Best of Both Worlds in Online Control: Competitive Ratio and Policy Regret

Gautam Goel GGOEL@BERKELEY.EDU
Simons Institute, UC Berkeley

Naman Agarwal NAMANAGARWAL@GOOGLE.COM
Google AI Princeton

Karan Singh KARANSINGH@CMU.EDU
Tepper School of Business, Carnegie Mellon University

Elad Hazan EHAZAN@PRINCETON.EDU
Google AI Princeton & Department of Computer Science, Princeton University

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

We consider the fundamental problem of online control of a linear dynamical system from two different viewpoints: regret minimization and competitive analysis. We prove that the optimal competitive policy is well-approximated by a convex parameterized policy class, known as a disturbance-action control (DAC) policies. Using this structural result, we show that several recently proposed online control algorithms achieve the best of both worlds: sublinear regret vs. the best DAC policy selected in hindsight, and optimal competitive ratio, up to an additive correction which grows sublinearly in the time horizon. We further conclude that sublinear regret vs. the optimal competitive policy is attainable when the linear dynamical system is unknown, and even when a stabilizing controller for the dynamics is not available *a priori*.

Keywords: Nonstochastic control, regret minimization, competitive ratio.

1. Introduction

The study of online optimization consists of two main research directions. The first is online learning, which studies regret minimization in games. A notable framework within this line of work is online convex optimization, where an online decision maker iteratively chooses a point in a convex set and receives loss according to an adversarially chosen loss function. The metric of performance studied in this research thrust is regret, or the difference between overall loss and that of the best decision in hindsight.

The second direction is that of competitive analysis in metrical task systems. In this framework, the problem setting is similar, but the performance metric is very different. Instead of regret, the objective is to minimize the competitive ratio, i.e. the ratio of the reward of the online decision maker to that associated with the optimal sequence of decisions made in hindsight. For this ratio to remain bounded, an additional penalty is imposed on movement costs, or changes in the decision.

While the goals of the two research directions are similar, the performance metrics are very different and lead to different algorithms and methodologies. These two separate methodologies

have recently been applied to the challenging setting of online control, yielding novel and exciting methods to the field of control of dynamical systems.

In the paper we unify these two disparate lines of work by establishing a connection between the two objectives. Namely, we show that the Gradient Perturbation Controller (GPC) minimizes regret against the policy that has the *optimal competitive ratio* for any Linear Time Invariant (LTI) dynamical system. The GPC algorithm hence gives the best of both worlds: sublinear regret, and optimal competitive ratio, in a single efficient algorithm.

Our main technical contribution is proving that the optimal competitive policy derived in Goel and Hassibi (2021) is well-approximated by a certain convex policy class, for which efficient online learning was recently established in the work of Agarwal et al. (2019). This implies that known regret minimization algorithms for online control can compete with this optimal competitive policy with vanishing regret.

This structural result has other important implications to online control, yielding new results: we show that sublinear regret can be attained vs. the optimal competitive policy even when the underlying dynamical system is unknown, and even when a stabilizing controller is not available.

1.1. Related work

Control of dynamical systems. Our study focuses on two types of algorithms for online control. The first class of algorithms enjoy sublinear regret for online control of dynamical systems; that is, whose performance tracks a given benchmark of policies up to a term which is vanishing relative to the problem horizon. Abbasi-Yadkori and Szepesvári (2011) initiated the study of online control under the regret benchmark for linear time-invariant (LTI) dynamical systems. Bounds for this setting have since been improved and refined in Dean et al. (2018); Mania et al. (2019); Cohen et al. (2019); Simchowitz and Foster (2020). We are interested in adversarial noise and perturbations, and regret in the context of online control was initiated in the study of *nonstochastic* control setting (Agarwal et al., 2019), that allows for adversarially chosen (e.g. non-Gaussian) noise and general convex costs that may vary with time. This model has been studied for many extended settings, see Hazan and Singh (2022) for a comprehensive survey.

Competitive control. Goel and Wierman (2019) initiated the study of online control with competitive ratio guarantees and showed that the Online Balanced Descent algorithm introduced in Chen et al. (2018) has bounded competitive ratio in a narrow class of linear systems. This approach to competitive control was extended in a series of papers (Goel et al., 2019; Shi et al., 2020). In recent work, Goel and Hassibi (2021) obtained an algorithm with optimal competitive ratio in general linear systems using H_∞ techniques; in this paper we show that the competitive control algorithm obtained in Goel and Hassibi (2021) is closely approximated by the class of DAC policies, and use this connection to obtain our "best-of-both-worlds" result.

Online learning and Online Convex Optimization (OCO). The regret minimization techniques that are the subject of this paper are based in the framework of online convex optimization, see Hazan (2019). Recent techniques in online nonstochastic control are based on extensions of OCO to the setting of loss functions with memory (Anava et al., 2015) and adaptive or dynamic regret (Hazan and Seshadhri, 2009; Zhang et al., 2018).

Competitive analysis of online algorithms and simultaneous bounds on competitive ratio and regret. Competitive analysis was introduced in Sleator and Tarjan (1985) and was first studied in the context of Metrical Task Systems (MTS) in Borodin et al. (1992); we refer to Borodin and

El-Yaniv (2005) for an overview of competitive analysis and online algorithms. A series of recent papers consider the problem of obtaining online algorithms with bounded competitive ratio and sublinear regret. In Andrew et al. (2013), it was shown no algorithm can simultaneously achieve both objectives in OCO with switching costs. On the other hand, Daniely and Mansour (2019) described an online algorithm for MTS with optimal competitive ratio and sublinear regret on every time interval.

2. Preliminaries

We consider the task of online control in linear time-invariant (LTI) dynamical systems. In this setting, the interaction between the learner and the environment proceeds as described next. At each time step, the learner incrementally observes the current state $x_t \in \mathbb{R}^m$ of the system, subsequently chooses a control input $u_t \in \mathbb{R}^n$, and consequently is subject to an instantaneous cost $c(x_t, u_t)$ defined via the quadratic cost function (we assume the existence of β, μ such that $\beta I \succeq Q, R \succeq \mu I$)

$$c(x, u) = x^\top Q x + u^\top R u.$$

As a consequence of executing the control input u_t, the dynamical system evolves to a subsequent state x_{t+1}, as dictated by the following linear system parameterized by the matrices $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{m \times n}$, bounded as $\|A\|, \|B\| \leq \kappa$, and the perturbation sequence $(w_t)_{t \in [T]}$.

$$x_{t+1} = A x_t + B u_t + w_t.$$

We assume without loss of generality that $x_1 = 0$. The learner does not directly observe the perturbations, or know of them in advance. We do not make any (e.g., distributional) assumptions on the perturbations, other than that they satisfy a point-wise bound $\|w_t\| \leq W$ for all times steps t. By the means of such interaction across T time steps, we ascribe an aggregate cost to the learner $\mathcal{A}$ as

$$J_T(\mathcal{A} | w_{1:T}) = \sum_{t=1}^{T} c(x_t, u_t).$$

2.1. Policy classes

Since the learner selects the control inputs adaptively upon observing the state, the behavior of a learner may be described by a (strictly causal) policy π, a mapping from the observed state sequence to the immediate action. We consider the following policy classes in the paper:

1. Π_{ALL} is the exhaustive set of T-length sequence of control inputs.
2. Π_{SC} is the class of all strictly causal policies, mapping the *heretofore* observed state sequence to the next action.
3. $\mathcal{K} \subset \mathbb{R}^{n \times m}$ is a class of linear state-feedback policies. Each member of this class is parameterized by some matrix $K \in \mathcal{K}$, and recommends the immediate action $u_t \stackrel{\text{def}}{=} K x_t$. Both the stochastic-optimal policy ($\mathcal{H}_2$-control) – Bayes-optimal for i.i.d. perturbations – and the robust policy ($\mathcal{H}_\infty$-control) – minimax-optimal for arbitrary perturbations – are linear state-feedback policies.

4. $\mathcal{M}$ is a class of disturbance-action controllers (DAC), defined below, that recommend actions as a linear transformation of the past few perturbations, rather than the present state.

A linear policy K is called stable if the spectral radius of $A + BK$ is strictly less than 1. Such policies ensure that the state sequence remains bounded under their execution. The notion of strong stability, introduced by Cohen et al. (2018), is a non-asymptotic characterization of the notion of stability defined as follows.

Definition 1 *A linear policy $K \in \mathbb{R}^{n\times m}$ is said to be (κ, γ)-strongly stable with respect to an LTI (A, B) if there exist matrices S, L satisfying $A + BK = SLS^{-1}$ such that*

$$\max\{1, \|K\|, \|S\|\|S^{-1}\|\} \leq \kappa \text{ and } \max\{1/2, \|L\|\} \leq 1 - \gamma.$$

A sufficient condition for the existence of a strongly stable policy is the strong controllability of the linear system (A, B), a notion introduced in Cohen et al. (2018). In words, strong controllability measures the minimum length and magnitude of control input needed to drive the system to any unit-sized state.

Let $\mathbb{K}$ be a fixed (κ, γ)-strongly stable linear policy for the discussion that follows. We will specify a particular choice for $\mathbb{K}$ in Section 3. We formally define a disturbance action controller below. The purpose of superimposing a stable linear policy $\mathbb{K}$ on top of the linear-in-perturbation terms is to ensure that the state sequence produced under the execution of a (possibly non-stationary) disturbance-action controller remains bounded.

Definition 2 *A disturbance-action controller (DAC), specified by a horizon H and parameters $M = \left(M^{[0]}, \ldots M^{[H-1]}\right) \in \mathbb{R}^{n\times m}$, chooses the action at the time t as*

$$u_t(M) \stackrel{def}{=} \mathbb{K}x_t + \sum_{i=1}^{H} M^{[i-1]} w_{t-i},$$

where x_t is state at time t, and $w_1, \ldots w_{t-1}$ are past perturbations.

Definition 3 *For any $H \in \mathbb{N}, \gamma < 1$, and $\theta \geq 1$, an (H, θ, γ)-DAC policy class is the set of all H-horizon DAC policies where $M = \left(M^{[0]}, \ldots M^{[H-1]}\right)$ satisfy $\|M^{[i]}\| \leq \theta(1 - \gamma)^i$ for all $i \in \{0, \ldots, H - 1\}$.*

2.2. Performance measures

This paper considers multiple criteria that may be used to assess the learner's performance. We introduce these below.

Let $w_{1:T}$ be the perturbation sequence the dynamics are subject to. Given the foreknowledge of this sequence, we define the following notions of optimal cost; note that these notions are *infeasible* in the sense that no online learner can match these on all instances.

1. $OPT_*(w_{1:T}) \stackrel{\text{def}}{=} \min_{u_{1:T}\in\Pi_{\text{ALL}}} J_T(u_{1:T}|w_{1:T})$ is the cost associated with the best sequence of control inputs given the perturbation sequence. No policy, causal or otherwise, can attain a cost smaller than $OPT_*(w_{1:T})$ on the the perturbation sequence $w_{1:T}$.

2. For any policy class Π, $OPT_\Pi(w_{1:T}) \stackrel{\text{def}}{=} \min_{\pi\in\Pi} J_T(\pi|w_{1:T})$ is the cost of the best policy in Π, subject to the perturbation sequence. Note that $OPT_*(w_{1:T}) = OPT_{\Pi_{\text{ALL}}}(w_{1:T})$.

With respect to these baselines, we define the following performance measures.

Competitive Ratio: The competitive ratio of an (online) learner $\mathcal{A}$ is the worst-case ratio of its cost to the optimal offline cost $OPT_*(w_{1:T})$ over all possible perturbation sequences.

$$\alpha_T(\mathcal{A}) \stackrel{\text{def}}{=} \max_{w_{1:T}} \frac{J_T(\mathcal{A}|w_{1:T})}{OPT_*(w_{1:T})}$$

The optimal competitive ratio is the competitive ratio of the best strictly causal controller.

$$\alpha_T^* \stackrel{\text{def}}{=} \min_{\mathcal{A}\in\Pi_{\text{SC}}} \alpha_T(\mathcal{A})$$

The infinite-horizon optimal competitive ratio $\alpha^* = \lim_{T\to\infty} \alpha_T^*$ is defined as the limiting optimal competitive ratio as the horizon extends to infinity, whenever it exists.

Regret: On any perturbation sequence $w_{1:T}$, given a policy class Π, the regret of an online learner $\mathcal{A}$ is assigned to be the excess aggregate cost incurred in comparison to that of the best policy in Π,

$$R_{T,\Pi}(\mathcal{A}|w_{1:T}) = J_T(\mathcal{A}|w_{1:T}) - OPT_\Pi(w_{1:T}).$$

The worst-case regret is defined as the maximum regret attainable over all perturbation sequences,

$$R_{T,\Pi}(\mathcal{A}) = \max_{w_{1:T}} R_{T,\Pi}(\mathcal{A}|w_{1:T}).$$

The two types of performance guarantees introduced above are qualitatively different in terms of the bound they espouse and the baseline they compare to. In particular:

Tighter bound for regret: A sub-linear regret guarantee implies that the average costs of the learner and the baseline asymptotically match, while even an optimal competitive-ratio bound promises an average cost at most a constant factor times that of the baseline.

Stronger baseline for competitive ratio: Competitive ratio measures performance relative to the optimal dynamic policy while regret measure performance relative to the best static policy from a (typically parametric) policy class.

2.3. Characterization of the optimal Competitive Ratio algorithm

The following explicit characterization of a strictly causal policy that achieves an optimal competitive ratio in the infinite-horizon setting was recently obtained in Goel and Hassibi (2021); this theorem shows that the competitive policy in the original system with state $x \in \mathbb{R}^m$ can be viewed as a state-feedback controller in a synthetic system with state $\xi \in \mathbb{R}^{2m}$.

Theorem 4 (Optimal Competitive Policy) *The strictly causal controller with an optimal infinite-horizon competitive ratio α^* is given by the policy $u_t = \widehat{K}\xi_t$, where $\widehat{K} \in \mathbb{R}^{n\times 2m}$ and the synthetic state $\xi \in \mathbb{R}^{2m}$ evolves according to the dynamics*

$$\xi_{t+1} = \widehat{A}\xi_t + \widehat{B}_u u_t + \widehat{B}_w \hat{w}_{t+1},$$

$$\textit{where } \widehat{A} = \begin{bmatrix} A & K\Sigma^{1/2} \\ 0 & 0 \end{bmatrix}, \quad \widehat{B}_u = \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad \widehat{B}_w = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad \widehat{w}_t = \Sigma^{-1/2} Q^{1/2} \nu_t.$$

The sequence ν_t is recursively defined as $\nu_{t+1} = (A - KQ^{1/2})\nu_t + w_t$ starting with $\nu_1 = 0$. Here the matrices $\widehat{K}, K, \Sigma$ (and auxiliary constants $P, \widetilde{B}, \widetilde{H}$ and $\widehat{P}$) satisfy

$$K = APQ^{1/2}\Sigma^{-1}, \quad \Sigma = I + Q^{1/2}PQ^{1/2}, \quad P = BB^\top + APA^\top - K\Sigma K^\top,$$

$$\widehat{K} = -(I_n + \widehat{B}_u^\top \widetilde{P}\widehat{B}_u)^{-1}\widehat{B}_u^\top \widetilde{P}\widehat{A}, \quad \widetilde{B} = \begin{bmatrix}\widehat{B}_u & \widehat{B}_w\end{bmatrix}, \quad \widetilde{H} = \begin{bmatrix} I & 0 \\ 0 & -\alpha^* I\end{bmatrix} + \widetilde{B}^\top \widehat{P}\widetilde{B},$$

$$\widetilde{P} = \widehat{P} - \widehat{P}\widehat{B}_w(-\alpha^* I_p + \widehat{B}_w^\top \widehat{P}\widehat{B}_w)^{-1}\widehat{B}_w^\top \widehat{P},$$

$$\text{and} \quad \widehat{P} = \begin{bmatrix} Q & Q^{1/2}\Sigma^{1/2}, \\ Q^{1/2}\Sigma^{1/2} & \Sigma\end{bmatrix} + \widehat{A}^\top \widehat{P}\widehat{A} - \widehat{A}^\top \widehat{P}\widetilde{B}\widetilde{H}^{-1}\widetilde{B}^\top \widehat{P}\widehat{A}.$$

Furthermore, let $\{x_t\}_{t=1}^T$ be the state sequence produced under the execution of such a policy. Then, the state sequence satisfies at all time t that $\xi_t = \begin{bmatrix} x_t - \nu_t \\ \widehat{w}_t\end{bmatrix}$.

Let $\widehat{K}_0 \in \mathbb{R}^{n\times m}$ be the sub-matrix induced by the first m columns of $\widehat{K}$. In general, the infinite-horizon optimal competitive ratio may not be finite. However, the stability of the associated filtering operation (i.e. $|\lambda_{\max}(A - KQ^{1/2})| < 1$) and the closed loop control system (i.e. $|\lambda_{\max}(A + B\widehat{K}_0)| < 1$) is sufficient to ensure the existence of this limit. We utilize the following bounds that quantify this.

Assumption 1 *$\widehat{K}_0$ is (κ, γ)-strongly stable with respect to the linear system (A, B), and $-K^\top$ is (κ, γ)-strongly stable with respect to the linear system $(A^\top, Q^{1/2})$. Also, $\|\widehat{K}\| \le \kappa$.*

We note that the above bounds are quantifications, and not strengthening, of the stability criterion. In particular, any stable controller is strongly stable for some $\kappa \ge 1, \gamma < 1$. Here, we use the same parameters to state the strong stability for both controllers, K and $\widehat{K}$, for convenience. Such a simplification is valid, since given (κ_1, γ_1)- and (κ_2, γ_2)-strongly stable controllers, the said controllers are also $(\max\{\kappa_1, \kappa_2\}, \max\{\gamma_1, \gamma_2\})$-strongly stable.

2.4. Low-regret algorithms

Deviating from the methodologies of optimal and robust control, Agarwal et al. (2019) propose considering an online control formulation in which the noise is adversarial, and thus the optimal controller is only defined in hindsight. This motivates different, online-learning based methods for the control task. Agarwal et al. (2019) proposed an algorithm called GPC (Gradient Perturbation Controller) and show the following theorem (which we restate in notation consistent with this paper), which for LTI systems shows that regret when compared against any strongly-stable policy scales at most as $O(\sqrt{T})$.

Theorem 5 *Given any κ, γ, let $\mathcal{K}(\kappa, \gamma)$ be the set of (κ, γ) strongly stable linear policies. There exists an algorithm $\mathcal{A}$ such that the following holds,*

$$J_T(\mathcal{A}|w_{1:T}) - \min_{K\in\mathcal{K}(\kappa,\gamma)} J_T(K|w_{1:T}) \le O(\sqrt{T}\log(T)).$$

Here $O(\cdot)$ contains polynomial factors depending on the system constants.

As can be observed from the analysis presented by Agarwal et al. (2019), the above regret guarantee holds not just against the set of strongly stable linear policies but also against the set of (H, θ, γ)-DAC policies. The regret bound has been extended to different interaction models such as unknown systems (Hazan et al., 2020b), partial observation (Simchowitz et al., 2020) and adaptive regret (Gradu et al., 2020). Furthermore the regret bound in this setting has been improved to logarithmic in T (Foster and Simchowitz, 2020b; Simchowitz, 2020).

3. Statement of results

3.1. Main Result

The central observation we make is that regret-minimizing algorithms subject to certain qualifications automatically achieve an optimal competitive ratio bound up to a vanishing average cost term.

Typical regret-minimizing online control algorithms (Agarwal et al., 2019; Hazan et al., 2020a) compete against the class of stable linear state-feedback policies. In general, neither the offline optimal policy (with cost OPT_*) nor the optimal competitive-ratio policy can be approximated by a linear policy (Goel and Hassibi, 2020). However, the algorithm proposed in Agarwal et al. (2019) and follow-up works that build on it also compete with a more-expressive class, that of disturbance-action policies (DACs). In Agarwal et al. (2019), this choice was made purely for computational reasons to circumvent the non-convexity of the cost associated with linear policies; in this work, however, we use the flexibility of DAC policies to approximate the optimal competitive policy.

More formally, we prove that we can find a DAC which generates a sequence of states and control actions which closely track the sequence of states and control actions generated by the optimal competitive policy by taking the history H of the DAC to be sufficiently large. This structural characterization of the competitive policy is sufficient to derive our best-of-both-worlds result, since a regret-minimizing learner competitive against an appropriately defined DAC class would also be competitive against the policy achieving an optimal competitive ratio, and hence achieve an optimal competitive ratio up to a residual regret term.

Theorem 6 (Optimal Competitive Policy is Approximately DAC) *Fix a horizon T and a disturbance bound W. For any $\varepsilon > 0$, set*

$$H = \log(1-\gamma/2)^{-1} \log\left(\frac{1088 W^2 \kappa^{11} \max(1, \beta^2)}{\gamma^4 \varepsilon} T\right), \qquad \theta = 2\kappa^2 \max(1, \beta^{1/2})$$

and define $\mathcal{M}$ be the set of (H, θ, γ)-DAC policies with stabilizing component $\mathbb{K} = \widehat{K}_0$. Let $\mathcal{A}$ be the algorithm with the optimal competitive ratio α^*. Then there exists a policy $\pi \in \mathcal{M}$ such that for any perturbations $w_{1:T}$ satisfying $\|w_t\| \leq W$, the cost incurred by π satisfies

$$J_T(\pi | w_{1:T}) < J_T(\mathcal{A} | w_{1:T}) + \varepsilon.$$

We now show that this result implies best-of-both-worlds.[1]

1. Detailed proofs of all the results in this section appear in the full version of this paper, available at https://arxiv.org/abs/2211.11219.

3.2. Best-of-Both-Worlds in Known Systems

We begin by considering the case when the learner knows the linear system (A, B). In this setting, both the regret and competitive ratio thus measure the additional cost imposed by not knowing the perturbations in advance. The result below utilizes the regret bounds against DAC policies and associated algorithms from Agarwal et al. (2019); Simchowitz (2020).

Theorem 7 (Best-of-both-worlds in online control (known system)) *Assuming (A, B) is known to the learner, there exists a constant*

$$R_T = \tilde{O}\left(\text{poly}(m, n, \beta, \kappa, \gamma^{-1})W^2 \times \min\left\{\sqrt{T}, \text{poly}(\mu^{-1})\,\text{polylog}\,T\right\}\right)$$

and a computationally efficient online control algorithm $\mathcal{A}$ which simultaneously achieves the following performance guarantees:

1. *(Optimal competitive ratio) The cost of $\mathcal{A}$ satisfies for any perturbation sequence $w_{1:T}$ that*

$$J_T(\mathcal{A}|w_{1:T}) < \alpha^* \cdot OPT_*(w_{1:T}) + R_T,$$

   where α^* is the optimal competitive ratio.

2. *(Low regret) The regret of $\mathcal{A}$ relative to the best linear state-feedback or DAC policy selected in hindsight grows sub-linearly in the horizon T, i.e. for all $w_{1:T}$, it holds*

$$J_T(\mathcal{A}|w_{1:T}) < \min_{\pi\in\mathcal{K}} J_T(\pi|w_{1:T}) + R_T \quad \textit{and} \quad J_T(\mathcal{A}|w_{1:T}) < \min_{\pi\in\mathcal{M}} J_T(\pi|w_{1:T}) + R_T.$$

3.3. Best-of-Both-Worlds in Unknown Systems

We now present the main results for online control of unknown linear dynamical system. The first theorem deals with the case when the learner has coarse-grained information about the linear system (A, B) in the form of access to a stabilizing controller $\mathbb{K}$. In general, to compute such a stable controller, it is sufficient to know (A, B) to some constant accuracy, as noted in Cohen et al. (2019). This theorem utilizes low-regret algorithms from Hazan et al. (2020a); Simchowitz (2020).

Theorem 8 *For a (k, κ)-strongly controllable linear dynamical system (A, B), there exists a constant*

$$R_T = \tilde{O}\left(\text{poly}(m, n, \beta, k, \kappa, \gamma^{-1})W^2 \times \min\left\{T^{2/3}, \text{poly}(\mu^{-1})\sqrt{T}\right\}\right)$$

and a computationally efficient online control algorithm $\mathcal{A}$ such that, when given access to a (κ, γ)-strongly stable initial controller $\mathbb{K}$, it guarantees

$$J_T(\mathcal{A}|w_{1:T}) < \min\{\alpha^* \cdot OPT_*(w_{1:T}), \min_{\pi\in\mathcal{K}} J_T(\pi|w_{1:T}), \min_{\pi\in\mathcal{M}} J_T(\pi|w_{1:T})\} + R_T.$$

When an initial stabilizing controller is unavailable, we make use of the "blackbox control" algorithm in Chen and Hazan (2021) to establish the next theorem.

Theorem 9 *For a (k, κ)-strongly controllable linear dynamical system (A, B), there exists a constant*

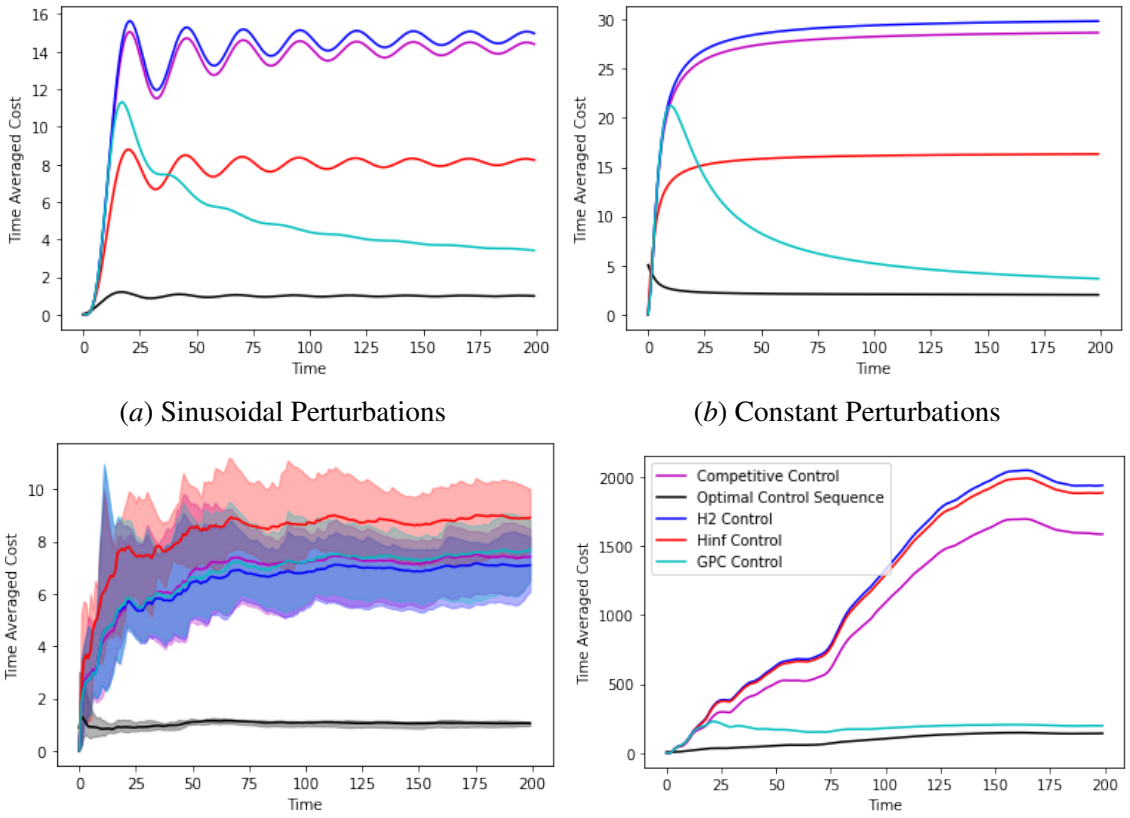
$$R_T = 2^{\text{poly}(m,n,\beta,k,\kappa,\gamma^{-1})} + \tilde{O}\left(\text{poly}(m, n, \beta, k, \kappa, \gamma^{-1})W^2 \times \min\left\{T^{2/3}, \text{poly}(\mu^{-1})\sqrt{T}\right\}\right)$$

and a computationally efficient online control algorithm $\mathcal{A}$ that guarantees

$$J_T(\mathcal{A}|w_{1:T}) < \min\{\alpha^* \cdot OPT_*(w_{1:T}), \min_{\pi\in\mathcal{K}} J_T(\pi|w_{1:T}), \min_{\pi\in\mathcal{M}} J_T(\pi|w_{1:T})\} + R_T.$$

4. Experiments

In Figure 1 we compare the performance of various controllers, namely, the H_2 controller, the H_∞ controller, the infinite horizon competitive controller from Goel and Hassibi (2021), the GPC controller from Agarwal et al. (2019) and the "clairvoyant offline" controller which selects the optimal-in-hindsight sequence of controls. We do this comparison on a two dimensional double integrator system with different noise sequences.[2] We confirm as the results of our paper suggest that the GPC controller attaining the best of both worlds guarantee is indeed the best performing controller and in particular matches and sometimes improves over the performance of competitive control.



(*a*) Sinusoidal Perturbations

(*b*) Constant Perturbations

(*c*) Gaussian Perturbations

(*d*) Gaussian Random Walk Perturbations

Figure 1: Relative performance of the linear-quadratic controllers in the double integrator system.

5. Conclusions, Open Problems and Limitations

We have proved that the optimal competitive policy in an LTI dynamical system is well-approximated by the class of Disturbance Action Control (DAC) policies. This implies that the Gradient Pertur-

2. Further experiment details along with more simulations on different systems can be found in the full version of the paper, available at https://arxiv.org/abs/2211.11219.

bation Control (GPC) algorithm and related approaches are able to attain sublinear regret vs. this policy, even when the dynamical system is unknown ahead of time. This is the first time that a control method is shown to attain both sublinear regret vs. a large policy class, and simultaneously a competitive ratio vs. the optimal dynamic policy in hindsight (up to a vanishing additive term). It remains open to extend our results to time varying and nonlinear systems, the recent methods of Minasyan et al. (2021); Gradu et al. (2020) are a potentially good starting point.

Acknowledgements

Elad Hazan gratefully acknowledges funding from NSF grant #1704860.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119, 2019.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. *Advances in Neural Information Processing Systems*, 28, 2015.

Lachlan Andrew, Siddharth Barman, Katrina Ligett, Minghong Lin, Adam Meyerson, Alan Roytman, and Adam Wierman. A tale of two metrics: Simultaneous bounds on competitiveness and regret. In *Conference on Learning Theory*, pages 741–763. PMLR, 2013.

Allan Borodin and Ran El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 2005.

Allan Borodin, Nathan Linial, and Michael E Saks. An optimal on-line algorithm for metrical task system. *Journal of the ACM (JACM)*, 39(4):745–763, 1992.

Niangjun Chen, Gautam Goel, and Adam Wierman. Smoothed online convex optimization in high dimensions via online balanced descent. In *Conference On Learning Theory*, pages 1574–1594. PMLR, 2018.

Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. In *Conference on Learning Theory*, pages 1114–1143. PMLR, 2021.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pages 1029–1038. PMLR, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309, 2019.

Amit Daniely and Yishay Mansour. Competitive ratio vs regret minimization: achieving the best of both worlds. In *Algorithmic Learning Theory*, pages 333–368. PMLR, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*, pages 3211–3221. PMLR, 2020a.

Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In Hal Daumé III and Aarti 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 2020b. URL https://proceedings.mlr.press/v119/foster20b.html.

Gautam Goel and Babak Hassibi. The power of linear controllers in lqr control. *arXiv preprint arXiv:2002.02574*, 2020.

Gautam Goel and Babak Hassibi. Competitive control. *arXiv preprint arXiv:2107.13657*, 2021.

Gautam Goel and Adam Wierman. An online algorithm for smoothed regression and lqr control. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2504–2513. PMLR, 2019.

Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. *arXiv preprint arXiv:2007.04393*, 2020.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, pages 393–400, 2009.

Elad Hazan and Karan Singh. Introduction to online nonstochastic control. *arXiv preprint arXiv:2211.09619*, 2022.

Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, pages 408–421. PMLR, 2020a.

Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In Aryeh Kontorovich and Gergely 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 2020b. URL https://proceedings.mlr.press/v117/hazan20a.html.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164, 2019.

Edgar Minasyan, Paula Gradu, Max Simchowitz, and Elad Hazan. Online control of unknown time-varying dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

Guanya Shi, Yiheng Lin, Soon-Jo Chung, Yisong Yue, and Adam Wierman. Online optimization with memory and competitive control. *Advances in Neural Information Processing Systems*, 33: 20636–20647, 2020.

Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18318–18329. Curran Associates, Inc., 2020.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

Daniel D Sleator and Robert E Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.

Andras Varga. Detection and isolation of actuator/surface faults for a large transport aircraft. In *Fault Tolerant Flight Control*, pages 423–448. Springer, 2010.

Lijun Zhang, Tianbao Yang, Zhi-Hua Zhou, et al. Dynamic regret of strongly adaptive methods. In *International conference on machine learning*, pages 5882–5891. PMLR, 2018.