Online Stackelberg Optimization via Nonlinear Control

William Brown W.BROWN@COLUMBIA.EDU
Columbia University

Christos Papadimitriou CHRISTOS@CS.COLUMBIA.EDU
Columbia University

Tim Roughgarden TIM.ROUGHGARDEN@GMAIL.COM
Columbia University, a16z Crypto

Editors: Shipra Agrawal and Aaron Roth

Abstract

In repeated interaction problems with adaptive agents, our objective often requires anticipating and optimizing over the space of possible agent responses. We show that many problems of this form can be cast as instances of online (nonlinear) control which satisfy *local controllability*, with convex losses over a bounded state space which encodes agent behavior, and we introduce a unified algorithmic framework for tractable regret minimization in such cases. When the instance dynamics are known but otherwise arbitrary, we obtain oracle-efficient $O(\sqrt{T})$ regret by reduction to online convex optimization, which can be made computationally efficient if dynamics are locally *action-linear*. In the presence of adversarial disturbances to the state, we give tight bounds in terms of either the cumulative or per-round disturbance magnitude (for *strongly* or *weakly* locally controllable dynamics, respectively). Additionally, we give sublinear regret results for the cases of unknown locally action-linear dynamics as well as for the bandit feedback setting. Finally, we demonstrate applications of our framework to well-studied problems including performative prediction, recommendations for adaptive agents, adaptive pricing of real-valued goods, and repeated gameplay against no-regret learners, directly yielding extensions beyond prior results in each case.

Keywords: Online convex optimization, online control, Stackelberg games, local controllability

1. Introduction

Machine learning problems involving strategic or adaptive agents are commonly framed as Stackelberg games, wherein the leader aims to commit to an optimal strategy in anticipation of the follower's best response. This approach has been effectively applied to challenges ranging from performative feature manipulation (Hardt et al., 2015; Dong et al., 2018; Perdomo et al., 2020; Jagadeesan et al., 2022b) and optimal pricing (Roth et al., 2015; Daskalakis and Syrgkanis, 2015; Nedelec et al., 2020) to resource allocation in security games (Blum et al., 2014; Balcan et al., 2015; Alcantara-Jiménez and Clempner, 2020) and learning in tabular games (Letchford et al., 2009; Peng et al., 2019; Lauffer et al., 2022; Collina et al., 2023), often with a regret minimization objective. Additionally, several of these settings have been independently extended to account for agents that may update their strategies gradually over time rather than optimally responding in each round (Zrnic et al., 2021a; Brown et al., 2022; Braverman et al., 2017; Deng et al., 2019; Brown et al., 2023). Despite their conceptual similarities, these problems have largely been approached as distinct areas of study, each with their own growing body of techniques. Our aim in this work is to offer a unifying perspective and algorithmic approach for problems of this form, through the lens of online control.

For the broad family of online "Stackelberg-style" optimization problems, the language of control is quite natural to adopt: we are navigating a dynamical system where states corresponding to agent strategies evolve as a function of our own actions, and where objectives which consider best-response stability can be expressed in terms of the stationary behavior of this system. Our results consider a general class of online control instances for representing such problems, which we introduce in Section 2, and in Section 3 we give a sequence of no-regret algorithms for these instances satisfying a range of robustness properties. In Section 4, we show that several online optimization problems involving adaptive agents, including variants of online performative prediction (as in Kumar et al. (2022)), online recommendations (as in Agarwal and Brown (2023)), adaptive pricing (as in Roth et al. (2015)), and learning in time-varying games (as in Anagnostides et al. (2023)) can be embedded in our framework and solved by our algorithms.

While there has been a great deal of recent progress in online linear control, yielding algorithms which can optimize over stabilizing linear policies even with general convex costs, adversarial disturbances, and unknown dynamics (Agarwal et al., 2019a; Simchowitz et al., 2020; Cassel et al., 2022; Minasyan et al., 2022), the required assumptions and regret benchmarks for these algorithms do not always type-check with the settings we are interested in. For the examples we consider, we will often wish to allow for nonlinear dynamics (e.g. encoding an agent's utility function) and explicitly bounded spaces (e.g. via projection into the simplex), and we will seek to compete with regret benchmarks which correspond to stable responses by the agent. Unfortunately, as we show in Proposition 4, the latter goal is incompatible with linear policies even under linear dynamics and in the absence of any disturbances: the performance of *every* linear policy can be $\Omega(T)$ worse than the best policy in the class of affine "state-targeting" policies.

In contrast, the orthogonal set of assumptions we identify enables tractable regret minimization even for *nonlinear* control problems and comports with the requirements of Stackelberg optimization across a wide range of settings, including the ability to compete with state-targeting policies. For convex and compact state and action spaces $\mathcal{X}$ and $\mathcal{Y}$, our first key assumption is that the dynamics $D(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ satisfy a notion of *local controllability*. While local controllability is well-studied for continuous-time and asymptotic control (Aoki, 1974; Kuhn and Wohltmann, 1989; Barbero-Liñán and Jakubczyk, 2013; Boscain et al., 2021), we are unaware of any prior applications to finite-time online optimization, and we adapt existing definitions to be appropriate for this setting. We say that $D(x, y)$ is *strongly* locally controllable if every state in a fixed-radius ball around y is reachable in a single round by an appropriate choice of x, and that $D(x, y)$ is *weakly* locally controllable if the reachable radius around y is allowed to vanish near the boundary of $\mathcal{Y}$. We also assume that our loss f_t in each round is determined (or well-approximated by) an adversarially-chosen convex function depending only on the state y_t.

When these conditions hold, we show in Theorem 5 that this is sufficient to obtain $O(\sqrt{T})$ regret with respect to the loss of the best fixed state, provided that dynamics are known and we have offline access to an oracle for non-convex optimization; the oracle call can be removed if dynamics are locally *action-linear*, i.e. given by (or locally well-approximated by) a function linear in x at each fixed y. If adversarial disturbances to the dynamics are present, our approach can be extended for both weakly (Theorem 8) and strongly (Theorem 9) locally controllable dynamics with additional regret scaling linearly in total disturbance magnitude, provided that each round's disturbance cannot be too large in the case of weak local controllability; we give lower bounds showing that each dependence on disturbance magnitude is tight. The aforementioned results all extend to the case where the dynamics (absent disturbances) are given by a known but time-dependent function

$D_t(x, y)$. If dynamics are unknown but time-invariant, and locally action-linear with appropriate regularity parameters, we obtain sublinear regret provided that a "near-stabilizing" action is known at $t = 1$. We additionally extend our approach to the bandit feedback setting, where we obtain $O(T^{3/4})$ regret. In Section 4 we show that each of the following, with appropriate assumptions, can be cast as a locally controllable instance with state-only convex surrogate losses:

- **Performative prediction:** Minimize prediction loss $\mathbb{E}_{z\sim p_t} f_t(x_t, z)$ for a classifier x_t, where the distribution p_t in each round is updated according to the prior classifier and distribution.
- **Adaptive recommendations:** Maximize the reward $f_t(i_t)$ when showing menus $K_t \subseteq [n]$ of size $k \ll n$ to an agent, whose choice $i_t \sim p(K_t, v_t)$ in each round depends on preferences which are influenced by choices in prior rounds (encoded in the "memory vector" v_t).
- **Adaptive pricing:** Maximize profit $\langle p_t, x_t\rangle$ - $c_t(x_t)$ for selling bundles of goods x_t to an agent at prices p_t and with costs c_t, where the agent's purchased bundle x_t is a function of their utility function, consumption rate, and existing reserves.
- **Repeated gameplay:** Maximize the reward $x_t^\top A_t y_t$ obtained from playing a sequence of time-varying games (A_t, B_t) against a no-regret learning agent.

In each case, application of our algorithms from Section 3 yields results which extend beyond the applicability regimes of prior work, such as by enabling relaxation of previous assumptions or a novel extension to adversarial or dynamic problem variants.

1.1. Related Work

Online control. Much of the recent progress in online control (Agarwal et al., 2019a,b; Cassel et al., 2022; Minasyan et al., 2022) considers linear systems with general convex losses, benchmarking against a class of ("strongly stable") fast-mixing linear policies introduced for linear-quadratic control (Cohen et al., 2018) by leveraging the framework of "OCO with memory" (Anava et al., 2014). Results have also been shown for nonlinear policy classes via neural networks (Chen et al., 2022), and for nonlinear dynamics with oracles in episodic settings (Kakade et al., 2020), via approximation with random Fourier features (Lale et al., 2021; Luo et al., 2022), via adaptive regret for time-varying linear systems (Gradu et al., 2022; Minasyan et al., 2022), and via dynamic regret over actions in terms of disturbance "attenuation" (Muthirayan and Khargonekar, 2022). For a further overview of online control and its historical context, see Hazan and Singh (2022). In contrast to the bulk of prior work in which states and actions are bounded implicitly via policy stability notions, we consider state and action spaces which are bounded explicitly, as enabled by nonlinearity in dynamics (e.g. via projection, or range decay of dynamics near the boundary). These works also view disturbances as intrinsic to the system, and account for their influence directly in regret benchmarks (the "optimal policy" will face the same sequence of disturbances in hindsight, regardless of state). Within the context of Stackelberg optimization where a fixed protocol largely determines an agent's strategy updates, we view the role of disturbances as more akin to adversarial *corruptions* as considered in reinforcement learning (Lykouris et al., 2021; Zhang et al., 2021); while we incur linear dependence, our regret benchmarks are agnostic to alternate counterfactual disturbance sequences.

Strategizing against learners. Initially formulated within the context of repeated auctions (Braverman et al., 2017), a recent line of work has considered the problem of optimizing long-run rewards in a repeated game against a no-regret learner across a range of tabular and Bayesian settings (Deng et al., 2019; Mansour et al., 2022; Brown et al., 2023; Zhang et al., 2023). While bounds on attainable reward have been known in terms of the Price of Anarchy (Blum et al., 2008; Hartline et al., 2015b), this sequence of results has highlighted important connections with Stackelberg equilibria: the Stackelberg value of the game is attainable on average against any no-regret learner, and it is the maximum attainable value against many common no-regret algorithms (such as no-swap learners, as shown by Deng et al. (2019)). This theme has emerged in other simultaneous learning settings as well; notably, Zrnic et al. (2021b) show that long-run outcomes in strategic classification are shaped by relative learning rates between parties, which can designate either as the Stackelberg leader.

Nested convex optimization. The technique of identifying convex structure nested inside a more general problem has been applied broadly across a range of online optimization settings (Neu and Olkhovskaya, 2021; Shen et al., 2023; Flokas et al., 2019). For repeated interaction problems involving an agent with unknown utility, such as optimal pricing, Roth et al. (2015) identify utility conditions under which the non-convex objective over prices becomes convex in the space of agent actions, and where explorability properties resembling local controllability hold, which enables convex optimization by locally learning agent preferences; this "revealed preferences" approach has also been applied to strategic classification (Dong et al., 2018). In recent work concerning recommendations for agents with history-dependent preferences (Agarwal and Brown, 2022, 2023), properties related to local controllability are leveraged to enable tractable optimization as well. We consider each of these settings as applications in Section 4.

2. Model and Preliminaries

Let $\mathcal{X}$ and $\mathcal{Y}$ be convex and compact subsets of Euclidean space, respectively denoting the action and state spaces, where we assume $\dim(\mathcal{X}) \geq \dim(\mathcal{Y})$. Further, we assume that $\mathcal{Y}$ contains a ball of radius r around the origin $\mathbf{0}$, and is contained in a ball of radius R around the origin.

An instance of our control problem consists of choosing a sequence of actions $\{x_t \in \mathcal{X}\}$ over T rounds, which will yield a sequence of states $\{y_t \in \mathcal{Y}\}$, and we will incur losses determined by adversarially chosen functions $\{f_t\}$. Let the initial state be $y_0 = \mathbf{0}$. In the basic version of our problem, upon choosing each x_t for rounds $t \in [T]$, we observe the state update to

$$y_t = D(x_t, y_{t-1}),$$

where $D : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ is an arbitrary continuous function which we refer to as the *dynamics* of our problem. We sometimes allow *disturbances* to the dynamics, where $y_t = D(x_t, y_{t-1}) + w_{t+1}$ for $\{w_t\}$ chosen adversarially. In some cases we allow *time-varying* dynamics $D : \mathcal{X} \times \mathcal{Y} \times [T] \to \mathcal{Y}$, where the dynamics in each round are denoted by $D_t(x_t, y_{t-1})$.

Here and in Section 3, we assume that our loss in round is given by $f_t(y_t)$, where each f_t is a L-Lipschitz convex function revealed after playing x_t; we relax these assumptions for some of our applications in Section 4, e.g. to allow dependence on x_t as well. We generally measure will performance with respect to the best fixed state, and the regret for an algorithm $\mathcal{A}$ yielding $\{y_t\}$ is

$$\mathrm{Reg}_T(\mathcal{A}) = \sum_{t=1}^{T} f_t(y_t) - \min_{y \in \mathcal{Y}} \sum_{t=1}^{T} f_t(y).$$

In Proposition 4, we relate this benchmark to the class of "state-targeting" policies, which can sometimes be expressed by affine functions, and we compare their performance to linear policies. Throughout, we use $\|\cdot\|$ to donate the Euclidean norm, and we let $\mathcal{B}_\epsilon(y) = \{\hat{y} : \|y - \hat{y}\| \leq \epsilon\}$ denote the norm ball of radius ϵ around y. We let $\Pi_{\mathcal{Y}}(\cdot)$ denote Euclidean projection into the set $\mathcal{Y}$; $\mathbf{u}_n$ denotes the uniform distribution over n items, and $\Delta(n)$ denotes the probability simplex.

2.1. Locally Controllable Dynamics

A number of properties under the name "local controllability" have been considered for various continuous-time and asymptotic control settings (Aoki, 1974; Kuhn and Wohltmann, 1989; Barbero-Liñán and Jakubczyk, 2013; Boscain et al., 2021), generally relating to the notion that all states in a neighborhood around a given state are reachable. We give two formulations of local controllability for our setting, which we take as properties of the dynamics D holding over all inputs.

Definition 1 (Weak Local Controllability) *For $\rho \in (0, 1]$, an instance $(\mathcal{X}, \mathcal{Y}, D)$ satisfies (weak) ρ-local controllability if for any $y \in \mathcal{Y}$ and $y^* \in \mathcal{B}_{\rho \cdot \pi(y)}(y)$, there is some x such that $D(x, y) = y^*$, where $\pi(y) = \min_{\hat{y} \in \mathrm{bd}(\mathcal{Y})} \|\hat{y} - y\|$ is the distance from y to the boundary of $\mathcal{Y}$.*

Definition 2 (Strong Local Controllability) *For $\rho > 0$, an instance $(\mathcal{X}, \mathcal{Y}, D)$ satisfies strong ρ-local controllability if for any $y \in \mathcal{Y}$ and $y^* \in \mathcal{B}_\rho(y) \cap \mathcal{Y}$, there is some x such that $D(x, y) = y^*$.*

We often refer to weak local controllability simply as local controllability. This property ensures that there is always some action x_t which results in the next state y_t staying fixed at y_{t-1}, as well as some action which moves the state to any point in a surrounding ball; in the weak case, the size of the reachable ball is allowed to decay as y_t approaches the boundary of $\mathcal{Y}$. The parameter ρ controls the speed at which we can navigate the state space: when $\rho = 1$ in the weak case (or $\rho \geq R$ in the strong case), we can always immediately reach some point on the boundary of $\mathcal{Y}$, yet for ρ close to zero we may only be able to move in a small neighborhood. Our results use local controllability to minimize regret over $\mathcal{Y}$ by reduction to online convex optimization. As we prove in Appendix A, up to a quantifier alternation which vanishes as ρ approaches 0, a property of this form is essentially necessary: competing with the best state y is impossible if we cannot remain in its neighborhood.

Proposition 3 *Suppose there is some $y \in \mathcal{Y}$ and values $\alpha, \beta > 0$ such that for all $\hat{y} \in \mathcal{B}_\alpha(y)$ and $x \in \mathcal{X}$, $D(x, \hat{y}) \notin \mathcal{B}_\beta(\hat{y})$. Then, there are losses such that $\mathrm{Reg}_T(\mathcal{A}) = \Omega(T)$ for any algorithm $\mathcal{A}$.*

2.2. States vs. Policies

While regret benchmarks in online control are typically expressed in terms of a reference class of policies, we note that there is a class of "state-targeting" policies which track the reward of fixed states (asymptotically, and up to the influence of disturbances), and which can be implemented if D is known; we maintain the formulation in terms of fixed states for clarity with respect to our motivations for Stackelberg optimization. Existing no-regret algorithms for online control typically compete with linear policies, and choose actions each round by implementing policies which are linear in multiple past states (as in e.g. Agarwal et al. (2019a)). Here, we show that all such policies can be arbitrarily suboptimal when compared to state-targeting policies, even for dynamics which are linear up to projection and with fixed convex losses over states, as they may yield actions and states which remain fixed at $\mathbf{0}$ in every round even if the optimal state is always immediately accessible under the dynamics. We prove Proposition 4 in Appendix A.

Proposition 4 *For an instance $(\mathcal{X}, \mathcal{Y}, D)$, let the class of state-targeting policies for $\hat{\mathcal{Y}} \subseteq \mathcal{Y}$ be given by $\mathcal{P}_{\hat{\mathcal{Y}}} = \{P_{\hat{y}} : \hat{y} \in \hat{\mathcal{Y}}\}$ where $P_{\hat{y}}(y) = \operatorname{argmin}_{\{x \in \mathcal{X} : D(x,y) \in \hat{\mathcal{Y}}\}} \|D(x,y) - \hat{y}\|^2$. Define the regret of a policy class $\mathcal{P}$ as*

$$\mathrm{Reg}_T(\mathcal{P}) = \min_{P \in \mathcal{P}} \left(\sum_{t=1}^{T} f_t(y_t) \right) - \min_{y \in \mathcal{Y}} \left(\sum_{t=1}^{T} f_t(y) \right),$$

where y_t is updated by playing P at each round. For any ρ-locally controllable instance, there is a set $\hat{\mathcal{Y}} \subseteq \mathcal{Y}$ for which $\mathrm{Reg}_T(\mathcal{P}_{\hat{\mathcal{Y}}}) = O(\sqrt{T\rho^{-1}})$. Further, for any class $\mathcal{P}_{\mathcal{K}}$ where each $K \in \mathcal{P}_{\mathcal{K}}$ is a matrix yielding actions $x_t = -Ky_{t-1}$, there is an instance where $\mathrm{Reg}_T(\mathcal{P}_{\mathcal{K}}) \geq \Omega(T)$ for $\rho = 1$.

If dynamics are linear up to projection with $D(x_t, y_{t-1}) = \Pi_{\mathcal{Y}}(By + Ax)$ for full-rank A, and $\dim(\mathcal{X}) = \dim(\mathcal{Y})$, note that $P_{\hat{y}}(y) = A^{-1}(\hat{y} - By)$ implements any $P_{\hat{y}}$ for sufficiently large $\mathcal{X}$.

3. No-Regret Algorithms for Locally Controllable Dynamics

Here we give a sequence of no-regret algorithms satisfying a range of robustness properties. Our primary algorithm NESTEDOCO, presented in Section 3.1, operates over known time-varying dynamics without disturbances and requires an offline non-convex optimization oracle, and we identify conditions in Section 3.2 which remove the oracle requirement. In Section 3.3 we give two algorithms, NESTEDOCO-BD and NESTEDOCO-UD, which allow adversarial disturbances to weakly and strongly locally controllable dynamics, respectively. In Section 3.4 we extend NESTEDOCO to accommodate unknown dynamics under appropriate regularity conditions (provided an initial "approximately stabilizing" action is known at $t = 1$), and in Section 3.5 we give an algorithm which obtains $O(T^{3/4})$ regret under bandit feedback.

3.1. Nonlinear Control via Online Convex Optimization

When dynamics satisfy local controllability and y_{t-1} is not too close to $\mathrm{bd}(\mathcal{Y})$, all points y_t in a ball around y_{t-1} are feasible with an appropriate x_t; this enables execution of an online convex optimization (OCO) algorithm over $\mathcal{Y}$ by playing the action x_t which yields a state update to the target y_t chosen at each iteration, computed via offline non-convex optimization. Here we assume that D is known and can be queried for any inputs, and that disturbances to the state are not present. We allow the dynamics to change over time, potentially as a function of previous actions x_s and losses f_s for $s < t$, provided that D_t can be determined in each round. We use Follow the Regularized Leader (FTRL) as our OCO subroutine (Shalev-Shwartz and Singer, 2006; Abernethy et al., 2008), yet we note that it may be substituted for any OCO algorithm whose per-round step size is guaranteed to be sufficiently small (such as OGD with a constant learning rate); statements of the FTRL algorithm and its key properties are provided in Appendix B. We instantiate FTRL over a contracted space $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$, calibrated to ensure that the minimum loss over $\tilde{\mathcal{Y}}$ is close to that for $\mathcal{Y}$, yet where each step of FTRL lies within the feasible region ensured by (weak) local controllability.

Theorem 5 *For a ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ without disturbances and with D_t known at each t, the regret of* NESTEDOCO *for convex L-Lipschitz losses $f_t : \mathcal{Y} \to \mathbb{R}$ is at most*

$$\mathrm{Reg}_T(\text{NESTEDOCO}) \leq 2L\sqrt{(1 + R(r\rho)^{-1})TG\gamma^{-1}}$$

with respect to any state $y^* \in \mathcal{Y}$, with T queries made to a non-convex optimization oracle.

The proof for Theorem 5 is given in Appendix C.

Algorithm 1 Nested Online Convex Optimization (NESTEDOCO).

Let $\psi : \mathcal{Y} \to \mathbb{R}$ be γ-strongly convex with $\operatorname{argmin}_y \psi(y) = \mathbf{0}$ and $\max_{y,y'} |\psi(y) - \psi(y')| \le G$
Let $\eta = (G\gamma)^{1/2}((1 + \frac{R}{r\rho})TL^2)^{-1/2}$
Let $\widetilde{\mathcal{Y}} = \{y : \frac{1}{1-\delta} y \in \mathcal{Y}\}$ for $\delta = \eta \frac{L}{r\rho\gamma}$
Initialize FTRL to run for T rounds over $\widetilde{\mathcal{Y}}$ with regularizer ψ and parameter η
for $t = 1$ to T **do**
  Let y^* be the point chosen by FTRL
  Use $\texttt{Oracle}(y_{t-1}, y^*)$ to compute $x_t = \operatorname{argmin}_x \|D_t(x, y_{t-1}) - y^*\|^2$
  Play action x_t
  Observe y_t and loss $f_t(y_t)$, update FTRL
end for

3.2. Efficient Updates for Action-Linear Dynamics

While NESTEDOCO requires no assumptions on the dynamics beyond local controllability, there are large classes of dynamics for which the oracle call can be removed. We say that dynamics are *action-linear* if $y_x = D(x, y)$ is linear in x, for $y_x \in \operatorname{int}(\mathcal{Y})$ (and arbitrary for $y_x \in \operatorname{bd}(\mathcal{Y})$).

Proposition 6 *For a ρ-locally controllable and action-linear instance $(\mathcal{X}, \mathcal{Y}, D)$, the per-round optimization problem for $\texttt{Oracle}(y_{t-1}, y^*)$ in* NESTEDOCO *is convex.*

Proof For $y = y_{t-1} \in \widetilde{\mathcal{Y}} \subseteq \operatorname{int}(\mathcal{Y})$, we have $D(x, y) = A_y \cdot x + b_y$ for some matrix A_y and vector b_y, and so we can solve $x_t = \operatorname{argmin}_{x \in \mathcal{X}} \|A_y \cdot x + b_y - y^*\|^2$ efficiently. ■

The class of action-linear dynamics is quite general, owing to the flexibility permitted by nonlinear parameterizations of (A_y, b_y) in terms of y; in Appendix D, we show that local controllability holds for multiple explicit families of instances when appropriate eigenvalue conditions are satisfied. We can further relax this condition to accommodate dynamics where action-linearity holds only *locally* in the neighborhood of stabilizing actions (i.e. actions x^* where $D(x^*, y) = y$).

Definition 7 (Locally Action-Linear Dynamics) *An instance $(D, \mathcal{X}, \mathcal{Y})$ is locally action-linear if, for any $y \in \operatorname{int}(\mathcal{Y})$, x^* such that $D(x^*, y) = y$, and x such that $D(x, y) \in \operatorname{int}(\mathcal{Y})$, the dynamics are given by $D(x, y) = A_y x + b_y + q_y(x)$, where A_y is a matrix and b_y is a vector, both with norms bounded by some absolute constant, where and $q_y : \mathcal{X} \to \mathbb{R}^{\dim(\mathcal{Y})}$ is any function where $\|q_y(x)\| \le C \|A_y(x - x^*)\|^{1+c}$ for some constants $C, c > 0$.*

By this condition, for any x in a sufficiently small neighborhood around x^*, the deviation of dynamics (and thus the resulting y_{t+1}) from action-linearity vanishes. Note that our algorithm always chooses a target y_t will always be near y_{t-1}; as such, these deviations from non-action-linearity can be modeled as *disturbances* with magnitude strictly less than our per-round step size $\|y_{t+1} - y_t\|$ (along with universal constant factors). The existence of an efficient implementation follows as a straightforward corollary of Theorem 8 in Section 3.3, which extends NESTEDOCO to accommodate bounded adversarial disturbances, as we can then select actions by disregarding the influence of q_y and only considering the local approximation $D(x, y) = A_y x + b_y$ at each state y (assuming that each decomposition between q_y and the action-linear component is known).

3.3. Adversarial Disturbances

Our algorithm NESTEDOCO can be extended to accommodate adversarial disturbances, where the state is updated as $y_t = D(x_t, y_{t-1}) + w_t$, with $\{w_t\}$ chosen adversarially. In the weak local controllability case, we show a sharp threshold effect in terms of whether or not $\|w_t\|$ is allowed to exceed the undisturbed distance from the boundary by a factor of $\frac{\rho}{1+\rho}$: if disturbances are bounded below this threshold, regret minimization remains feasible with a tight $\Theta(E)$ dependence on the total disturbance magnitude, yet if disturbances may exceed this, no sublinear regret rate is attainable even for a *constant* total disturbance magnitude. When ρ is small, an adversary can push us to the boundary faster than we can "undo" past disturbances, causing our feasible range to decay.

Theorem 8 (Bounded Disturbances for Weak Local Controllability) *For any $\rho \in (0, 1]$, suppose that a sequence of adversarial disturbances w_t for a ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ satisfies $\sum_{t=1}^{T} \|w_t\| \leq E$ and $\|w_t\| \leq \frac{\rho-\alpha\rho}{1+\rho} \cdot \pi\left(D(x_t, y_{t-1})\right)$, for some $\alpha \in \mathbb{R}$. If $\alpha > 0$, there is an algorithm* NESTEDOCO-BD *with regret for convex Lipschitz losses f_t bounded by*

$$\mathrm{Reg}_T(\textsc{NestedOCO-BD}) \leq O\left(\sqrt{T \cdot (\alpha\rho)^{-1}} + E\right),$$

and there is an instance where any algorithm $\mathcal{A}$ obtains $\mathrm{Reg}_T(\mathcal{A}) = \Omega(E)$. If $\alpha < 0$, there is an instance such that any algorithm $\mathcal{A}$ obtains $\mathrm{Reg}_T(\mathcal{A}) \geq \Omega(T)$ even when $E = O(1)$.

The maximum disturbance bound can be removed when dynamics are strongly locally controllable, as the ensured feasible range of the dynamics does not vanish at the boundary of the state space. For such instances, we can minimize regret (with tight $O(E \cdot \rho^{-1})$ dependence) even if disturbances are only implicitly bounded by the state space diameter (which is at least ρ, without loss of generality).

Theorem 9 (Unbounded Disturbances for Strong Local Controllability) *For any $\rho > 0$ and strongly ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ with disturbances w_t satisfying $\sum_{t=1}^{T} \|w_t\| \leq E$, there is an algorithm* NESTEDOCO-UD *with regret for convex Lipschitz losses f_t bounded by*

$$\mathrm{Reg}_T(\textsc{NestedOCO-UD}) \leq O\left(\sqrt{T} + E \cdot \rho^{-1}\right),$$

and there is an instance where any algorithm $\mathcal{A}$ obtains $\mathrm{Reg}_T(\mathcal{A}) \geq \Omega\left(E \cdot \rho^{-1}\right)$.

In each case, our lower bounds in terms of E hold for the same constants obtained by our algorithms, and our algorithms obtain the stated regret guarantees even when E is not known in advance. We present the algorithms and analysis for each theorem in Appendix E; both operate by tracking deviations from an idealized trajectory without disturbances, and calibrating parameters to preserve sufficient reachability margin for applying corrections towards this trajectory in each round. The lower bounds both proceed by considering an instance with a fixed target state y^* and losses which track the distance from y^*, along with an adversary whose goal is to maximize this distance by selecting disturbances which push the current state away from y^*.

3.4. Unknown Dynamics

Up until this point, we have assumed that the dynamics D can be queried arbitrarily in each round. While this has required minimal assumptions on D beyond local controllability, accommodation of

unknown dynamics is often desired in online control (Cassel et al., 2022; Minasyan et al., 2022) and for several of our applications (Roth et al., 2015; Agarwal and Brown, 2023). Here we give conditions under which regret minimization can be implemented without advance knowledge of D by an algorithm PROBINGOCO, which maintains continuously-updating local linear approximations of D near y_t across rounds. Crucially, we assume that D is time-invariant and locally action-linear with sufficiently small Lipschitz parameters, and that for the initial state y_0 some *near-stabilizing* action x_1 is known, i.e. $\|D(x_1, y_0) - y_0\| \leq \epsilon$, for some $\epsilon = o(\sqrt{T})$.

Theorem 10 *For any ρ-locally controllable and time-invariant instance $(D, \mathcal{X}, \mathcal{Y})$ which satisfies local action-linearity and appropriate Lipschitz conditions, there is an algorithm* PROBINGOCO *with* $\mathrm{Reg}_T(\text{PROBINGOCO}) \leq O(\sqrt{T})$ *for convex Lipschitz losses f_t and unknown dynamics D, provided that at $t = 1$ we are given some x_1 such that $\|D(x_1, y_0) - y_0\| = o(\sqrt{T})$.*

We state PROBINGOCO and prove Theorem 10 in Appendix F, along with additional details on the regularity and near-stability assumptions. The crux of our analysis, beyond that from our previous results, hinges on being able to maintain and update local linear approximations of D throughout our optimization which are sufficiently accurate to allow us to discard the effects of both learned representation errors and action non-linearity from $q_y(x)$ as bounded disturbances. We implement each update from our nested regret minimization algorithm as a series of $O(\dim(\mathcal{X}))$ steps involving small near-orthogonal perturbations to our targets y_t, which we then use to update our local estimate for D.

3.5. Bandit Feedback

We can extend our approach from NESTEDOCO to accommodate bandit feedback for convex losses by replacing FTRL with the FKM algorithm (Flaxman et al., 2004) and appropriately recalibrating parameters. FKM obtains $O(T^{3/4})$ regret, which is the best currently-known bound for bandit convex optimization without additional assumptions (e.g. strong convexity), and we obtain an analogous bound here for nested optimization. We note that this extension to bandit feedback can again be applied for any algorithm with a small per-round step-size bound, though this property does not hold for algorithms which sample from larger sets to reduce variance of gradient estimators (e.g. those from Abernethy et al. (2008); Hazan and Levy (2014)).

Theorem 11 *For any ρ-locally controllable instance $(D, \mathcal{X}, \mathcal{Y})$, there is an oracle-efficient algorithm* NESTEDBCO *with expected regret bounded by*

$$\mathrm{Reg}_T(\text{NESTEDBCO}) = O\left(nRLT^{3/4}(r\rho)^{-1}\right)$$

for L-Lipschitz convex losses f_t under bandit feedback.

We present the NESTEDBCO algorithm and prove Theorem 11 in Appendix G.

4. Applications for Online Stackelberg Optimization

We give several applications of our framework to online Stackelberg problems involving strategic or adaptive agents, each cast as an instance of online control with nonlinear dynamics where local controllability holds, and where our objectives are well-approximated by convex surrogate losses only

over the state. Each application extends prior work by either allowing for more relaxed assumptions, unifying distinct problem instances, or giving a novel formulation to account for dynamic and adversarial behavior; analysis and comparison to related work is contained in Appendices H-K.

4.1. Online Performative Prediction

Performative Prediction was introduced by Perdomo et al. (2020) to capture settings in which the data distribution may shift as a function of the classifier itself. We consider the online formulation of Performative Prediction introduced in Kumar et al. (2022) as an instance of online convex optimization with unbounded memory, which we extend to accommodate a *stateful* variant of the problem (as in Brown et al. (2022)) in which the update to the distribution is a function of both the classifier and the current distribution itself. Let $\mathcal{X} \subseteq \mathbb{R}^n$ denote our space of classifiers, and let p_0 be the initial distribution over $\mathbb{R}^n$. When a classifier x_t is deployed, the distribution is updated to

$$p_t = (1-\theta)p_{t-1} + \theta\, \mathcal{D}(x_t, y_{t-1})$$

where $\mathcal{D}(x_t, y) = A(x_t, y_{t-1}) + \xi$, for a random variable $\xi \in \mathbb{R}^n$ with mean μ and covariance Σ, and with $y_t = A(x_t, y_{t-1})$, where A satisfies ρ-local controllability for some $\rho > 0$ and appropriate smoothness notions. We also assume there is some linear $s : \mathcal{X} \to \mathcal{Y}$ such that $A(x, y) = s(x)$ if $y = s(x)$. We then receive loss $\tilde{f}_t(x_t, p_t) = \mathbb{E}_{z\sim p_t}[f_t(x_t, z)]$, where each f_t is convex and Lipschitz.

This generalizes the model of Kumar et al. (2022), in which $A(x, y) = A \in \mathbb{R}^{n\times n}$ is taken to be a fixed matrix; there, ρ-local controllability is satisfied for some $\rho > 0$ provided that A is nonsingular. Their aim is to compete with the best fixed classifier by running regret minimization over $\mathcal{X}$. Here we run NESTEDOCO over $\mathcal{Y}$, taken over the range of s, which allows us to compete against the best fixed classifier as well by the properties of s; while the classifiers x_t we play will generally not result in stabilizing points of A, their excess loss compared to each $s^{-1}(y_t)$ is bounded.

Theorem 12 (Regret Minimization for Performative Prediction) *For any $\theta > 0$, the dynamics for Online Performative Prediction are ρ-locally controllable, and* NESTEDOCO *obtains regret $O(\sqrt{T(\rho^{-1} + \theta^{-1})})$ with respect to the best fixed classifier.*

4.2. Adaptive Recommendations

Online interactions with economic agents of various types are ubiquitous, and the resulting control problems tend to be manifestly nonlinear; here we treat two diverse examples from this space. The Adaptive Recommendations problem, as introduced by Agarwal and Brown (2022), is about providing menu recommendations repeatedly to an agent, whose choice distribution is a function of their past selections, while the controller's reward in each round depends on adversarial losses over the choice. In each round $t \in [T]$, we show the agent a (possibly randomized) menu K_t containing k (out of n) items, and the agent's instantaneous choice distribution conditioned on seeing K_t is

$$p_t(i; K_t, v_{t-1}) = \begin{cases} \frac{s_i(v_{t-1})}{\sum_{j\in K_t} s_j(v_{t-1})} & i \in K_t \\ 0 & i \notin K_t \end{cases}$$

where each $s_i : \Delta(n) \to [\lambda, 1]$ is the agent's *preference scoring function* for item i, for some $\lambda > 0$, taking as input the agent's *memory vector* $v \in \Delta(n)$. The memory vector updates each round as

$$v_t = (1-\theta_t)v_{t-1} + \theta_t p_t,$$

where $\theta_t \in [\theta, 1]$ for $\theta > 0$ is a possibly time-dependent update speed, and we receive loss $f_t(p_t)$, where each f_t is convex and L-Lipschitz. Note that the set of feasible choice distributions when considering all menu distributions $x_t \in \Delta(\binom{n}{k})$ depends on the memory vector v_t. The regret benchmark considered by Agarwal and Brown (2022) is the intersection of all such sets, denoted the "everywhere instantaneously-realizable distribution" set $\text{EIRD} = \cap_{v \in \Delta} \text{IRD}(v)$, where $\text{IRD}(v)$ is the "instantaneously realizable distribution" set for v, given as the convex hull of the choice distributions $p(K_t)$ resulting from each menu $K_t \in [\binom{n}{k}]$ when v is the memory vector. It is shown that the set is non-empty when λ is not too small, and algorithms which minimize regret with respect to any distribution in EIRD are given in Agarwal and Brown (2022) and Agarwal and Brown (2023) under varying assumptions regarding the scoring functions and update speed.

While the prior work considers a bandit version of the problem with unknown dynamics, here we consider a full-feedback deterministic variant of the problem for simplicity, which further allows us to circumvent barriers posed by uncertainty Agarwal and Brown (2022, 2023) and relax structural assumptions (e.g. on θ_t or s_i). We can cast this as an instance of our framework by taking $\mathcal{X} = \Delta(\binom{n}{k})$ and $\mathcal{Y} = \text{EIRD}$, where D expresses updates to the memory vector. We assume $v_0 = \mathbf{u}_n$, and we reparameterize to run our algorithm over $\Delta(n)$. We optimize surrogate losses $f_t^*(v_t)$, and bound excess regret from $f_t(p_t)$.

Theorem 13 (Regret Minimization over EIRD) *For $\lambda > \frac{k-1}{n-1}$, the dynamics for Adaptive Recommendations over* EIRD *are θ-locally controllable, and* NESTEDOCO *obtains regret $O(\sqrt{T\theta^{-1}})$.*

In Agarwal and Brown (2023), a property for scoring functions is considered which enables regret minimization over a potentially much larger set of distributions than EIRD. A scoring function $s_i : \Delta(n) \to [\frac{\lambda}{\sigma}, 1]$ is said to be (σ, λ)-*scale-bounded* for $\sigma > 1$ if, for all $v \in \Delta(n)$, we have that

$$\sigma^{-1}((1-\lambda)v_i + \lambda) \leq s_i(v) \leq \sigma((1-\lambda)v_i + \lambda).$$

The set considered is the ϕ-*smoothed simplex* $\Delta^\phi(n) = \{(1-\phi)v + \phi\mathbf{u}_n : v \in \Delta(n)\}$, for $\phi = \Theta(k\lambda\sigma^2)$, where it is shown that $\text{IRD}(v)$ contains a ball around v for $v \in \Delta^\phi(n)$. We take $\mathcal{Y} = \Delta^\phi(n)$, which satisfies local controllability, and optimize over $f_t^*(v_t)$ with NESTEDOCO.

Theorem 14 (Regret Minimization over $\Delta^\phi(n)$) *For (σ, λ)-scale-bounded scoring functions s_i, for any $\lambda > 0$ and $\sigma > 1$, the dynamics for Adaptive Recommendations over $\Delta^\phi(n)$ are $\Omega(\theta\lambda\phi)$-locally controllable, and* NESTEDOCO *obtains regret $O(\sqrt{T(\theta\lambda\phi)^{-1}})$.*

4.3. Adaptive Pricing

Here we consider an Adaptive Pricing problem for real-valued goods, formulated as a dynamic extension of the setting of Roth et al. (2015) where purchase history and consumption affect demand. In each round we set per-unit price vectors $p_t \in \mathbb{R}^n_+$, and an agent buys some bundle of goods $x_t \in \mathbb{R}^n_+$, which results in us obtaining a reward $\langle p_t, x_t \rangle - c_t(x_t)$, where our production cost function c_t at each round is convex and L_c-Lipschitz, and may be chosen adversarially.

Departing from Roth et al. (2015), we consider an agent who maintains goods reserves $y_{t-1} \in \mathbb{R}^n_{\geq 0}$ and consumes an adversarially chosen fraction $\theta_t \in [\theta, 1]$ of every good's reserve at each round (for some $\theta > 0$). The agent then chooses a bundle x_t to maximize their utility $g(p_t, x_t, y_t) = v(y_t) - \langle p_t, x_t \rangle$, where $y_t = (1-\theta_t)y_{t-1} + x_t$ is their updated reserve bundle. We make several

regularity assumptions on the agent's valuation function $v : \mathbb{R}_+^n \to \mathbb{R}_+$, all of which are satisfied by several classically studied utility families (which we discuss in Appendix 4.3). Notably, we assume that v is strictly concave and increasing, and homogeneous; the range is bounded under rationality.

Our aim will be to set prices which allow us to compete with the best *stable reserve policy*, e.g. against any pricing policy where the agent maintains the same reserve bundle $y_t = y^*$ at each round for some y^* regardless of θ_t. We take an appropriate convex set of such bundles as our state space, for which we show that local controllability holds. Observe that to induce a purchase of $x_t = \theta_t y_{t-1}$, it suffices to set prices $p_t = \nabla v(y_{t-1})$, as we then have that $\nabla_{x_t}(v((1-\theta_t)y_{t-1}+x_t)-\langle p_t, x_t\rangle) = \mathbf{0}$. By homogeneity of v, we also have that $\langle \nabla v(y_t), \theta_t y_t\rangle = \theta_t k \cdot v(y_t)$ for some k, and we show that optimization via the concave surrogate rewards

$$f_t^*(y_t) = \theta_t k \cdot v(y_t) - c_t(\theta_t y_t)$$

will closely track our true rewards $f_t(p_t, x_t) = \langle p_t, x_t\rangle - c_t(x_t)$. While neither our true nor surrogate rewards will be Lipschitz, we extend NESTEDOCO to obtain sublinear regret over Hölder continuous losses by appropriately calibrating our step size (which may be of independent interest).

Theorem 15 (Regret Minimization over Stable Reserve Policies) *For any $\theta > 0$, the dynamics for Adaptive Pricing can are θ-locally controllable, and* NESTEDOCO *obtains regret $o(T\theta^{-1})$ with respect to the best stable reserve policy.*

4.4. Steering Learners in Online Games

A recent line of work (Deng et al., 2019; Mansour et al., 2022; Brown et al., 2023) explores maximizing rewards in a repeated game against a no-regret learner, and Anagnostides et al. (2023) study of no-regret dynamics in time-varying games. We consider these questions in unison, and aim to optimize reward against a no-regret learner for game matrices chosen adversarially and online.

Consider adversarial sequences of two-player $m \times n$ bimatrix games (A_t, B_t), where $m > n$; we assume that the convex hull of the rows of each B_t contains the unit ball. As Player A, we choose strategies $x_t \in \Delta(m)$ each round to maximize our reward against Player B, who chooses their strategies $y_t \in \Delta(n)$ according to a no-regret algorithm (in particular, online projected gradient descent). The game (A_t, B_t) is only revealed after both players have chosen strategies for round t. Our aim here is to illustrate the feasibility of *steering* the opponent's trajectory, and so we consider games where Player A's reward is predominantly a function only of Player B's actions. We assume that $\|xA_t - xA_t^*\| \leq \delta_t$ for any $x \in \Delta(m)$, where each A_t^* is a matrix with identical rows, and that per-round changes to B_t are bounded, with $\|xB_t - xB_{t-1}\| \leq \epsilon_t$ for any $x \in \Delta(m)$. We measure the regret of an algorithm $\mathcal{A}$ with respect to *any* profile $(x, y) \in \Delta(m) \times \Delta(n)$, where

$$\text{Reg}_T(\mathcal{A}) = \max_{(x,y)\in\Delta(m)\times\Delta(n)} \sum_{t=1}^{T} xA_t y - x_t A_t y_t.$$

When Player B plays OGD with step size $\theta = \Theta(T^{-1/2})$, their strategy updates each round as

$$y_{t+1} = \Pi_{\Delta(n)}\left(y_t + \theta(x_t B_t)\right),$$

with $y_1 = \mathbf{u}_n$, and yields regret $O(\sqrt{T})$ for Player B with respect to any $y \in \Delta(n)$ for the loss sequence $\{x_t B_t : t \in [T]\}$. To cast this in our framework, we consider $\Delta(n) = \mathcal{Y}$ as our state

space, where we select actions x_{t-1} to induce desired updates to y_t and optimize over the surrogate losses $\{\mathbf{u}_m A_t^* y_t : t \in [T]\}$. While we do not see B_t prior to choosing each x_t, we view our update errors from instead selecting an action in terms of the dynamics resulting from B_{t-1} as adversarial disturbances and run NESTEDOCO-UD, as the dynamics are strongly locally controllable.

Theorem 16 (Regret Minimization in Online Games) *For $\theta = \Theta(T^{-1/2})$, repeated play against* OGD *in online $m \times n$ games can be cast as a θ-strongly locally controllable instance of online control with nonlinear dynamics, for which* NESTEDOCO-UD *obtains regret $O(\sqrt{T}+\sum_t(\delta_t+\epsilon_t))$.*

References

Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Annual Conference Computational Learning Theory*, 2008. URL https://api.semanticscholar.org/CorpusID:8547150.

Arpit Agarwal and William Brown. Diversified recommendations for agents with adaptive preferences. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a75db7d2ee1e4bee8fb819979b0a6cad-Paper-Conference.pdf.

Arpit Agarwal and William Brown. Online recommendations for agents with discounted adaptive preferences, 2023.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M. Kakade, and Karan Singh. Online control with adversarial disturbances, 2019a.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/78719f11fa2df9917de3110133506521-Paper.pdf.

Shipra Agrawal, Yiding Feng, and Wei Tang. Dynamic pricing and learning with bayesian persuasion, 2023.

Saba Ahmadi, Avrim Blum, and Kunhe Yang. Fundamental bounds on online strategic classification, 2023.

Guillermo Alcantara-Jiménez and Julio B. Clempner. Repeated stackelberg security games: Learning with incomplete state information. *Reliability Engineering System Safety*, 195:106695, 2020. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2019.106695. URL https://www.sciencedirect.com/science/article/pii/S0951832019304478.

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 736–749, 2022.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On the convergence of no-regret learning dynamics in time-varying games, 2023.

Oren Anava, Elad Hazan, and Shie Mannor. Online convex optimization against adversaries with memory and application to statistical arbitrage, 2014.

Masanao Aoki. Local Controllability of a Decentralized Economic System1. *The Review of Economic Studies*, 41(1):51–63, 01 1974. ISSN 0034-6527. doi: 10.2307/2296398. URL https://doi.org/10.2307/2296398.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, page 61–78, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334105. doi: 10.1145/2764468.2764478. URL https://doi.org/10.1145/2764468.2764478.

M. Barbero-Liñán and B. Jakubczyk. Second order conditions for optimality and local controllability of discrete-time systems, 2013.

Avrim Blum, MohammadTaghi Hajiaghayi, Katrina Ligett, and Aaron Roth. Regret minimization and the price of total anarchy. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 373–382, 2008.

Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/cc1aa436277138f61cda703991069eaf-Paper.pdf.

Ugo Boscain, Daniele Cannarsa, Valentina Franceschi, and Mario Sigalotti. Local controllability does imply global controllability, 2021.

Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Selling to a no-regret buyer. *CoRR*, abs/1711.09176, 2017. URL http://arxiv.org/abs/1711.09176.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world, 2022.

William Brown, Jon Schneider, and Kiran Vodrahalli. Is learning in games good for the learners?, 2023.

Asaf Cassel, Alon Cohen, and Tomer Koren. Efficient online linear control with stochastic convex costs and unknown dynamics, 2022.

Xinyi Chen, Edgar Minasyan, Jason D. Lee, and Elad Hazan. Provable regret bounds for deep online learning and control, 2022.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. *CoRR*, abs/1806.07104, 2018. URL http://arxiv.org/abs/1806.07104.

Natalie Collina, Eshwar Ram Arunachaleswaran, and Michael Kearns. Efficient stackelberg strategies for finitely repeated games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 643–651, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

Constantinos Daskalakis and Vasilis Syrgkanis. Learning in auctions: Regret is hard, envy is easy. *CoRR*, abs/1511.01411, 2015. URL http://arxiv.org/abs/1511.01411.

Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations, 2022.

Yuan Deng, Jon Schneider, and Balusubramanian Sivan. Strategizing against no-regret learners, 2019.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, page 55–70, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358293. doi: 10.1145/3219166.3219193. URL https://doi.org/10.1145/3219166.3219193.

Zhe Feng, Okke Schrijvers, and Eric Sodomka. Online learning for measuring incentive compatibility in ad auctions. *CoRR*, abs/1901.06808, 2019. URL http://arxiv.org/abs/1901.06808.

Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *CoRR*, cs.LG/0408007, 2004. URL http://arxiv.org/abs/cs.LG/0408007.

Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320, 03 2016. ISSN 0033-362X. doi: 10.1093/poq/nfw006. URL https://doi.org/10.1093/poq/nfw006.

Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games, 2019.

Jason Gaitonde, Jon M. Kleinberg, and Éva Tardos. Polarization in geometric opinion dynamics. In Péter Biró, Shuchi Chawla, and Federico Echenique, editors, *EC '21: The 22nd ACM Conference on Economics and Computation, Budapest, Hungary, July 18-23, 2021*, pages 499–519. ACM, 2021.

Negin Golrezaei, Adel Javanmard, and Vahab S. Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *CoRR*, abs/2002.11137, 2020. URL https://arxiv.org/abs/2002.11137.

Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics, 2022.

Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic classification. *CoRR*, abs/1506.06980, 2015. URL http://arxiv.org/abs/1506.06980.

Jason Hartline, Vasilis Syrgkanis, and Eva Tardos. No-regret learning in bayesian games. *Advances in Neural Information Processing Systems*, 28, 2015a.

Jason D. Hartline, Vasilis Syrgkanis, and Éva Tardos. No-regret learning in repeated bayesian games. *CoRR*, abs/1507.00418, 2015b. URL http://arxiv.org/abs/1507.00418.

Elad Hazan. Introduction to online convex optimization, 2021.

Elad Hazan and Kfir Levy. Bandit convex optimization: Towards tight bounds. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Elad Hazan and Karan Singh. Introduction to online nonstochastic control, 2022.

Jan Hazla, Yan Jin, Elchanan Mossel, and Govind Ramnarayan. A geometric model of opinion polarization. *CoRR*, abs/1910.05274, 2019.

Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems, 2022a.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. *CoRR*, abs/2202.00628, 2022b. URL https://arxiv.org/abs/2202.00628.

Liyan Jia, Lang Tong, and Qing Zhao. An online learning approach to dynamic pricing for demand response, 2014.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control, 2020.

Yash Kanoria and Hamid Nazerzadeh. Dynamic reserve prices for repeated auctions: Learning from bids. *CoRR*, abs/2002.07331, 2020. URL https://arxiv.org/abs/2002.07331.

H. Kuhn and H.-W. Wohltmann. Controllability of economic systems under alternative expectations hypotheses—the discrete case. *Computers Mathematics with Applications*, 18(6):617–628, 1989. ISSN 0898-1221. doi: https://doi.org/10.1016/0898-1221(89)90112-0. URL https://www.sciencedirect.com/science/article/pii/0898122189901120.

Raunak Kumar, Sarah Dean, and Robert D. Kleinberg. Online convex optimization with unbounded memory, 2022.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 757–762, 2021. doi: 10.1109/CDC45484.2021.9683670.

Niklas Lauffer, Mahsa Ghasemi, Abolfazl Hashemi, Yagiz Savas, and Ufuk Topcu. No-regret learning in dynamic stackelberg games, 2022.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *Algorithmic Game Theory*, 2009. URL https://api.semanticscholar.org/CorpusID:1795572.

Wenhao Luo, Wen Sun, and Ashish Kapoor. Sample-efficient safe learning for online nonlinear control with control barrier functions, 2022.

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3242–3245. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/lykouris21a.html.

Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against learners in bayesian games, 2022.

Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *J. ACM*, 54(5):22–es, oct 2007. ISSN 0004-5411. doi: 10.1145/1284320.1284321. URL https://doi.org/10.1145/1284320.1284321.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/33e75ff09dd601bbe69f351039152189-Paper.pdf.

John Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. *CoRR*, abs/2102.08570, 2021. URL https://arxiv.org/abs/2102.08570.

Edgar Minasyan, Paula Gradu, Max Simchowitz, and Elad Hazan. Online control of unknown time-varying dynamical systems, 2022.

Jamie Morgenstern and Tim Roughgarden. Learning simple auctions. *CoRR*, abs/1604.03171, 2016. URL http://arxiv.org/abs/1604.03171.

Marco Mussi, Gianmarco Genalti, Alessandro Nuara, Francesco Trovò, Marcello Restelli, and Nicola Gatti. Dynamic pricing with volume discounts in online settings, 2022.

Deepan Muthirayan and Pramod P. Khargonekar. Online learning robust control of nonlinear dynamical systems, 2022.

Thomas Nedelec, Clement Calauzenes, Vianney Perchet, and Noureddine El Karoui. Robust stackelberg buyers in repeated auctions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1342–1351. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/nedelec20a.html.

Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10407–10417. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5631e6ee59a4175cd06c305840562ff3-Paper.pdf.

Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:92982174.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. *CoRR*, abs/2002.06673, 2020. URL https://arxiv.org/abs/2002.06673.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos, 2022.

Aaron Roth, Jonathan R. Ullman, and Zhiwei Steven Wu. Watch and learn: Optimizing from revealed preferences feedback. *CoRR*, abs/1504.01033, 2015. URL http://arxiv.org/abs/1504.01033.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *J. ACM*, 62(5), nov 2015. ISSN 0004-5411. doi: 10.1145/2806883. URL https://doi.org/10.1145/2806883.

Shai Shalev-Shwartz and Yoram Singer. Online learning meets optimization in the dual. In *Proceedings of the 19th Annual Conference on Learning Theory*, COLT'06, page 423–437, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540352945. doi: 10.1007/11776420_32. URL https://doi.org/10.1007/11776420_32.

Lingqing Shen, Nam Ho-Nguyen, and Fatma Kılınç-Karzan. An online convex optimization-based framework for convex bilevel optimization. *Mathematical Programming*, 198(2):1519–1582, 04 2023. ISSN 1436-4646. doi: 10.1007/s10107-022-01894-5. URL https://doi.org/10.1007/s10107-022-01894-5.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. *CoRR*, abs/2001.09254, 2020. URL https://arxiv.org/abs/2001.09254.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2011.12.028. URL https://www.sciencedirect.com/science/article/pii/S0022000012000281. JCSS Special Issue: Cloud Computing 2011.

Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Alexander Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Steering no-regret learners to optimal equilibria, 2023.

Xuezhou Zhang, Yiding Chen, Jerry Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. *CoRR*, abs/2106.06630, 2021. URL https://arxiv.org/abs/2106.06630.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2, 04 2003.

Tijana Zrnic, Eric Mazumdar, S. Shankar Sastry, and Michael I. Jordan. Who leads and who follows in strategic classification? *CoRR*, abs/2106.12529, 2021a. URL https://arxiv.org/abs/2106.12529.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15257–15269. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/812214fb8e7066bfa6e32c626c2c688b-Paper.pdf.

Appendix A. Omitted Proofs for Section 2

Proof of Proposition 3. Without loss of generality, assume $\alpha \leq \beta/2$ and that T is even. Let $f_t = \|y_t - y\|$ for each t. Consider any round t where $y_{t-1} \in B_\alpha(y)$; then, for all actions x_t, we have that $y_t \notin \mathcal{B}_\alpha(y)$, as $\mathcal{B}_\alpha(y) \subseteq \mathcal{B}_\beta(y_{t-1})$; as such, we incur loss $f_t(y_t) \geq \alpha$ in round t. Now suppose $y_{t-1} \notin B_\alpha(y)$; then, we must have incurred loss at least $f_{t-1}(y_{t-1}) \geq \alpha$ in round $t-1$. As losses are non-negative, our total loss is at least $\alpha T/2$, as loss α is incurred at least every other round; given that the best fixed state $y^* = y$ incurs total loss 0, we have that $\mathrm{Reg}_{\mathcal{A}}(T) = \Omega(T)$ for any algorithm $\mathcal{A}$. ■

Proof of Proposition 4. We begin by observing that for instances $(\mathcal{X}, \mathcal{Y}, D)$, the class of state-targeting policies contains a policy which obtains the reward of the best fixed state up to $O(\sqrt{T\rho^{-1}})$, for sufficiently large T. Consider the set $\hat{\mathcal{Y}} = \{y^* \in \mathcal{Y} : \pi(y^*) \geq (T\rho)^{-1/2}\}$. Note that the reward of any $y \in \mathcal{Y}$ is matched by some $y^* \in \hat{\mathcal{Y}}$ up to $O(\sqrt{T\rho^{-1}})$ for any fixed inner radius r, outer radius R, and Lipschitz constant L. For any such y^*, note that under the policy P_{y^*} when starting at $y_0 = 0$, the distance between y_t and y^* in each round t is updated to at most:

$$\|y_t - y^*\| \leq \max\left(0, \rho \cdot \pi(y_{t-1})\right).$$

It is straightforward to see that $\hat{\mathcal{Y}}$ is convex, and so our state y_t will never leave $\hat{\mathcal{Y}}$ on its path to y^*; as such, we reach y^* within $O(\sqrt{T\rho^{-1}})$ rounds, after which point our reward exactly tracks that of y^*. For some $y^* \in \hat{\mathcal{Y}}$, this yields a regret for P_{y^*} of at most $O(\sqrt{T\rho^{-1}})$ to the best fixed state in $\mathcal{Y}$.

Next, consider an instance where $\mathcal{X}$ and $\mathcal{Y}$ are both the unit ball in $\mathbb{R}^n$. With $y_0 = 0$, let the dynamics be given by

$$y_t = \Pi_{\mathcal{Y}}\left(y_{t-1} + x_t\right).$$

Observe that this satisfies ρ-local controllability for any $\rho \leq 1$, as a ball of radius $\pi(y_{t-1})$ is always feasible around y_{t-1}. Let each loss $f_t = \|y - p\|^2$, for some $p \neq 0$. Immediately we can see that any matrix policy $K \in \mathcal{P}_{\mathcal{K}}$ has regret $\Omega(T)$, as the action $x_t = 0$ will be played in each round. ■

Appendix B. Follow the Regularized Leader

Here we state the FTRL algorithm and several of its key properties; see e.g. Hazan (2021) for proofs of Propositions 17 and 18.

Algorithm 2 Follow the Regularized Leader (FTRL)

Choose a time horizon T, step size η, and γ-strongly convex regularizer $\psi : \mathcal{Y} \to \mathbb{R}$
Let $y_1 = \mathrm{argmin}_{y \in \mathcal{Y}} \psi(y)$
for $t = 1$ to T **do**
  Play y_t and observe loss $f_t(y_t)$
  Set $\nabla_t = \nabla f_t(y_t)$
  Set $y_{t+1} = \mathrm{argmin}_{y \in \mathcal{Y}} \left(\eta \cdot \sum_{s=1}^{t} \nabla_s^\top y + \psi(y)\right)$
end for

Proposition 17 *For a γ-strongly convex regularizer $\psi : \mathcal{Y} \to \mathbb{R}$ where $|\psi(y) - \psi(y')| \leq G$ for all $y, y' \in \mathcal{Y}$, and for convex L-Lipschitz losses $f_1, \ldots, f_T$, the regret of* FTRL *is bounded by*

$$\mathrm{Reg}_T(\mathrm{FTRL}) \leq \eta \frac{TL^2}{\gamma} + \frac{G}{\eta}.$$

Proposition 18 *Any pair of points y_t and y_{t+1} chosen by* FTRL *satisfies $\|y_{t+1} - y_t\| \leq \eta \frac{L}{\gamma}$.*

Appendix C. Analysis for NESTEDOCO

Proof of Theorem 5. First we show that any point chosen by FTRL will be feasible under local controllability, by induction. It is straightforward to see that $\tilde{\mathcal{Y}}$ is convex and $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$; further, any $y \in \tilde{\mathcal{Y}}$ is bounded away from $\mathrm{bd}(\mathcal{Y})$. By the definition of $\tilde{\mathcal{Y}}$, we have that $y = (1-\delta)y'$ for some $y' \in \mathcal{Y}$. Recall that $\mathcal{B}_r(\mathbf{0}) \subseteq \mathcal{Y}$, and note that $\mathcal{B}_{\delta r}(y) = \{y + \delta\hat{y} : \hat{y} \in \mathcal{B}_r(\mathbf{0})\}$. Let y'' be any point in $\mathcal{B}_r(\mathbf{0})$. By convexity of $\mathcal{Y}$, we then have that any point $(1-\delta)y' + \delta y''$ lies in $\mathcal{Y}$, and so for any $y \in \tilde{\mathcal{Y}}$ we have that $\mathcal{B}_{r\delta}(y) \subseteq \mathcal{Y}$. Each y_{t-1} lies in $\tilde{\mathcal{Y}}$, and so we have that $\pi(y_{t-1}) \geq r\delta$; as such, any point y_t in $\mathcal{B}_{r\delta\rho}(y_{t-1}) \subseteq \mathcal{B}_{\rho\cdot\pi(y_{t-1})}(y_{t-1})$ is feasible. Given that $\eta\frac{L}{\gamma} \leq r\delta\rho$, by Proposition 18 we have that $y_t \in \mathcal{B}_{r\delta\rho}(y_{t-1})$ in each round for the chosen point. Each action will be selected by solving for

$$\operatorname*{argmin}_{x_t \in \mathcal{X}} \|D(x_t, y_{t-1}) - y^*\|^2$$

via a call to $\texttt{Oracle}(y_{t-1}, y^*)$. Each call is guaranteed to have a solution which achieves an objective of 0 where $D(x_t, y_{t-1}) = y^*$ for some $y^* \in \mathcal{B}_{\rho\cdot\pi(y_{t-1})}(y_{t-1})$ by local controllability, yielding an exact state update to $y_t = y^*$ as we assume $\texttt{Oracle}$ can solve arbitrary non-convex minimization problems. To bound the regret, first note that for any $y^* \in \mathcal{Y}$, we have

$$\sum_{t=1}^{T} f_t(y_t) \leq \eta \frac{TL^2}{\gamma} + \frac{G}{\eta} + \sum_{t=1}^{T} f_t((1-\delta)y^*)$$

by Proposition 17, as $(1-\delta)y^* \in \tilde{\mathcal{Y}}$ for any $y^* \in \mathcal{Y}$. Then, observe that for any $y^* \in \mathcal{Y}$, we have that

$$\begin{aligned}\sum_{t=1}^{T} f_t((1-\delta)y^*) &\leq \sum_{t=1}^{T} \left(f_t(y^*) + L\|\delta y^*\|\right) \\ &\leq \sum_{t=1}^{T} \left(f_t(y^*) + \delta LR\right).\end{aligned}$$

Combining the previous claims, we have that

$$\begin{aligned}\sum_{t=1}^{T} f_t(y_t) - f_t(y^*) &\leq \delta TLR + \eta\frac{TL^2}{\gamma} + \frac{G}{\eta} \\ &= \eta\left(1 + \frac{R}{r\rho}\right)\frac{TL^2}{\gamma} + \frac{G}{\eta} \\ &= 2\sqrt{\frac{(1+\frac{R}{r\rho})TGL^2}{\gamma}}\end{aligned}$$

upon setting $\delta = \eta \frac{L}{r\rho\gamma}$ and $\eta = \sqrt{\frac{G\gamma}{(1+\frac{R}{r\rho})TL^2}}$, which yields the theorem. ■

Appendix D. Examples and Analysis for Action-Linear Dynamics

As a simple yet general example of dynamics which are both action-linear and locally controllable, consider update rules in which a step is taken by applying a nonsingular matrix transformation to the action, where the matrix can be parameterized by the state, with projection back into $\mathcal{Y}$ if necessary.

Example 1 *Let both $\mathcal{X}$ and $\mathcal{Y}$ be given by the unit ball $\mathcal{B}_1(\mathbf{0})$ in $\mathbb{R}^n$. For any fixed y, let the updates from $D(x, y)$ be given by*

$$D(x, y) = \Pi_{\mathcal{Y}} \left(y + A_y \cdot x \right),$$

where each A_y is a square matrix with minimum absolute eigenvalue $|\lambda_n(A_y)| \geq \pi(y) \cdot \rho$ for some $\rho > 0$. Then, the instance $(\mathcal{X}, \mathcal{Y}, D)$ is action-linear and satisfies ρ-local controllability.

Proof for Example 1. It is straightforward to see that $D(x, y)$ is action-linear. To show ρ-local controllability, let y^* be any point in $\mathcal{B}_{\rho \cdot \pi(y)}(y)$. It suffices to show that there is some $x^* \in \mathcal{X}$ such that $A_y \cdot x^* = y^* - y$. As A_y is non-singular, we can solve for $x^* = A_y^{-1}(y^* - y)$, where $\|y^* - y\| \leq \rho \cdot \pi(y)$ and $\left|\lambda_1(A_y^{-1})\right| \leq \frac{1}{\rho \cdot \pi(y)}$, and so we have that $x^* \in \mathcal{B}_1(\mathbf{0}) = \mathcal{X}$. ■

We can also extend this to include state-parameterized generalizations of any linear system governed by nonsingular matrices over a bounded-radius state space (for a sufficiently large action space).

Example 2 *Let $\mathcal{Y}$ be given by the radius-R ball $\mathcal{B}_R(\mathbf{0})$ in $\mathbb{R}^n$, and let $\mathcal{X} = \mathcal{B}_{cR}(\mathbf{0})$. For any fixed y, let the updates from $D(x, y)$ be given by*

$$D(x, y) = \Pi_{\mathcal{Y}} \left(K_y \cdot y + A_y \cdot x \right),$$

where both K_y and A_y are square matrices. For any y, let $M_y = K_y - I$, and suppose we take c large enough such that $c \cdot |\lambda_n(A_y)| \geq |\lambda_1(M_y)| + \pi(y) \cdot \rho$ for some $\rho > 0$. Then, the instance $(\mathcal{X}, \mathcal{Y}, D)$ is action-linear and satisfies ρ-local controllability.

Proof for Example 2. Here, again it is evident that $D(x, y)$ is action-linear, and so it suffices to show that there is some $x^* \in \mathcal{X}$ such that

$$\begin{aligned} K_y \cdot y + A_y \cdot x^* &= y + M_y \cdot y + A_y \cdot x^* \\ &= y^* \end{aligned}$$

for any y^* in $\mathcal{B}_{\rho \cdot \pi(y)}(y)$. As in the proof for Example 1, we have that $\|M_y \cdot y\| \leq R \cdot |\lambda_1(M_y)|$, and for large enough c there is some x^* such that $A_y \cdot x^* = \hat{y}$ for any $\hat{y}$ where $\|\hat{y}\| \leq R \cdot |\lambda_1(M_y)| + \pi(y) \cdot \rho$. Thus, any point $y^* \in \mathcal{B}_{R \cdot |\lambda_1(M_y)| + \pi(y) \cdot \rho}(y + M_y \cdot y)$ is feasible by some x^*, which contains the ball $\mathcal{B}_{\pi(y) \cdot \rho}(y)$. ■

Appendix E. Algorithms for Adversarial Disturbances

E.1. NESTEDOCO-BD and Proofs for Theorem 8

We show that it is possible simulate NESTEDOCO over the undisturbed states $\hat{y}_t$ under the assumption that the dynamics are in $\alpha\rho$-locally controllable for some $\alpha \in (0, 1)$ while retaining sufficient range in the feasible region around y_t to correct for the disturbance w_{t-1} from the previous round. Here, the oracle call for computing x_t in each round is updated to consider the true state y_{t-1}.

Algorithm 3 NESTEDOCO with Adversarial Disturbances (NESTEDOCO-BD).

Initialize NESTEDOCO for T rounds over $(\mathcal{X}, \mathcal{Y}, D)$ for $\alpha\rho$-locally controllable dynamics
for $t = 1$ to T **do**
  Let $\hat{y}_t$ be the target state chosen by NESTEDOCO
  Use $\texttt{Oracle}(y_{t-1}, \hat{y}_t)$ to compute $x_t = \operatorname{argmin}_{x \in \mathcal{X}} \|D(x, y_{t-1}) - \hat{y}_t\|^2$
  Play action x_t.
  Observe disturbed state $y_t = \hat{y}_t + w_t$ and loss $f_t(y_t)$.
  Update NESTEDOCO with state $\hat{y}_t$ and loss $f_t(\hat{y}_t)$.
end for

Theorem 8 follows directly from Theorems 19, 20, and 21. Intuitively, when the per-round disturbance magnitude is at most $\frac{\rho - \alpha\rho}{1+\rho} \cdot \pi\left(D(x_t, y_{t-1})\right)$, one can calibrate NESTEDOCO for the case of $\alpha\rho$-locally controllable dynamics and maintain sufficient "slack" to correct for the previous round's disturbance in every round. When disturbances exceed $\frac{\rho}{1+\rho} \cdot \pi\left(D(x_t, y_{t-1})\right)$, an adversary can continually push the state towards the boundary of $\mathcal{Y}$, which may require vanishing disturbance magnitude as rounds progress due to the limited range promised by local controllability near the boundary.

Theorem 19 *For a ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ with convex losses $f_t : \mathcal{Y} \to \mathbb{R}$ and adversarial disturbances w_t where $\|w_t\| \leq \frac{\rho - \alpha\rho}{1+\rho} \cdot \pi\left(D(x_t, y_{t-1})\right)$ and $\sum_{t=1}^{T} \|w_t\| \leq E$, the regret of* NESTEDOCO-BD *with respect to the reward of any state is bounded by*

$$\mathrm{Reg}_T(\text{NESTEDOCO-BD}) \leq O\left(\sqrt{T \cdot (\alpha\rho)^{-1}} + E\right),$$

with T queries made to an oracle for non-convex optimization.

Proof We show by induction that each call to $\texttt{Oracle}(y_{t-1}, \hat{y}_t)$ yields a feasible action x_t satisfying $\hat{y}_t = D(x_t, y_{t-1})$. This is immediate for $t = 1$, and suppose this holds up to some round $t - 1$, where we have that $y_{t-1} = \hat{y}_{t-1} + w_{t-1}$. Given that NESTEDOCO selects actions under $\alpha\rho$-local controllability, we can bound

$$\|\hat{y}_t - \hat{y}_{t-1}\| \leq \alpha\rho \cdot \pi(\hat{y}_{t-1}).$$

Further, the magnitude of the disturbance w_{t-1} is bounded by

$$\|w_{t-1}\| \leq \frac{\rho - \alpha\rho}{1+\rho} \cdot \pi(\hat{y}_{t-1}),$$

yielding that

$$\begin{aligned}\|\hat{y}_t - y_{t-1}\| &\leq \|\hat{y}_t - \hat{y}_{t-1} - w_{t-1}\| \\ &\leq \left(\alpha\rho + \frac{\rho - \alpha\rho}{1+\rho}\right) \cdot \pi(\hat{y}_{t-1}). && (y_{t-1} = w_{t-1} + \hat{y}_{t-1})\end{aligned}$$

As such, we have that

$$\begin{aligned}\rho \cdot \pi(y_{t-1}) &\geq \rho\left(1 - \frac{\rho - \alpha\rho}{1+\rho}\right) \cdot \pi(\hat{y}_{t-1}) \\ &= \rho\left(\alpha + \frac{1-\alpha}{1+\rho}\right) \cdot \pi(\hat{y}_{t-1}),\end{aligned}$$

and so by ρ-local controllability some feasible action x_t exists, as $\hat{y}_t$ lies in $\mathcal{B}_{\rho\cdot\pi(y_{t-1})}$. The regret bound for NESTEDOCO holds over the states $\hat{y}_t$, and so we can bound the total regret of NESTEDOCO-BD with respect to any $y^* \in \mathcal{Y}$ as:

$$\begin{aligned}\sum_{t=1}^{T} f_t(y_t) - f_t(y^*) &\leq \sum_{t=1}^{T} f_t(\hat{y}_t) - f_t(y^*) + L\|y_t - \hat{y}_t\| \\ &\leq \text{Reg}_T(\text{OEN-FTRL}) + L\sum_{t=1}^{T}\|w_t\| && \text{(Thm. 5)} \\ &\leq 2\sqrt{\frac{(1+\frac{R}{r\alpha\rho})TGL^2}{\gamma}} + LE.\end{aligned}$$

■

We show that the dependence on E is tight up to the constant. Note that we we can obtain regret $O(\sqrt{T\cdot(\alpha\rho)^{-1}}) + LE$ in the following instance via NESTEDOCO-BD.

Theorem 20 (Regret Lower Bound for Bounded Disturbances) *Suppose for any $\alpha > 0$ and $\rho \in (0,1]$ an adversary can choose w_t with $\|w_t\| \leq \frac{\rho-\alpha\rho}{1+\rho}\cdot\pi\left(D(x_t, y_{t-1})\right)$, where $\sum_{t=1}^{T}\|w_t\| = E$ for any E. There is a ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ with L-Lipschitz convex losses f_t such that any algorithm $\mathcal{A}$ obtains regret $\text{Reg}_T(\mathcal{A}) \geq \max(LE, \frac{\rho-\alpha\rho}{1+\rho}TL)$.*

Proof Consider any norm $\|\cdot\|$ over $\mathbb{R}^n$. Let $\mathcal{Y}$ be the unit ball $B_1(\mathbf{0})$, and let each $f_t(y_t) = L\|y_t\|$. Consider any action space $\mathcal{X}$ and dynamics D where ρ-local controllability exactly characterizes the range of D, i.e. for any y and y', there is some x such that $D(x,y) = y'$ if and only if $y' \in \mathcal{B}_{\rho\cdot\pi(y)}(x,y)$.

First, note that $\pi(y) = 1 - \|y\|$ for any $y \in \mathcal{Y}$. In each round t , suppose an algorithm plays an action x_t at state y_{t-1} which yields an target undisturbed update $\hat{y} = D(x_t, y_{t-1})$. The adversary can then choose any w_t satisfying $\|w_t\| \leq \frac{\rho-\alpha\rho}{1+\rho}\cdot(1-\|\hat{y}_t\|)$; suppose each w_t is given by

$$w_t = \hat{y}_t \cdot \frac{\frac{\rho-\alpha\rho}{1+\rho}\cdot(1-\|\hat{y}_t\|)}{\|\hat{y}_t\|}$$

if $\hat{y}_t$ is non-zero, and an arbitrary vector w_t with $\|w_t\| = \frac{\rho-\alpha\rho}{1+\rho}$ if $\hat{y}_t = \mathbf{0}$. This satisfies the disturbance norm bound, and further yields $y_t = \hat{y}_t + w_t$, where for non-zero $\hat{y}$ we have

$$y_t = \hat{y}_t \cdot \left(1 + \frac{\frac{\rho-\alpha\rho}{1+\rho} \cdot (1 - \|\hat{y}_t\|)}{\|\hat{y}_t\|}\right)$$

and thus for any $\hat{y}$,

$$\begin{aligned}\|y_t\| &\geq \|\hat{y}_t\| + \frac{\rho - \alpha\rho}{1+\rho} \cdot (1 - \|\hat{y}_t\|) \\ &\geq \frac{\rho - \alpha\rho}{1+\rho},\end{aligned}$$

yielding a loss $f_t(y_t) \geq L \cdot \frac{\rho-\alpha\rho}{1+\rho}$ at a disturbance cost of $\|w_t\| = \frac{\rho-\alpha\rho}{1+\rho}(1 - \|\hat{y}_t\|)$. Assuming the adversary continues this strategy in each round until any disturbance budget $E = \sum_{t=1}^{T} \|w_t\|$ is exhausted, this yields a regret for any algorithm of at least

$$\mathrm{Reg}_T(\mathcal{A}) \geq \min\left(LE, \frac{\rho - \alpha\rho}{1+\rho} TL\right),$$

as $y^* = \mathbf{0}$ obtains total loss 0. ■

The disturbance upper bound is indeed necessary for ρ-locally controllable dynamics. We show a sharp threshold effect at $\frac{\rho}{1+\rho} \cdot \pi(D(x_t, y_{t-1}))$, wherein an adversary who is allowed to exceed this limit by any amount can force an algorithm to incur linear regret even with only a constant budget. Note that for any $\rho \in (0, 1]$ and $\alpha < 0$, there is some $\beta \in [0, 1)$ such that $\frac{\rho-\alpha\rho}{1+\rho} \geq \frac{\rho}{1+\beta\rho}$.

Theorem 21 *Suppose an adversary can choose any state disturbances w_t with $\|w_t\| \leq \frac{\rho}{1+\beta\rho} \cdot \pi\left(D(x_t, y_{t-1})\right)$, for any $\rho \in (0, 1]$ and any $\beta \in [0, 1)$. Then, there is a ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ with convex losses f_t such that any algorithm $\mathcal{A}$ obtains regret $\mathrm{Reg}_T(\mathcal{A}) = \Theta(T)$ even if $\sum_{t=1}^{T} \|w_t\| = O(1)$.*

Proof Consider any instance $(\mathcal{X}, \mathcal{Y}, D)$ where ρ-local controllability exactly characterizes the range of D, i.e. for any y and y', there is some x such that $D(x, y) = y'$ if and only if $y' \in \mathcal{B}_{\rho \cdot \pi(y)}(x, y)$.

Let $d_t = \pi(y_t)$ for each round. Beginning at any round t, suppose the adversary observes an action x_t which yields an update $\hat{y}_t = D(x_t, y_{t-1})$. Let $z_t = \mathrm{argmin}_{y \in \mathrm{bd}(\mathcal{Y})} \|y - \hat{y}_t\|$, and suppose the adversary chooses the disturbance:

$$w_t = \operatorname*{argmin}_{w : \|w\| \leq \frac{\rho}{1+\beta\rho} \cdot \pi(\hat{y}_t)} \|\hat{y}_t + w_t - z_t\|.$$

This forces y_t closer to the boundary at each round, regardless of the choice of x_t:

$$\begin{aligned}d_t &= \left(1 - \frac{\rho}{1+\beta\rho}\right) \cdot \pi(\hat{y}_t) \\ &\leq \left(1 + \rho - \frac{\rho}{1+\beta\rho} - \frac{\rho^2}{1+\beta\rho}\right) d_{t-1} && (\pi(\hat{y}_t) \leq (1+\rho) d_{t-1}) \\ &\leq \frac{1 + \beta\rho + \beta\rho^2 - \rho^2}{1+\beta\rho} d_{t-1} \\ &\leq \left(1 - \frac{(1-\beta)\rho^2}{1+\beta\rho}\right) d_{t-1},\end{aligned}$$

where $\pi(\hat{y}_t) \leq (1+\rho)d_{t-1}$ holds by our assumption on $D(x, y)$. Assuming the adversary applies a disturbance w_t selected as above in each round $t \leq T$, we have that

$$d_t \leq \left(1 - \frac{(1-\beta)\rho^2}{1+\beta\rho}\right)^t \cdot d_0,$$

where the magnitude of each disturbance is bounded by

$$\begin{aligned}\|w_t\| &\leq \frac{\rho+\rho^2}{1+\beta\rho} d_{t-1} \\ &\leq \frac{\rho+\rho^2}{1+\beta\rho}\left(1 - \frac{(1-\beta)\rho^2}{1+\beta\rho}\right)^{t-1} \cdot d_0,\end{aligned}$$

where we take the initial state distance to the boundary $d_0 = \pi(y_0)$ to be a constant bounded away from zero. This yields that the sum of disturbance magnitudes $E = \sum_{t=1}^{T} \|w_t\|$ is at most:

$$\begin{aligned}\sum_{t=1}^{T} \|w_t\| &\leq d_0 \frac{\rho+\rho^2}{1+\beta\rho} \cdot \sum_{t=1}^{T}\left(1 - \frac{(1-\beta)\rho^2}{1+\beta\rho}\right)^{t-1} \\ &\leq d_0 \cdot \frac{\rho+\rho^2}{(1-\beta)\rho^2} \\ &= O(1).\end{aligned}$$

Now suppose that the loss at each round is given by $f_t(y_t) = \|y_t - y_0\|$. Then, our regret with respect to y_0 is at least:

$$\begin{aligned}\sum_{t=1}^{T} f_t(y_t) - f_t(y_0) &\leq \sum_{t=1}^{T} d_0 - d_t \\ &\leq d_0\left(T - \sum_{t=1}^{T} \frac{(1-\beta)\rho^2}{1+\beta\rho}\right) \\ &\leq d_0\left(T - \frac{1 - \frac{(1-\beta)\rho^2}{1+\beta\rho}}{\frac{(1-\beta)\rho^2}{1+\beta\rho}}\right) \\ &\leq d_0\left(T - \frac{1+\beta\rho}{(1-\beta)\rho^2}\right) \\ &= \Theta(T).\end{aligned}$$

■

Together, the previous three theorems yield Theorem 8.

E.2. NESTEDOCO-UD and Proofs for Theorem 9

We can remove the bound on the maximum disturbance for strongly locally controllable instances, as the feasible update sets do not vanish at the boundary of $\mathcal{Y}$. Recall that an instance $(\mathcal{X}, \mathcal{Y}, D)$

satisfies strong ρ-local controllability for $\rho > 0$ if, for any $y \in \mathcal{Y}$ and $y^* \in \mathcal{B}_\rho(y) \cap \mathcal{Y}$, there is some x such that $D(x, y) = y^*$. We assume without loss of generality that $\rho \leq 2R$, where R is the radius of $\mathcal{Y}$.

Intuitively, our algorithm tracks the target state which would be chosen by FTRL in the absence of all disturbances (by recording the loss counterfactual loss rather than the one truly experienced), and always seeks to minimize distance to that state.

Algorithm 4 NESTEDOCO with Unbounded Disturbances (NESTEDOCO-UD).

Initialize FTRL for T rounds over $\mathcal{Y}$ with step size $\eta = \sqrt{\frac{G\gamma}{TL^2}}$.
for $t = 1$ to T **do**
  Let $\hat{y}_t$ be the target state chosen by FTRL.
  Use $\texttt{Oracle}(y_{t-1}, \hat{y}_t)$ to compute $x_t = \operatorname{argmin}_{x \in \mathcal{X}} \|D(x, y_{t-1}) - \hat{y}_t\|^2$.
  Play action x_t.
  Observe disturbed state $y_t = D(x_t, y_{t-1}) + w_t$ and loss $f_t(y_t)$.
  Update FTRL with state $\hat{y}_t$ and loss $f_t(\hat{y}_t)$.
end for

Theorem 22 *For a strongly ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ with convex losses $f_t : \mathcal{Y} \to \mathbb{R}$ and adversarial disturbances w_t where $\sum_{t=1}^T \|w_t\| \leq E$, the regret of* NESTEDOCO-UD *is bounded by*

$$\text{Reg}_T(\text{NESTEDOCO-UD}) \leq O\left(\sqrt{T} + E \cdot \rho^{-1}\right)$$

with respect to the reward of any state, with T queries made to an oracle for non-convex optimization.

Proof We begin by bounding the total state error $\sum_{t=1}^t \|y_t - \hat{y}_t\|$ across rounds. First, note that for any fixed $\rho > 0$, and any desired $\alpha \in (0, 1)$, we have that $\eta \frac{L}{\gamma} \leq \rho\alpha$ for sufficiently large T, as $\eta \frac{L}{\gamma} = \sqrt{\frac{G}{T\gamma}}$; we assume this holds for any given choice of α, and so we have that $\|\hat{y}_{t+1} - \hat{y}_t\| \leq \rho\alpha$ by Proposition 18. For a total disturbance budget E, we separately consider disturbances w_t depending on whether or not the accumulated disturbance error up to w_t is driven to 0 in the next round. Define W_+ and W_- as:

$$W_+ = \{w_t : D(x_{t+1}, y_t) \neq \hat{y}_{t+1}\}$$

and

$$W_- = \{w_t : D(x_{t+1}, y_t) = \hat{y}_{t+1}\}$$

with $E_+ = \sum_{w_t \in W_+} \|w_t\|$ and $E_- = \sum_{w_t \in W_-} \|w_t\|$. First, observe that at each round t corresponding to $w_t \in W_-$, given that $\|\hat{y}_{t+1} - y_t\| \leq \rho$ we have that $\|w_t\| = \|y_t - \hat{y}_t\| \leq (1 + \alpha)\rho$, as $\|\hat{y}_{t+1} - \hat{y}_t\| \leq \alpha\rho$. As such, we have that

$$\begin{aligned}\sum_{t:w_t \in W_-} f_t(y_t) - f_t(\hat{y}_t) &\leq \sum_{t:w_t \in W_-} L \|y_t - \hat{y}_t\| \\ &\leq (1 + \alpha) L E_-.\end{aligned}$$

Next, consider any $w_t \in W_+$. As our instance is strongly ρ-locally controllable, we must have that $\|\hat{y}_{t+1} - y_t\| > \rho$, as otherwise there would some feasible action x_{t+1} which would be selected that would yield $w_t \in W_-$. Since $\|\hat{y}_{t+1} - \hat{y}_t\| \leq \alpha\rho$, it then must be the case that $\|w_t\| = \|y_t - \hat{y}_t\| > (1-\alpha)\rho$, and so we can bound the number of disturbances in W_+ as:

$$|W_+| \leq \frac{E_+}{(1-\alpha)\rho}.$$

Assuming a maximal distance $\|\hat{y}_t - y_t\| = 2R$ for each round t corresponding to some $w_t \in W_+$, this yields

$$\begin{aligned}\sum_{t:w_t \in W_+} f_t(y_t) - f_t(\hat{y}_t) &\leq \sum_{t:w_t \in W_+} L\,\|y_t - \hat{y}_t\| \\ &\leq \frac{2LRE_+}{(1-\alpha)\rho}\end{aligned}$$

We can assume α is small enough to yield $\frac{2R}{\rho} \geq (1+\alpha)\cdot(1-\alpha)$, and so we have

$$\sum_{t=1}^{T} f_t(y_t) - f_t(\hat{y}_t) \leq \frac{2LRE}{(1-\alpha)\rho}.$$

The regret bound for FTRL holds over the states $\hat{y}_t$, and so we can bound the total regret of NESTEDOCO-BD with respect to any $y^* \in \mathcal{Y}$ as:

$$\begin{aligned}\sum_{t=1}^{T} f_t(y_t) - f_t(y^*) &\leq \sum_{t=1}^{T} f_t(\hat{y}_t) - f_t(y^*) + \sum_{t=1}^{T} f_t(y_t) - f_t(\hat{y}_t) \\ &\leq \eta\frac{TL^2}{\gamma} + \frac{G}{\eta} + \frac{2LRE}{(1-\alpha)\rho} && \text{(Prop. 17)} \\ &\leq 2\sqrt{\frac{TGL^2}{\gamma}} + \frac{2LRE}{(1-\alpha)\rho}.\end{aligned}$$

■

Theorem 23 (Regret Lower Bound for Unbounded Disturbances) *Suppose an adversary can choose any state disturbances w_t with $\sum_{t=1}^{T}\|w_t\| = E$. For any $\rho \in (0,1]$, there is a strongly ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ with convex losses f_t such that any algorithm $\mathcal{A}$ obtains regret* $\text{Reg}_T(\mathcal{A}) = \min(\frac{2LRE}{\rho}, 2TLR)$.

Proof Let $\mathcal{Y} = [-R, R]$ for any $R > 0$ and let $f_t(y_t) = -Ly_t + LR$ for each y. Suppose strong ρ-local controllability exactly characterizes the range of D, i.e. for any $y, y' \in \mathcal{Y}$ there is some x such that $D(x, y) = y'$ if and only if $|y - y'| \leq \rho$. Consider an adversary who chooses disturbances w_t in each round such that $y_t = -R$ until their disturbance budget E is exhausted. This requires a disturbance of magnitude at most $R + \rho$ for w_1, as we assume $y_0 = 0$, and at most ρ in subsequent rounds, and thus the adversary can force any algorithm to remain at $y_t = -R$ for $(E - R)\rho^{-1}$ rounds.

As such, any algorithm must incur loss of at least $2LR(E-R)\rho^{-1}$ across these rounds, and further must incur average loss LR over the subsequent $2R\rho^{-1}$ rounds (if T is not yet reached), for an additional loss of $2LR^2\rho^{-1}$, as they can only decrease per-round loss by $L\rho$ given the restriction on the range of D. As the optimal state $y^* = R$ obtains loss 0, the total regret is at least:

$$\sum_{t=1}^{T} f_t(y_t) - f_t(y^*) \geq \min\left(\frac{2LRE}{\rho}, 2TLR\right).$$

■

Together, the previous two theorems yield Theorem 9. Note that for both algorithms it remains computationally efficient to optimize over action-linear dynamics, as the constraint that $D(x, y_{t-1}) \in \mathcal{Y}$ can be encoded as a convex contraint over $\mathcal{X}$.

Appendix F. Unknown Dynamics: Analysis for PROBINGOCO

Algorithm 5 Probing Online Convex Optimization (PROBINGOCO).

Let $n = \dim(\mathcal{X})$, let $y_0 = \mathbf{0}$, and let $x_1 \in \mathcal{X}$ such that $\|D(x_1, y_0) - y_0\| \leq \epsilon = o(\sqrt{T})$
Initialize NESTEDOCO-BD to run over $\mathcal{Y}$ for $T/(2n+1)$ rounds
Run ESTIMATE for $2n+1$ rounds:
Play x_1
for $i = 1$ to n **do**
  Play $x_1 + \epsilon \cdot e_i$
  Play $x_1 - \epsilon \cdot e_i$
end for
Solve for estimates $(\hat{A}_y, \hat{b}_y)$ which are consistent with with the previous $2n+1$ observed state updates, up to error $O(\epsilon)$
for $t = 2n+1$ to T **do**
  Let $t^* = t$
  Using $(\hat{A}_y, \hat{b}_y)$, target $y = y_{t^*}$
  Let y^* be the next point chosen by NESTEDOCO-BD
  for $i = 1$ to n **do**
    Using $(\hat{A}_y, \hat{b}_y)$, target $y = y_{t^*} + \frac{2i-1}{2n}(y^* - y_{t^*}) + \epsilon \cdot e_i$
    Using $(\hat{A}_y, \hat{b}_y)$, target $y = y_{t^*} + \frac{2i}{2n}(y^* - y_{t^*}) - \epsilon \cdot e_i$
  end for
  Update estimates $(\hat{A}_y, \hat{b}_y)$, solving for values which are consistent with the previous $2n+1$ observed state updates, up to error $O(\epsilon)$
end for

Proof of Theorem 10 Assume the following hold for $D(x, y)$ at each y:

- $D(x, y) = A_y \cdot x + b_y + y + q_y(x)$, for a function $q_y : \mathcal{X} \to \mathbb{R}^n$;
- A_y has a largest absolute eigenvalue bounded by an absolute constant, smallest absolute eigenvalue bounded away from 0, and is L_α-Lipschitz in the matrix ℓ_2 norm;

- b_y has a norm bounded by an absolute constant, and is L_β-Lipschitz;
- $\|q_y(x)\| \le \epsilon$ for any x such that $\|A_y \cdot x + b_y - y\| = O(\sqrt{T})$.

In the neighborhood of any y^*, observe that playing $x = A_y^{-1}(y^* - y - b_y)$ yields an update to $y^* + w_\epsilon$, where the error term w_ϵ has magnitude bounded linearly in terms of the neighborhood size as well as polynomial in the relevant constants. We assume sufficiently small values of ϵ, L_α, and L_β (whose relative bounds may trade off with each other, and in general will be inverse-polynomial in problem parameters other than T) to bound the error of this process in accordance with the requirements of Theorem 8, as well as to ensure that estimation error for $(\hat{A}_y, \hat{b}_y)$ is uniformly bounded for all $t \le T$. Given $\epsilon = o(\sqrt{T})$, this yields estimation error terms $w_t \le C\sqrt{T}$ in each round, for small enough C to obtain the obtain the desired regret bound. ■

Appendix G. Bandit Feedback: Analysis for NESTEDBCO

We first state the FKM algorithm and its bounds for regret and per-round step size.

Algorithm 6 FKM (Flaxman et al., 2004)

Input: decision set $\mathcal{K}$ containing $\mathbf{0}$, set $v_1 = \mathbf{0}$, parameters $\eta, \tilde{\delta}$.
Let $v_1 \in \text{int}(\mathcal{K})$ such that $\nabla \mathcal{R}(v_1) = 0$,
for $t = 1$ to T **do**
  Draw $u_t \in \mathbb{S}$ uniformly, set $y_t = v_t + \tilde{\delta} u_t$
  Play y_t, observe loss $f_t(y_t)$, set $g_t = \frac{n}{\tilde{\delta}} f_t(y_t) u_t$
  Update $v_{t+1} = \Pi_{\mathcal{K}_{\tilde{\delta}}}[v_t - \eta g_t]$, where $\mathcal{K}_{\tilde{\delta}} = \{(1-\tilde{\delta})v : v \in \mathcal{K}\}$
end for

Proposition 24 (Flaxman et al. (2004)) *For L-Lipschitz convex losses and a domain $\mathcal{K}$ with diameter $2R$ which contains a ball of radius r around the origin,* FKM *obtains expected regret*

$$\text{Reg}_T(\text{FKM}) \le \eta \frac{n^2}{\tilde{\delta}^2} T + \frac{4R^2}{\eta r^2} + \frac{8\tilde{\delta} RLT}{r},$$

with each point y_t contained in $\mathcal{K}$. Further, each pair of consecutive points y_t, y_{t+1} chosen by FKM *satisfies $\|y_{t+1} - y_t\| \le 2\tilde{\delta} + \frac{\eta n L}{\tilde{\delta}}$.*

The NESTEDBCO algorithm is essentially equivalent to NESTEDOCO, replacing FTRL with FKM and recalibrating parameters.

Proof of Theorem 11. Following the proof of Theorem 5, to apply the bound of FKM to our setting (along with excess regret at most δLR per round from contracting $\mathcal{Y}$ to $\widetilde{\mathcal{Y}}$), the key step is to show that each point selected by FKM is feasible under weakly locally controllable dynamics over $\widetilde{\mathcal{Y}}$, i.e. $\|y_{t+1} - y_t\| \le r\delta\rho$. Let $\tilde{\delta} = \frac{1}{T^{1/4}} = r\delta\rho/4$, and let $\eta = \frac{R}{2nrLT^{3/4}}$. Assume for simplicity that $r \le 1$ and $T^{1/4} \ge \frac{R}{r}$. When instantiating FKM over $\widetilde{\mathcal{Y}}$ with parameters η and $\tilde{\delta}$, by Proposition 24

Algorithm 7 Nested Bandit Convex Optimization (NESTEDBCO).

Let $\tilde{\delta} = \frac{1}{T^{1/4}} = r\delta\rho/4$, let $\eta = \frac{R}{2nrLT^{3/4}}$
Let $\widetilde{\mathcal{Y}} = \{y : \frac{1}{1-\delta} y \in \mathcal{Y}\}$
Initialize FKM to run for T rounds over $\widetilde{\mathcal{Y}}$ with parameters $\eta, \tilde{\delta}$
for $t = 1$ to T **do**
  Let y^* be the point chosen by FKM
  Use $\texttt{Oracle}(y_{t-1}, y^*)$ to compute $x_t = \operatorname{argmin}_x \|D_t(x, y_{t-1}) - y^*\|^2$
  Play action x_t
  Observe y_t and loss $f_t(y_t)$, update SCRIBLE
end for

we then have

$$\begin{aligned}
\|y_{t+1} - y_t\| &\le 2\tilde{\delta} + \frac{\eta n L}{\tilde{\delta}} \\
&\le r\delta\rho/2 + \left(\frac{R}{2nrLT^{3/4}}\right)\frac{nL}{\tilde{\delta}} \\
&\le r\delta\rho/2 + \tilde{\delta}/2 \\
&\le r\delta\rho,
\end{aligned}$$

and so each selected point is feasible. This allows us to bound our regret by

$$\begin{aligned}
\mathrm{Reg}_T(\text{NESTEDBCO}) &= \mathrm{Reg}_T(\text{FKM}) + \delta LRT \\
&= \eta\frac{n^2}{\tilde{\delta}^2}T + \frac{4R^2}{\eta r^2} + \frac{8\tilde{\delta}LRT}{r} + \delta LRT \\
&= \eta\frac{16n^2}{r^2\delta^2\rho^2}T + \frac{4R^2}{\eta r^2} + 2\delta\rho LRT + \delta LRT && (\tilde{\delta} = r\delta\rho/4) \\
&\le 16\eta n^2 T^{3/2} + \frac{4R^2}{\eta r^2} + \frac{12LRT^{3/4}}{r\rho} && (\delta = \tfrac{4}{r\rho T^{1/4}}, r \le 1) \\
&\le \frac{16nLRT^{3/4}}{r} + \frac{12LRT^{3/4}}{r\rho} && (\eta = \tfrac{R}{2nrLT^{3/4}}) \\
&= O\left(nRLT^{3/4}(r\rho)^{-1}\right).
\end{aligned}$$

■

Appendix H. Background and Proofs for Section 4.1: Performative Prediction

H.1. Background

Introduced by Perdomo et al. (2020), the Performative Prediction problem captures settings in which the data distribution for which a classifier is deployed may shift as a function of the classifier itself, notably including strategic classification Hardt et al. (2015) as well as problems related to reinforcement learning and causal inference. While a number of extensions of strategic classification to

online settings have been considered Dong et al. (2018); Zrnic et al. (2021b); Ahmadi et al. (2023), the bulk of the literature on performative prediction considers settings with a fixed loss function and distribution "update map" Perdomo et al. (2020); Miller et al. (2021); Jagadeesan et al. (2022b); Mendler-Dünner et al. (2020); Piliouras and Yu (2022); Brown et al. (2022), where the update map may sometimes depend on the current distribution (as in the Stateful Performative Prediction setting of Brown et al. (2022)). For the *location-scale* family of update maps introduced by Miller et al. (2021) (and additionally explored by Jagadeesan et al. (2022b) from a regret minimization perspective), which yields a convex "performative risk" objective function, a formulation of Online Performative Prediction is given by Kumar et al. (2022) as an application of online convex optimization with unbounded memory, in which the classification loss function may change over time and the distribution updates may occur gradually.

Here, we generalize the problem formulation of Kumar et al. (2022) to also accommodate notions of statefulness similar to that in Brown et al. (2022). In particular, the instances we consider will resemble location-scale maps when restricting attention only the performatively stable classifiers for each distribution, yet the update effect of a non-stable classifier may be distribution-dependent and nonlinear, provided that the update map satisfies local controllability (viewing classifiers as actions and distributions as states) and mild regularity properties (e.g. invertibility and Lipschitz conditions).

H.2. Model

In the setting of Online Performative Prediction we consider, as formulated by Kumar et al. (2022), in each round $t \in [T]$ we deploy some classifier x_t, and observe samples from some distribution p_t, which may change dynamically as a function of the history of interactions. Here, we take $\mathcal{X} \subseteq \mathbb{R}^n$ as our space of classifiers, e.g. representing weight vectors for regression, which we assume is bounded and convex. The initial data distribution is given by some distribution p_0 over $\mathbb{R}^n$. In each round, upon deploying a classifier x_t, the distribution is updated according to

$$p_t = (1-\theta)p_{t-1} + \theta\, \mathcal{D}(x_t, y_{t-1}),$$

for $\theta \in (0, 1]$, where $\mathcal{D}(x_t, y_{t-1})$ is the distribution *update map* taking as input our classifier x_t and some representation of the *state* $y \in \mathcal{Y}$, where we assume $\mathcal{Y} \subseteq \mathbb{R}^n$ is convex, contains $\mathcal{B}_r(\mathbf{0})$, is bounded with radius R, and that $y_0 = 0$. We make the following assumptions on $\mathcal{D}$.

Assumption 1 *We assume the distribution update map $\mathcal{D}(x, y)$ operates as follows:*

- *$\mathcal{D}(x, y) = A(x, y) + \xi$, with $A : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$,*
- *ξ is a random variable in $\mathbb{R}^n$ with mean μ and covariance Σ,*
- *$A(x, y)$ satisfies ρ-local controllability and has an inverse action mapping $X(y, y^*)$ where*

  $$A(X(y, y^*), y) = y^*,$$

  defined over feasible pairs, which is L_y-Lipschitz in y (when feasibility of y^* holds), and
- *There is a linear invertible function $s : \mathcal{X} \to \mathcal{Y}$ such that $A(x, y) = s(x)$ if $y = s(x)$, where $s^{-1} : \mathcal{Y} \to \mathcal{X}$ is S-Lipschitz.*

Further, $A(x, y)$ is known and ξ can be sampled freely.

The inverse action mapping assumption simply enforces that classifiers need not change drastically to have the same update effect under small changes to the state. The final assumption imposes a linear structure over *performatively stable* classifiers (i.e. classifiers for which the resulting distribution will remain fixed under $\mathcal{D}$, as formulated by Perdomo et al. (2020)), but we note that the distribution may update in an arbitrarily nonlinear fashion (subject to the other conditions) when x_t is not a performatively stable classifier for the distribution induced by the previous state y_{t-1}. The ability to accommodate a state component is reminiscent of prior work involving notions of statefulness in performative prediction such as Brown et al. (2022). Our setting generalizes that of Kumar et al. (2022), in which the map A is taken to be a fixed matrix. For any nonsingular matrix A there is immediately a linear map $s(x) = A^{-1}x$, and local controllability can be defined in terms of the largest and smallest absolute eigenvalues of A (as a special case of our Example 1 with a fixed matrix). We view the nonsingularity assumption (and invertibility in the more general case) as fairly mild, as it amounts to assuming that the distribution map can depend on all parameters of classifier without any necessary (linear) dependency structure imposed, and that no two classifiers are equivalent only to the population but not the optimizer (as otherwise one could simply reduce dimensionality of $\mathcal{X}$). However, even in the case where A is singular, we note that this issue is resolvable augmenting the state representation y_t to incorporate the choice of free classifier parameters which affect loss but not distribution updates (e.g. by adding a vector w_t to y_t which is orthogonal to the range of A and linear in x_t). We assume invertibility here for simplicity, and we take $\mathcal{Y}$ to be simply be given by the range of s over $\mathcal{X}$. At each round t, some scoring function $f_t(x, z)$ is chosen adversarially, and our loss is then given by

$$\tilde{f}_t(x_t, p_t) = \mathop{\mathbb{E}}_{z\sim p_t}[f_t(x_t, z)].$$

We assume each f_t is convex and L_z-Lipschitz in both x and z, and that $p_0 = y_0 + \xi$. We measure our regret with respect to the best performatively stable classifier, i.e. the loss of any classifier as if were held constant indefinitely as the distribution updates. We define our regret as follows:

$$\mathrm{Reg}_T(\mathcal{A}) = \max_{x^*} \sum_{t=1}^{T} \tilde{f}_t(x_t, p_t) - \tilde{f}_t(x^*, \mathcal{D}(x^*, s(x^*)))$$

Here, the role of $s(x^*)$ captures the convergence of the distribution to a stable point, resulting from taking the limit of the distribution update rule as t grows large.

As in many of the applications we consider, here our loss is determined both by our action (the classifier) and the state (in terms of the distribution). Our approach for casting Online Performative Prediction as an instance of online nonlinear control in our framework will be to define appropriate surrogate convex losses which depend only on the state, over which we run NESTEDOCO. Here, these will correspond to losses only over the updated distribution component $\mathcal{D}(x_t, y_{t-1})$, which we show closely track our true incurred loss.

H.3. Analysis

For each round t, define the surrogate loss $f_t^*(y)$ as:

$$f_t^*(y) = \mathop{\mathbb{E}}_{z\sim y_t+\xi}\left[f_t(s^{-1}(y), z)\right].$$

Lemma 25 *Each $f_t^*(y)$ is convex and $(1+S)L_z$-Lipschitz in y.*

Proof Consider any individual sample $v \sim \xi$. We can then view $g(y) = (s^{-1}(y), y+v)$ as a vector-valued function which is $(1+S^*)$-Lipschitz. The function $f_t(g(y))$ is a L_z-Lipschitz and convex function of this linear function of y, and thus $f_t(s^{-1}(y), y+v)$ is convex and $(1+S^*)L_z$-Lipschitz in y. The function $f_t^*(y)$ is an average of such functions, taken over the expectation of ξ, and thus is convex and $(1+S^*)L_z$-Lipschitz in y as well. ■

Observe that $f_t^*(y) = \tilde{f}_t(s^{-1}(y), \mathcal{D}(s^{-1}(y), y)$. We will run NESTEDOCO for these losses over the ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, A)$, where we can track the current state $y_t = A(x_t, y_{t-1})$ at each step as a function of our past actions given knowledge of A, and can compute gradients of $f_t^*(y_t)$ to arbitrary desired precision by sampling from ξ. This will yield the regret bound from Theorem 5 with respect to the surrogate losses, and the key challenge will be to analyze our error between the true and surrogate losses.

Lemma 26 *For any round t we have that*

$$\tilde{f}_t(x_t, p_t) - f_t^*(y_t) \leq (1-\theta)^h M + \frac{\eta L_z(1+S)}{\gamma} \cdot \left(L_y + \frac{1-\theta}{\theta}\right)$$

Proof For any $h < t$, the loss of x_t over the distribution $y_{t-h} + \xi = \mathcal{D}(x_{t-h}, y_{t-h-1})$ can be expressed as

$$\hat{f}_t(x_t, y_{t-h}) = \mathop{\mathbb{E}}_{z \sim \xi + y_{t-h}} \left[f_t(x_t, z)\right],$$

which is convex and L_z-Lipschitz in both parameters when taking the expectation over ξ. For round t in isolation, using the inverse action mapping bound and the bound on $\|y_t - y_{t-1}\|$ from Proposition 18 we have that

$$\begin{aligned}
\hat{f}_t(x_t, y_t) - f_t^*(y_t) &= \hat{f}_t(x_t, y_t) - \hat{f}_t(s^{-1}(y_t), y_t) \\
&= \hat{f}_t(X(y_{t-1}, y_t), y_t) - \hat{f}_t(X(y_t, y_t), y_t) \\
&\leq \frac{\eta L_y L_z}{\gamma},
\end{aligned}$$

and further for previous states that

$$\hat{f}_t(x_t, y_{t-h}) - f_t^*(y_t) = (L_y + h)\frac{\eta L_z(1+S)}{\gamma}.$$

We can decompose the distribution p_t into updates from past rounds as

$$p_t = (1-\theta)^t p_0 + \sum_{h=0}^{t-1} \theta(1-\theta)^h \, \mathcal{D}(x_{t-h}, y_{t-h-1})$$

which then yields a loss discrepancy of at most

$$\begin{aligned}
\tilde{f}_t(x_t, p_t) - f_t^*(y_t) &\leq (1-\theta)^t f_t(x_t, p_0) + \frac{\eta L_z(1+S)}{\gamma} \left(\sum_{h=0}^{t-1} \theta(1-\theta)^h (L_y + h)\right) \\
&\leq \frac{\eta L_z(1+S)}{\gamma} \cdot \left(L_y + \frac{1-\theta}{\theta} + (1-\theta)^t\right)
\end{aligned}$$

between the true and surrogate loss for round t. ■

We can now bound the cumulative regret of NESTEDOCO for the problem.

Theorem 27 *For any $\theta > 0$, when Assumption 1 holds for the distribution update rule, Online Performative Prediction can be cast as a ρ-locally controllable instance of online control with nonlinear dynamics, for which* NESTEDOCO *obtains regret*

$$\mathrm{Reg}_T(\textsc{NestedOCO}) \leq 2\sqrt{\frac{(1 + L_y + \frac{R}{r\rho} + \frac{2-\theta}{\theta})TGL_z^2(1+S)^2}{\gamma}}$$

with respect to the best performatively stable classifier classifier.

Proof Combining the previous results with Theorem 5, we have that for any $x^* \in \mathcal{X}$ our regret is at most

$$\begin{aligned}\sum_{t=1}^{T} \tilde{f}_t(x_t, p_t) - \tilde{f}_t(\mathcal{D}(x^*, s(x^*))) &\leq \sum_{t=1}^{T} \hat{f}_t(y_t) - \tilde{f}_t(x^*, \mathcal{D}(x^*, s(x^*))) + \sum_{t=1}^{T} \tilde{f}_t(x_t, p_t) - f_t^*(y_t) \\ &\leq \eta\left(1 + L_y + \frac{2-\theta}{\theta} + \frac{R}{r\rho}\right)\frac{TL_z(1+S)}{\gamma} + \frac{G}{\eta} \\ &= 2\sqrt{\frac{(1 + L_y + \frac{R}{r\rho} + \frac{2-\theta}{\theta})TGL_z^2(1+S)^2}{\gamma}}\end{aligned}$$

upon setting $\eta = \sqrt{\frac{G\gamma}{(1+L_y+\frac{R}{r\rho}+\frac{2-\theta}{\theta})TL_z^2(1+S)^2}}$. ■

Theorem 12 follows directly from Theorem 27. For Online Performative Prediction, in the full generality of the setting considered, the per-round optimization problem may not be convex, in which case we make use of the non-convex optimization oracle access for NESTEDOCO. However, in each of the following applications we show that the action selection step can indeed be implemented efficiently without imposing additional restrictions on the dynamics.

Appendix I. Background and Proofs for Section 4.2: Adaptive Recommendations

I.1. Background

Motivated by problems involving preference dynamics and feedback loops in recommendation systems (see e.g.Flaxman et al. (2016)), a number of recent works Hazla et al. (2019); Gaitonde et al. (2021); Dean and Morgenstern (2022); Jagadeesan et al. (2022a); Agarwal and Brown (2022, 2023) have explored models of repeated recommendation where given to an agent whose preferences or opinions evolve over time. Several of these models Hazla et al. (2019); Dean and Morgenstern (2022); Jagadeesan et al. (2022a) consider population-level effects for settings where a single recommendation is given each round and consumers (or producers) update their behavior according to linear dynamics. Nonlinear preference dynamics with *menus* of recommendations for a single agent are considered in Agarwal and Brown (2022, 2023), where the aims to minimize regret for adversarial losses over the agent's choices. The Adaptive Recommendations formulation of Agarwal and

Brown (2022) somewhat resembles the "Dueling Bandits" setting of Yue et al. (2012), where $k > 1$ actions are chosen in each round, yet where preferences can now evolve dynamically as a function of the history rather than remaining fixed. Whereas Agarwal and Brown (2022, 2023) study a bandit formulation of the problem with unknown preference dynamics, here we consider a full-feedback model with known dynamics, allowing for relaxed structural assumptions (on the agent's "memory horizon" and "preference scoring functions") at the cost of stronger informational assumptions, while maintaining the overall dynamics of the problem.

I.2. Model

Here, we are tasked with repeatedly recommending menus of content to an agent. Out of a universe of n elements (e.g. video channels, clothing items), we show a subset of size k (denoted K_t) to the agent in each round, for T total rounds. The agent chooses one item $i \in K_t$ from the menu, according to a distribution in terms of their *preferences*, which are a function of their selection history. Conditioned on being shown a menu K_t, the agent's choice distribution has positive mass only on the k items $i \in K_t$. The agent's representation of their selection history is given by their *memory vector* $v_t \in \Delta(n)$, and choices are determined by their *preference scoring functions* $s_i : \Delta(n) \to [\lambda, 1]$ for each i, which map the agent's memory vector to relative preference scores for each item. The menu we show to the agent may be chosen from some distribution $x_t \in \Delta(\binom{n}{k})$, and for each $K_t \in [\binom{n}{k}]$ the agent's menu-conditional distribution $p_t(\cdot; K_t, v_{t-1}) \in \Delta(n)$ is proportional to the scores $s_i(v_t)$ for items in K_t, given as

$$p_t(i; K_t, v_{t-1}) = \frac{s_i(v_{t-1})}{\sum_{j \in K_t} s_j(v_{t-1})}$$

for each $i \in K_t$, with $p_t(j; K_t, v_{t-1}) = 0$ for $j \notin K_t$. The joint item choice distribution, considering both random selection of a menu K_t according to x_t, and the agent's choice from K_t, is given by

$$p_t(\cdot; x_t, v_{t-1}) = \sum_{K_t \in \binom{n}{k}} x_t(K_t) \cdot p_t(\cdot; K_t, v_{t-1})$$

which we may denote simply by the vector $p_t \in \Delta(n)$, or as a function $p_t(x_t)$. In contrast to prior work, here we consider a deterministic variant of the problem as an illustration of the flexibility of our framework for online nonlinear control. In particular, we assume that the agent's memory vector v_t updates according to its expectation over p_t as

$$v_t = (1 - \theta_t)v_{t-1} + \theta_t p_t,$$

where $\theta_t \in [\theta, 1]$ is the per-round update speed, and we assume that the agent's scoring functions s_i are known. We receive convex and L-Lipschitz losses $f_t(p_t)$ in each round in terms of the agent's choices, over which we aim to minimize regret with respect to some distribution set $\mathcal{Y} \subseteq \Delta(n)$.

The prior work (Agarwal and Brown, 2022, 2023) has considered two particular subsets of $\Delta(n)$ as regret benchmarks. We show that both can be cast as locally controllable instances of online control, and further, we make use of local controllability to give a general characterization of convex sets $\mathcal{Y} \subseteq \Delta(n)$ over which sublinear regret is attainable. We recall some key definitions and results from (Agarwal and Brown, 2022, 2023).

Definition 28 (Instantaneously Realizable Distributions) *The set of instantaneously realizable distributions at a memory vector $v \in \Delta(n)$ is given by*

$$\mathrm{IRD}(v) = \mathrm{convhull}\left\{p(\cdot; K, v) : K \in \left[\binom{n}{k}\right]\right\}.$$

Each such set $\mathrm{IRD}(v_{t-1})$ corresponds to the feasible distributions p_t, given the agent's scoring functions and memory v_{t-1}. It is shown by Agarwal and Brown (2023) that each IRD sets can be directly characterized in terms of the ratios between target frequencies and scores.

Proposition 29 (Menu Times for IRD Agarwal and Brown (2023)) *Given a memory vector $v \in \Delta(n)$ and target distribution $p \in \Delta(n)$, let the menu time μ_i for item i be given by*

$$\mu_i = \frac{k \cdot \frac{p(i)}{s_i(v)}}{\sum_{j=1}^{n} \frac{p(j)}{s_j(v)}},$$

where $\sum_{i=1}^{n} \mu_i = k$. Then, $p \in \mathrm{IRD}(v)$ if and only if $\mu_i \leq 1$ for each $i \in [n]$.

We recall the prior benchmark sets considered, and the corresponding assumptions which yield feasibility of regret minimization. We state informal analogues of the prior results as translated to our setting, which we then show formally below.

Definition 30 (Everywhere Instantaneously Realizable Distributions) *The set of everywhere instantaneously realizable distributions is given by*

$$\mathrm{EIRD} = \bigcap_{v \in \Delta(n)} \mathrm{IRD}(v).$$

Proposition 31 (Corollary of Agarwal and Brown (2022)) *If $\lambda \geq \frac{k}{n} + \frac{k}{n(n-1)}$, then* EIRD *is non-empty, and there is a $o(T)$ regret algorithm with respect to any distribution $p \in$* EIRD.

Distributions $p_t \in$ EIRD are always feasible regardless of v_{t-1} by an appropriate choice of x_t, but EIRD may be quite small in relation to $\Delta(n)$. Under stronger assumptions for each s_i, a potentially much larger set becomes feasible as a regret benchmark.

Definition 32 (ϕ-Smoothed Simplex) *The ϕ-smoothed simplex $\Delta^{\phi}(n)$ for $\phi \in [0, 1]$ is given by*

$$\Delta^{\phi}(n) = \{(1 - \phi)v + \phi \mathbf{u}_n : v \in \Delta(n)\}$$

Definition 33 (Scale-Bounded Functions) *A scoring function $s_i : \Delta(n) \to [\frac{\lambda}{\sigma}, 1]$ is said to be (σ, λ)-scale-bounded for $\sigma > 1$ and $\lambda > 0$ if, for all $v \in \Delta(n)$, we have that*

$$\sigma^{-1}((1 - \lambda)v_i + \lambda) \leq s_i(v) \leq \sigma((1 - \lambda)v_i + \lambda).$$

For such functions, each score $s_i(v)$ cannot be too far from item i's weight in memory, and it is shown that $\mathrm{IRD}(v)$ contains a ball around v for each $v \in \Delta^{\phi}(n)$, for an appropriate choice of ϕ.

Proposition 34 (Corollary of Agarwal and Brown (2023)) *If each s_i is (σ, λ)-scale-bounded, then there is a $o(T)$ regret algorithm with respect to any distribution $p \in \Delta^{\phi}(n)$, for $\phi = \Theta(k\lambda\sigma^2)$.*

We extend these results to general convex benchmark sets $\mathcal{Y} \subseteq \Delta(n)$, where we can characterize the feasibility of regret minimization via local controllability using the menu times μ_i. When ρ-local controllability holds over a set $\mathcal{Y}$, we can minimize regret via NESTEDOCO using surrogate losses $f_t^*(v_t)$, which closely track our true losses $f_t(p_t)$.

I.3. Analysis

We make use of the menu time quantities μ_i for a memory vector v and target distribution p to translate our notion of local controllability to the Adaptive Recommendations setting. Let $\mathcal{Y}$ be any convex subset of $\Delta(n)$, let $\mathcal{X} = \Delta(\binom{n}{k})$, where the dynamics $D_t(x_t, v_{t-1})$ are given by

$$D_t(x_t, v_{t-1}) = (1-\theta_t)v_{t-1} + \theta_t p_t(x_t).$$

Note that $D_t(x_t, v_{t-1})$ is action-linear in x_t, and thus we can solve for x_t efficiently (in terms of $\dim(\mathcal{X}) = O(n^k)$); further, there is a construction given in Agarwal and Brown (2023) for removing exponential dependence on k when computing menu distributions. We consider $\mathcal{Y}$ as an $(n-1)$-dimensional subset of $\mathbb{R}^n$, where we define the the ball $\mathcal{B}_\rho(v)$ of radius ρ around a point $v \in \mathcal{Y}$ as:

$$\mathcal{B}_\rho(v) = \{p \in \Delta(n) : \|p - v\| \leq \rho\}.$$

Theorem 35 *An instance of Adaptive Recommendations $(\mathcal{X}, \mathcal{Y}, D)$ satisfies $\rho\theta$-local controllability if, for any $v \in \mathcal{Y}$ and $p \in \mathcal{B}_{\rho\cdot\pi(v)}$, we have that*

$$\frac{(k-1)p(i)}{s_i(v)} \leq \sum_{j\neq i}^{n} \frac{p(j)}{s_j(v)}$$

for every $i \in [n]$.

This follows immediately from Proposition 31 and the definition of local controllability, which can analogously extend to strong local controllability. We can use this formulation to unify the feasibility analysis for each of the previously considered sets.

Lemma 36 *For $\lambda \geq \frac{k-1}{n-1} + \epsilon$ and $\epsilon \geq 0$, the* EIRD *set contains a ball of radius $\rho = \Theta(\frac{\epsilon}{nk+\epsilon})$ around $\mathbf{u}_n$, and any instance $(\mathcal{X}, \text{EIRD}, D)$ satisfies θ-local controllability.*

Proof For any $v \in \Delta(n)$, $i \in [n]$, and $p \in \mathcal{B}_\rho(\mathbf{u}_n)$ we have $p(i) \leq \frac{1}{n} + \frac{\rho\sqrt{2}}{2}$ and $s_i(v) \geq \frac{k-1}{n-1} + \epsilon$, yielding that

$$\frac{(k-1)p(j)}{s_j(v)} \leq \frac{1 + \frac{\rho n\sqrt{2}}{2}}{\frac{n}{n-1} + \frac{\epsilon n}{k-1}},$$

and over all items $j \neq i$ (with $s_j(v) \leq 1$) we have

$$\sum_{j\neq i}^{n} \frac{p(j)}{s_j(v)} \geq 1 - \frac{1}{n} - \frac{\rho\sqrt{2}}{2}.$$

Observe that the bounds for each term are equalized at $\frac{n-1}{n}$ when $\rho = \epsilon = 0$, and so $\mathbf{u}_n \in$ EIRD whenever $\lambda \geq \frac{k-1}{n-1}$. We can specify $\epsilon(\rho)$ in terms of ρ to maintain equality, and thus inclusion of

$p \in \text{EIRD}$. Taking $\epsilon(\rho)$ in terms of ρ as

$$\begin{aligned}\epsilon(\rho) &= \frac{\rho n(k-1)}{\frac{2(n-1)}{\sqrt{2}n} - \rho} \\ &= \frac{\frac{\rho n(k-1)\sqrt{2}}{2}}{\left(1 - \frac{1}{n} - \frac{\rho\sqrt{2}}{2}\right)} \\ &= (k-1)\left(\frac{\frac{1}{n} + \frac{\rho\sqrt{2}}{2}}{1 - \frac{1}{n} - \frac{\rho\sqrt{2}}{2}} - \frac{1}{n-1}\right)\end{aligned}$$

gives us that

$$\frac{1}{n-1} + \frac{\epsilon(\rho)}{k-1} \geq \frac{\frac{1}{n} + \frac{\rho\sqrt{2}}{2}}{1 - \frac{1}{n} - \frac{\rho\sqrt{2}}{2}}$$

for $\rho \geq 0$, and so we maintain that $p \in \text{EIRD}$. Inverting, we have

$$\rho(\epsilon) = \frac{\epsilon\frac{2(n-1)}{\sqrt{2}n}}{n(k-1) + \epsilon}$$

as the radius of a ball around $\mathbf{u}_n$ contained in EIRD. To see that EIRD is θ-locally controllable, consider any v_{t-1} and v^* in EIRD where $v^* \in \mathcal{B}_{\pi(v_{t-1})}(v_{t-1})$, and let $v_t = (1-\theta_t)v_{t-1} + \theta_t v^*$. By playing an action distribution x_t which induces $p_t(x_t) = v^*$, the memory vector is then updated to v_t. This is feasible for any $v_t \in \mathcal{B}_{\theta\cdot\pi(v_{t-1})}(v_{t-1})$, as each corresponds to some $v^* \in \mathcal{B}_{\pi(v_{t-1})}(v_{t-1})$. ■

We remark that for the EIRD set, if losses are given over p_t rather than v_t, one can define dynamics which directly consider the state to simply be the induced distribution p_t in each round, which satisfies strong local controllability with any $p_t \in \text{EIRD}$ feasible at each round; in general, we consider dynamics to view the memory vector as the state, as the feasible updates p_t are a function of v_t. Such is the case for the ϕ-smoothed simplex, for which we can state an analogous local controllability result.

Lemma 37 *If each s_i is (σ, λ)-scale-bounded, then any instance $(\mathcal{X}, \Delta^\phi(n), D)$ over the ϕ-smoothed simplex for $\phi = \Theta(k\lambda\sigma^2)$ satisfies $\Omega(\theta\lambda\phi)$-local controllability.*

Proof The following lemma from Agarwal and Brown (2023) shows that a ball of distributions around any memory vector $v \in \Delta^\phi(n)$ is feasible under $\text{IRD}(v)$.

Lemma 38 (IRD for Scale-Bounded Preferences Agarwal and Brown (2023)) *Let each s_i be (σ, λ)-scale-bounded with $\sigma \leq \sqrt{4(n-1)/k}$, and let $v \in \Delta^\phi(n)$ be a vector in the ϕ-smoothed simplex, for $\phi \geq \Theta k\lambda\sigma^2$. Then, $p \in \text{IRD}(v)$ for any vector $p \in \mathcal{B}_{\lambda\phi}(v) \cap \Delta^\phi(n)$.*

Let $d = \min(\lambda\phi, \pi(v_{t-1})) \leq \lambda\phi\pi(v_{t-1})$ for any v_{t-1} in $\Delta^\phi(n)$. Any $v^* \in \mathcal{B}_d(v_{t-1})$ then is contained in $\text{IRD}(v_{t-1})$, and so playing x_t such that $p_t(x_t) = v^*$ yields an update to $v_t = (1 -$

$\theta_t)v_{t-1} + \theta v^*$, which is feasible for any $v_t \in \mathcal{B}_{d\theta}(v_{t-1})$, and so $\Omega(\theta\lambda\phi)$-local controllability holds. ■

For any such set $\mathcal{Y}$ which yields locally controllable dynamics for the instance $(\mathcal{X}, \mathcal{Y}, D)$, we can minimize regret over $\mathcal{Y}$ via NESTEDOCO, where we optimize with respect to the surrogate losses $f_t^*(v_t)$. Note that for our regret benchmark of the best per-round instantaneously distribution in $\mathcal{Y}$, any fixed vector v^* which is instantaneously targeted across all rounds yields an item distribution $p_t = v^*$ in each round, and so $f_t^*(v^*) = f_t(p^*)$. We assume that y_0 is bounded inside $\mathcal{Y}$ (which typically will hold for $y_0 = \mathbf{u}_n$).

Theorem 39 *For any ρ-locally controllable instance $(\mathcal{X}, \mathcal{Y}, D)$ of Adaptive Recommendations with update speed $\theta > 0$, running* NESTEDOCO *over the surrogate losses $f_t^*(v_t)$ yields regret*

$$\text{Reg}_T(\text{NESTEDOCO}) \leq 2\sqrt{\frac{(2 + \frac{R}{r\rho} + \frac{1}{\theta})TGL^2}{\gamma}}$$

with respect to the true losses $f_t(p_t)$ over $\mathcal{Y}$.

Proof Beyond applying the regret bound for NESTEDOCO from Theorem 5, the key step here is to bound surrogate loss errors as:

$$\begin{aligned}
\sum_{t=1}^{T} f_t(p_t) - f_t(v^*) &\leq \sum_{t=1}^{T} f_t^*(v_t) - f_t(v^*) + \sum_{t=1}^{T} f_t(v_t) - f_t(p_t) \\
&\leq \eta\left(1 + \frac{R}{r\rho}\right)\frac{TL^2}{\gamma} + \frac{G}{\eta} + \sum_{t=1}^{T} f_t(v_t) - f_t\left(\frac{v_t - (1-\theta_t)v_{t-1}}{\theta_t}\right) \\
&\leq \eta\left(1 + \frac{R}{r\rho}\right)\frac{TL^2}{\gamma} + \frac{G}{\eta} + \sum_{t=1}^{T} f_t(v_t) - f_t\left(v_{t-1} + \frac{v_t - v_{t-1}}{\theta_t}\right) \\
&\leq \eta\left(1 + \frac{R}{r\rho}\right)\frac{TL^2}{\gamma} + \frac{G}{\eta} + L\left(1 + \frac{1}{\theta}\right)\sum_{t=1}^{T}\|v_t - v_{t-1}\| \\
&\leq \eta\left(2 + \frac{R}{r\rho} + \frac{1}{\theta}\right)\frac{TL^2}{\gamma} + \frac{G}{\eta} \\
&= 2\sqrt{\frac{(2 + \frac{R}{r\rho} + \frac{1}{\theta})TGL^2}{\gamma}}
\end{aligned}$$

upon setting $\eta = \sqrt{\frac{G\gamma}{(2+\frac{R}{r\rho}+\frac{1}{\theta})TL^2}}$, which yields the theorem. ■

Theorems 13 and 14 follow from Theorem 39, as well as from Lemmas 36 and 37, respectively.

Appendix J. Background and Proofs for Section 4.3: Adaptive Pricing

J.1. Background

While there is a large literature on designing online mechanisms for pricing discrete goods via auctions (Mehta et al., 2007; Kanoria and Nazerzadeh, 2020; Golrezaei et al., 2020; Morgenstern and

Roughgarden, 2016; Feng et al., 2019; Braverman et al., 2017), there is comparatively little work related to online pricing problems for real-valued goods. Most work for such problems to date requires strong assumptions on valuation functions, often either assuming linearity (Jia et al., 2014) or additivity (Agrawal et al., 2023), or requiring approximability via discretization (Mussi et al., 2022). Here, we introduce a novel formulation for an Adaptive Pricing problem which builds on the myopic-demand fixed-cost setting of Roth et al. (2015), which we extend to accommodate adversarial *consumption rates* for the agent (which affect demand, as a function of the agent's *reserves*) as well as adversarial production costs. As in Roth et al. (2015), our setting can accommodate general convex (increasing) production cost functions and concave (increasing) valuations for the agent, provided that valuations additionally are homogeneous; to our knowledge, this encompasses a much wider class of valuations and costs than considered by any prior work on no-regret dynamic pricing for real-valued goods.

J.2. Model

In each round t, an agent (the *consumer*) begins with goods reserves $y_{t-1} \in \mathbb{R}^n_{\geq 0}$ (with $y_0 = \mathbf{0}$), then consumes an adversarially chosen fraction $\theta_t \in [\underline{\theta}, 1]$ of each good simultaneously (e.g. corresponding to their rate of manufacturing downstream items, using the goods as components), updating their reserves to $(1-\theta_t)y_{t-1}$. We (the *producer*) show the consumer some vector $p_t \in \mathbb{R}^n_+$ of per-unit prices for each good, and the consumer purchases some bundle of goods x_t. The consumer's valuation function for reserves of goods is given by $v : \mathbb{R}^n_+ \to \mathbb{R}_+$, and their selection of $x_t = x^*(p_t, \theta_t, y_{t-1})$ is given by

$$x^*(p_t, \theta_t, y_{t-1}) = \operatorname*{argmax}_{x \in \mathbb{R}^n_+} v(x + (1-\theta_t)y_{t-1}) - \langle p_t, x \rangle.$$

We later discuss behavior of x^* when the argmax is undefined; it will suffice for us to only consider price vectors for which it is defined. This updates the consumer's reserves to $y_t = x_t + (1-\theta_t)y_{t-1}$. Upon seeing the consumer's purchased bundle x_t, we receive their payment $\langle p_t, x_t \rangle$ minus our production cost $c_t(x_t) : \mathbb{R}^n_+ \to \mathbb{R}_+$, where c_t is adversarially chosen. Our utility is then given by

$$f_t(p_t, x_t) = \langle p_t, x_t \rangle - c_t(x_t).$$

We make the following assumptions on production costs c_t and the consumer's valuation v.

Assumption 2 (Production Costs) *We assume that for each c_t, the following hold over $\mathbb{R}^n_+$:*

- *c_t is non-negative, convex, and L_c-Lipschitz,*
- *$\lim_{\epsilon \to 0} c_t(\epsilon \cdot \mathbf{1}) \leq C_0$ for some $C_0 \geq 0$, and*
- *$c_t(x) \geq \phi \|x\| + C_0$ for some $\phi > 0$.*

Further, each c_t is revealed prior to setting prices p_{t+1}.

Assumption 3 (Consumer Valuations) *We assume that the following hold over some set $\mathcal{Y} \subseteq \mathbb{R}^n_+$:*

- *v is non-negative, continuous, and differentiable,*
- *v is strictly concave and increasing,*

- *v is (λ, β)-Hölder continuous for some $\lambda \geq 1$ and $\beta \in (0, 1]$, i.e.*

$$\left|v(y) - v(y')\right| \leq \lambda \left\|y - y'\right\|^{\beta},$$

  and

- *v is homogeneous of degree k for some $k \in (0, 1)$, i.e. $v(by) = b^k v(y)$ for any $b > 0$.*

Further, v is known to the producer.

Given the concavity assumption, we note that it is without loss of generality to assume that $k \in (0, 1)$ for the homogeneity parameter. There are several well-studied valuation families which satisfy these properties for an appropriate set $\mathcal{Y}$; see Roth et al. (2015) for proofs of each example.

Example 3 (Constant Elasticity of Substitution (CES)) *Valuations of the form*

$$v(y) = \left(\sum_{i=1}^{n} \alpha_i y_i^{\kappa}\right)^{\beta},$$

with each $\alpha_i, \kappa, \beta > 0$ and $\kappa, \beta\kappa < 1$, are Hölder continuous, differentiable, strictly concave, non-decreasing, and homogeneous over a convex set in $\mathbb{R}^n_+$.

Example 4 (Cobb-Douglas) *Valuations of the form*

$$v(y) = \prod_{i=1}^{n} y_i^{\alpha_i},$$

with $\alpha_i > 0$ and $\sum_{i=1}^{n} \alpha_i < 1$ are Hölder continuous, differentiable, strictly concave, non-decreasing, and homogeneous over a convex set in $\mathbb{R}^n_+$.

We initially assume that Assumption 3 holds over all of $\mathbb{R}^n_+$, but will restrict our attention to the set $\mathcal{Y} \subseteq \mathbb{R}^n_+$ of bundles where $v(y) \geq \phi \|y\|$ for each $y \in \mathcal{Y}$, and we note that our results can be extended to arbitrary downward-closed convex sets (where $by \in \mathcal{Y}$ for any $y \in \mathcal{Y}$ and $b \in (0, 1]$). In Section J.3 we that show Assumptions 2 and 3 yield several important properties which enable optimization via our framework. We show a unique mapping between price vectors and bundle purchases (for any fixed reserves and consumption rate), that restricting attention to $\mathcal{Y}$ is justified under rationality constraints, and that $\mathcal{Y}$ is convex.

Further, there is some price vector which yields a reserve update to any $y_t \in \mathcal{Y}$ in a neighborhood around y_{t-1}, yielding local controllability. Crucially, we show that there are concave surrogate rewards $f_t^*(y_t)$ which will closely track our true rewards $f_t(p_t, x_t)$, leveraging the following property of homogeneous functions.

Proposition 40 (Euler's Theorem for Homogeneous Functions) *A continuous and differentiable function $v : \mathcal{Y} \to \mathbb{R}_+$ is homogeneous of degree k if and only if*

$$\langle \nabla v(y), y \rangle = k \cdot v(y).$$

We run NESTEDOCO directly over these concave surrogate rewards (by inverting the sign of each), where each p_t can be computed efficiently in terms of y_{t-1} and θ_t, and we show that the surrogate reward distance from our true rewards is bounded. While our rewards will not be Lipschitz over $\mathcal{Y}$ in general, we show that appropriately calibrating our step size yields sublinear regret with dependence on the Hölder continuity parameters. We measure our regret with respect to the set of *stable reserve policies*, i.e. pricing policies where y_t remains constant.

Definition 41 (Regret for Stable Reserve Policies) *Let $\mathcal{P}_{\mathcal{Y}} = \{P_y : y \in \mathcal{Y}\}$ be the set of stable reserve policies, where for any y_{t-1} and θ_t satisfying $(1-\theta_t)y_{t-1} \leq y^*$, playing prices computed by a policy $p_t = P_y^*(y_{t-1}, \theta)$ yields*

$$(1-\theta_t)y_{t-1} + x^*(p_t, \theta_t, y_{t-1}) = y^*.$$

It is straightforward to see that any $P_y^* \in \mathcal{P}_{\mathcal{Y}}$ maintains the invariant that $y_t = y^*$, provided that some such p_t is always feasible.

J.3. Analysis

We show a series of results establishing the key conditions allowing us to formulate this problem as a locally controllable instance of online nonlinear control. We first show that any positive bundle is the unique optimal purchase for some positive price vector.

Lemma 42 *For any reserves $y_{t-1} \in \mathbb{R}^n_{\geq 0}$, consumption rate $\theta_t \in [\theta, 1]$, and vector $y_t \in \mathbb{R}^n_+$ where $y_t > (1-\theta_t)y_{t-1}$ elementwise, the bundle $x_t = y_t - (1-\theta_t)y_{t-1}$ is the unique solution to*

$$x_t = x^*(p_t, \theta_t, y_{t-1})$$

for prices $p_t = \nabla v(y_t)$.

Proof Recall that the consumer's bundle choice is given by

$$x^*(p_t, \theta_t, y_{t-1}) = \operatorname*{argmax}_{x \in \mathbb{R}^n_+} v(x + (1-\theta_t)y_{t-1}) - \langle p_t, x \rangle.$$

Note that $v((1-\theta_t)y_{t-1} + x) - \langle p_t, x \rangle$ is strictly concave in x for any $x \in \mathbb{R}^n_+$, as the gradients

$$\nabla_x v((1-\theta_t)y_{t+1} + x) = \nabla_{y_t} v(y_t)$$

are preserved at each point $y_t = (1-\theta_t)y_{t+1} + x$, and subtracting the linear function $\langle x, p_t \rangle$ does not affect strict concavity. We also have that $p_t \in \mathbb{R}^n_+$ for prices $p_t = \nabla v(y_t)$, as v is strictly concave and non-decreasing. This yields that $v((1-\theta_t)y_{t-1} + x) - \langle p_t, x \rangle$ has a unique global maximum at $x_t = y_t - (1-\theta_t)y_{t-1}$, as $\nabla_x(v((1-\theta_t)y_{t+1} + x) - \langle p_t, x \rangle) = \mathbf{0}$. ■

As such, the argmax for $x^*(p_t, \theta_t, y_{t-1})$ is unique whenever $p_t = \nabla v(y)$ for some $y \in \mathbb{R}^n_+$. We let $p^*(x_t; y_{t-1}, \theta_t) = \nabla v((1-\theta_t)y_{t-1} + x_t)$ denote this price vector which induces a purchase of x_t. For any other price vector p, the maximizing bundle x_t either approaches a point on the boundary of $\mathbb{R}^n_+$, or grows unboundedly. We restrict our attention to bundles contained in $\mathbb{R}^n_+$, and show that the issue of unboundedness is resolved by rationality considerations for the producer. We characterize the per-round rewards of stable reserve policies as concave functions of $y \in \mathbb{R}^n_+$, and show that the optimal such policy corresponds to some state $y^* \in \mathcal{Y}$, where $\mathcal{Y}$ is convex and bounded.

Lemma 43 *The round-t reward of a stable reserve policy P_y corresponding to any $y \in \mathbb{R}^n_+$ is given by a strictly concave function*

$$f_t(P_y) = \theta_t k \cdot v(y) - c_t(\theta_t y).$$

Proof We first note that we can maintain $y_t = y$ in every round by Lemma 42, as $y_0 = \mathbf{0}$ and $(1-\theta_t)y < y$. As such, a bundle $x_t = \theta_t y$ is purchased in each round at prices $\nabla v(y)$, and our reward is given by

$$\begin{aligned} f_t(P_y) &= f_t(p^*(\theta_t y; y, \theta_t), \theta_t y) \\ &= \langle \nabla v(y), \theta_t y \rangle - c_t(\theta_t y) \\ &= \theta_t k \cdot v(y) - c_t(\theta_t y), \end{aligned}$$

where the final step follows from Proposition 40 for homogeneous functions. The function $\theta_t k \cdot v(y)$ is strictly concave, which is preserved upon subtracting the convex function $c_t(\theta_t y)$. ■

Lemma 44 *The set $\mathcal{Y} = \{y \in \mathbb{R}^n_+ : v(y) \geq \phi \|y\|\}$ is convex.*

Proof Consider any two points $y, y' \in \mathcal{Y}$, and let $y'' = ay + (1-a)y'$ for any $a \in [0,1]$. Recall that $y^* \in \mathbb{R}^n_+$ belongs to $\mathcal{Y}$ if and only if $v(y^*) \geq \phi \|y^*\|$. By concavity of v, we have that

$$\begin{aligned} v(y'') &= v(ay + (1-a)y') \\ &\geq av(y) + (1-a)v(y') \\ &\geq \phi \|ay\| + \phi \left\|(1-a)y'\right\| \\ &\geq \phi \left\|ay + (1-a)y'\right\| \\ &= \phi \left\|y''\right\| \end{aligned}$$

and so $y'' \in \mathcal{Y}$, yielding convexity of $\mathcal{Y}$. ■

Lemma 45 *For any $z \in \mathbb{R}^n_+$ where $z \notin \mathcal{Y}$, there is some $y \in \mathcal{Y}$ such that $f_t(P_y) \geq f_t(P_z)$ for any θ_t and c_t.*

Proof Consider some $z \notin \mathcal{Y}$ such that $v(z) = \psi \|z\|$, for $\psi < \phi$, and let $y = \left(\frac{\psi}{\phi}\right)^{1/k} z$. By homogeneity of v, we have that $v(y) = \frac{\phi}{\psi} v(z) = \phi \|z\|$, and so $y \in \mathcal{Y}$ as $\|z\| > \|y\|$. For any round with costs c_t and consumption rate θ_t we then have that:

$$\begin{aligned} f_t(P_y) - f_t(P_z) &= \theta_t k \left(v(y) - v(z)\right) - c_t(\theta_t y) + c_t(\theta_t z) \\ &= \theta_t k \left(\frac{\psi}{\phi} - 1\right) \psi \|z\| - c_t(\theta_t y) + c_t(\theta_t z) && \text{(homogeneity of } v\text{)} \\ &\geq \theta_t k \left(\frac{\psi}{\phi} - 1\right) \psi \|z\| + \theta_t \phi \|z - y\| && \text{(lower bound and convexity of } c_t\text{)} \\ &\geq \theta_t k \left(\frac{\psi}{\phi} - 1\right) \psi \|z\| + \theta_t \left(1 - \left(\frac{\psi}{\phi}\right)^{1/k}\right) \phi \|z\| \\ &\geq \theta_t \left(1 - \frac{\psi}{\phi}\right) \phi \|z\| - \theta_t \left(1 - \frac{\psi}{\phi}\right) \psi \|z\| && (k, \tfrac{\psi}{\phi} < 1) \\ &> 0. && (\phi > \psi) \end{aligned}$$

■

Thus the optimal P_y for any cost and consumption sequence corresponds to some $y \in \mathcal{Y}$. We can also bound the radius of $\mathcal{Y}$.

Lemma 46 *Let $V = \max_{y \in \mathbb{R}^n_+ : \|y\|=1} v(y)$. Then, for every $y \in \mathcal{Y}$ we have that*

$$\|y\| \leq \left(\frac{V}{\phi}\right)^{\frac{1}{1-k}}.$$

Proof Let $y^* = \operatorname{argmax}_{y:\|y\|=1} v(y)$, where we have $v(y^*) = V$. Consider the vector by^* for any $b > 0$. By homogeneity of v, we have that

$$\begin{aligned} v(by^*) &= b^k v(y^*) \\ &= b^k V. \end{aligned}$$

For any $b > \left(\frac{V}{\phi}\right)^{\frac{1}{1-k}}$ we have that

$$\begin{aligned} v(by^*) &= \frac{b}{b^{1-k}} \cdot V \\ &\leq b\phi, \end{aligned}$$

where $\|by^*\| > b$ and thus $by^* \notin \mathcal{Y}$. This holds for all vectors with norm b, as any such vector z will have at most $b^k V$ by homogeneity, which yields the result. ■

The previous result also implies that $by \in \mathcal{Y}$ for any $b < 1$ and $y \in \mathcal{Y}$. We assume that $V > \phi$, which is without loss of generality as we may otherwise take ϕ to be smaller artificially; we assume ϕ is small enough to ensure that $\mathcal{Y}$ contains a ball $\mathcal{B}_1(y_1)$ of radius 1 around some $y_1 \in \mathcal{Y}$, and we let $R = \left(\frac{V}{\phi}\right)^{\frac{1}{1-k}}$. We consider the dynamics to be given by

$$D_t(p_t, y_{t-1}) = (1 - \theta_t) y_{t-1} + x^*(p_t, \theta_t, y_{t-1}).$$

We let $\mathcal{Z} = \mathbb{R}^n_+$ denote our action space of price vectors; while dynamics here are not action-linear, we can still compute our desired action $p_t = \nabla v(y_t)$ efficiently, as we assume we have knowledge of v. While the dynamics depend on θ_t, our choice of action p_t depends only on the target update y_t to the consumer's reserves, by Lemma 42. Further, upon observing x_t, we can solve for θ_t as

$$\theta_t = 1 - \frac{y_t - x_t}{y_{t-1}}$$

for purposes of representing our surrogate losses, which are given by

$$f_t^*(y_t) = \theta_t k \cdot v(y) - c_t(\theta_t y).$$

We now show that the dynamics satisfy local controllability.

Lemma 47 (Local Controllability) *The instance $(\mathcal{Z}, \mathcal{Y}, D_t)$ satisfies θ-local controllability for each round t.*

Proof We show that θ-local controllability holds over all of $\mathbb{R}^n_+$, which implies θ-local controllability over $\mathcal{Y}$ as each distance $\pi(y_{t-1})$ while the feasible update region remains the same. By Lemma 42, any update where $y_t \geq (1-\theta_t)y_{t-1}$ elementwise is feasible. Each $\pi(y_{t-1})$ over $\mathbb{R}^n_+$ is simply the minimum element of y_t, which we denote here by m. Each element of y_{t-1} is decreased by at least θm, and so any y_t in the ℓ_∞ ball of radius $\theta m = \theta\pi(y_{t-1})$, and thus the ℓ_2 ball of radius $\theta\pi(y_{t-1})$, is feasible. ■

We are now ready to analyse the regret of NESTEDOCO for the problem. The remaining key issues to resolve will be the errors between our true and surrogate rewards f_t and f_t^*, as well as the lack of Lipschitz continuity for our rewards. We will make use of more general formulations of the guarantees of FTRL, (see e.g. Hazan (2021)).

Proposition 48 *For a γ-strongly convex regularizer $\psi : \mathcal{Y} \to \mathbb{R}$ where $|\psi(y) - \psi(y')| \leq G$ for all $y, y' \in \mathcal{Y}$, and for convex losses $f_1, \ldots, f_T$, the regret of* FTRL *is bounded by*

$$\mathrm{Reg}_T(\mathrm{FTRL}) \leq \sum_{t=1}^{T}(g_t(y_t) - g_t(y_{t+1})) + \frac{G}{\eta},$$

where $g_t(y) = \langle \nabla_t f_t(y_t), y\rangle$ and $g_t(y_t) - g_t(y_{t+1}) \geq \frac{\gamma}{\eta}\|y_{t+1} - y_t\|^2$.

We show that this implies a regret bound for (λ, β)-Hölder continuous convex losses, recovering the λ-Lipschitz bounds when $\beta = 1$.

Theorem 49 *For (λ, β)-Hölder continuous convex losses,* FTRL *with obtains regret bounded by*

$$\mathrm{Reg}_T(\mathrm{FTRL}) \leq T\lambda\left(\frac{\eta\lambda}{\gamma}\right)^{\beta/(2-\beta)} + \frac{G}{\eta}$$

and chooses points which satisfy $\|y_{t+1} - y_t\| \leq \left(\frac{\eta\lambda}{\gamma}\right)^{1/(2-\beta)}$ in each round.

Proof For (λ, β)-Hölder continuous convex losses f_t, we have that

$$\begin{aligned} g_t(y_t) - g_t(y_{t+1}) &= \langle \nabla_t f_t(y_t), y_t - y_{t+1}\rangle \\ &= \langle \nabla_t f_t(y_t), (2y_t - y_{t+1}) - y_t\rangle \\ &\leq f_t(2y_t - y_{t+1}) - f_t(y_t) \end{aligned}$$

by convexity of f_t, where $\|(2y_t - y_{t+1}) - y_t\| = \|y_t - y_{t+1}\|$, and so

$$g_t(y_t) - g_t(y_{t+1}) \leq \lambda\|y_t - y_{t+1}\|^\beta$$

by Hölder continuity. Combining with the lower bound on $g_t(y_t) - g_t(y_{t+1})$ from Proposition 48 gives us that

$$\frac{\gamma}{\eta}\|y_{t+1} - y_t\|^2 \leq g_t(y_t) - g_t(y_{t+1}) \leq \lambda\|y_t - y_{t+1}\|^\beta$$

and thus

$$g_t(y_t) - g_t(y_{t+1}) \leq \lambda\left(\frac{\eta\lambda}{\gamma}\right)^{\beta/(2-\beta)},$$

yielding a regret bound of

$$\mathrm{Reg}_T(\mathrm{FTRL}) \le T\lambda \left(\frac{\eta\lambda}{\gamma}\right)^{\beta/(2-\beta)} + \frac{G}{\eta}$$

with per-round distance at most $\|y_{t+1} - y_t\| \le \left(\frac{\eta\lambda}{\gamma}\right)^{1/(2-\beta)}$. ■

We note that the concave surrogate rewards $f_t^*(y_t)$ are a sum of a $(k\lambda, \beta)$-Hölder continuous function and a $(L_c, 1)$-Hölder continuous (i.e. Lipschitz) function; we assume that each function is (L, β)-Hölder continuous with $L = k\lambda + L_c$, which is sufficient for for large enough T as we will have $\|y_t - y_{t-1}\| \le 1$ and thus $\|y_t - y_{t-1}\| \le \|y_t - y_{t-1}\|^\beta$. We use a similar analysis to bound the error between true and surrogate rewards, yielding our regret bound for NESTEDOCO.

Theorem 50 *The regret of* NESTEDOCO *with respect to the stable reserve policies $\mathcal{P}_\mathcal{Y}$ is bounded by*

$$\mathrm{Reg}_T(\mathrm{NESTEDOCO}) \le 2L \left(\frac{G}{\gamma}\right)^{\beta/2} \left(T\left(3 + \left(\frac{R}{\theta}\right)^\beta\right)\right)^{(2-\beta)/2}.$$

Proof We reparameterize to treat the bundle y_1 where $\mathcal{B}_1(y_1) \subseteq \mathcal{Y}$ as the origin, and assume the choice of regularizer has y_1 as its minimum. By Theorem 5, for any step size and $\delta > 0$ such that $\|y_t - y_{t-1}\| \le \delta\theta$, running NESTEDOCO for the θ-locally controllable instance $(\mathcal{Z}, \mathcal{Y}, D)$ over the surrogate rewards f_t^*, with inradius 1 and radius R, obtains

$$\begin{aligned}
\sum_{t=1}^T f_t^*(y^*) - \sum_{t=1}^T f_t^*(y_t) &\le TL(\delta R)^\beta + TL\left(\frac{\eta L}{\gamma}\right)^{\beta/(2-\beta)} + \frac{G}{\eta} \\
&\le TL\left(1 + \left(\frac{R}{\theta}\right)^\beta\right)\left(\frac{\eta L}{\gamma}\right)^{\beta/(2-\beta)} + \frac{G}{\eta} \\
&\le 2L\left(\frac{G}{\gamma}\right)^{\beta/2}\left(T\left(1 + \left(\frac{R}{\theta}\right)^\beta\right)\right)^{(2-\beta)/2} \\
&\triangleq \mathrm{Reg}_T(f^*)
\end{aligned}$$

for any $y^* \in \mathcal{Y}$, upon setting $\delta = \frac{1}{\theta}\left(\frac{\eta\lambda}{\gamma}\right)^{1/(2-\beta)}$ and $\eta = \left(\frac{G}{KT}\right)^{(2-\beta)/2}$, where

$$K^* = L\left(1 + \left(\frac{R}{\theta}\right)^\beta\right)\left(\frac{L}{\gamma}\right)^{\beta/(2-\beta)}.$$

Note that the surrogate rewards exactly track the true rewards when a stable reserve policy P_{y^*} is played, and so our regret with respect to the best stable reserve policy P_{y^*} is at most

$$\begin{aligned}
\sum_{t=1}^{T} f_t(P_{y^*}) - \sum_{t=1}^{T} f_t(y_t) &\leq \text{Reg}_T(f^*) + \sum_{t=1}^{T} f_t^*(y_t) - f_t(p_t, x_t) \\
&\leq \text{Reg}_T(f^*) + \sum_{t=1}^{T} \langle \nabla v(y_t), \theta y_t - x_t \rangle - c_t(\theta y_t) + c_t(x_t) \\
&\leq \text{Reg}_T(f^*) + \sum_{t=1}^{T} (1-\theta_t) \left(\langle \nabla v(y_t), y_{t-1} - y_t \rangle + L \left\| y_t - y_{t-1} \right\| \right) \\
&\qquad\qquad (x_t = (1-\theta_t) y_{t-1}) \\
&\leq \text{Reg}_T(f^*) + \sum_{t=1}^{T} \left(\langle \nabla v(y_t), y_t - (2y_t - y_{t-1}) \rangle + L \left\| y_t - y_{t-1} \right\| \right) \\
&\leq \text{Reg}_T(f^*) + \sum_{t=1}^{T} v(y_t) - v(2y_t - y_{t-1}) + L \left\| y_t - y_{t-1} \right\| \\
&\qquad\qquad \text{(concavity of } v) \\
&\leq \text{Reg}_T(f^*) + \sum_{t=1}^{T} 2L \left\| y_t - y_{t-1} \right\|^{\beta} \qquad \text{(Hölder, } \left\| y_t - y_{t-1} \right\| \leq 1) \\
&\leq \text{Reg}_T(f^*) + 2TL \left(\frac{\eta L}{\gamma} \right)^{\beta/(2-\beta)} \\
&\leq 2L \left(\frac{G}{\gamma} \right)^{\beta/2} \left(T \left(3 + \left(\frac{R}{\theta} \right)^{\beta} \right) \right)^{(2-\beta)/2}
\end{aligned}$$

upon updating K^* to K as

$$K = L \left(3 + \left(\frac{R}{\theta} \right)^{\beta} \right) \left(\frac{L}{\gamma} \right)^{\beta/(2-\beta)},$$

which yields the theorem. ■

Theorem 15 follows directly from Theorem 50.

Appendix K. Background and Proofs for Section 4.4: Steering Learners

K.1. Background

While much of the literature related to no-regret learning in general-sum games considers either rates of convergence to (coarse) correlated equilibria Blum et al. (2008); Anagnostides et al. (2022) or welfare guarantees for such equilibria Roughgarden (2015); Hartline et al. (2015a), a recent line of work Braverman et al. (2017); Deng et al. (2019); Mansour et al. (2022) has considered the question of *optimizing* one's reward when playing against a no-regret learner. A target benchmark which has emerged for this problem is the value of the *Stackelberg* equilibrium of a game (the

optimal mixed strategy to "commit to", assuming an opponent best responds), which was shown by attainable by Deng et al. (2019) against any no-regret algorithm and optimal in many cases (e.g. for no-swap learners), both up to $o(T)$ terms, and further which may yield higher reward for the optimizer than (coarse) correlated equilibria.

We show a class of instances for which the problem for optimizing reward against a learner playing according to gradient descent can be formulated as a locally controllable instance of online nonlinear control with adversarial perturbations and surrogate state-based losses. The simplest non-trivial instances we consider are those where the optimizer's reward is a function only of the learner's actions (i.e. all rows of their reward matrix are identical), and the optimization problem amounts to *steering* the learner to a desired strategy via one's choice of actions. Additionally, we allow the game matrices to change over time, which has not been substantially considered in prior work to our knowledge. We require that the learner's matrices do not change too quickly (which we model as adversarial disturbances to dynamics), and the optimizer's matrices can change arbitrarily provided that they remain close to *some* row-identical matrix (which we model as imprecision in our surrogate loss function).

K.2. Model

Here we are tasked with playing a sequence of bimatrix games against a no-regret learning opponent, where the game matrices may change adversarially in each round. We assume the following properties hold for the adversarial sequence of games.

Assumption 4 *For a sequence $\{(A_t, B_t) : t \in [T]\}$ of $m \times n$ bimatrix games, with $m > n$:*

- *Each entry of A_t and B_t lies in $[-\frac{L}{2\sqrt{n}}, \frac{L}{2\sqrt{n}}]$*
- *the convex hull of the of the rows of each B_t contains the unit ball in $\mathbb{R}^n$,*
- *$\|xA_t - xA_t^*\| \leq \delta_t$ for any $x \in \Delta(m)$, where each row of A_t^* is identical, and*
- *$\|xB_t - xB_{t-1}\| \leq \epsilon_t$ for any $x \in \Delta(m)$.*

Each game (A_t, B_t) is revealed after Players A and B commit to their respective strategies $x_t \in \Delta(m)$ and $y_t \in \Delta(n)$. Observe that due to the first property, for any $z \in \mathcal{B}_1(\mathbf{0})$, there is some $x \in \Delta(m)$ such that $xB = z$. By the second property, we have that $xA_t^* = x'A_t^*$ for any $x, x' \in \Delta(m)$.

We recall the Online Gradient Descent algorithm with convex losses ℓ_t from Zinkevich (2003).

Algorithm 8 Online Gradient Descent (OGD)

Input: Convex set $\mathcal{Y} \subseteq \mathbb{R}^n$, initial point $y_1 \in \mathcal{Y}$, and step sizes $\theta_1, \ldots, \theta_T$.
for $t = 1$ to T **do**
  Play y_t and observe loss $\ell_t(y_t)$.
  Set $\nabla_t = \nabla \ell_t(y_t)$.
  Set $y_{t+1} = \Pi_{\mathcal{Y}}(y_t - \theta_t \nabla_t) = \operatorname{argmin}_{y \in \mathcal{Y}} \|y_t - \theta_t \nabla_t - y\|$.
end for

Proposition 51 (Zinkevich (2003)) *For differentiable convex losses $\ell_t : \mathcal{Y} \to \mathbb{R}$, with $\theta_{t+1} \leq \theta_t$ for each $t \leq T$, then for all $y^* \in \mathcal{Y}$ the regret of* OGD *is bounded by*

$$\sum_{t=1}^{T} \ell_t(y_t) - \ell_t(y^*) \leq \frac{2R_B^2}{\theta_T} + \sum_{t=1}^{T} \frac{\theta_t}{2} \|\nabla_t\|^2,$$

where R_B is the radius of $\mathcal{Y}$. If $\|\nabla_t\| \leq G_B$ and $\theta_t = \frac{2R_B}{G_B\sqrt{T}}$ for all $t \leq T$, we have that

$$\sum_{t=1}^{T} \ell_t(y_t) - \ell_t(y^*) \leq 2R_B G_B \sqrt{T}.$$

We assume that Player B plays according to OPGD in our setup, with $y_1 = \mathbf{u}_n$ and $\theta = \frac{R_B}{G_B\sqrt{T}}$. At each round t, we (Player A) choose some mixed strategy $x_t \in \Delta(n)$, and Player B plays some mixed strategy $y_t \in \Delta(n)$. Utilities for each player are given by the game (A_t, B_t) as

$$u_t^A(x_t, y_t) = x_t A_t y_t;$$
$$u_t^B(x_t, y_t) = x_t B_t y_t.$$

Note that the loss gradient $-\nabla u_t^B(x_t, y_t)$ each round for Player B (for negative utilities) is given by

$$\nabla_t = -x_t B,$$

and so their mixed strategy is updated at each round according to

$$y_t = \Pi_{\Delta(n)} \left(y_{t-1} + \theta(x_{t-1} B_{t-1}) \right).$$

Our utility is given by $x_t A_t y_t = \mathbf{u}_n A_t^* y_t + x_t(A_t - A_t^*)y_t$, as x_t does not affect rewards from A_t^*. We benchmark the regret of an algorithm $\mathcal{A}$ against the optimal profile $(x, y) \in \Delta(m) \times \Delta(n)$:

$$\text{Reg}_T(\mathcal{A}) = \max_{(x,y) \in \Delta(m) \times \Delta(n)} \sum_{t=1}^{T} x A_t y - x_t A_t y_t.$$

Note that the per-round average utility for the maximizing (x, y) is at least as high as that obtained by the Stackelberg equilibrium of the average game $\left(\sum_t \frac{A_t}{T}, \sum_t \frac{B_t}{T}\right)$, as for this objective one can choose both players' strategies without restriction. We remark that finding the Stackelberg equilibrium for any fixed game (A_t^*, B_t) in our setting, where A_t^* has identical rows, is straightforward: it suffices to optimize over $[n]$, as any fixed action $j \in [n]$ is a best response to some $x \in \Delta(m)$ by our assumption on the rows of B_t, and as our rewards are only a function of Player B's strategy y. However, we are not aware of any prior work which enables competing with the average-game Stackelberg value against a learning opponent when games arrive online.

K.3. Analysis

We first show that the problem can be formulated via known, strongly θ-locally controllable dynamics with adversarial disturbances. As B_t changes slowly between rounds, we can run NESTEDOCO-UD with disturbances representing the error resulting from assuming that B_t does not change from B_{t-1}.

Lemma 52 *Given the knowledge available prior to selecting x_t, updates for y_{t+1} can be expressed via known action-linear dynamics $(\mathcal{X}, \mathcal{Y}, D_t)$ which satisfy strong θ-local controllability, and with adversarial disturbances w_t satisfying $\sum_{t=1}^{T} \|w_t\| \leq \theta \sum_{t=1}^{T} \epsilon_t$.*

Proof First, note that we can compute Player B's current strategy y_t, as it is a function only of games and strategies up to round $t-1$, all of which are observable. Given the update rule for OGD, we can formulate the dynamics $D_t(x_t, y_t)$ update as

$$\begin{aligned} D_t(x_t, y_t) &= \Pi_{\Delta(n)}\left(y_t + \theta(x_t B_t)\right) \\ &= \Pi_{\Delta(n)}\left(y_t + \theta(x_t B_{t-1}) + \theta(x_t(B_t - B_{t-1}))\right) \\ &= \Pi_{\Delta(n)}\left(y_t + \theta(x_t B_{t-1})\right) + w_t \end{aligned}$$

where w_t represents the error from assuming $B_t = B_{t-1}$. by standard properties of Euclidean projection, and the change bound on B_t, we have that $\|w_t\| \leq \|\theta(x_t(B_t - B_{t-1}))\| \leq \theta\epsilon_t$. Further, the update is action-linear (up to projection, prior to w_t).

To see that D_t satisfies strong θ-local controllability, we recall that the convex hull of the rows of B_{t-1} contain the unit ball, and so for any y^* in $\mathcal{B}_\theta(y_t) \cap \Delta(n)$ there is some $x_t \in \Delta(m)$ such that $\theta(x_t B_{t-1}) = y^* - y_t$. ■

At round each round t, our loss is given by $f_t(x_t, y_t) = -x_t A_t y_t$. There are two barriers to running our algorithm. First, the update for y_t is determined by x_{t-1} and not x_t, yet we do not see A_{t-1} prior to selecting x_{t-1}, which would be required to take the appropriate step following f_{t-1}. Second, the loss depends on x_t in addition to y_t. To address both issues, we instead run NESTEDOCO-UD with surrogate losses $\tilde{f}_t(\tilde{y}_t) = -\mathbf{u}_n A_{t-1} y_t$, with action rounds relabeled to account for the fact that x_{t-1} influences the step for y_t (which does not change the behavior of the algorithm). We set $A_0 = \mathbf{0}_{m,n}$.

Theorem 53 *Repeated play against an opponent using* OGD *with step size $\theta = \Theta(T^{-1/2})$ in a sequence of games (A_t, B_t) satisfying Assumption 4 can be cast as an instance of online control with strongly θ-locally controllable dynamics, for which the regret of* NESTEDOCO-UD *is at most*

$$\mathrm{Reg}_T(\text{NESTEDOCO-UD}) \leq O\left(\sqrt{T} + \sum_{t=1}^{T}(\delta_t + \epsilon_t)\right),$$

with efficient per-round computation.

Proof We first analyze regret with respect to the surrogate losses $\tilde{f}_t(y_t)$. To run NESTEDOCO-UD for $\alpha > 0$, it suffices to calibrate the step size for the internal FTRL instance such that $\eta\frac{L}{\gamma} \leq \theta\alpha$. Given that rewards are bounded in $[-\frac{L}{2\sqrt{n}}, \frac{L}{2\sqrt{n}}]$, we have that each $x_t B_t y_t$ is $\frac{L}{\sqrt{n}}$-Lipschitz for the ℓ_1 norm, and thus L-Lipschitz for the ℓ_2 norm, so we can take $G_B = L$. Further, the ℓ_2 radius of $\Delta(n)$ is $R_B = \sqrt{2}/2$, and so we have that

$$\theta = \sqrt{\frac{2}{L^2 T}}.$$

Then, for a strongly θ-locally controllable instance with total perturbation bound $\sum_{t=1}^{T} \|w_t\| \le E$, we obtain the regret bound

$$\mathrm{Reg}_T(\textsc{NestedOCO-UD}) \le \eta \frac{TL^2}{\gamma} + \frac{G}{\eta} + \frac{2LRE}{(1-\alpha)\theta} \qquad \text{(Thm. 22)}$$

for any

$$\eta \le \min\left(\sqrt{\frac{G\gamma}{L^2 T}}, \alpha\sqrt{\frac{2}{T}}\right).$$

By Lemma 52, we can efficiently run NESTEDOCO-UD over the surrogate losses $\tilde{f}_t$ and bound regret with respect to any $y^* \in \mathcal{Y}$ as:

$$\sum_{t=1}^{T} \tilde{f}_t(y_t) - \tilde{f}_t(y^*) \le \eta \frac{TL^2}{\gamma} + \frac{G}{\eta} + \frac{\sqrt{2}L \cdot \sum_{t=1}^{T} \epsilon_t}{1-\alpha}.$$

Further, we can bound the error from the surrogate losses as

$$\begin{aligned}
\sum_{t=1}^{T} f_t(x_t, y_t) - \tilde{f}_t(y_t) &= \sum_{t=1}^{T} f_t(x_t, y_t) - f_{t-1}(\mathbf{u}_n, y_t) \\
&\le \frac{L}{2\sqrt{n}} + \sum_{t=1}^{T-1} f_t(x_t, y_t) - f_t(\mathbf{u}_n, y_{t+1}) \\
&\qquad\qquad (f_0(\mathbf{u}_n, y_1) = 0,\ f_T(x_T, y_T) \le \tfrac{L}{2\sqrt{n}}) \\
&\le \frac{L}{2\sqrt{n}} + \eta \frac{TL^2}{\gamma} + \sum_{t=1}^{T-1} x_t (A_t - A_t^*) y_t \qquad \text{(Prop. 18)} \\
&\le \frac{L}{2\sqrt{n}} + \eta \frac{TL^2}{\gamma} + \sum_{t=1}^{T} \delta_t, \qquad \text{(Assumption 4, Cauchy-Schwarz)}
\end{aligned}$$

and likewise, for any $(x^*, y^*) \in \Delta(m) \times \Delta(n)$ we can bound

$$\begin{aligned}
\sum_{t=1}^{T} \tilde{f}_t(y^*) - f_t(x^*, y^*) &\le -f_T(x^*, y^*) - \sum_{t=1}^{T-1} x^* (A_t - A_t^*) y^* \\
&\le \frac{L}{2\sqrt{n}} + \sum_{t=1}^{T} \delta_t.
\end{aligned}$$

Combining the previous results, we have that for any $(x^*, y^*) \in \Delta(m) \times \Delta(n)$, the regret of NESTEDOCO-UD with respect to the true losses is bounded by

$$\begin{aligned}\sum_{t=1}^{T} f_t(x_t, y_t) - f_t(x^*, y^*) &\leq \sum_{t=1}^{T} \tilde{f}_t(\tilde{y}_t) - \tilde{f}_t(y^*) + \sum_{t=1}^{T} f_t(x_t, y_t) - \tilde{f}_t(y_t) + \sum_{t=1}^{T} \tilde{f}_t(y^*) - f_t(x^*, y^*) \\ &\leq \eta \frac{2TL^2}{\gamma} + \frac{G}{\eta} + \frac{L}{\sqrt{n}} + 2\sum_{t=1}^{T} \delta_t + \frac{\sqrt{2}L \cdot \sum_{t=1}^{T} \epsilon_t}{1-\alpha} \\ &\leq 3 \cdot \max\left(\sqrt{\frac{TGL^2}{\gamma}}, \sqrt{\frac{T}{2\alpha^2}}\right) + \frac{L}{\sqrt{n}} + 2\sum_{t=1}^{T} \delta_t + \frac{\sqrt{2}L \cdot \sum_{t=1}^{T} \epsilon_t}{1-\alpha}\end{aligned}$$

for any $\alpha \in (0, 1)$, which yields the theorem. ■

Theorem 16 follows directly from Theorem 53.