Online Performance Optimization of Nonlinear Systems: A Gray-Box Approach

Zhiyu He [1] **Michael Muehlebach** [2] **Saverio Bolognani** [1] **Florian Dörfler** [1]

Abstract

We propose a gray-box controller to optimize the performance of a nonlinear system in an online manner. This is motivated by the observation that model-based and model-free approaches own complementary benefits in sample efficiency and optimality in the presence of inaccurate models. To achieve the best of both worlds, our controller incorporates approximate model information into model-free updates via adaptive convex combinations. Further, it leverages real-time outputs of the system and iteratively adjusts control inputs. We quantify conditions on the quality of approximate models that render the gray-box approach preferable to model-based or model-free approaches. We characterize the performance of our controller via dynamic regret in a constrained, time-varying setting, and highlight how the regret scales with the number of iterations, the problem dimension, and the cumulative effect of inaccurate models.

1. Introduction

Online decision-making encompasses optimization, control, and learning subject to changing, unknown performance metrics and environments (Cesa-Bianchi & Lugosi, 2006; Hazan, 2022), which has found broad applications in learning systems (e.g., classifiers and recommender systems) and infrastructure networks (e.g., power grids and communication networks).

At the heart of online decision-making is an iterative feedback loop between the decision-maker and the environment (or the adversary), whereby a decision-maker commits to a decision at each iteration, incurs a loss selected by the environment, and adjusts her decision accordingly. In general, arbitrary variations of the loss are allowed from one iteration to the next, as long as the loss is bounded or a related compactness assumption on decision variables is satisfied. However, in many applications, there exist *in vivo* dynamics with latent states that couple the loss over iterations or specify the environmental feedback via an input-output relationship of a dynamical system. Dynamics bring the role of system models to the forefront of online decision-making.

[1]Automatic Control Laboratory, ETH Zürich, 8092 Zürich, Switzerland [2]Max Plack Institute for Intelligent Systems, Tübingen 72076, Germany. Correspondence to: Zhiyu He <zhiyhe@ethz.ch>.

Workshop on Foundations of Reinforcement Learning and Control at the 41st International Conference on Machine Learning, Vienna, Austria. Copyright 2024 by the author(s).

1.1. Related Work

Online optimization studies sequentially choosing decisions in the face of unknown, streaming problems. Since the objective can be arbitrarily set at every time step, the realistic goal is usually not achieving optimality, but instead minimizing the *regret*, i.e., the cumulative gap against the benchmark sequence of decisions in hindsight (Hazan, 2022).

Online control focuses on regulating a dynamical system while optimizing cumulative stage costs. In contrast to online optimization, online control tackles the situation where the costs are temporally dependent due to dynamics. In the realm of linear systems, a plethora of online linear control policies have been developed to handle unknown dynamics (Dean et al., 2018; Simchowitz & Foster, 2020), adversarial noises (Agarwal et al., 2019), as well as convex, time-varying costs (Cassel et al., 2022). These policies enjoy sublinear regrets against the optimal linear policy in hindsight (Tsiamis et al., 2023). However, in the nonlinear world, the studies of the structure and performance guarantee of online controllers are still nascent (Kakade et al., 2020; Boffi et al., 2021).

Feedback optimization is an emerging paradigm for optimizing the performance of nonlinear dynamical systems (Hauswirth et al., 2024; Simonetto et al., 2020). In contrast to online optimization, the streaming objective depends on both the decision (i.e., the input) and the corresponding output of the system, see Figure 1. Compared to online control, feedback optimization directly searches for the optimal decision vector rather than being confined to the class of linear policies. Moreover, the goal is achieving optimal *steady-state* operations (i.e., when the state and the output reach fixed values given a constant input) instead of optimizing total costs that include transients. The core insight is to implement optimization iterations as feedback controllers, which measure outputs and iteratively adjust

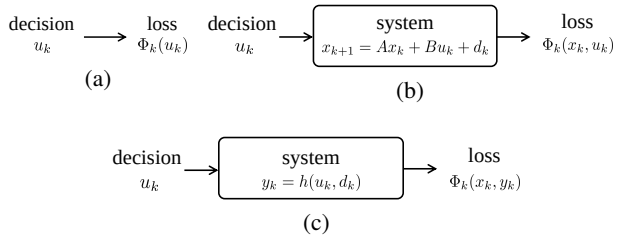

Figure 1. The block diagrams illustrate (a) online optimization, (b) online control, and (c) feedback optimization. There are fundamental differences between online control and feedback optimization, for instance in the way decisions are taken and the overall goal, see also Section 1.1.

inputs to drive the dynamical system to an optimal steady state. This closed-loop structure brings salient properties, e.g., stability, constraint satisfaction, and adaption to non-stationary environments. It also proves effective in the real world (Ortmann et al., 2023).

Model-based and model-free feedback optimization. The manifold benefits of feedback optimization rely on the knowledge of the *input-output sensitivity* of the system, which measures how a change in the input causes the corresponding change in the output. This requirement stems from using the chain rule to form the gradient-based adjustment direction of the controller. In terms of complex, large-scale, and poorly known systems, the lack of knowledge of accurate sensitivities can result in sub-optimality, constraint violation, or instability (Hauswirth et al., 2024). To address this issue, two streams of strategies have been explored.

One stream is *model-based*, in that the key model information (i.e., sensitivity) is learned from offline data (e.g., through data-driven representations in the behavioral framework (Bianchin et al., 2024)) or online interactions (e.g., through recursive estimation based on streaming data (Picallo et al., 2022)). Nonetheless, if the sensitivity is not learned fast and accurately enough, the iterations may suffer from considerable sub-optimality, see Section VI in (He et al., 2024). Another stream is *model-free* (or *black-box*), which avoids learning sensitivities altogether. To this end, zeroth-order optimization schemes are employed to construct stochastic gradient estimates from function evaluations, thus circumventing the need for sensitivity information (Chen et al., 2020; Tang et al., 2023; He et al., 2024). Overall, the stochasticity of gradient estimates may affect the convergence rate, thereby increasing the sample complexity of model-free feedback optimization compared to its model-based counterparts.

Gray-box pipelines. Model-based and model-free approaches own complementary benefits in sample efficiency and provable optimality in the presence of inaccurate models. Thus, it is promising to develop *gray-box* approaches to achieve the best of both worlds. The power of gray-box pipelines has been showcased in various problems, e.g., reinforcement learning, predictive control, and stabilization. Some methods rely on model-based pipelines and introduce model-free, learning-based blocks for inference or improvement (Achterhold et al., 2023; Ma et al., 2023). Others augment model-free pipelines with model-based priors (e.g., prior mean or a model-based policy as a warm start (Qu et al., 2021)) or utilize synthetic data generated from transition models to enhance model-free training (Janner et al., 2019). Further, recent works on learning-augmented control (Li et al., 2023) combine a model-based (albeit sub-optimal) policy with a black-box, machined-learned policy. However, it is unclear how gray-box approaches can be designed and proven useful in the context of performance optimization of nonlinear systems.

1.2. Motivations and Contributions

We aim to optimize the performance of a nonlinear system subject to a time-varying objective. Given the complementary benefits of model-based and model-free pipelines (see Section 1.1), we pursue a gray-box approach that utilizes approximate model information, thereby achieving the best of both worlds. In our context, such information refers to the approximate input-output sensitivity of the system, which can be obtained through prior knowledge, first-principle models, or estimation (Ortmann et al., 2023; Ma et al., 2023). Further, we are interested in the following questions related to performance measures. How to quantify the conditions favoring gray-box approaches over model-based or model-free methods, and vice versa? How to establish provable improvement using the same performance measure for all approaches? All these questions call for a new algorithmic design and analysis.

Our main contributions are summarized as follows.

- We propose a gray-box feedback optimization controller to optimize the performance of a nonlinear system subject to time-varying objectives and input constraints. When interconnected with the system, this controller uses real-time outputs and iteratively adjusts inputs by adaptively combining model-based inexact gradients from approximate sensitivities and model-free gradient estimates.
- We characterize the conditions on the quality of input-output sensitivities that render the gray-box controller preferable. These conditions endow the proposed controller with flexibility, such that it can fully exploit approximate models of different levels of accuracy by adjusting combination coefficients.
- We quantify the overall performance through dynamic regret, which focuses on the cumulative gap against the optimal benchmark. The gray-box controller overcomes the issue of sub-optimality, which is experienced

by model-based controllers due to the error accumulation of approximate models. The gray-box controller can also exploit sensitivities of lower quality (e.g., with bounded errors) to improve sample efficiency compared to model-free approaches.

The rest of this article is organized as follows. In Section 2, we provide the problem of interest. Section 3 presents the design of the gray-box controller. The performance guarantee in a time-varying and constrained setting is established in Section 4. We perform numerical evaluations in Section 5. Finally, Section 6 concludes this article.

2. Problem Formulation

Consider a system abstracted by its nonlinear steady-state input-output map $h : \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^q$

$$y = h(u, d), \tag{1}$$

where $u \in \mathbb{R}^p$ is the input, $y \in \mathbb{R}^q$ is the output, and $d \in \mathbb{R}^r$ is the unknown exogenous disturbance. We consider an online setting with changing disturbances $\{d_k\}_{k\in\mathbb{N}}$. At time k, we aim to find an input u that optimizes the input-output performance of the system (1) induced by u and d_k, i.e.,

$$\begin{aligned} \min_{u\in\mathbb{R}^p, y\in\mathbb{R}^q} \quad & \Phi_k(u, y) \\ \text{s.t. } & y = h(u, d_k), \\ & u \in \mathcal{U}. \end{aligned} \tag{2}$$

In problem (2), $\Phi_k : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is the objective at time $k \in \mathbb{N}$, and $\mathcal{U} \subset \mathbb{R}^p$ is the constraint set for u.

The constrained, time-varying problem (2) reflects the specifications of changing objectives, variable disturbances, and input constraints. We will leverage feedback control by iteratively adjusting the input u after the output y encoding d_k is measured and the current objective Φ_k is revealed. Let $u_k^* \in \mathbb{R}^p$ be an optimal point of problem (2) at time k. The goal is to generate control inputs that are competitive with the sequence of time-varying optimal solutions $\{u_k^*\}_{k\in\mathbb{N}}$.

Let $\tilde{\Phi}_k(u) \triangleq \Phi_k(u, h(u, d_k))$ be the reduced objective at time k. Some assumptions are as follows.

Assumption 2.1. The map $h(u, d)$ is differentiable with respect to u.

Assumption 2.2. The set $\mathcal{U}$ is a compact, convex set with diameter $D > 0$, i.e., $\forall u_1, u_2 \in \mathcal{U}, \|u_1 - u_2\| \le D$.

Assumption 2.3. The function $\tilde{\Phi}_k(u)$ is convex, L_k-smooth, and M_k-Lipschitz with respect to u. The function $\Phi_k(u, y)$ is $M_{\Phi,k}$-Lipschitz with respect to y. Moreover, $\{L_k\}$, $\{M_k\}$, and $\{M_{\Phi,k}\}$ are bounded.

Let $\mathcal{U}_\tau \triangleq \{u + \tau v | u \in \mathcal{U}, v \in \mathbb{B}_p\}$ denote a set inflated from $\mathcal{U}$ by a limited range $\tau\mathbb{B}_p$, where $\tau > 0$ is an expansion coefficient, and $\mathbb{B}_p$ is the closed unit ball in $\mathbb{R}^p$. The following assumption specifies the boundedness of $\tilde{\Phi}_k$.

Assumption 2.4. The function $\tilde{\Phi}_k(u)$ is uniformly bounded, i.e., $\exists G \ge 0, \exists \tau > 0, \forall u \in \mathcal{U}_\tau, \forall k \in \mathbb{N}, |\tilde{\Phi}_k(u)| \le G$.

Assumptions 2.2-2.4 are typical in the literature, e.g., (Jadbabaie et al., 2015; Zhao et al., 2021; Hazan, 2022). Assumption 2.2 also implies that the norm of any point in $\mathcal{U}$ is bounded, i.e., $\exists \overline{D} \ge 0, \forall u \in \mathcal{U}, \|u\| \le \overline{D}$. Assumption 2.4 is related to Assumptions 2.2 and 2.3, because a continuous function defined on a compact set is bounded.

A tempting solution to problem (2) is to directly use numerical optimization solvers. Nonetheless, solvers require the explicit expression of the map h and the exact values of the disturbances $\{d_k\}$. These requirements can be hard to satisfy when complex systems and unknown disturbances are involved. Hence, we pursue a feedback optimization controller that utilizes real-time output measurements to optimize the dynamic behavior of (1).

Model-based feedback optimization controllers (Picallo et al., 2022) learn and use the input-output sensitivity $\nabla_u h(u, d_k)$ of the system (1) and iteratively update inputs by leveraging the gradient of $\tilde{\Phi}_k(u)$. After invoking the chain rule, their update rule reads as

$$\begin{aligned} u_{k+1} = \mathrm{Proj}_{\mathcal{U}} \Big(u_k - \eta \big(& \nabla_u \Phi_k(u_k, y_k) \\ & + \nabla_u h(u_k, d_k)^\top \nabla_y \Phi_k(u_k, y_k) \big) \Big), \end{aligned} \tag{3}$$

where $\eta > 0$ is a step size, y_k is the output of the system (1) at time $k \in \mathbb{N}$, and $\mathrm{Proj}_{\mathcal{U}}(\cdot)$ denotes the projection to the constraint set $\mathcal{U}$. Model-free controllers (Chen et al., 2020; Tang et al., 2023; He et al., 2024) bypass the information on sensitivities and rely purely on stochastic exploration. Their trade-offs in sample efficiency and solution accuracy are discussed in Section 1.1. In contrast, we will merge approximate sensitivities (obtained, among others, through prior knowledge or recursive estimation) into model-free updates, thus achieving the best of both worlds.

3. Design of the Running Gray-Box Controller

For a given input u_k at time $k \in \mathbb{N}$, suppose that we have access to an approximate input-output sensitivity $\hat{H}_k \in \mathbb{R}^{q\times p}$ of (1), which differs from the true sensitivity $H_k \triangleq \nabla_u h(u_k, d_k)$. Note that H_k indicates how a change in u_k will cause the change in the output of (1). Such an approximate $\hat{H}_k$ can be obtained through prior knowledge, first-principle models (Ortmann et al., 2023; Ma et al., 2023), or online learning and estimation (Picallo et al., 2022).

Our proposed feedback controller iteratively adjusts inputs by using real-time output measurements. The update direction is constructed by adaptively fusing an inexact gradient

from the approximate sensitivity $\hat{H}_k$ and a gradient estimate based on stochastic exploration. The update rules are

$$w_{k+1} = \mathrm{Proj}_{\mathcal{U}}(w_k - \eta\hat{\phi}_k), \tag{4a}$$
$$\hat{\phi}_k = \alpha_k\hat{\phi}_{k,1} + (1-\alpha_k)\hat{\phi}_{k,2}, \tag{4b}$$
$$\hat{\phi}_{k,1} = \nabla_u\Phi_k(u_k, y_k) + \hat{H}_k^\top\nabla_y\Phi_k(u_k, y_k), \tag{4c}$$
$$\hat{\phi}_{k,2} = \frac{pv_k}{\delta}\big(\Phi_k(u_k, y_k) - \Phi_{k-1}(u_{k-1}, y_{k-1})\big), \tag{4d}$$
$$u_{k+1} = w_{k+1} + \delta v_{k+1}, \tag{4e}$$

where w_k is a candidate solution, $\mathrm{Proj}_{\mathcal{U}}(\cdot)$ denotes the projection to the constraint set $\mathcal{U}$, $\eta > 0$ is a step size, $\alpha_k \in [0,1]$ is a convex combination coefficient, $\delta \in (0,\tau)$ is a smoothing parameter, p is the size of the input, and $v_0, \ldots, v_{k+1} \sim U(\mathbb{S}_{p-1})$ are i.i.d. random variables sampled from the unit sphere $\mathbb{S}_{p-1}$.

In the iterative update our running gray-box controller (4) merges two directions. The first one (i.e., $\hat{\phi}_{k,1}$ in (4c)) is an inexact gradient constructed from $\hat{H}_k$, whereas the second (i.e., $\hat{\phi}_{k,2}$ in (4d)) is a stochastic gradient estimate. The controller subsequently performs a projection (see (4a)) to the constraint set $\mathcal{U}$, thus ensuring that the candidate solution w_{k+1} satisfies the constraint. Finally, the solution is perturbed by δv_{k+1} to form the input u_{k+1} (see (4e)), and this input is applied to (1). Our controller (4) uses the latest information at time k (i.e., the partial gradients and values of the current objective $\tilde{\Phi}_k$) to adapt to the variation of problem (2).

Remark 3.1. While the candidate solution w_k lies in $\mathcal{U}$, the input u_k in the transient stage may violate the constraint. If we need strict constraint satisfaction, we can project in (4a) onto a deflated set $(1-\kappa)\mathcal{U}$ as (Zhao et al., 2021), where $\kappa > 0$.

For problem (2), model-based controllers purely using $\{\hat{H}_k\}_{k\in\mathbb{N}}$ (i.e., (4) with $\alpha_k = 1, \forall k \in \mathbb{N}$) are favorable provided that $\hat{H}_k$ is a sufficiently accurate estimate of H_k, or more specifically,

$$\epsilon_{H,k} \triangleq \|\hat{H}_k - H_k\| \le \frac{\bar{\epsilon}'}{(k+1)^\theta}, \quad \theta \ge \frac{1}{4}, \quad k \in \mathbb{N}, \tag{5}$$

where $\epsilon_{H,k}$ is the error of $\hat{H}_k$ compared to H_k, and $\bar{\epsilon}' > 0$ is a specified error bound. The intuition is that the resulting cumulative error $\sum_{k=1}^T \epsilon_{H,k}$ will be of an order lower than $\mathcal{O}(T^{\frac{3}{4}})$, which leads to a salient guarantee in the sense of dynamic regret (Hazan, 2022; Jadbabaie et al., 2015; Zhao et al., 2021). A detailed explanation is referred to Section 4. Nonetheless, (5) may not always hold due to various issues, e.g., noisy measurements or nonlinear dynamics. In such a case, our proposed gray-box approach (4) is favorable. We analyze two cases of $\hat{H}_k$ and show how to select the combination coefficient α_k in (4b).

Case 1: Approximate Sensitivity with a Bounded Error

In many applications, we construct approximate sensitivities based on prior knowledge or first-principle models (Ortmann et al., 2023; Ma et al., 2023), which are then fixed during online operation. This practice corresponds to the case where a bounded error exists between the approximate sensitivity $\hat{H}_k$ and the ground-truth H_k, i.e.,

$$\epsilon_{H,k} = \|\hat{H}_k - H_k\| \le \bar{\epsilon}, \quad k \in \mathbb{N}, \tag{6}$$

where $\bar{\epsilon} > 0$. In this case, we select a constant $C > 0$ and use the following vanishing combination coefficient

$$\alpha_k = \min\Big\{\frac{C}{(k+1)^{\frac{1}{4}}}, 1\Big\}, \quad k \in \mathbb{N}. \tag{7}$$

The rationale of (7) is to tune the scaled cumulative error $\sum_{k=1}^T \alpha_k\epsilon_{H,k}$, thereby regulating the order of the dynamic regret incurred by the proposed controller. Moreover, C is concerned with the trade-off between the convergence rate and the cumulative error. A large C implies more emphasis on the model-based direction $\hat{\phi}_{k,1}$ in the initial stage. While this emphasis contributes to a fast response, it may cause an increase in the magnitude of the cumulative error.

Case 2: Asymptotically Accurate Sensitivity

Online estimation techniques can be incorporated to generate increasingly accurate estimates of sensitivities based on the trajectory of the system (1). However, we may not always learn sensitivities sufficiently fast due to measurement errors, lack of covariance data, nonlinear dynamics, etc. For instance, the estimation error (5) of the sensitivity may decrease as

$$\epsilon_{H,k} = \|\hat{H}_k - H_k\| \le \frac{\bar{\epsilon}'}{(k+1)^\theta}, \ \ \theta \in \Big(0, \frac{1}{4}\Big), \ \ k \in \mathbb{N}, \tag{8}$$

where $\bar{\epsilon}' > 0$. When (8) arises, we utilize the following vanishing coefficient

$$\alpha_k = \min\Big\{\frac{C'}{(k+1)^{\frac{1}{4}-\theta}}, 1\Big\}, \quad k \in \mathbb{N}. \tag{9}$$

Remark 3.2. In practice, we can establish bounds (6) and (8) (with high probability) through finite-sample analysis based on concentration inequalities (Tsiamis et al., 2023).

4. Performance Certificates

We analyze the interconnection of the system (1) with our running gray-box controller (4). Such an interconnection leads to a *closed loop*, in that the controller measures the output and then generates the input.

We first introduce a lemma that gives an upper bound on the expected distance between the direction $\hat{\phi}_k$ used by our gray-box controller (4) and the true gradient $\nabla\tilde{\Phi}_k(w_k)$. Recall

that $\epsilon_{H,k} = \|\hat{H}_k - H_k\|$ is the error of the approximate sensitivity $\hat{H}_k$ compared to the true sensitivity H_k.

Lemma 4.1. *Let Assumptions 2.1,2.2-2.4 hold. The update rule (4) ensures that*

$$\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_k - \nabla\tilde{\Phi}_k(w_k)\|] \leq \alpha_k M_{\Phi,k}\epsilon_{H,k} + (1-\alpha_k)\frac{2pG}{\delta} + L_k\delta\Big(\alpha_k + (1-\alpha_k)\frac{p}{2}\Big).$$

Proof. See Appendix A.1. □

We proceed to offer our performance certificate, i.e., dynamic regret (Hazan, 2022; Jadbabaie et al., 2015; Zhao et al., 2021), which is a common measure for decision-making in non-stationary environments. It fits our context involving time-varying objectives and disturbances.

Specifically, we focus on the cumulative difference between the objective values evaluated at the candidate solutions $\{w_k\}_{k=1}^T$ and those at the optimal points $\{u_k^*\}_{k=1}^T$. To capture the variation of (2), we introduce the path length $C_T \triangleq \sum_{k=1}^T \|u_k^* - u_{k-1}^*\|$, which accumulates the shifts between two consecutive optimal points (Hazan, 2022). The following theorem characterizes the dynamic regret incurred by the closed-loop system.

Theorem 4.2. *Suppose that Assumptions 2.1,2.2-2.4 hold. Consider approximate sensitivities $\{\hat{H}_k\}_{k=1}^T$ that satisfy (5), (6), or (8). Let $\eta = 1/p^{\frac{2}{3}}T^{\frac{3}{4}}$ and $\delta = \min(p^{\frac{1}{3}}/T^{\frac{1}{4}}, \tau)$. The closed-loop interconnection of the system (1) and the gray-box controller (4) incurs the following dynamic regret*

$$\begin{aligned}\mathrm{Reg}_T^d &\triangleq \sum_{k=1}^T \big(\mathbb{E}_{v_{[T]}}[\tilde{\Phi}_k(w_k)] - \tilde{\Phi}_k(u_k^*)\big) \\ &= \mathcal{O}\left(p^{\frac{2}{3}}T^{\frac{3}{4}}(C_T+1) + \sum_{k=1}^T \alpha_k\epsilon_{H,k}\right). \end{aligned} \tag{10}$$

Furthermore, given $\hat{H}_k$ that satisfies (6) (or (8)) and $\{\alpha_k\}_{k\in\mathbb{N}}$ are designed as in (7) (respectively, in (9)), this closed-loop interconnection ensures that

$$\mathrm{Reg}_T^d = \mathcal{O}\Big(p^{\frac{2}{3}}T^{\frac{3}{4}}(C_T+1)\Big). \tag{11}$$

Proof. See Appendix A.2. □

The order of the dynamic regret (10) is determined by two major parts. The first part is proportional to the path length C_T, and the second part reflects the error accumulation due to $\hat{H}_k$. Recall that model-based controllers correspond to (4) with $\alpha_k = 1, \forall k \in \mathbb{N}$ and $\delta = 0$. For those controllers, when $\hat{H}_k$ satisfies (6) or (8), the second part (i.e., $\sum_{k=1}^T \alpha_k\epsilon_{H,k}$) will be of the orders of $\mathcal{O}(T)$ or $\mathcal{O}\big(T^{1-\theta}\big)$, respectively, where $0 < \theta < \frac{1}{4}$. With the same choice of η as in Theorem 4.2, the corresponding orders of Reg_T^d become $\mathcal{O}(p^{\frac{2}{3}}T^{\frac{3}{4}}(C_T+1)+T)$ and $\mathcal{O}(p^{\frac{2}{3}}T^{\frac{3}{4}}(C_T+1)+T^{1-\theta})$. In contrast, our gray-box controller (4) allows tuning $\{\alpha_k\}_{k=1}^T$ to arrive at (11). We will further illustrate the difference in the magnitude of Reg_T^d between the gray-box controller and the model-free controller (i.e., (4) with $\alpha_k = 0, \forall k \in \mathbb{N}$) in Section 5.

Remark 4.3. If the path length C_T is known in advance, then the order of Reg_T^d in (10) can be refined to $\mathcal{O}\big(p^{\frac{2}{3}}T^{\frac{3}{4}}\sqrt{C_T+1} + \sum_{k=1}^T \alpha_k\epsilon_{H,k}\big)$ by choosing the step size $\eta = \sqrt{C_T+1}/p^{\frac{2}{3}}T^{\frac{3}{4}}$. For model-based controllers using accurate H_k, we recover an expected regret bound of the order of $\mathcal{O}(\sqrt{T}(C_T+1))$ by selecting $\eta = 1/\sqrt{T}$.

Our controller perturbs the candidate solution w_k with an exploration noise δv_k and applies the input u_k to the system, see (4e). The difference between the objective values $\tilde{\Phi}_k(u_k)$ and $\tilde{\Phi}_k(w_k)$ can increase the magnitude of the dynamic regret when the inputs $\{u_k\}_{k=1}^T$ are compared with the optimal points $\{u_k^*\}_{k=1}^T$. The following corollary, however, states that the order of the dynamic regret is unchanged.

Corollary 4.4. *Let the conditions of Theorem 4.2 hold. The closed-loop interconnection incurs*

$$\sum_{k=1}^T \big(\mathbb{E}_{v_{[T]}}[\tilde{\Phi}_k(u_k)] - \tilde{\Phi}_k(u_k^*)\big) = \mathcal{O}\Big(p^{\frac{2}{3}}T^{\frac{3}{4}}(C_T+1)\Big).$$

Proof. See Appendix A.3. □

5. Numerical Evaluations

We evaluate the performance of our proposed gray-box controller (4) when applied to the following nonlinear dynamical system

$$\begin{aligned} x_{k+1} &= Ax_k + B_1u_k + B_2\sin(u_k) + Ed_x, \\ y_k &= Cx_k + Dd_y, \end{aligned} \tag{12}$$

where $x \in \mathbb{R}^{20}$, $u \in \mathbb{R}^{10}$, and $y \in \mathbb{R}^5$ denote the state, input, and output, respectively, and $d_x, d_y \in \mathbb{R}^5$ are disturbances. We draw the elements of the system matrices in (12) from the normal distribution. We further scale A to let its spectral radius be 0.05, i.e., the dynamics are quickly contracting. When the system (12) evolves to a steady state given $u_k = u, \forall k \in \mathbb{N}$, its input-output map is

$$\begin{aligned} y &= h'(u, d_x, d_y) \\ &\triangleq C(I-A)^{-1}(B_1 + B_2\sin(u) + Ed_x) + Dd_y. \end{aligned}$$

The input-output sensitivity matrix H of (12) at u is

$$H = C(I-A)^{-1}(B_1 + B_2\,\mathrm{diag}(\cos(u))),$$

where $\mathrm{diag}(\cos(u))$ is a diagonal matrix with its diagonal entries given by $\cos(u)$.

We aim to optimize the steady-state input-output performance of the system (12) as characterized by

$$\begin{aligned} \min_{u\in\mathbb{R}^{10}, y\in\mathbb{R}^5} \ & \Phi(u,y) = u^\top M_1 u + m_2^\top u + \|y\|^2 \\ \text{s.t. } & y = h'(u, d_x, d_y), \\ & \underline{u} \le u \le \bar{u}, \end{aligned} \tag{13}$$

where $\underline{u} \in \mathbb{R}^{10}$ and $\bar{u} \in \mathbb{R}^{10}$ denote the lower bound and the upper bound on u, respectively, and they are generated from the multivariate normal distribution. Moreover, the equality constraint in (13) corresponds to the steady-state map $h'(u, d_x, d_y)$ of the system (12). Every 5×10^3 iteration, m_2 and the positive definite M_1 in the objective are regenerated from normal distributions, and the disturbances d_x, d_y are regenerated from uniform distributions.

Hence, problem (13) is time-varying with input constraints. Though the objective in (13) is nonconvex in u because of the nonlinear term $\sin(u)$ in the map h', we obtain the comparator sequence $\{u_k^*\}$ by calling the fmincon function of MATLAB.

We compare the closed-loop interconnection of the system (12) with various controllers: model-based feedback optimization controllers with projection (3) and using the accurate sensitivity H_k or the inexact sensitivity $\hat{H}$, the controller with sensitivity learning based on $\hat{H}$ (Picallo et al., 2022), the model-free controller in (He et al., 2024), and our proposed gray-box controller (4) using $\hat{H}$. Specifically, the approximate sensitivity $\hat{H}$ is a perturbed version of $C(I-A)^{-1}B$, and the element-wise relative error is not more than 30%. We set the corresponding step sizes as $\eta = 7.5 \times 10^{-5}$. The model-free controller (He et al., 2024) experiences divergence with this step size, and, therefore, we select $\eta = 5 \times 10^{-5}$ in this case. The smoothing parameter is $\delta = 10^{-3}$. We set $C = 1$ in the rule (7) adopted by the gray-box controller to tune α_k.

Figure 2 illustrates the evolutions of the time-averaged dynamic regrets (i.e., Reg_T^d/T) incurred by such closed-loop interconnections. The direct use of the approximate sensitivity $\hat{H}$ diminishes solution accuracy. Nonetheless, by suitably incorporating this information, the gray-box controller achieves a better performance compared to the model-free controller and the controller with sensitivity learning. Further, for the considered iteration range it closely matches the benchmark with the exact sensitivity.

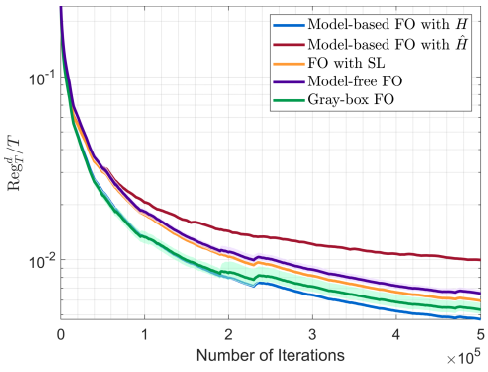

Figure 2. Comparison of different controllers when interconnected with the system (12) to solve the time-varying problem (13). In the legend "FO" and "SL" stand for "feedback optimization" and "sensitivity learning", respectively.

6. Conclusion

We proposed a gray-box feedback optimization controller to optimize the input-output performance of a nonlinear system. This controller merges the approximate input-output sensitivities of the system into model-free updates via an adaptive convex combination. We quantified the accuracy conditions of the sensitivities that render the gray-box approach preferable, and we provided design guidelines for setting combination coefficients therein. We demonstrated that the gray-box controller exploits approximate sensitivities for sample efficiency, and that it circumvents error accumulation and ensures solution accuracy. Future directions include leveraging other forms of prior knowledge or model information, tackling output constraints via dualization, as well as analyzing the interplay between model-free control and online identification.

References

Achterhold, J., Tobuschat, P., Ma, H., Buechler, D., Muehlebach, M., and Stueckler, J. Black-box vs. gray-box: A case study on learning table tennis ball trajectory prediction with spin and impacts. In *Learning for Dynamics and Control Conference*, pp. 878–890, 2023.

Agarwal, N., Bullins, B., Hazan, E., Kakade, S., and Singh, K. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pp. 111–119, 2019.

Bianchin, G., Vaquero, M., Cortés, J., and Dall'Anese, E. Online stochastic optimization for unknown linear systems: Data-driven controller synthesis and analysis. *IEEE Transactions on Automatic Control*, 69(7):4411–4426, 2024.

Boffi, N. M., Tu, S., and Slotine, J.-J. E. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control Conference*, pp. 471–483, 2021.

Cassel, A. B., Cohen, A., and Koren, T. Efficient online linear control with stochastic convex costs and unknown

dynamics. In *Conference on Learning Theory*, pp. 3589–3604, 2022.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, New York, NY, 2006.

Chen, Y., Bernstein, A., Devraj, A., and Meyn, S. Model-free primal-dual methods for network optimization with application to real-time optimal power flow. In *American Control Conference*, pp. 3140–3147, 2020.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.

Gao, X., Jiang, B., and Zhang, S. On the information-adaptive variants of the ADMM: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018.

Hauswirth, A., He, Z., Bolognani, S., Hug, G., and Dörfler, F. Optimization algorithms as robust feedback controllers. *Annual Reviews in Control*, 57, 2024. Art. no. 100941.

Hazan, E. *Introduction to online convex optimization*. MIT Press, Princeton, NJ, 2022.

He, Z., Bolognani, S., He, J., Dörfler, F., and Guan, X. Model-free nonlinear feedback optimization. *IEEE Transactions on Automatic Control*, 69(7):4554–4569, 2024.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online optimization: Competing with dynamic comparators. In *International Conference on Artificial Intelligence and Statistics*, pp. 398–406, 2015.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Li, T., Yang, R., Qu, G., Lin, Y., Wierman, A., and Low, S. H. Certifying black-box policies with stability for nonlinear control. *IEEE Open Journal of Control Systems*, 2:49–62, 2023.

Ma, H., Büchler, D., Schölkopf, B., and Muehlebach, M. Reinforcement learning with model-based feedforward inputs for robotic table tennis. *Autonomous Robots*, 47 (8):1387–1403, 2023.

Ortmann, L., Rubin, C., Scozzafava, A., Lehmann, J., Bolognani, S., and Dörfler, F. Deployment of an online feedback optimization controller for reactive power flow optimization in a distribution grid. In *IEEE PES Innovative Smart Grid Technologies Europe*, 2023.

Picallo, M., Ortmann, L., Bolognani, S., and Dörfler, F. Adaptive real-time grid operation via online feedback optimization with sensitivity estimation. *Electric Power Systems Research*, 212, 2022. Art. no. 108405.

Qu, G., Yu, C., Low, S., and Wierman, A. Exploiting linear models for model-free nonlinear control: A provably convergent policy gradient approach. In *60th IEEE Conference on Decision and Control*, pp. 6539–6546, 2021.

Simchowitz, M. and Foster, D. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning*, pp. 8937–8948, 2020.

Simonetto, A., Dall'Anese, E., Paternain, S., Leus, G., and Giannakis, G. B. Time-varying convex optimization: Time-structured algorithms and applications. *Proceedings of the IEEE*, 108(11):2032–2048, 2020.

Tang, Y., Ren, Z., and Li, N. Zeroth-order feedback optimization for cooperative multi-agent systems. *Automatica*, 148, 2023. Art. no. 110741.

Tsiamis, A., Ziemann, I., Matni, N., and Pappas, G. J. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

Zhao, P., Wang, G., Zhang, L., and Zhou, Z.-H. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(1):5562–5606, 2021.

A. Appendix

A.1. Proof of Lemma 4.1

An upper bound on the expected distance is

$$\begin{aligned}
&\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_k - \nabla\tilde{\Phi}_k(w_k)\|] \\
&\quad= \mathbb{E}_{v_{[k]}}[\|\alpha_k\hat{\phi}_{k,1} + (1-\alpha_k)\hat{\phi}_{k,2} - \nabla\tilde{\Phi}_k(w_k)\|] \\
&\quad\le \alpha_k \underbrace{\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,1} - \nabla\tilde{\Phi}_k(w_k)\|]}_{①} \\
&\qquad+ (1-\alpha_k)\underbrace{\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,2} - \nabla\tilde{\Phi}_k(w_k)\|]}_{②},
\end{aligned} \tag{14}$$

where we used the triangle inequality. For term ① in (14),

$$\begin{aligned}
① &\overset{(s.1)}{\le} \mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,1} - \nabla\tilde{\Phi}_k(u_k)\|] \\
&\quad+ \mathbb{E}_{v_{[k]}}[\|\nabla\tilde{\Phi}_k(u_k) - \nabla\tilde{\Phi}_k(w_k)\|] \\
&\overset{(s.2)}{\le} \mathbb{E}_{v_{[k]}}[\|\epsilon_k\|] + L_k\delta \overset{(s.3)}{\le} M_{\Phi,k}\epsilon_{H,k} + L_k\delta,
\end{aligned} \tag{15}$$

where (s.1) is obtained by adding and subtracting $\nabla\tilde{\Phi}_k(u_k)$ and using the triangle inequality; (s.2) utilizes the expression $\epsilon_k = \hat{\phi}_{k,1} - \nabla\tilde{\Phi}_k(u_k)$, the assumption that $\tilde{\Phi}_k(u)$ is L_k-smooth, and (4e); (s.3) follows from

$$\|\epsilon_k\| = \|(H_k' - \hat{H}_k)^\top\nabla_y\Phi_k(u_k, y_k)\| \le \epsilon_{H,k}M_{\Phi,k}. \tag{16}$$

Let $\tilde{\Phi}_{k,\delta} : \mathbb{R}^p \to \mathbb{R}$ be the smooth approximation of the objective $\tilde{\Phi}_k$ at time k, see also (Gao et al., 2018). Term ② in (14) satisfies

$$\begin{aligned}
② &\le \mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,2} - \nabla\tilde{\Phi}_{k,\delta}(w_k)\|] \\
&\quad+ \mathbb{E}_{v_{[k]}}[\|\nabla\tilde{\Phi}_{k,\delta}(w_k) - \nabla\tilde{\Phi}_k(w_k)\|] \\
&\overset{(s.1)}{\le} \underbrace{\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,2} - \nabla\tilde{\Phi}_{k,\delta}(w_k)\|]}_{③} + \frac{L_kp\delta}{2},
\end{aligned} \tag{17}$$

where (s.1) follows from Lemma 4.1 in (Gao et al., 2018), because $\tilde{\Phi}_k$ is L_k-smooth. For term ③ in (17), we have

$$③ \overset{(s.1)}{\le} \sqrt{\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,2} - \nabla\tilde{\Phi}_{k,\delta}(w_k)\|^2]} \overset{(s.2)}{\le} \sqrt{\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,2}\|^2]},$$

where (s.1) uses the inequality $\forall a \in \mathbb{R}^p, \mathbb{E}^2[\|a\|] \le \mathbb{E}[\|a\|^2]$, and (s.2) holds since $\mathbb{E}_{v_{[k]}}[\hat{\phi}_{k,2}] = \nabla\tilde{\Phi}_{k,\delta}(w_k)$ and the variance of $\hat{\phi}_{k,2}$ is not greater than its second moment. The upper bound on $\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,2}\|^2]$ is

$$\begin{aligned}
&\mathbb{E}_{[k]}[\|\hat{\phi}_{k,2}\|^2] \\
&\overset{(s.1)}{=} \frac{p^2}{\delta^2}\mathbb{E}_{v_{[k]}}[|\tilde{\Phi}_k(w_k+\delta v_k) - \tilde{\Phi}_{k-1}(w_{k-1}+\delta v_{k-1})|^2] \\
&\le \frac{2p^2}{\delta^2}\left(\mathbb{E}_{v_{[k]}}\left[|\tilde{\Phi}_k(w_k+\delta v_k)|^2 + |\tilde{\Phi}_{k-1}(w_{k-1}+\delta v_{k-1})|^2\right]\right) \\
&\overset{(s.2)}{\le} \frac{4p^2G^2}{\delta^2},
\end{aligned} \tag{18}$$

where (s.1) holds since $\forall v_k \sim U(\mathbb{S}_{p-1}), \|v_k\| = 1$; (s.2) uses the boundedness of $\tilde{\Phi}_k$, i.e., $\forall u \in \mathcal{U}_\sigma, k \in \mathbb{N}, |\tilde{\Phi}_k(u)| \le G$. Therefore, ③ $\le \frac{2pG}{\delta}$. We plug this bound into (17). Then, we combine the upper bounds on terms ① and ② with (14) and arrive at Lemma 4.1.

A.2. Proof of Theorem 4.2

Because the optimal point u_k^* lies in $\mathcal{U}$, we know from (4a) and the Pythagorean theorem that

$$\begin{aligned}
\|w_{k+1} - u_k^*\|^2 &\le \|w_k - \eta\hat{\phi}_k - u_k^*\|^2 \\
&= \|w_k - u_k^*\|^2 - 2\eta\hat{\phi}_k^\top(w_k - u_k^*) + \eta^2\|\hat{\phi}_k\|^2.
\end{aligned}$$

We rearrange terms and obtain

$$\hat{\phi}_k^\top(w_k - u_k^*) \le \frac{\eta}{2}\|\hat{\phi}_k\|^2 + \frac{\|w_k - u_k^*\|^2 - \|w_{k+1} - u_k^*\|^2}{2\eta}. \tag{19}$$

Moreover, we know from (4b) that

$$\begin{aligned}
&\mathbb{E}_{v_{[k]}}[\hat{\phi}_k^\top(w_k - u_k^*)] \\
&= \alpha_k\underbrace{\mathbb{E}_{v_{[k]}}[\hat{\phi}_{k,1}^\top(w_k - u_k^*)]}_{①} + (1-\alpha_k)\underbrace{\mathbb{E}_{v_{[k]}}[\hat{\phi}_{k,2}^\top(w_k - u_k^*)]}_{②}.
\end{aligned} \tag{20}$$

For term ① in (20),

$$\begin{aligned}
① &= \mathbb{E}_{v_{[k]}}[\nabla\tilde{\Phi}_k(w_k)^\top(w_k - u_k^*)] \\
&\quad+ \mathbb{E}_{v_{[k]}}[(\hat{\phi}_{k,1} - \nabla\tilde{\Phi}_k(w_k))^\top(w_k - u_k^*)] \\
&\overset{(s.1)}{\ge} \mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(w_k)] - \tilde{\Phi}_k(u_k^*) - D(M_{\Phi,k}\epsilon_{H,k} + L_k\delta),
\end{aligned}$$

where (s.1) utilizes the convexity of $\tilde{\Phi}_k$, the Cauchy-Schwarz inequality, the inequality $\|w_k - u_k^*\| \le D$ (see Assumption 2.2), and the bound (15). For term ② in (20),

$$\begin{aligned}
② &\overset{(s.1)}{=} \mathbb{E}_{v_{[k-1]}}\left[\mathbb{E}_{v_k}[\hat{\phi}_{k,2}^\top(w_k - u_k^*)|v_{[k-1]}]\right] \\
&\overset{(s.2)}{=} \mathbb{E}_{v_{[k-1]}}\left[\mathbb{E}_{v_k}[\hat{\phi}_{k,2}|v_{[k-1]}]^\top(w_k - u_k^*)\right] \\
&\overset{(s.3)}{=} \mathbb{E}_{v_{[k-1]}}\left[\nabla\tilde{\Phi}_{k,\delta}(w_k)^\top(w_k - u_k^*)\right] \\
&\overset{(s.4)}{\ge} \mathbb{E}_{v_{[k]}}\left[\tilde{\Phi}_{k,\delta}(w_k) - \tilde{\Phi}_{k,\delta}(u_k^*)\right] \\
&= \mathbb{E}_{v_{[k]}}\left[\tilde{\Phi}_{k,\delta}(w_k) - \tilde{\Phi}_k(w_k)\right] + \mathbb{E}_{v_{[k]}}\left[\tilde{\Phi}_k(w_k) - \tilde{\Phi}_k(u_k^*)\right] \\
&\quad+ \mathbb{E}_{v_{[k]}}\left[\tilde{\Phi}_k(u_k^*) - \tilde{\Phi}_{k,\delta}(u_k^*)\right] \\
&\overset{(s.5)}{\ge} \mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(w_k)] - \tilde{\Phi}_k(u_k^*) - L\delta^2,
\end{aligned}$$

where (s.1) uses the tower rule; (s.2) holds since $w_k - u_k^*$ is measurable with respect to $v_{[k-1]}$; (s.3) follows from

Lemma 4.1 in (Gao et al., 2018) and the independence of $\Phi_{k-1}(u_{k-1}, y_{k-1})$ and v_k; (s.4) uses the convexity of $\tilde{\Phi}_{k,\delta}$ and the independence of $\tilde{\Phi}_{k,\delta}(w_k)$ and v_k; (s.5) also follows from (Gao et al., 2018). We incorporate the above lower bounds into (20), combine it with (19), and telescope the inequality to obtain

$$\sum_{k=1}^{T}\Big(\mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(w_k)]-\tilde{\Phi}_k(u_k^*)\Big)\le \underbrace{\frac{\eta}{2}\sum_{k=1}^{T}\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_k\|^2]}_{①}$$
$$+\underbrace{\frac{1}{2\eta}\sum_{k=1}^{T}\left(\mathbb{E}_{v_{[k]}}[\|w_k-u_k^*\|^2]-\mathbb{E}_{v_{[k]}}[\|w_{k+1}-u_k^*\|^2]\right)}_{②}$$
$$+D\sum_{k=1}^{T}\alpha_k(M_{\Phi,k}\epsilon_{H,k}+L_k\delta)+\sum_{k=1}^{T}(1-\alpha_k)L\delta^2. \quad (21)$$

For term ① in (21), we have

$$①\le\sum_{k=1}^{T}\Big(2\alpha_k^2\,\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,1}\|^2]+2(1-\alpha_k)^2\,\mathbb{E}_{v_{[k]}}[\|\hat{\phi}_{k,2}\|^2]\Big)$$
$$\overset{\text{(s.1)}}{\le}\sum_{k=1}^{T}\Big(4\alpha_k^2(M_k^2+M_{\Phi,k}^2\epsilon_{H,k}^2)+8(1-\alpha_k)^2\frac{p^2G^2}{\delta^2}\Big).$$

In (s.1), we use (16) and $\|\hat{\phi}_{k,1}\|^2=\|\nabla\tilde{\Phi}_k(u_k)-\epsilon_k\|^2\le 2M_k^2+2\|\epsilon_k\|^2$. We also utilize the upper bound (18). Furthermore, term ② in (21) satisfies

$$②\le\mathbb{E}_{v_{[k]}}[\|w_1\|^2]-2\,\mathbb{E}_{v_{[k]}}[w_1^\top u_1^*]+2\,\mathbb{E}_{v_{[k]}}[w_{T+1}^\top u_T^*]$$
$$+2\sum_{k=1}^{T-1}\mathbb{E}_{v_{[k]}}[w_{k+1}^\top(u_k^*-u_{k+1}^*)]$$
$$\overset{\text{(s.1)}}{\le}5\overline{D}^2+2\overline{D}\sum_{k=1}^{T-1}\|u_k^*-u_{k+1}^*\|,$$

where (s.1) uses the Cauchy-Schwarz inequality and the fact that $\forall u\in\mathcal{U},\|u\|\le\overline{D}$, see also the discussion below Assumption 2.4. By incorporating the above bounds into (21) and invoking the parametric conditions, we have

$$\mathrm{Reg}_T^d\le\overline{D}p^{\frac{2}{3}}T^{\frac{3}{4}}C_T+D\sum_{k=1}^{T}\alpha_k\epsilon_{H,k}M_{\Phi,k}$$
$$+\underbrace{2\eta\sum_{k=1}^{T}\alpha_k^2M_k^2+4\eta\sum_{k=1}^{T}(1-\alpha_k)^2\frac{p^2G^2}{\delta^2}+\frac{5\overline{D}^2}{2\eta}}_{\sim\mathcal{O}\left(p^{\frac{2}{3}}T^{\frac{3}{4}}\right)}$$
$$+\underbrace{D\sum_{k=1}^{T}\alpha_kL_k\delta+\sum_{k=1}^{T}(1-\alpha_k)L\delta^2+2\eta\sum_{k=1}^{T}\alpha_k^2M_{\Phi,k}^2\epsilon_{H,k}^2}_{\sim o(p^{\frac{2}{3}}T^{\frac{3}{4}})}.$$

Therefore, (10) holds. Furthermore, when $\hat{H}_k$ satisfies (6) and $\{\alpha_k\}_{k=1}^T$ are set by (7), we have

$$\sum_{k=1}^{T}\alpha_k\epsilon_{H,k}\le\sum_{k=1}^{T}\frac{C\bar{\epsilon}}{(k+1)^{\frac{1}{4}}}\le C\bar{\epsilon}\int_0^T\frac{1}{(x+1)^{\frac{1}{4}}}\mathrm{d}x$$
$$=\frac{4}{3}C\bar{\epsilon}[(T+1)^{\frac{3}{4}}-1].$$

We can perform a similar derivation when $\hat{H}_k$ satisfies (8) and $\{\alpha_k\}_{k=1}^T$ are given by (9). Hence, the order (11) of the dynamic regret Reg_T^d is proved.

A.3. Proof of Corollary 4.4

We leverage the following decomposition

$$\mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(u_k)]-\tilde{\Phi}_k(u_k^*)$$
$$=\mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(u_k)]-\mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(w_k)]+\mathbb{E}_{v_{[T]}}[\tilde{\Phi}_k(w_k)]-\tilde{\Phi}(u_k^*)$$
$$\overset{\text{(s.1)}}{\le}M_{\Phi,k}\delta+\mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(w_k)]-\tilde{\Phi}(u_k^*),$$

where (s.1) follows from the inequality $\forall a\in\mathbb{R},a\le|a|$, the property that $\tilde{\Phi}_k$ is $M_{\Phi,k}$-Lipschitz (see Assumption 2.3), as well as $\|v_k\|=1$. Hence,

$$\sum_{k=1}^{T}\Big(\mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(u_k)]-\tilde{\Phi}_k(u_k^*)\Big)$$
$$\le\delta\sum_{k=1}^{T}M_{\Phi,k}+\sum_{k=1}^{T}\Big(\mathbb{E}_{v_{[k]}}[\tilde{\Phi}_k(w_k)]-\tilde{\Phi}_k(u_k^*)\Big).$$

Since $\{M_{\Phi,k}\}$ is bounded and $\delta\le p^{\frac{1}{3}}/T^{\frac{1}{4}}$, the order of $\delta\sum_{k=1}^{T}M_{\Phi,k}$ is $\mathcal{O}(p^{\frac{1}{3}}T^{\frac{3}{4}})$. We combine this result with (11) and prove the corollary.