation controller (3)-(5) enjoys sublinear regret against the best fixed policy $\underline{M}^*$ in hindsight, and thereby sublinear regret against an expressive class of controllers, see Hazan and Singh [2023] for an overview of results.

2.2 Problem setting: Uncertain system and safety constraints

While the presented control scheme has been extended in many directions, for example to bandit loss functions Sun et al. [2023] and partial observations Simchowitz et al. [2020], one challenge for the application of GPC to safety-critical systems is the adherence to input and state constraints in the face of model uncertainty. In this work, we consider a setting where the true system parameters (A, B) are unknown, and only an input-state trajectory $\{u_t, x_t\}_{t=0}^{T_D}$ is available. Furthermore, we restrict actions to a set

$$u_t \in \mathcal{U} = \{u \in \mathbb{R}^{n_u} \mid G_{\mathcal{U}} u \le g_{\mathcal{U}}\} \quad \forall t \ge 0. \tag{7}$$

and subject the state to polytopic safety constraints

$$x_t \in \mathcal{S} = \{x \in \mathbb{R}^{n_x} \mid G_{\mathcal{S}} x \le g_{\mathcal{S}}\} \quad \forall t \ge 1 \tag{8}$$

where both $\mathcal{U}$ and $\mathcal{S}$ are known user-specified convex compact sets that contain the origin. In order to render the problem of safety tractable, we assume that there exists a state feedback controller that can keep the system safe from initial state $x_0 = 0$ no matter which disturbances are chosen by the adversary.

Definition 3 (Safe control policies)**.** *A control policy is called safe if it generates inputs $u_t \in \mathcal{U}$ for which the state of* (1) *satisfies $x_t \in \mathcal{S}$ for all time $t \ge 0$.*

Assumption 1. *There exists K_{safe} such that given $x_0 = 0$, the state feedback $u_t = K_{\mathrm{safe}} x_t$ is safe for all disturbance realizations $w_t \in \mathcal{W}$ and all time $t \ge 0$.*

Remark 1. *On first glance, Assumption 1 may seem restrictive, but note that 1) we do not have access to K_{safe}, and 2) the application of $u_t = K_{\mathrm{safe}} x_t$ may incur high costs without the disturbance feedback, whose addition can in turn cause a loss of safety. Informally, Assumption 1 guarantees that the disturbances in $\mathcal{W}$ are not too large compared to the set of safe states $\mathcal{S}$.*

3 Safe nonstochastic control

A guarantee of safety constraints (8) during operation requires that the control input u_t, or more specifically the control parameters $\underline{M}_t$ are always chosen such that $x_{t+1} \in \mathcal{S}$. Since most non-stochastic control algorithms in the literature already include projections into a set of parameters $\mathcal{M}$ for regret guarantees, it is a natural adaptation to enforce input and constraint satisfaction by similarly projecting into the set of safe control parameters $\mathcal{M}_{\mathrm{safe},t}$ and applying $u_t(\Pi_{M_{\mathrm{safe},t}} M_{\mathrm{opt},t})$ instead of $u_t(\underline{M}_{\mathrm{opt}})$ for control. The resulting algorithm presented in this paper is shown in Algorithm 1.

Algorithm 1 Safe online optimal control

Identification phase:

Collect data $(u_t, x_t)_{t=0}^{T_{\text{ini}}}$.
Construct set of unfalsified models Ω as in (9).
Compute (κ, γ)-strongly stable state feedback gain K for all Ω as in Lemma 1.
Choose a nominal model $(\hat{A}, \hat{B}) \in \Omega$.

Control phase:

for each time step $t = T_{\text{ini}}, \dots T_{\text{ini}} + T$ **do**
  Record state x_t and construct latest disturbance estimate $\hat{w}_{t-1} := x_t - (\hat{A}x_t + \hat{B}u_t)$.
  Receive cost function and update policy $\underline{M}_{t,\text{opt}} = \underline{M}_{t-1} - \eta_t \nabla l_t(\underline{M}_{t-1})$.
  Project to closest safe policy $\underline{M}_t = \Pi_{\mathcal{M}_{\text{safe},t}} \underline{M}_{t,\text{opt}}$.
  Apply control $u_t(\underline{M}_t)$ (3)

3.1 From data to a set of models

Instead of identifying one best-fit system, we consider the set of models (A, B) that agree with (may have produced) the given or recorded input-state data. Let $\mathbb{Z}_{i,j} = \{i, i+1, .., j\}$. Given an input-state data trajectory $\{x_t, u_t\}_{t=0}^{T}$ resulting from the application of T arbitrary inputs to system (1), and assuming that the unknown disturbances $\{w_t\}_{t=0}^{T-1}$ were always in the known set $\mathcal{W}$, the resulting set of *consistent* or *unfalsified* models is given by

$$\Omega_{[0,T]} = \{[A \quad B] \in \mathbb{R}^{n_x \times (n_x + n_u)} \mid x_{t+1} - [A \quad B] \begin{bmatrix} x_t \\ u_t \end{bmatrix} \in \mathcal{W},\, t = 0, \dots, T-1\}. \tag{9}$$

The set $\Omega_{[0,T]}$ inherits convexity and closedness from $\mathcal{W}$, and can be directly constructed in half-space representation by reorganizing the inequality constraints that represent $\mathcal{W}$. If the data trajectory is sufficiently informative (see Lemma 2 in the Appendix), then $\Omega_{[0,T]}$ is also bounded and may be described as convex hull of its vertices $\Omega = \text{conv}(\{[A_i\, B_i]\}_{i=1}^{N_v})$. As a representation of model uncertainty, Ω behaves nicely: First, as new data streams in, new constraints are added to the set and therefore updates never increase the uncertainty set in size. Second, crucially, as long as the assumed disturbance bound $\mathcal{W}$ holds, Ω always contains the true data-generating system matrices by construction, and every statement that holds for all models $[A\, B]$ inside Ω necessarily holds for the actual unknown system. Since Ω is defined by input-state data (and the disturbance bound $\mathcal{W}$) alone, these statements can be inferred directly from data. In this work, we will use the set of models Ω to construct constraints on the control parameters with which safety can be guaranteed.

3.2 From a set of models to safety constraints

In this work, safety is defined as constraints in input and state space. In order to derive a set of safe control parameters, we need to map the state space constraints $x_t \in \mathcal{S}$ into constraints on the policy parameters $\underline{M}_t$. As intermediate mapping, we may consider the space of inputs since v_t spans all of $\mathcal{R}^m$ in the sense that any desired safe input u can be reproduced by some choice of control parameters $\underline{M}$ such that $u = Kx + v(\underline{M}_t)$. The challenge is that the constraints on $\underline{M}$ 1) need to be recursively feasible, i.e., the state is only steered to where constraint satisfaction remains possible, 2) need to consider all possible models in Ω need to be considered, and 3) should not be conservative but restrict the space of parameters as little as possible.

Consider the set of models Ω containing the true system and the disturbance bound $\mathcal{W}$. The state evolution of the unknown system (1) satisfies the inclusion

$$x_{t+1} \in \Omega \begin{bmatrix} x_t \\ u_t \end{bmatrix} \oplus \mathcal{W}, \quad \Omega \begin{bmatrix} x_t \\ u_t \end{bmatrix} = \{[A \quad B] \begin{bmatrix} x_t \\ u_t \end{bmatrix} \mid [A \quad B] \in \Omega\} \tag{10}$$

where $\oplus$ denotes the Minkowski set addition. With the Minkowski (Pontryagin) set difference $\ominus$, we can reformulate the above into a sufficient condition on the state and input at the current time step for satisfaction of safety constraints at the next time step,

$$\Omega \begin{bmatrix} x_t \\ u_t \end{bmatrix} \in \mathcal{S} \ominus \mathcal{W} \implies x_{t+1} \in \mathcal{S}. \tag{11}$$

Remark 2. *Non-emptiness of $\mathcal{S} \ominus \mathcal{W}$ is covered by Assumption 1 since the existence of K_{safe} implies that the the safe set $\mathcal{S}$ is specified large enough to contain the disturbance set, i.e., $\mathcal{S} \supseteq \mathcal{W}$ and consequently $\mathcal{S} \ominus \mathcal{W} \neq \emptyset$.*

In order to guarantee that the left-hand-side of (11) remains feasible during operation, we construct a (maximal) robust control invariant subset of $\mathcal{S}$ [Blanchini, 1999, Rawlings et al., 2017].

Definition 4. *A set $\mathcal{X}$ is robust control invariant (RCI) for dynamics $x_{t+1} = Ax_t + Bu_t + w_t$, $w_t \in \mathcal{W}$, if for all $x \in \mathcal{X}$ there exists $u \in \mathcal{U}$ such that $Ax + Bu + w \in \mathcal{X}$ for all $w \in \mathcal{W}$.*

A maximal RCI subset of a the safe state set $\mathcal{S}$ is a set that contains all other RCI subsets of $\mathcal{S}$. The maximal RCI subset is well defined since the set property of robust control invariance is closed under the union. For the present discrete-time linear dynamics, maximal RCI sets are computed via recursive erosion and expansion [Blanchini and Miani, 2015], see Appendix for details.

In the following, let $\mathcal{X}$ be the maximal subset of $\mathcal{S}$ that is RCI for all models in Ω. That is, let $\mathcal{X}$ be such that

$$(\forall x \in \mathcal{X})(\exists u \in \mathcal{U})\ \Omega \begin{bmatrix} x \\ u \end{bmatrix} \in \mathcal{X} \ominus \mathcal{W}. \tag{12}$$

Since Ω, $\mathcal{S}$ and $\mathcal{W}$ are compact and convex polytopes, so is $\mathcal{X}$ and we can write $\mathcal{X} = \{x \in \mathbb{R}^n \mid G_{\mathcal{X}} x \le g_{\mathcal{X}}\}$. Similarly, define $\mathcal{X} \ominus \mathcal{W} = \{x \in \mathbb{R}^{n_x} \mid G_{\mathcal{X}} x \le g_{\mathcal{X} \ominus \mathcal{W}}\}$.

Remark 3. *For the true system, $\mathcal{X}$ is nonempty by Assumption 1 and contains the origin.*

Substituting the DAC policy (3) into (12) and reformulating based on vertices of Ω leads to linear constraints on the control parameters

$$G_{\mathcal{X}} B_i \underline{M} \hat{\underline{w}}_t \le g_{\mathcal{X} \ominus \mathcal{W}} - G_{\mathcal{X}}(A_i + B_i K) x_t \quad i = 1, \dots, N_v, \tag{13}$$

$$G_{\mathcal{U}} \underline{M} \hat{\underline{w}}_t \le g_{\mathcal{U}} - G_{\mathcal{U}} K x_t,, \tag{14}$$

which define a convex constraint set $\mathcal{M}(\hat{\underline{w}}_t, x_t) = \{\underline{M} \in \mathbb{R}^{n_u \times L n_x} \mid (13),(14) \text{ are satisfied}\}$ that is parameterized by the past estimated disturbances in $\hat{\underline{w}}_t$ and the current state x_t.

Remark 4. *Note that $\mathcal{M}(\hat{\underline{w}}_t, x_t)$ also depends on the chosen state feedback gain K, which is however constant throughout. If K is safe as per Definition 3 (such that $K = K_{\mathrm{safe}}$ from Assumption 1), then $\underline{M}_t = 0$ is a safe parameter choice for all time and $\{0\}$ is a common subset of all sets $\mathcal{M}(\hat{\underline{w}}_t, x_t)$ with $x_t \in \mathcal{X}$, $\hat{w}_{t-i} \in \mathcal{W}$. In general, the set of control parameters which are safe for all possible states and disturbances is the intersection of all such $\mathcal{M}(\hat{\underline{w}}, x)$. In other words, more restrictive but fixed safety constraints as in [Li et al., 2021, 2024] are recovered by minimizing the RHS* (13),(14) *of over all $x_t \in \mathcal{X}$ and requiring the inequalities to hold for all $\hat{\underline{w}}_t = [1, \hat{w}_{t-1}^{\mathrm{T}}, \dots, \hat{w}_{t-L}^{\mathrm{T}}]^{\mathrm{T}}$ with $\hat{w}_{t-i} \in \mathcal{W}$.*

3.3 Theoretical guarantees

If at each time step t, the control parameters $\underline{M}_t$ are projected into $\mathcal{M}_{\mathrm{safe},t} = \mathcal{M}(\hat{\underline{w}}_t, x_t)$, we may guarantee safety as shown in the following result.

Lemma 1 (Recursive feasibility)**.** *Let $x_t \in \mathcal{X}$. Then $\mathcal{M}_{\mathrm{safe},t} \neq \emptyset$ and any choice of control parameters $\underline{M}_t \in \mathcal{M}_{\mathrm{safe},t}$ leads to a nonempty constraint set in the next time step, $\mathcal{M}_{t+1,\mathrm{safe}} \neq \emptyset$.*

Proof. Since $\mathcal{X}$ is RCI, $\mathcal{M}_{\mathrm{safe},t} \neq \emptyset$ for all $x \in \mathcal{X}$ by construction. Moreover, any choice of control parameters $\underline{M}_t \in \mathcal{M}_{\mathrm{safe},t}$ leads to $x_{t+1} \in \mathcal{X}$. Since $\mathcal{X} \subseteq \mathcal{F}$, the safety condition (12) is feasible for all states in $\mathcal{X}$. □

Theorem 1 (Constraint satisfaction)**.** *Assume that the model uncertainty in Ω is small enough such that $\mathcal{X} \neq \emptyset$. Then, for any $T \ge 0$ and all possible disturbance sequences $w_{0:T} \in \mathcal{W}$, the proposed control strategy in Algorithm 1 is safe in the sense of Definition 3.*

Proof. Since $x_0 = 0$, we have $x_0 \in \mathcal{X}$ as long as $\mathcal{X} \neq \emptyset$. By Lemma 1 the set of control parameters $\mathcal{M}_{\mathrm{safe},0}$ is not empty and any choice $\underline{M}_0 \in \mathcal{M}_{\mathrm{safe},0}$ satisfies input constraints by construction of $\mathcal{M}_{\mathrm{safe},0}$ and leads to a next state $x_1 \in \mathcal{X}$. Since $\mathcal{X} \subseteq \mathcal{S}$, the next state x_1 is safe. Safety for all time follows by induction. □

Remark 5. *Besides safety with certainty, the proposed approach based on SMI offers another distinct advantage: Online adaptation of the above safety constraints is trivial. With every new data triple* (x_{t-1}, u_{t-1}, x_t)*,* Ω *may be updated by adding the constraints representing* $x_t - Ax_{t-1} - Bu_{t-1} \in \mathcal{W}$*. Since* $\Omega_t \subseteq \Omega_{t-1}$*, a newly computed maximal RCI set* $\mathcal{X}_t$ *will always contain the prior version,* $\mathcal{X}_t \supseteq \mathcal{X}_{t-1}$ *and the constraints on the policy parameters are relaxed without loss of recursive feasibility. If computational time is an issue, update computations may happen asynchronously by computing an update of* $\mathcal{X}$ *on a batch of new data and injecting it into the control algorithm once the computation is finished. By contrast, a similarly easy adaptation is not possible if error bounds around a least square estimate replace the set of models: the error bound of LSE decreases with more data, but the change of the estimate itself may cause the new set of models to not be contained in the prior one. Consequently, a careful transition between updates is necessary Li et al. [2024], which is not the case in the proposed approach.*

If the gradient perturbation controller presented in Section 2 runs with an approximate model $(\hat{A}, \hat{B})$, the only difference between the loss simulation (5) and the approximation in (6) is that an additional error is introduced due to the model error. By also bounding this additional error, setting an appropriate learning rate, and restricting control parameters $\underline{M}_t$ to a special set, the gradient perturbation controller for uncertain systems achieves sublinear regret with respect to the class of state feedback controllers $u_t = Kx_t$ Hazan et al. [2020], linear dynamical controllers Simchowitz [2020], or disturbance action policies with fixed control parameters Chen and Hazan [2021]. In order to recover similar regret guarantees with the additional projection to safety, we too require a (κ, γ)-strongly stabilizing controller as in Definition 2. In the foll we show how such a controller may be constructed for all models in Ω.

Synthesis of a strongly stabilizing controller A sufficient condition for stability $\rho(A + BK) < 1$ of all models $(A, B) \in \Omega$ is given by existence of a common quadratic Lyapunov function $V(x) = x^{\mathrm{T}}Px$ for all hypothetical closed loop systems $x^+ = (A + BK)x$, $(A, B) \in \Omega$. Computationally, this check requires solving a finite system of linear matrix inequalities (LMI) in a semi-definite program. Since regret bounds in the literature depend on the notion of (κ, γ)-strong stability, we provide a semi-definite program for the direct synthesis of a $(\sqrt{c}, \gamma)$-strongly stable controller with specified rate $\gamma < 1$ and minimal constant $\sqrt{c}$ in the following. The idea is to combine a bound on the norm powers of A_{K} based on the positive invariance of Lyapunov sublevel sets [Ahiyevich et al., 2018] with the fact since $\rho(rA) = r\rho(A)$ for any matrix A, stability of $\frac{1}{\gamma}A_{\mathrm{K}}$ (i.e., $\rho(\frac{1}{\gamma}A_{\mathrm{K}}) \leq 1$) implies $\rho(A_{\mathrm{K}}) \leq \gamma$. Recall that $\Omega = \operatorname{conv}\{[A_i \quad B_i]\}_{i=1}^{N_v}$.

Proposition 1. *Choose a desired spectral radius* $0 \leq r < 1$ *and let* (c, Z, Y) *be the solution of*

$$\underset{c,Z,Y}{\text{minimize}} \quad c \tag{15}$$

$$\text{subject to} \quad I_n \preceq Z \preceq cI_n, \tag{16}$$

$$\begin{bmatrix} rZ & A_iZ + B_iY \\ * & rZ \end{bmatrix} \succ 0 \quad \forall i = 1, \dots, N_v. \tag{17}$$

Then the controller $K = YZ^{-1}$ *is* $(\sqrt{c}, r)$*-strongly stable for all* $(A, B) \in \Omega$*.*

Please see Appendix for the proof.

On regret bounds with safety constraints The presented algorithm allows to run a safe variant of GPC with any nominal model $(\hat{A}, \hat{B}) \in \Omega$, for example chosen via LSE and projection or as Chebyshev center of Ω. The computation of a strongly stable controller in Proposition 1 allows for a recovery of GPC regret bounds in literature, as long as the safety constraints are not active. The presented design of safety constraints restricts control parameters as little as possible. In fact, it was motivated by the following Proposition.

Proposition 2. *Every causally safe control policy (without foreknowledge of* w_t*) needs to keep the state in the maximal RCI subset* $\mathcal{X}_{\max} \subseteq \mathcal{S}$*.*

Proof. If starting from x_t there exists an input sequence that keeps the state inside $\mathcal{S}$ for all possible disturbance sequences $w_{t:}$ and all time, then the resulting state trajectory would be part of the maximal RCI subset of $\mathcal{S}$. Since x_t is not, the proof follows by contradiction. □

In other words, enforcing the state to stay within the maximal RCI subset of $\mathcal{S}$ does not lead to a meaningful change of regret bounds if the comparator class is restricted to causally safe policies. In the continuation of this work, we are interested in formalizing regret bounds in case of active safety constraints.

4 Better policy gradients by adaptive initial state - An MPC perspective

In essence, the gradient perturbation controller presented above takes decisions that minimize the loss of model-based predictions. Recall the definition of the loss l_t based on a model rollout. If instead of updating parameter towards the minimzer of the loss function, the policy parameters were chosen directly as the minimzer in each time step, the scheme could be interpreted as parameterized model predictive control (MPC): At each time step, choose the policy parameters parameterized by solving the finite-time optimal control problem (OCP) $\underline{M}_t^* = \operatorname{argmin}_{\underline{M}} l_t(\underline{M})$, where compared to classical MPC formulations the costs act only on the terminal state. In other words, GPC tries to emulate a parameterized MPC by always updating the parameters towards the MPC solution. As such, MPC lends itself as analysis tool for GPC and existing results in MPC may carry over. One difference between classical MPC formulations and the present nonstochastic control version defined by l_t comes from the fact that in MPC, the simulation (or rollout) is interpreted as *prediction*, instead of *loss approximation in hindsight*. As such, the initial state in (5) would be updated to the current state at each time step, i.e., set to $x_{0|t} = x_t$.

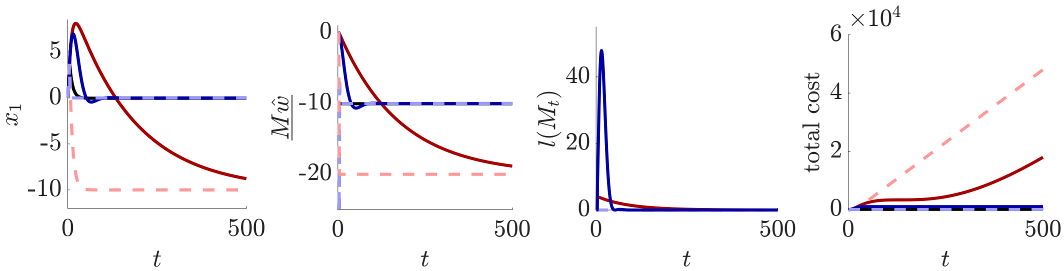
Note that with $x_{0|t} = 0$, the optimal solution $\underline{M}_t^*$ depends only on the current cost function c_t and the past disturbances $\hat{w}_{-L:H-1}$. Imagine the case where the cost function is fixed and the disturbances are constant or very slowly time-varying (compared to the update rate of GPC). Then, $\underline{M}_t^*$ is constant and GPC converges quickly to fixed parameters, representing a very simple constant policy. If instead, the initial state of (5) was set to $x_{0|t} = x_t$, the OCP would implicity represent a linear affine map from x_t to $\underline{M}_t^*$ [Goulart et al., 2006], with the map being parameterized by the disturbances. As a consequence, GPC with varying initial state (for the loss simulation) could still influence the dynamics.

A pathological example for the gradient perturbation controller Consider a simple integrator system with constant disturbance where the first and second component of the state may denote the position and velocity of a point mass, control inputs change the velocity, and the disturbance represents unknown changes in acceleration and velocity in between time steps,

$$x_{t+1} = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix} x_t + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_t + \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{18}$$

Let $K = [-1 \quad -1]$ stabilize the system, imagine the objective is to keep the point mass at the origin, and let the learner's system model be exact so that the resulting predictions (loss simulation) used to compute the gradient are exact. Since the estimated disturbances in $\underline{\hat{w}}_t$ are constant, so is any DAP $\underline{M}\underline{\hat{w}}_t$ and we choose a minimal disturbance memory of $L = 1$ without loss of generality. For ease of exposition, set the horizon to $H = 2$. In this simple setting, we would expect GPC to perform quite well. However, it does not, as seen in Figure 1 (a), where the position x_1 tends to -10 instead of zero. As shown by the behavior of the associated MPC algorith, this is not an issue of convergence, but of a loss function disconnected to the problem at hand. Figure 1 (c),(d) shows the disconnect between loss, which tends to zero, and cumulative costs, which grow unbounded. GPC takes gradient steps that minimize $\bar{x}_{2|t,1}^2 = (0.1v + 2)^2$ and converges to a constant input $\underline{M}\underline{\hat{w}} = -20$. The resulting steady state $x_\infty = (A + BK)x_\infty + [1 \ -20]^\mathrm{T}$ is $x_\infty = [-10 \ -10]^\mathrm{T}$. With larger horizons H, the steady state error of GPC shrinks, but only tends to zero for the maximal choice $H = t$, i.e., if the full horizon is taken into account. For example a horizon of $H = 50$ leads to a steady state $[-0.0127, -10]^\mathrm{T}$.

If the loss simulation instead starts at the current state $x_{0|t} = x_t$, the steady state error vanishes and MPC even beats the best fixed DAP $\underline{M}*$ computed in hindisight (and denoted by Opt). If x_t is accounted for in the loss, GPC minimizes $x_{2|t,1} = ([1 \quad 0]\, A^2 z_0 + 0.1v + 2)^2 = ([0.9 \quad 0.1]\, x_{0|t} + 0.1v + 2)^2$ and no longer tends to a constant input, but towards an affine linear state feedback $v_t = -[9 \quad 1]\, x_t - 20$ under which the steady state $x_\infty = [0 \ -10]^\mathrm{T}$ incurs zero cost. Regret against the best fixed DAP in hindisght is not only sublinear, but bounded.



(a) Position of the mass. (b) Policy $v_t = \underline{M_t w}$ (c) Simulated loss. (d) Cumulative costs.

GPC — MPC — Opt — GPC var ini — MPC var ini

Figure 1: A simple pathological example of the basic nonstochastic control algorithm (OGD) as proposed in the literature. GPC's loss tends to zero while the costs do not. With varying initial condition (var ini), the costs tend to zero.

A generalization Considering the MPC variants lets us generalize this example. In the following, consider constant disturbances $w_t = w$ and fixed costs $c(x, u)$ with minimizing steady state $(x^*, u^*) = \text{argmin}\, c(x, u)$ such that $x^* = A_{\mathrm{K}} x^* + Bu^* + w$. Assume that x^* is reachable in H time steps and that $x_{H|t} = x^*$ is the terminal state of the solution trajectory to the OCP such that $x_{H|t} = A_{\mathrm{K}}^H x_{0|t} + S_{H-1} Bv + S_{H-1} w$, where $S_{H-1} = I + A_{\mathrm{K}} + \ldots + A_{\mathrm{K}}^{H-}$. At every time step t, solving the OCP with $x_{0|t} = 0$ leads to a constant input v_t where

$$Bv_t = S_{H-1}^{-1} x^* - w, \quad x_{t+1} = A_{\mathrm{K}} x_t + S_{H-1}^{-1} x^*. \tag{19}$$

The state thus converges, since A_{K} is stable, but setting $x_t = x_{t+1} = x_\infty$ leads to

$$x_\infty = (I - A_{\mathrm{K}})^{-1} S_{H-1}^{-1} x^* = (I - A_{\mathrm{K}})^{-1} (I - A_{\mathrm{K}})(I - A_{\mathrm{K}}^H) x^* = (I - A_{\mathrm{K}}^H) x^* \tag{20}$$

so that x_t only converges (close) to x^* for very large horizons H where $A_{\mathrm{K}}^H \approx 0$.

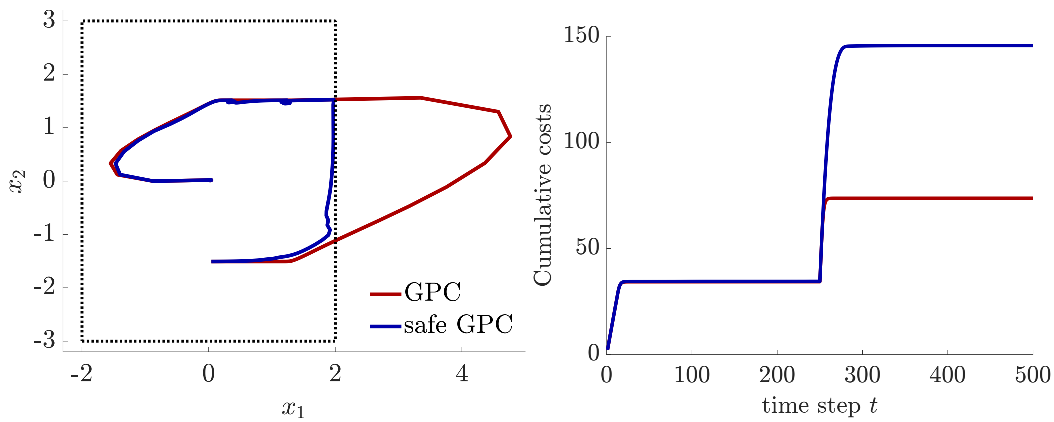
This is different in the case where the initial state is updated to the current state, $x_{0|t} = x_t$.

Proposition 3. *Consider constant disturbances $w_t = w$ and assume the predicted terminal state satisfies $x_{H|t} = x^*$ for all $t \geq 0$. Then the closed-loop dynamics induced by MPC with $x_{0|t} = x_t$ are stable and x_t converges to x^*.*

The technical proof of Proposition 3 is in the Appendix. We note here that with the change of initial state in the OCP, the first (optimal) predicted state $x_{1|t}$ is the actual next state x_{t+1}. So that if the state ever converges, i.e., $x_t = x_{t+1}$, we had $x_{1|t} = x_t$ which implies $x_{k+1|t} = x_{k|t}$ (since the inputs $v_{k|t}$ are constant) so that $x_{H|t} == \ldots = x_t$ which implies $x_t = x^*$ by assumption. In short, the state can *only* converge to the optimal state. As a consequence of Proposition 3, GPC with varying initial state chases an optimal policy that achieves bounded $O(1)$ regret, instead of one that induces a steady-state error.

5 Simulation Example

Consider the numerical example of a linearized DC-DC converter from Section V.B in [Lorenzen et al., 2016], where $\boldsymbol{A} = \begin{bmatrix} 1 & 0.0075 \\ -0.143 & 0.996 \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} 4.798 \\ 0.115 \end{bmatrix}$, the state is subject to constraints $|x_1| \leq 2$, $|x_2| \leq 3$, and the disturbance is bounded as $\|w\|_\infty \leq 0.2$. We let $\mathcal{U} = \{u \in \mathbb{R} \mid |u| \leq 4\}$ and generate an input-state data trajectory of length $T_{\text{Data}} = 15$ starting from zero initial state with inputs and disturbances sampled uniformly from $\mathcal{U}$ and $\mathcal{W}$, respectively. After building the set of models Ω from the data, we solve (15) with $r = 0.6$ and receive a controller $K = [-0.33\, 0.78]$ that is $(8.6, 0.6)$-strongly stable for all models in Ω. We choose the Chebyshev center of Ω as nominal model $(\hat{A}, \hat{B})$, set $H = 10$, $L = 1$, pick a learning rate $\eta = 0.1$ and transition to the control phase of Algorithm 1. During the control phase, we let the disturbance vary at constant rate from zero to $[0.2\, 0.2]^{\mathrm{T}}$ and back to zero over $T = 500$ time steps. The cost functions are defined as $c_t(x_t, u_t) = (x_{t,2} - x_{t,2}^*)^2$ where $x_{t,2}^* = 1.5$ for the first 250 time steps and $x_{t,2}^* = -1.5$ for the



(a) State space

(b) Cumulative costs

Figure 2: Behavior of safe GPC (blue) and GPC without state and input constraints (red). Both methods first steer to and stabilize the state at the optimal $x_2 = 1.5$ in the first 250 time steps, and the optimal -1.5 in the second 250 time steps. Since safe GPC needs to adhere to the state constraints on x_1, it takes more time steps to transition and suffers higher cost along the way.

last 250 time steps. Recall that both the disturbances and future cost functions are unknown to the control algorithm. Figure 2 shows the resulting trajectories and cumulative costs for the proposed safe nonstochastic control algorithm running with varying initial state as proposed in Section 4. For comparison, the equivalent nonstochastic control algorithm without safety constraints is also shown. In the transition from $x_2 = 1.5$ to $x_2 = -1.5$, high values of x_1 are necessary. As seen in Figure 2(a), the proposed algorithms satisfies the safety constraints with virtually no conservatism.

6 Conclusion

This work addressed the challenge of ensuring safety in online nonstochastic control for linear systems with unknown parameters. By leveraging a data-driven robust control approach based on set membership identification, we derived non-conservative constraint sets for policy parameters and constructed a strongly stabilizing controller. In contrast to existing works, both safety and strong stability are guaranteed for all unfalsified models and hold with certainty. In simulation, we demonstrated that our approach can effectively maintain system safety and performance from data alone. By integrating principles from model predictive control, we ensured recursive feasibility of the safety constraints and showed how updating the initial state of policy gradient rollouts effectively eliminates steady-state errors under constant or slowly varying disturbances. Beyond the above, this work left certain questions unanswered. First and foremost, we left a formal regret bound against an expressive class of causally safe policies open for future work. We hypothesize that sublinear regret against an expressive class of noncausally safe policies is unattainable in general, since a policy with foreknowledge of future disturbances may lead the state outside of the maximal RCI set and rely on the disturbances to stay safe.

The MPC perspective also poses new questions. What role would intermediate costs play if applied to policy gradient rollouts? Moreover, if rollouts are interpreted as predictions, could not a learned disturbance model, instead of simply the last few disturbance estimates, be included in policy gradient rollouts without losing convexity? The lessons also go in the other direction, as most works in robust MPC either consider nominal predictions without disturbances, implicitly hoping that disturbances average out over time, or defend against the worst case, as in min-max MPC. As a consequence, these algorithms perform poorly if disturbances are constant or slowly-time-varying, a setting which nonstochastic control (with varying initial states) handles gracefully. Another exciting connection to explore is that of nonstochastic control and real-time iterative MPC [Gros et al., 2020], where at each time step, the (sub-)optimal input sequence is computed by updating the prior solution, instead of recomputing anew. Overall, this work highlights the potential of combining online convex optimization-based policy search with robust and predictive control techniques to achieve both safety and performance in real-world control systems.

References

N. Agarwal, B. Bullins, E. Hazan, S. Kakade, and K. Singh. Online control with adversarial disturbances. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 111–119. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agarwal19c.html.

U. M. Ahiyevich, S. E. Parsegov, and P. S. Shcherbakov. Upper bounds on peaks in discrete-time linear systems. *Automation and Remote Control*, 79(11):1976–1988, Nov 2018. ISSN 1608-3032. doi: 10.1134/S0005117918110036. URL https://doi.org/10.1134/S0005117918110036.

J. Berberich, J. Köhler, M. A. Müller, and F. Allgöwer. Robust constraint satisfaction in data-driven mpc. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1260–1267. IEEE, 2020.

A. Bisoffi, C. De Persis, and P. Tesi. Controller design for robust invariance from noisy data. *IEEE Transactions on Automatic Control*, 68(1):636–643, 2023. doi: 10.1109/TAC.2022.3170373.

F. Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, 1999. ISSN 0005-1098. doi: https://doi.org/10.1016/S0005-1098(99)00113-2. URL https://www.sciencedirect.com/science/article/pii/S0005109899001132.

F. Blanchini and S. Miani. *Invariant sets*, pages 121–191. Springer International Publishing, Cham, 2015. ISBN 978-3-319-17933-9. doi: 10.1007/978-3-319-17933-9_4. URL https://doi.org/10.1007/978-3-319-17933-9_4.

X. Chen and E. Hazan. Black-box control for linear dynamical systems. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1114–1143. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/chen21c.html.

A. Cohen, A. Hasidim, T. Koren, N. Lazic, Y. Mansour, and K. Talwar. Online linear quadratic control. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1029–1038. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/cohen18b.html.

D. Foster and M. Simchowitz. Logarithmic regret for adversarial online control. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3211–3221. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20b.html.

P. J. Goulart, E. C. Kerrigan, and J. M. Maciejowski. Optimization over state feedback policies for robust control with constraints. *Automatica*, 42(4):523–533, 2006. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2005.08.023. URL https://www.sciencedirect.com/science/article/pii/S0005109806000021.

S. Gros, M. Zanon, R. Quirynen, A. Bemporad, and M. Diehl. From linear to nonlinear mpc: bridging the gap via the real-time iteration. *International Journal of Control*, 93(1):62–80, 2020. doi: 10.1080/00207179.2016.1222553.

E. Hazan and K. Singh. Introduction to online nonstochastic control, 2023.

E. Hazan, S. Kakade, and K. Singh. The nonstochastic control problem. In A. Kontorovich and G. Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 408–421. PMLR, 08 Feb–11 Feb 2020. URL https://proceedings.mlr.press/v117/hazan20a.html.

Y. Li, S. Das, and N. Li. Online optimal control with affine constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8527–8537, May 2021. doi: 10.1609/aaai.v35i10.17035. URL https://ojs.aaai.org/index.php/AAAI/article/view/17035.

Y. Li, J. Yu, L. Conger, and A. Wierman. Learning the uncertainty sets for control dynamics via set membership: A non-asymptotic analysis, 2023.

Y. Li, S. Das, J. Shamma, and N. Li. Safe adaptive learning-based control for linear quadratic regulators constraints, 2024.

X. Liu, Z. Yang, and L. Ying. Online nonstochastic control with adversarial and static constraints. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22277–22288. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/liu23at.html`.

M. Lorenzen, F. Dabbene, R. Tempo, and F. Allgöwer. Constraint-tightening and stability in stochastic model predictive control. *IEEE Transactions on Automatic Control*, 62(7):3165–3177, 2016.

M. Lorenzen, M. Cannon, and F. Allgöwer. Robust MPC with recursive model update. *Automatica*, 103:461–471, 2019.

A. Martin, L. Furieri, F. Dörfler, J. Lygeros, and G. Ferrari-Trecate. Regret optimal control for uncertain stochastic systems, 2023.

E. Minasyan, P. Gradu, M. Simchowitz, and E. Hazan. Online control of unknown time-varying dynamical systems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15934–15945. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/856b503e276cc491e7e6e0ac1b9f4b17-Paper.pdf`.

M. Nonhoff, E. Dall'Anese, and M. A. Müller. Online convex optimization for robust control of constrained dynamical systems, 2024.

J. B. Rawlings, D. Q. Mayne, and M. Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, Madison, Wisconsin, 2 edition, 2017. ISBN 9780975937730.

T. Sarkar and A. Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5610–5618. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/sarkar19a.html`.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

M. Simchowitz. Making non-stochastic control (almost) as easy as stochastic. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018. URL `https://proceedings.mlr.press/v75/simchowitz18a.html`.

M. Simchowitz, K. Singh, and E. Hazan. Improper learning for non-stochastic control. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3320–3436. PMLR, 09–12 Jul 2020. URL `https://proceedings.mlr.press/v125/simchowitz20a.html`.

Y. J. Sun, S. Newman, and E. Hazan. Optimal rates for bandit nonstochastic control. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21908–21919. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/45591d6727f0e127295f8d16adba6b23-Paper-Conference.pdf`.

J. Teutsch, S. Kerz, D. Wollherr, and M. Leibold. Sampling-based stochastic data-driven predictive control under data uncertainty. *arXiv preprint arXiv:2402.00681*, 2024.

H. J. Van Waarde, J. Eising, M. K. Camlibel, and H. L. Trentelman. The informativity approach: To data-driven analysis and control. *IEEE Control Systems Magazine*, 43(6):32–66, 2023. doi: 10.1109/MCS.2023.3310305.

A. Wagenmaker and K. Jamieson. Active learning for identification of linear dynamical systems. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3487–3582. PMLR, 09–12 Jul 2020. URL `https://proceedings.mlr.press/v125/wagenmaker20a.html`.

H. Zhou and V. Tzoumas. Safe non-stochastic control of linear dynamical systems. In *IEEE Conference on Decision and Control (CDC)*, 2023. URL `https://arxiv.org/pdf/2308.12395.pdf`.

A Appendix

Variants of the following lemma are well known in the system identification literature. The following version is adapted from Bisoffi et al. [2023].

Lemma 2. *The set of consistent models Ω is convex and closed. It is bounded if and only if its generating data satisfies* $\operatorname{rank} \begin{bmatrix} x_0 & \cdots & x_{N-1} \\ u_0 & \cdots & u_{N-1} \end{bmatrix} = n_x + n_u$.

In practice, the rank condition of Lemma 2 is easily satisfied by long enough trajectories with random inputs.

The following is a classical result in control due to Lyapunov.

Proposition 4. *A system $x_{t+1} = Ax_t$ is stable in the sense that $\lim_{t\to\infty} x_t = 0$ if and only if there exists $P \succ 0$ such that*

$$P - A^{\mathrm{T}} P A \succ 0. \tag{21}$$

Proposition 4 implies the existence of a scalar *Lyapunov* function $V(x) = x^{\mathrm{T}} P x$ which attains its minimum at the origin ($V(x) > 0$ for all $x \neq 0$ and $V(0) = 0$) and descents with time ($V(x_{t+1}) < V(x_t)$) until $x_t = 0$ since for all $x_t \neq 0$ the condition 21 guarantees

$$V(x_{t+1}) - V(x_t) = x_t^{\mathrm{T}} A^{\mathrm{T}} P A x_t - x_t P x_t = x_t^{\mathrm{T}} (A^{\mathrm{T}} P A - P) x_t < 0. \tag{22}$$

Informally, this implies $\lim_{t\to\infty} V(x_t) = V(x_{t\to\infty}) = \min_x V(x) = V(0)$ and the state tends to the origin.

Construction of the maximal RCI subset A maximal RCI subset $\mathcal{X}$ of $\mathcal{F}_x$ can be constructed by recursion [Blanchini and Miani, 2015], where the idea is to first set $\mathcal{X}_0 = \mathcal{F}_x$ and iteratively compute $\mathcal{X}_{k+1}$ as the set of all states from which $\mathcal{X}_k$ can be surely reached (for all disturbances in $\mathcal{W}$). That is, $\mathcal{X}_{k+1}$ contains all states for which there exists an admissible input which drives the nominal state (without disturbance) $A_* x + B_* u$ into $\mathcal{X}_k \ominus \mathcal{W}$,

$$\mathcal{X}_{k+1} = \operatorname{proj}_{1:n_x} \{ z \in \operatorname{col}(\mathcal{X}_k, \mathcal{U}) \mid \Omega z \in \mathcal{X}_k \ominus \mathcal{W} \}. \tag{23}$$

Crucially, $x \in \mathcal{X}_{k+1}$ guarantees the existence of *one* input that drives *all* models of Ω into $\mathcal{X}_k \ominus \mathcal{W}$ and may be computed similar to $\mathcal{F}_x$ above based on vertices of Ω, yielding again a convex polytope. Note that $\mathcal{X}_{k+1} \subseteq \mathcal{X}_k$ by construction. As soon as $\mathcal{X}_{k+1} = \mathcal{X}_k$ the computation is stopped and $\mathcal{X} := \mathcal{X}_k$ is RCI for the true system following (12).

Proof of Proposition 1 The proof makes use the well-known fact that sublevel sets of Lyapunov functions are positive invariant, which we formally define next before proving the result.

Definition 5. *A set $\mathcal{X}$ is positive invariant for dynamics $x_{t+1} = f(x_t)$ if $f(x) \in \mathcal{X}$ for all $x \in \mathcal{X}$.*

Lemma 3. *Consider a system $x_{t+1} = f(x_t)$ and let $V(x)$ be a Lyapunov function such that $V(0) = 0$, $V(x) > 0 \,\forall x \neq 0$, and $V(f(x)) \leq V(x) \,\forall x \in \mathbb{R}^{n_x}$. Then any sublevel set $\mathcal{E}_c = \{x \in \mathbb{R}^{n_x} \mid V(x) \leq c, c \geq 0\}$ of $V(x)$ is positive invariant for dynamics $x_{t+1} = f(x_t)$.*

Proof. We first show that condition 17 implies stability of of

$$x_{t+1} = \frac{1}{r}(A + BK) x_t. \tag{24}$$

By the Schur complement, it is equivalent to $rZ \succ 0$ and $rZ - (\hat{A}^{(i)}Z + \hat{B}^{(i)}Y)^{\mathrm{T}}(rZ)^{-1}(\hat{A}^{(i)}Z + \hat{B}^{(i)}Y) \succ 0$. Multiplying from left and right by Z^{-1} yields

$$rZ^{-1} - (\hat{A}^{(i)} + \hat{B}^{(i)}YZ^{-1})^{\mathrm{T}}\frac{1}{r}Z^{-1}(\hat{A}^{(i)} + \hat{B}^{(i)}YZ^{-1}) \succ 0. \tag{25}$$

Substituting $K = YZ^{-1}$ and $P = Z^{-1}$ and dividing by r leads to

$$P - \frac{1}{r}(\hat{A}^{(i)} + \hat{B}^{(i)}K)^{\mathrm{T}}P\frac{1}{r}(\hat{A}^{(i)} + \hat{B}^{(i)}K) \succ 0 \tag{26}$$

so that $V(x) = x^{\mathrm{T}}Px$ is a Lyapunov function certifying stability for each closed loop system $x_{t+1} = \frac{1}{r}(\hat{A}^{(i)} + \hat{B}^{(i)}K)x_t$ by Proposition 4. By convexity of Ω, stability of the models at the vertices of Ω implies stability for all models in Ω, which in turn implies $\rho(\frac{1}{r}(A + BK)) \leq 1$ or equivalently $\rho(A + BK) \leq r$ for all $(A, B) \in \Omega$.

Since $V(x_{t+1}) \leq V(x_t)$ the ellipsoidal sublevel set $\mathcal{V} = \{x \in \mathbb{R}^{n_x} \mid x^{\mathrm{T}}Px \leq 1\}$ of $V(x)$ is positive invariant for system 24, i.e., $x_0 \in \mathcal{V} \implies r^{-t}(A + BK)^t x_0 \in \mathcal{V}$ for all time $t \geq 0$. Multiplying all sides in condition 16 by $P = Z^{-1}$ yields $P \preceq I$ and $I \preceq cP$ which in turn implies $x^{\mathrm{T}}x \leq x^{\mathrm{T}}Px$ and $x^{\mathrm{T}}Px \leq \frac{1}{c}x^{\mathrm{T}}x$. As a consequence, whenever $x^{\mathrm{T}}x \leq 1$ then $x^{\mathrm{T}}Px \leq 1$ and thus $\mathcal{V}$ contains the unit norm ball $\mathcal{B}_1 = \{x \in \mathbb{R}^{n_x} \mid \|x\| = \sqrt{x^{\mathrm{T}}x} \leq 1\}$. Additionally, whenever $x^{\mathrm{T}}Px \leq 1$ then $\frac{1}{c}x^{\mathrm{T}}x \leq 1$ (and equivalently $x^{\mathrm{T}}x \leq c$) so that $\mathcal{V}$ is contained in a ball around the origin with radius $\sqrt{c}$ denoted by $\mathcal{B}_{\sqrt{c}} = \{x \in \mathbb{R}^{n_x} \mid \|x\| \leq \sqrt{c}\}$. By positive invariance of $\mathcal{V}$ then, $x \in \mathcal{B}_1 \subseteq \mathcal{V}$ implies $r^{-t}(A + BK)^t x \in \mathcal{V} \subseteq \mathcal{B}_{\sqrt{c}}$ for all time $t \geq 0$. In other words

$$\|x\| \leq 1 \implies \frac{1}{r^t}\|(A + BK)^t x\| \leq \sqrt{c} \iff \sup_{\|x\| \leq 1} \frac{\|(A + BK)^t x\|}{\|x\|} \leq \sqrt{c}r^t \tag{27}$$

which proofs the last part of the result by definition of the induced matrix norm. □

Proof of Proposition 3

Proof. Rearranging $x^* = x_{H|t} = A_{\mathrm{K}}^H x_{0|t} + S_{H-1}Bv + S_{H-1}w$ with $x_{0|t} = x_t$ leads to

$$Bv_t = S_{H-1}^{-1}(x^* - A_{\mathrm{K}}^H x_t) - w, \quad x_{t+1} = (A_{\mathrm{K}} - S_{H-1}^{-1}A_{\mathrm{K}}^H)x_t + S_{H-1}^{-1}x^*. \tag{28}$$

Note $(A_{\mathrm{K}} - S_{H-1}^{-1}A_{\mathrm{K}}^H) = S_{H-1}^{-1}(S_{H-1}A_{\mathrm{K}} - A_{\mathrm{K}}^H)$ and

$$S_{H-1}A_{\mathrm{K}} - A_{\mathrm{K}}^H = (I + \ldots + A_{\mathrm{K}}^{H-1})A_{\mathrm{K}} - A_{\mathrm{K}}^H = A_{\mathrm{K}} + \ldots + A_{\mathrm{K}}^{H-1} = S_{H-1} - I \tag{29}$$

so that the dynamics induced by MPC can be rewritten as

$$x_{t+1} = (I - S_{H-1}^{-1})x_t + S_{H-1}^{-1}x^*. \tag{30}$$

Letting $x_t = x_{t+1} = x_\infty$ immediately leads to $S_{H-1}^{-1}x_\infty = S_{H-1}^{-1}x^*$ which implies $x_\infty = x^*$ since S_{H-1}^{-1} has full rank. It remains to show that x_t actually converges, i.e., the closed-loop dynamics (30) are stable. Let λ be an eigenvalue of A_{K} such that $A_{\mathrm{K}}v = \lambda v$ for some $v \in \mathbb{R}^{n_x}$. Then $S_{H-1}v = (1 + \lambda + \ldots + \lambda^{H-1})v$ so that $(1 + \lambda + \ldots + \lambda^{H-1})^{-1}$ is an eigenvalue of S_{H-1}^{-1} and $1 - (1 + \lambda + \ldots + \lambda^{H-1})^{-1} = \frac{\lambda - \lambda^H}{1 - \lambda^H}$ is an eigenvalue of the closed-loop dynamics (30). Since $\frac{\lambda - \lambda^H}{1 - \lambda^H} \in [0, \lambda)$ for all $H \in \mathbb{N}$ we have $\rho(I - S_{H-1}^{-1}) < \rho(A_{\mathrm{K}})$ and the closed-loop dynamics are stable by (strong) stability of A_{K}. □