Handling Heterogeneous Curvatures in Bandit LQR Control

Yu-Hu Yan [1 2] **Jing Wang** [1 2] **Peng Zhao** [1 2]

Abstract

We investigate online Linear Quadratic Regulator (LQR) with bandit feedback and semi-adversarial disturbances. Previous works assume costs with *homogeneous* curvatures (i.e., with a uniform strong convexity lower bound), which can be hard to satisfy in many real scenarios and prohibits adapting to true curvatures for better performance. In this paper, we initiate the study of bandit LQR control with *heterogeneous* cost curvatures, aiming to strengthen the algorithm's adaptivity. To achieve this, we reduce the problem to bandit convex optimization with memory via a "with-history" reduction to avoid hard-to-control truncation errors. Then we provide a novel analysis for an important *stability* term that appeared in both regret and memory, using *Newton decrement* developed in interior-point methods. The analysis enables us to guarantee memory-related terms introduced in the reduction and also provide a simplified analysis for handling heterogeneous curvatures in bandit convex optimization. Finally, we achieve interpolated guarantees that can not only recover existing bounds for convex and quadratic costs but also attain new implications for cases of corrupted and decaying quadraticity.

1. Introduction

There have been extensive studies on Linear Quadratic Regulator (LQR) and the more general Linear Quadratic Gaussian (LQG) (Bellman, 1954; Kalman, 1960; Abbasi-Yadkori and Szepesvári, 2011; Lewis et al., 2012; Cohen et al., 2018; Dean et al., 2018; Mania et al., 2019). Specifically, LQR considers controlling the following linear dynamical system:

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \boldsymbol{\xi}_t,$$

where $\mathbf{x}_t, \mathbf{u}_t$ denote the state and action, A, B are the system transition matrices, and $\boldsymbol{\xi}_t$ is the disturbance. The system evolves for T rounds, and at the t-th round, the learner selects an action $\mathbf{u}_t$ and observes the next state $\mathbf{x}_{t+1}$. The goal of LQR is to minimize the following cumulative cost:

$$\min_{\pi\in\Pi} \sum_{t=1}^{T} c_t(\mathbf{x}_t^\pi, \mathbf{u}_t^\pi) \triangleq \sum_{t=1}^{T} {\mathbf{x}_t^\pi}^\top Q_t \mathbf{x}_t^\pi + \sum_{t=1}^{T} {\mathbf{u}_t^\pi}^\top R_t \mathbf{u}_t^\pi,$$

where $c_t(\cdot,\cdot)$ represents the quadratic cost, parameterized by the time-varying, positive semi-definite matrices Q_t, R_t. Here, $\mathbf{x}_t^\pi, \mathbf{u}_t^\pi$ denote the state and action generated by a specific policy π, chosen from a policy class Π. In the online version of LQR, the learner aims to minimize the game-theoretic *policy regret* (Dekel et al., 2012), which depicts the excess cumulative cost against the best policy in hindsight:

$$\textsc{Reg}_T^{\mathrm{C}} \triangleq \sum_{t=1}^{T} c_t(\mathbf{x}_t, \mathbf{u}_t) - \min_{\pi\in\Pi} \sum_{t=1}^{T} c_t(\mathbf{x}_t^\pi, \mathbf{u}_t^\pi), \qquad (1.1)$$

where $\mathbf{x}_t, \mathbf{u}_t$ are produced by the online algorithm.

In this work, we investigate online LQR in the case where the system transition matrices A, B are known, the learner only obtains a scalar cost in each round (also known as *bandit* feedback in online learning), and the disturbances are *semi-adversarial* — a mixture of stochastic and adversarial parts. In this setup, Sun et al. (2023) obtained an optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound, up to logarithmic factors. Moreover, when the cost functions are generally convex, this problem intersects with recent studies in *online non-stochastic control* (Hazan and Singh, 2022). Notably, the pioneering work (Agarwal et al., 2019) attained an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound through a novel reduction to online convex optimization with memory (Anava et al., 2015). The most related studies to ours, particularly in the bandit setup, are those by (Gradu et al., 2020; Cassel and Koren, 2020). The authors obtained an $\widetilde{\mathcal{O}}(T^{3/4})$ regret bound for Lipschitz and convex cost functions with adversarial disturbances, with Cassel and Koren (2020) further achieving an improved $\widetilde{\mathcal{O}}(T^{2/3})$ result for smooth cost functions.

The latest work for bandit LQR control (Sun et al., 2023) has achieved an $\widetilde{\mathcal{O}}(\alpha_{\min}^{-1}\sqrt{T})$ regret, where $\alpha_{\min}$ is the minimum cost curvatures, i.e., $Q_t, R_t \succeq \alpha_t I$ and $\alpha_{\min} = \min_t \alpha_t$. While optimal in T, it presents unfavorable curvature considerations for two reasons: *(i)* the learner must have

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China [2]School of Artificial Intelligence, Nanjing University, China. Correspondence to: Peng Zhao <zhaop@lamda.nju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1: A summary of our results for bandit LQR control with heterogeneous curvatures. Below, α_t is the curvature of the t-th quadratic cost (i.e., $Q_t, R_t \succeq \alpha_t I$). Our general results are given in Theorem 3 and we provide four special cases: pure convexity, pure quadraticity, corruptions in quadraticity, and decaying quadraticity, to demonstrate the adaptivity and robustness of our results. We use $\mathbb{1}_{[\cdot]}$ to denote the indicator function and $\mathcal{T} \subseteq [T]$ for the rounds with no quadraticity. Notably, for either Lipschitz or smooth cost functions, the results in four special cases can be achieved by one single algorithm. We distinguish the previous results by using a gray background color.

	Convexity ($\alpha_t = 0$)	**Quadraticity** ($\alpha_t = \alpha > 0$)	**Corrupted Quadraticity** ($\alpha_t = \alpha \cdot \mathbb{1}_{t \notin \mathcal{T}}$)	**Decaying Quadraticity** ($\alpha_t = t^{-\gamma}$)
Lipschitz Functions	$\widetilde{\mathcal{O}}(T^{3/4})$ (Gradu et al., 2020)	N/A	N/A	N/A
	$\widetilde{\mathcal{O}}(T^{3/4})$ [Corollary 1]	$\widetilde{\mathcal{O}}(T^{2/3})$ [Corollary 1]	(when $\lvert\mathcal{T}\rvert = T^{8/9}$) $\widetilde{\mathcal{O}}(T^{2/3})$ [Corollary 3]	$\begin{cases}\widetilde{\mathcal{O}}(T^{2/3+\gamma/3}), & \gamma \in [0, 1/4] \\ \widetilde{\mathcal{O}}(T^{3/4}), & \gamma \in (1/4, 1]\end{cases}$ [Corollary 5]
Smooth Functions	$\widetilde{\mathcal{O}}(T^{2/3})$ (Cassel and Koren, 2020)	$\widetilde{\mathcal{O}}(\sqrt{T})$ (Sun et al., 2023)	N/A	N/A
	$\widetilde{\mathcal{O}}(T^{2/3})$ [Corollary 2]	$\widetilde{\mathcal{O}}(\sqrt{T})$ [Corollary 2]	(when $\lvert\mathcal{T}\rvert = T^{3/4}$) $\widetilde{\mathcal{O}}(\sqrt{T})$ [Corollary 4]	$\begin{cases}\widetilde{\mathcal{O}}(T^{1/2+\gamma/2}), & \gamma \in [0, 1/3] \\ \widetilde{\mathcal{O}}(T^{2/3}), & \gamma \in (1/3, 1]\end{cases}$ [Corollary 6]

access to the curvature lower bound $\alpha_{\min}$, which can be hard to obtain before an algorithm initializes. This requirement restricts the algorithm's adaptivity to true curvatures (i.e., $\alpha_t \geq \alpha_{\min}$) for better performance; and *(ii)* when there are corruptions in the cost functions, e.g., there exist generally convex costs in the function sequence, algorithms designed for quadratic costs become unfeasible since $\alpha_{\min} = 0$. And algorithms for convex costs will significantly degrade the performance — e.g., the regret will become from $\widetilde{\mathcal{O}}(\sqrt{T})$ to $\widetilde{\mathcal{O}}(T^{2/3})$ for smooth costs. Therefore, a natural question arises: *Is it possible design an algorithm that is adaptive to heterogeneous curvatures for better performance and robust to (possibly) corrupted cost functions?*

Motivated by the question above, we focus on heterogeneous curvatures in bandit LQR control. As far as we know, the study of heterogeneous curvatures in control has been largely unexplored, with the exception of the work of Muthirayan et al. (2022). Specifically, they study the full information feedback, where the learner attains complete knowledge of costs. By contrast, we focus on the more challenging bandit feedback, which is common in many real-world applications where getting adequate feedback is hard.

Results. In this work, we achieve interpolated results that are adaptive to the true curvatures of costs. Our results can recover existing bounds and imply new ones in certain cases. Specifically, for Lipschitz costs, we obtain interpolated bounds between $\widetilde{\mathcal{O}}(T^{3/4})$ for convex functions (Gradu et al., 2020; Cassel and Koren, 2020) and $\widetilde{\mathcal{O}}(T^{2/3})$ for quadratic functions, a novel result in bandit LQR control. For smooth costs, we achieve interpolated results between $\widetilde{\mathcal{O}}(T^{2/3})$ (Cassel and Koren, 2020) and $\widetilde{\mathcal{O}}(\sqrt{T})$ (Sun et al., 2023). Moreover, our results also imply meaningful guarantees in the intermediate cases where corrupted or decaying quadraticity exists. For instance, our results can maintain the desired $\widetilde{\mathcal{O}}(\sqrt{T})$ bound for smooth costs even when $\mathcal{O}(T^{3/4})$ in T functions are not quadratic. And for Lipschitz costs, the desired $\widetilde{\mathcal{O}}(T^{2/3})$ bound is attainable even when $\mathcal{O}(T^{8/9})$ of the cost functions are not quadratic, thereby greatly enhancing our method's robustness. Table 1 summarizes our results.

Techniques. Many online decision-making tasks can be reduced to online learning with memory, an online model capturing the impact of past decisions in the present. Prior studies used a truncation-based reduction with easy-to-control truncation errors. However, heterogeneous curvatures with bandit feedback requires the usage of self-concordant barriers (Nesterov and Nemirovskii, 1994) as the regularizer in online update, which are inherently unbounded near the domain boundary and will thus make truncation errors hard-to-control. To avoid it, we adopt a *"with-history"* reduction scheme proposed by recent studies (Sun et al., 2023), which admits a lossless reduction. Consequently, we address the reduced problem within a *non-oblivious* adversary setup.

In both the regret and memory analysis, a key *stability term*, which captures the switching of the decisions based on certain local measures, is important. Luo et al. (2022) conducted initial studies on this term in Bandit Convex Optimization (BCO) with heterogeneous curvatures, using the proof argument by contradiction and some local stability analysis of self-concordant barriers. In this work, we further identify the importance of this term in the memory analysis and provide a simple analysis for it using *Newton decrement* developed in interior-point methods (Nesterov and Nemirovskii, 1994). This enables us to also provide a simplified regret analysis for BCO with heterogeneous curvatures. Our main technical finding is given in Theorem 1.

Organization. The rest of the paper is structured as follows. Section 2 provides preliminaries of the problem setup

and previous works on handling heterogeneous curvatures. Section 3 presents our method and key analysis for bandit LQR control with heterogeneous curvatures. Finally, Section 4 concludes the work. Due to page limits, most proofs are deferred to appendices.

2. Preliminaries

In this section, we introduce some preliminary knowledge, including our problem setup and assumptions in Section 2.1, and the latest progress on handling heterogeneous curvatures in online convex optimization in Section 2.2.

2.1. Problem Setup

In this work, we investigate online Linear Quadratic Regulator (LQR) control with bandit feedback, partial observations (sometimes referred to as LQG control in literature), and semi-adversarial disturbances. Concretely, we consider controlling the following linear dynamical system:

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \boldsymbol{\xi}_t, \quad \mathbf{y}_t = C\mathbf{x}_t + \mathbf{e}_t, \tag{2.1}$$

where $A \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{x}}}$, $B \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{u}}}$, $C \in \mathbb{R}^{d_{\mathbf{y}} \times d_{\mathbf{x}}}$ are known system transition matrices. Here, $\mathbf{x}_t \in \mathbb{R}^{d_{\mathbf{x}}}$ and $\mathbf{u}_t \in \mathbb{R}^{d_{\mathbf{u}}}$ represent the state and action respectively, while $\mathbf{y}_t \in \mathbb{R}^{d_{\mathbf{y}}}$ is a partial observation of the state. The learner can only observe a *bandit* feedback, i.e., a scalar value of $c_t(\mathbf{y}_t, \mathbf{u}_t)$, without access to the full function. Notably, same as Cassel and Koren (2020); Gradu et al. (2020); Sun et al. (2023), we consider the problem within an *oblivious* setup — the cost functions and disturbances are chosen by the environments in advance before the algorithm starts.

In the following, we list the assumptions used in this work.

Assumption 1 (Stablility)**.** The system is stable, i.e., the spectral radius $\rho(A) < 1$.

Assumption 2 (Disturbance)**.** The disturbances $\{\boldsymbol{\xi}_t, \mathbf{e}_t\}_{t=1}^T$ are semi-adversarial: $\boldsymbol{\xi}_t = \boldsymbol{\xi}_t^{\text{adv}} + \boldsymbol{\xi}_t^{\text{sto}}$ and $\mathbf{e}_t = \mathbf{e}_t^{\text{adv}} + \mathbf{e}_t^{\text{sto}}$, where $\mathbb{E}[\boldsymbol{\xi}_t^{\text{sto}}] = \mathbb{E}[\mathbf{e}_t^{\text{sto}}] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\xi}_t^{\text{sto}} \boldsymbol{\xi}_t^{\text{sto}\top}] \succeq \text{Var}_{\boldsymbol{\xi}} \cdot I$, $\mathbb{E}[\mathbf{e}_t^{\text{sto}} \mathbf{e}_t^{\text{sto}\top}] \succeq \text{Var}_{\mathbf{e}} \cdot I$. Besides, $\|\boldsymbol{\xi}_t\|_2, \|\mathbf{e}_t\|_2 \le W$.

Assumption 3 (Cost)**.** The cost $c_t(\cdot, \cdot)$ is quadratic:

$$c_t(\mathbf{y}, \mathbf{u}) = \mathbf{y}^\top Q_t \mathbf{y} + \mathbf{u}^\top R_t \mathbf{u},$$

non-negative, α_t-strongly convex and β_c-smooth:

$$Q_t, R_t \succeq \alpha_t I, \quad \nabla^2 c_t(\cdot, \cdot) \preceq \beta_c I,$$

and Lipschitz: for all $(\mathbf{y}, \mathbf{u}), (\mathbf{y}', \mathbf{u}') \in \mathbb{R}^{d_{\mathbf{y}} + d_{\mathbf{u}}}$

$$|c_t(\mathbf{y}, \mathbf{u}) - c_t(\mathbf{y}', \mathbf{u}')| \le L_c R_c \|(\mathbf{y} - \mathbf{y}', \mathbf{u} - \mathbf{u}')\|_2,$$

where $R_c \triangleq \max\{\|(\mathbf{y}, \mathbf{u})\|_2, 1\}$.

The assumptions are common in the literature. Specifically, Assumption 1 can be extended to strongly stabilizable systems, which are unstable but can be stabilized by a linear controller, due to the reduction proposed in Appendix A of Cassel et al. (2022). And Assumption 2 is typically considered when dealing with strongly convex cost functions (Simchowitz et al., 2020; Sun et al., 2023). In Assumption 3, while we provide a detailed list of various functional properties, not all of them are invoked for each result. The related assumptions will be specified as needed.

Notations. For clarity, we use bold symbols (e.g., $\mathbf{x}$) within the control context and italic symbols (e.g., w or $\boldsymbol{w}$) within the online learning context. For simplicity, we define $x_{a:b} \triangleq \sum_{i=a}^b x_i$ and $x_{[a:b]} \triangleq (x_a, \dots, x_b)$ for variable x. A function is described as α-quadratic when it is quadratic with $Q_t, R_t \succeq \alpha I$. We denote by $\|w\|_A \triangleq \sqrt{w^\top A w}$ a local norm for any w and positive semi-definite matrix A.

2.2. Handling Heterogeneous Curvature

In online convex optimization (Hazan, 2016), the learner submits a decision w_t inside a convex compact set $\mathcal{W} \subseteq \mathbb{R}^d$, aiming to minimize the regret (Cesa-Bianchi and Lugosi, 2006) defined on convex loss functions $h_t : \mathcal{W} \mapsto \mathbb{R}$, i.e.,

$$\text{Reg}_T \triangleq \sum_{t=1}^T h_t(w_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^T h_t(w). \tag{2.2}$$

The problem of handling heterogeneous curvatures can be traced back to the seminal work of Adaptive Online Gradient Descent (AOGD) (Bartlett et al., 2007), which studies σ_t-strongly convex loss function $h_t(\cdot)$ and achieves adaptive results to heterogeneous curvatures of loss functions.

The key idea of AOGD is implementing online gradient descent (Zinkevich, 2003) on *regularized* functions $\{\widetilde{h}_t(\cdot)\}_{t=1}^T$, where $\widetilde{h}_t(w) \triangleq h_t(w) + \frac{\lambda_t}{2}\|w\|_2^2$ for any $w \in \mathcal{W}$, with λ_t as the regularization coefficient. Intuitively, AOGD chooses a larger λ_t to add more curvatures to the functions for faster rates. As a price, this method requires a delicate balance between the regularization term, dominated by $\mathcal{O}(1/(\sigma_{1:t} + \lambda_{1:t}))$, and a bias term of $\mathcal{O}(\lambda_t)$. As a consequence, AOGD interpolates between the $\mathcal{O}(\sqrt{T})$ regret for convex functions (Zinkevich, 2003) and the $\mathcal{O}(\log T)$ guarantee for strongly convex functions (Hazan et al., 2007).

Addressing heterogeneous curvatures with bandit feedback is more challenging and was recently solved by Luo et al. (2022). Similar to AOGD, their method also optimizes the regularized loss functions. Differently, to handle the bandit feedback, they use Follow-the-Regularized-Leader (FTRL) (Abernethy et al., 2008) with self-concordant barriers and a gradient estimator with shrinking sampling, following Hazan and Levy (2014). The algorithmic details will be illuminated again in Section 3.2.1. In their work,

the key challenge comes from a hard-to-analyze *local-norm* term of $\|\nabla\psi(w)\|_{\nabla^{-2}\psi(w)}$, where $\psi(\cdot)$ is the FTRL regularizer. To handle this, they operate in a *lifted* domain $\boldsymbol{\mathcal{W}} \triangleq \{\boldsymbol{w} = (w, 1) \mid w \in \mathcal{W}\}$ and construct a corresponding normal barrier in $\boldsymbol{\mathcal{W}}$. This method enables bounding the above term through the advantageous properties of the normal barrier, inspired by Lee et al. (2020).

Self-concordant barriers, which becomes infinity near the domain boundary, are commonly used in interior-point methods for convex optimization (Nesterov and Nemirovskii, 1994). For any closed convex set, there always exists a corresponding self-concordant barrier. And a normal barrier in the lifted domain can always be constructed from a self-concordant barrier of the original domain. Interested readers can refer to Appendix A for formal definitions and properties about self-concordant and normal barriers.

To conclude, Luo et al. (2022) achieved interpolated results between $\widetilde{\mathcal{O}}(T^{2/3})$ (Saha and Tewari, 2011) for convex functions and $\widetilde{\mathcal{O}}(\sqrt{T})$ (Hazan and Levy, 2014) for strongly convex ones when functions are smooth, as well as interpolation between $\widetilde{\mathcal{O}}(T^{3/4})$ (Flaxman et al., 2005) and $\widetilde{\mathcal{O}}(T^{2/3})$ for Lipschitz functions. Furthermore, they investigated cases where the cost functions are mixtures of convex and strongly convex functions or have a decaying curvature coefficient, and achieved meaningful guarantees therein.

Our work falls under the topic of *universal* online learning, where the learner aims to design a single method capable of handling online functions with potentially different properties. Existing studies consider two kinds of universality: the curvatures are known but heterogeneous (Bartlett et al., 2007; Luo et al., 2022), or homogeneous but unknown (van Erven and Koolen, 2016; Zhang et al., 2022; Yan et al., 2023). While there are extensive studies on both threads with full information feedback, existing research with bandit feedback has only made progress on known but heterogeneous curvatures due to limited information. Our problem also falls in this category. Exploring homogeneous but unknown curvatures with bandit feedback remains an important future direction for investigation.

3. Our Method

This section proposes our method for handling heterogeneous curvatures in bandit LQR control. Section 3.1 reduces the problem to BCO with memory (switching cost). Section 3.2 handles heterogeneous curvatures in the reduced online learning problem. Finally, Section 3.3 applies the results back to the control setup.

3.1. "With-History" Reduction to BCO with Memory

In this part, we reduce bandit LQR control to BCO with memory via a with-history reduction scheme, where the memory can be further transformed into the stability analysis between consecutive decisions. The reduction builds on existing progress in handling partial observations and bandit feedback in the control problem (Cassel and Koren, 2020; Simchowitz et al., 2020; Sun et al., 2023).

Initially, we introduce the notion of "Nature's y" following Simchowitz et al. (2020), which is intuitively an external observation of the cumulative impact of disturbances.

Definition 1. *Nature's y (denoted by $\mathbf{y}^{\mathrm{nat}}$) is the observation without any action on the system. In system (2.1), in the t-th round, given disturbances $\boldsymbol{\xi}_{1:t}, \mathbf{e}_{1:t}$, Nature's y is defined as $\mathbf{y}_t^{\mathrm{nat}} \triangleq \mathbf{e}_t + \sum_{i=1}^{t-1} CA^{i-1}\boldsymbol{\xi}_{t-i}$.*

With Nature's y, we introduce the Disturbance-Response Policy (DRP) (Simchowitz et al., 2020), which is effective for handling partial observations.

Definition 2. *Given Nature's $\mathbf{y}_{t-m+1:t}^{\mathrm{nat}}$, a disturbance-response policy, parameterized by an m-length tuple of matrices $M = (M^{[0]}, \ldots, M^{[m-1]})$, chooses the action as $\mathbf{u}_t(M) = \sum_{i=0}^{m-1} M^{[i]}\mathbf{y}_{t-i}^{\mathrm{nat}}$. The DRP policy class is defined as $\mathcal{M} \triangleq \{M \mid \sum_{i=0}^{m-1} \|M^{[i]}\|_{\mathrm{op}} \le R\}$.*

Next we reduce the problem to BCO with memory. Importantly, due to the property of Nature's y, it holds that $\mathbf{y}_t = \mathbf{y}_t^{\mathrm{nat}} + \sum_{i=1}^{t-1} G^{[i]}\mathbf{u}_{t-i}(M_{t-i})$, where $G^{[i]} \triangleq CA^{i-1}B$ is the Markov operator of the system (2.1). Consequently, the cost $c_t(\mathbf{y}_t, \mathbf{u}_t)$ can be reinterpreted as a function F_t of the policies, i.e., $F_t(M_{[1:t]})$. This reformulation leads to a problem with *unbounded* memory, which has a lower bound of $\Omega(T)$ (Cesa-Bianchi et al., 2013) and thus hard to handle.

To address this, previous works (Agarwal et al., 2019; Simchowitz et al., 2020) proposed a *truncation-based* reduction, which artificially erases the effect of actions of more than H rounds before. The reduction imports a truncation error, which can be easily handled in previous setups. For illustration, we decompose observation $\mathbf{y}_t$ as follows:

$$\mathbf{y}_t = \mathbf{y}_t^{\mathrm{nat}} + \underbrace{\sum_{i=1}^{H} G^{[i]}\mathbf{u}_{t-i}(M_{t-i})}_{\textsc{Hist-I}} + \underbrace{\sum_{i=H+1}^{t-1} G^{[i]}\mathbf{u}_{t-i}(M_{t-i})}_{\textsc{Hist-II}},$$

where HIST-I includes the most recent H actions, i.e., $M_{[t-H:t-1]}$, and HIST-II comprises actions that are more than H steps away, i.e., $M_{[1:t-H-1]}$. Given the system's stability, previous works focus only on HIST-I, truncate HIST-II, define a truncated observation $\widetilde{\mathbf{y}}_t \triangleq \mathbf{y}_t^{\mathrm{nat}} + (\textsc{Hist-I})$, and finally obtain a loss function with bounded history, i.e., $f_t(M_{[t-H:t]}) \triangleq c_t(\widetilde{\mathbf{y}}_t, \mathbf{u}_t)$, with an ignorable truncation error overhead. By doing this, they can reduce the problem to BCO with bounded memory.

However, this truncated-based reduction will *fail* in our problem. This is mainly because the truncation error, i.e.,

$f_t(M_{[t-H:t]}) \neq c_t(\mathbf{y}_t, \mathbf{u}_t)$, can be greatly enlarged when incorporating self-concordant barriers (necessary for handling heterogeneous curvatures with bandit feedback), a concept in interior-point methods (Nesterov and Nemirovskii, 1994). Barrier functions are inherently unbounded near the domain boundary and will make the bias hard to control, thus making the truncation-based reduction fail for our purposes.

In this work, to avoid the hard-to-control truncation errors, we leverage a *"with-history"* reduction scheme (Sun et al., 2023), which allows a *lossless* reduction to BCO with memory. Intuitively, instead of truncating HIST-II, the method integrates it into the cost function to define a with-history function. A formal definition is provided below.

Definition 3. *Given the Markov operator G of system (2.1) and a cost function $c_t(\cdot,\cdot)$, the with-history loss function $f_t : \mathcal{M}^{H+1} \mapsto \mathbb{R}$ is defined as* $f_t(N_0, \ldots, N_H) \triangleq$

$$c_t\left(\mathbf{y}_t^{\text{nat}} + \sum_{i=1}^{H} G^{[i]}\mathbf{u}_{t-i}(N_{H-i}) + (\text{HIST-II}), \mathbf{u}_t(N_H)\right).$$

Accordingly, we define $\widetilde{f}_t(N) \triangleq f_t(N, \ldots, N)$.

The with-history function enjoys a notable lossless property: $f_t(M_{[t-H:t]}) = c_t(\mathbf{y}_t, \mathbf{u}_t)$, effectively avoiding the hard-to-control truncation errors mentioned earlier. However, the lossless property comes with a price: the with-history function $f_t(\cdot)$ can depend on past decisions $M_{[1:t-H-1]}$, making the reduced online learning problem a *non-oblivious* setup. Handling a non-oblivious adversary is typically challenging in BCO (Flaxman et al., 2005). The key difficulty is that the comparator $w^\star \in \arg\min_{w\in\mathcal{W}} \sum_{t\in[T]} h_t(w)$ becomes a random variable as $h_t(\cdot)$ may depend on the algorithm's past decisions, requiring analysis specially designed for the non-oblivious setup. Fortunately, in bandit control, the compared policy $M^\star \in \arg\min_{M\in\mathcal{M}} \sum_{t=1}^{T} c_t(\mathbf{y}_t(M), \mathbf{u}_t(M))$ is determined by the cost functions, hence *deterministic* since the costs are obliviously chosen. This characteristic avoids the undesired randomness and enables us to handle the reduced non-oblivious online learning problem.

Consequently, we reduce the problem to BCO with memory, up to constant errors. The regret over the with-history losses is defined and further decomposed as

$$\begin{aligned}
&\text{REG}_T^{\text{M}} \triangleq \sum_{t=H'}^{T} f_t(M_{[t-H:t]}) - \sum_{t=H'}^{T} \widetilde{f}_t(M^\star) \\
&\leq \underbrace{\sum_{t=1}^{T}\left(\widetilde{f}_t(M_t) - \widetilde{f}_t(M^\star)\right)}_{\text{UNARY-REG}} + \underbrace{\sum_{t=1}^{T}\left(f_t(M_{[t-H:t]}) - \widetilde{f}_t(M_t)\right)}_{\text{MEMORY}},
\end{aligned}$$

where $H' \triangleq H + m$ is the effective memory length. In the last line, the first term (the unary regret) is the standard regret defined on the unary function $\{\widetilde{f}_t(\cdot)\}_{t=1}^T$. And the second term (the memory term) is essential in control problems and other online decision-making tasks, e.g., in online Markov decision process (Even-Dar et al., 2009; Zhao et al., 2022). Next we analyze these two terms respectively.

Optimizing the unary regret necessitates $\nabla\widetilde{f}_t(M_t)$, which can be estimated using the value of $\widetilde{f}_t(M_t)$ (Flaxman et al., 2005). However, the learner only obtains a scalar value of $f_t(M_{[t-H:t]})$, i.e., $c_t(\mathbf{y}_t, \mathbf{u}_t)$, which leads to a mismatch between $(M_{t-H}, \ldots, M_t)$ and $(M_t, \ldots, M_t)$. To this end, we adopt a lazy-update method (Cassel and Koren, 2020). Specifically, at the t-th round, the learner draws a random bit $b_t \sim \text{Bern}(1/H)$, where $\text{Bern}(1/H)$ is a Bernoulli distribution with parameter $1/H$. Once $b_t \prod_{i=1}^{H-1}(1-b_{t-i}) = 1$, indicating that $M_{t-H} = \cdots = M_t$, the learner estimates the gradient using $f_t(M_{[t-H:t]})$. While the updates occur over a smaller time horizon rather than the entire one, the unary regret will be only $3H$ times larger. The typical choice of $H = \Theta(\log T)$ will not affect the final bound.

It can be verified that the Lipschitzness and smoothness of cost functions can be inherited to the with-history functions, which is formally stated in Lemma 8 in Appendix B. When the cost function $c_t(\cdot,\cdot)$ is L_c-Lipschitz, $f_t(\cdot)$ will be coordinate-wise L_f-Lipschitz, and the memory term can be bounded in the following way:

$$\text{MEMORY} \leq \frac{1}{2} L_f H^2 \sum_{t\in\mathcal{S}} (\vartheta_t + 2\delta_t), \tag{3.1}$$

where $\bar{M}_t = \mathbb{E}[M_t]$ denotes the policy without randomness, $\vartheta_t \triangleq \|\bar{M}_{t+1} - \bar{M}_t\|_{\text{F}}$ measures the switching costs, and $\delta_t \triangleq \|M_t - \bar{M}_t\|_{\text{F}}$ represents the variance. When $\widetilde{f}_t(\cdot)$ is L_f-Lipschitz and β_f-smooth, the upper bound becomes

$$\text{MEMORY} \leq \frac{1}{2} H^2 \sum_{t\in\mathcal{S}} (L_f\vartheta_t + \beta_f\vartheta_t^2 + 6\beta_f\delta_t^2). \tag{3.2}$$

To conclude, we reduce the problem to BCO with Switching Cost (BCO-SC), by upper-bounding the memory with the *switching* of policies (Cassel and Koren, 2020, Theorem 9).

3.2. Heterogeneous Curvature with Switching Cost

In this part, we address the BCO-SC (switching cost) problem with heterogeneous curvatures. For clarity, we frame the problem within the online convex optimization setup, as introduced in Section 2.2. Denoting by $\mathbb{E}[\text{REG}_T]$ the expectation of (2.2), we investigate the following regret of

$$\mathbb{E}[\text{REG}_T] + \sum_{t=2}^{T} \|w_t - \bar{w}_t\|_2 + \sum_{t=2}^{T} \|\bar{w}_t - \bar{w}_{t-1}\|_2 \tag{3.3}$$

Algorithm 1 Subroutine of Luo et al. (2022)

Input: bandit value $h_t(w_t)$, curvature σ_t, last-round decision $\bar{\boldsymbol{w}}_t$, step size η_{t+1}, regularization coefficient λ_t

1: Estimate gradient $\boldsymbol{g}_t \triangleq d(h_t(w_t) + \frac{\lambda_t}{2}\|w_t\|_2^2)\boldsymbol{H}_t^{1/2}\boldsymbol{s}_t$
2: Update as $\bar{\boldsymbol{w}}_{t+1} = \arg\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}} \boldsymbol{F}_{t+1}(\boldsymbol{w})$
3: Compute $\boldsymbol{H}_{t+1} \triangleq \nabla^2\Psi(\bar{\boldsymbol{w}}_{t+1}) + \eta_{t+1}(\sigma_{1:t} + \lambda_{0:t})\boldsymbol{I}$
4: Draw $\boldsymbol{s}_{t+1}$ randomly from $\mathbb{S}^{d+1} \cap (\boldsymbol{H}_{t+1}^{-1/2}\boldsymbol{e}_{d+1})^{\perp}$
5: Perturb $\boldsymbol{w}_{t+1} = (w_{t+1}, 1) = \bar{\boldsymbol{w}}_{t+1} + \boldsymbol{H}_{t+1}^{-1/2}\boldsymbol{s}_{t+1}$

for Lipschitz functions, and for smooth functions, we study

$$\mathbb{E}[\textsc{Reg}_T] + \sum_{t=2}^{T}\|w_t - \bar{w}_t\|_2^2 + \sum_{t=2}^{T}\|\bar{w}_t - \bar{w}_{t-1}\|_2 + \sum_{t=2}^{T}\|\bar{w}_t - \bar{w}_{t-1}\|_2^2, \quad (3.4)$$

according to the reduction in (3.1) and (3.2). For clarity, we use italic symbols (e.g., w) in the original domain and bold italic symbols (e.g., $\boldsymbol{w}$) in the lifted domain.[1] Next we review the method of Luo et al. (2022) in Section 3.2.1 and provide our key and simple analysis in Section 3.2.2.

3.2.1. Review of Luo et al. (2022)

This work studies heterogeneous curvatures in the BCO setup. The algorithm runs in a lifted domain $\boldsymbol{\mathcal{W}} \triangleq \{\boldsymbol{w} = (w, 1) \mid w \in \mathcal{W} \subseteq \mathbb{R}^d\}$. We describe the subroutine of the t-th round in Algorithm 1. Specifically, after submitting decision w_t and receiving bandit feedback $h_t(w_t)$ as well as the curvature σ_t, the learner constructs an unbiased gradient estimator in Line 1, where $(h_t(w_t) + \frac{\lambda_t}{2}\|w_t\|_2^2)$ is an observation of the regularized function, $\boldsymbol{s}_t$ is a uniformly unit vector for exploration, and $\boldsymbol{H}_t$ is a perturbation matrix. In Line 2, an FTRL algorithm updates the decision as $\bar{\boldsymbol{w}}_{t+1} = \arg\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}} \boldsymbol{F}_{t+1}(\boldsymbol{w})$, where

$$\boldsymbol{F}_{t+1}(\boldsymbol{w}) \triangleq \sum_{s=1}^{t}\left(\boldsymbol{g}_s^{\top}\boldsymbol{w} + \frac{\sigma_s}{2}\|\boldsymbol{w} - \bar{\boldsymbol{w}}_s\|_2^2\right) \quad (\textsc{Approx})$$
$$+ \sum_{s=1}^{t}\frac{\lambda_s}{2}\|\boldsymbol{w} - \bar{\boldsymbol{w}}_s\|_2^2 + \frac{\lambda_0}{2}\|\boldsymbol{w}\|_2^2 \quad (\textsc{Reglr-I})$$
$$+ \frac{1}{\eta_{t+1}}\Psi(\boldsymbol{w}). \quad (\textsc{Reglr-II})$$

Here, Approx is the approximation of the original function $h_t(\cdot)$ using the gradient estimator $\boldsymbol{g}_t$, Reglr-I is a regularization term following the same spirit of AOGD (Bartlett et al., 2007), and Reglr-II is the FTRL regularizer, where η_{t+1} is a non-increasing step size and Ψ is a normal barrier in the lifted domain. In Line 3, the learner constructs the perturbation matrix $\boldsymbol{H}_{t+1}$, which intuitively uses the curvature of the domain and the function for effective exploration. In Line 4, $\boldsymbol{w}^{\perp}$ represents space orthogonal to $\boldsymbol{w}$. Finally, in Line 5, the learner perturbs the updated decision.

Finally, we illuminate the parameter configurations of Algorithm 1 (step sizes $\{\eta_t\}_{t=1}^T$ and regularization coefficients $\{\lambda_t\}_{t=1}^T$). Specifically, for Lipschitz functions, we follow Algorithm 2 of Luo et al. (2022) and set them as follows:

$$\eta_t = d^{-4/3}(L_f + 1)^{2/3}\left(\frac{1}{\sigma_{1:t-1} + \lambda_{0:t-1}} + \frac{1}{T}\right)^{1/3}, \quad \lambda_t = \frac{d^{2/3}(L_f + 1)^{2/3}}{(\sigma_{1:t} + \lambda_{0:t})^{1/3}}, \quad (3.5)$$

where L_f denotes the coordinate-wise Lipschitzness of $f_t(\cdot)$. And for smooth functions, we follow Algorithm 1 of Luo et al. (2022) and set η_t and λ_t as follows:

$$\eta_t = \frac{1}{2d}\sqrt{\frac{\beta_f + 1}{\sigma_{1:t-1} + \lambda_{0:t-1}} + \frac{1}{T}}, \lambda_t = \frac{d\sqrt{\beta_f + 1}}{\sqrt{\sigma_{1:t} + \lambda_{0:t}}}, \quad (3.6)$$

where β_f represents the smoothness of $f_t(\cdot)$.

3.2.2. Our Analysis

In this part, we consider BCO-SC with heterogeneous curvatures for our purpose. To this end, we provide a simple *half-page proof* for an essential stability term in both regret and memory analysis, using Newton decrement developed in interior-point methods (Nesterov and Nemirovskii, 1994). Notably, Luo et al. (2022) has conducted some initial analysis on this term in the regret part using the proof argument by contradiction and some local stability analysis of self-concordant barriers. Our analysis can significantly simplify their two-page proof (please refer to Lemma 17 therein).

To better illuminate our key technique, we first give some basic knowledge of Newton decrement.

Definition 4. *For a self-concordant function f defined on $\mathcal{W}$,[2] for any $w \in \mathrm{int}(\mathcal{W})$, the Newton decrement is defined as $\lambda(w, f) \triangleq \|\nabla f(w)\|_{\nabla^{-2} f(w)}$.*

Newton decrement vanishes exactly at the minimizer $w^\star$ of f in the interior of $\mathcal{W}$, denoted by $\mathrm{int}(\mathcal{W})$, and can be considered as an observable measure of the proximity of any w to $w^\star$. Specifically, it exhibits the following property.

Proposition 1. *For a self-concordant barrier f, if the Newton decrement $\lambda(w, f) \le 1/2$, then it follows that $\|w - w^\star\|_{\nabla^2 f(w)} \le 2\lambda(w, f)$, where $w^\star = \arg\min_w f(w)$.*

In the following, we highlight a key stability term of $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$, which is essential in both the regret and

[1] w_t and $\bar{w}_t$ (in online learning context) correspond to M_t and $\bar{M}_t$ (in control context) respectively.

[2] The self-concordant function is a more general notion than self-concordant barriers — self-concordant barrier are self-concordant functions, but not vice versa.

memory analysis: *(i)* in the standard FTRL regret analysis, e.g., in Chapter 7 of Orabona (2019), an important term is $\langle \boldsymbol{g}_t, \bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1} \rangle$. Given the local boundedness of the gradient estimator, i.e., $\|\boldsymbol{g}_t\|_{\boldsymbol{H}_t^{-1}} \leq \mathcal{O}(d)$, it is crucial to analyze $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$; *(ii)* the switching cost terms — $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_2$ and $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_2^2$ in (3.3) and (3.4) are closely related to $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$, with the only difference being the kinds of norms, showing that the stability term is also essential in the memory analysis.

Below, we present our main technical contribution, a novel stability analysis for $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$, with a simple half-page proof using Newton decrement.

Theorem 1 (Key Technical Result)**.** *Using $\{\eta_t\}_{t=1}^T$ as step sizes, as long as $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 1/2$ is satisfied, Algorithm 1 guarantees the stability term as*

$$\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t} \leq 4d\eta_t + 2\eta_1 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\sqrt{\bar{\nu}},$$

where $\bar{\nu}$ denotes the normal barrier coefficient.

Proof. To begin with, we observe that

$$\nabla^2 \boldsymbol{F}_{t+1}(\boldsymbol{w}) \succeq \frac{1}{\eta_{t+1}} \nabla^2 \Psi_t(\boldsymbol{w}) \succeq \frac{1}{\eta_1} \nabla^2 \Psi_t(\boldsymbol{w}) \tag{3.7}$$

for any $\boldsymbol{w} \in \boldsymbol{\mathcal{W}}$, where $\Psi_t(\boldsymbol{w}) \triangleq \Psi(\boldsymbol{w}) + \frac{\eta_t \lambda_0}{2}\|\boldsymbol{w}\|_2^2 + \eta_t \sum_{s=1}^{t-1} \frac{\sigma_s + \lambda_s}{2}\|\boldsymbol{w} - \bar{\boldsymbol{w}}_s\|_2^2$ such that $\boldsymbol{F}_{t+1}$ can be rewritten as $\boldsymbol{F}_{t+1}(\boldsymbol{w}) = \sum_{s=1}^t \boldsymbol{g}_s^\top \boldsymbol{w} + \frac{1}{\eta_{t+1}}\Psi_{t+1}(\boldsymbol{w})$ for any $\boldsymbol{w} \in \boldsymbol{\mathcal{W}}$. The first step is due to the definitions of $\boldsymbol{F}_{t+1}$ and Ψ_t and the second step is because the step sizes are non-increasing (i.e., $\eta_1 \geq \ldots \geq \eta_t$). With (3.7), it holds that

$$\begin{aligned}
&\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t} = \|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\nabla^2 \Psi_t(\bar{\boldsymbol{w}}_t)} \\
&\leq \sqrt{\eta_t}\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\nabla^2 \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t)} \\
&\leq 2\sqrt{\eta_t}\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}),
\end{aligned} \tag{3.8}$$

where the first step is due to $\boldsymbol{H}_t \triangleq \nabla^2 \Psi(\bar{\boldsymbol{w}}_t) + \eta_t(\sigma_{1:t-1} + \lambda_{0:t-1})\boldsymbol{I} = \nabla^2 \Psi_t(\bar{\boldsymbol{w}}_t)$, the second step is by (3.7), and the last step is because of the aforementioned property of Newton decrement, as long as $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 1/2$. This condition will be verified in Lemma 1.

Next, Newton decrement $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1})$ can be bounded by

$$\begin{aligned}
&\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \triangleq \|\nabla \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t)\|_{\nabla^{-2} \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t)} \\
&\leq \underbrace{\|\boldsymbol{g}_t\|_{\nabla^{-2} \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t)}}_{\text{TERM (A)}} + \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\nabla \Psi(\bar{\boldsymbol{w}}_t)\|_{\nabla^{-2} \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t)}}_{\text{TERM (B)}},
\end{aligned}$$

which is due to $\nabla \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t) = \nabla \boldsymbol{F}_t(\bar{\boldsymbol{w}}_t) + \boldsymbol{g}_t + (1/\eta_{t+1} - 1/\eta_t)\nabla \Psi(\bar{\boldsymbol{w}}_t)$ and $\nabla \boldsymbol{F}_t(\bar{\boldsymbol{w}}_t) = \boldsymbol{0}$ since $\bar{\boldsymbol{w}}_t$ minimizes $\boldsymbol{F}_t$. Then TERM (A) can be bounded as

$$\begin{aligned}
\text{TERM (A)} &\triangleq \|\boldsymbol{g}_t\|_{\nabla^{-2} \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t)} \overset{(3.7)}{\leq} \sqrt{\eta_t}\|\boldsymbol{g}_t\|_{\nabla^{-2} \Psi_t(\bar{\boldsymbol{w}}_t)} \\
&= \sqrt{\eta_t}\|\boldsymbol{g}_t\|_{\boldsymbol{H}_t^{-1}} \leq 2d\sqrt{\eta_t},
\end{aligned}$$

where the second step is due to $\nabla^2 \boldsymbol{F}_{t+1}(\boldsymbol{w}) \succeq \frac{1}{\eta_t}\nabla^2 \Psi_t(\boldsymbol{w})$, the third step is because of the definition of $\boldsymbol{H}_t$, and the last step holds due to the property of the gradient estimator.

As for TERM (B), it holds that

$$\|\nabla \Psi(\bar{\boldsymbol{w}}_t)\|_{\nabla^{-2} \boldsymbol{F}_{t+1}(\bar{\boldsymbol{w}}_t)} \leq \sqrt{\eta_1}\|\nabla \Psi(\bar{\boldsymbol{w}}_t)\|_{\nabla^{-2} \Psi(\bar{\boldsymbol{w}}_t)} \leq \sqrt{\eta_1 \bar{\nu}},$$

where the first step is due to $\nabla^2 \boldsymbol{F}_{t+1}(\boldsymbol{w}) \succeq \frac{1}{\eta_1}\nabla^2 \Psi(\boldsymbol{w})$ for any $\boldsymbol{w} \in \boldsymbol{\mathcal{W}}$, and the last step is due to the property of the normal barrier, whose coefficient is denoted by $\bar{\nu}$. Putting everything together finishes the proof. □

The intuition of the proof is relating the local norm $\boldsymbol{H}_t$ to $\nabla^2 \boldsymbol{F}_{t+1}(\cdot)$, where $\boldsymbol{F}_{t+1}$ denotes the function that $\bar{\boldsymbol{w}}_{t+1}$ minimizes. By doing so, we upper-bound it by the Newton decrement $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1})$ using Proposition 1, which exhibits some nice properties making the analysis much easier.

By aggregating Theorem 1 over T rounds, we get $\sum_t \|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t} \leq \mathcal{O}(\eta_{T+1}^{-1} + \sum_t \eta_t)$, a standard result in FTRL regret analysis, e.g., in Chapter 7.3 of Orabona (2019). This can be easily handled, and thus leads to a simplified regret analysis for BCO with heterogeneous curvatures. Due to space constraints, we defer the regret guarantees and proofs to Lemma 9 and Lemma 10 in Appendix C.1.

Notably, Theorem 1 hinges on the condition $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 1/2$. In Lemma 1 below, we show that this condition can be satisfied by a proper initialization of the regularization coefficient λ_0 and requiring the time horizon T to be larger than a certain constant. The proof is deferred to Appendix C.2.

Lemma 1. *Denoting by $\bar{\nu}$ the normal barrier coefficient, $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 1/2$ holds under the following conditions:*

- *(i) $h_t(\cdot)$ is L_f-Lipschitz, the time horizon $T \geq T_0 = 2^{13} d^4 (L_f + 1)^2 (1 + \sqrt{\bar{\nu}})^6$, and $\lambda_0 = T_0$;*
- *(ii) $h_t(\cdot)$ is β_f-smooth, the time horizon $T \geq T_0 = 128 d^2 (1 + \sqrt{\bar{\nu}})^4$, and $\lambda_0 = (\beta_f + 1)T_0$.*

Consequently, we obtain $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t} \leq \sqrt{\eta_1} = \mathcal{O}(1)$.

Upon bounding $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$, we analyze the switching-cost terms (i.e., $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_2$ and $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_2^2$) by relating $\boldsymbol{H}_t$ to the identity matrix $\boldsymbol{I}$. Consequently, we find that the switching costs can be perfectly absorbed into the regret analysis, validating the stability of Algorithm 1. By doing this, we achieve adaptive guarantees to heterogeneous curvatures in BCO-SC. Theorem 2 concludes the results, and the corresponding proof is deferred to Appendix C.3.

Theorem 2. *Suppose the function $h_t(\cdot)$ is σ_t-strongly convex, and denote by $\{\lambda_t^\star\}_{t=1}^T$ the optimal regularization coefficients. When $h_t(\cdot)$ is L_f-Lipschitz continuous, Algorithm 1 can upper-bound the BCO-SC regret (3.3) by*

$$\widetilde{\mathcal{O}}\left(\inf_{\lambda_1^\star, \ldots, \lambda_T^\star} \left\{T^{1/3} + \lambda_{1:T-1}^\star + \sum_{t=1}^{T-1} (\sigma_{1:t} + \lambda_{0:t}^\star)^{-1/3}\right\}\right).$$

When $h_t(\cdot)$ is β_f-smooth, Algorithm 1 can upper-bound the BCO-SC regret (3.4) by

$$\widetilde{\mathcal{O}}\left(\inf_{\lambda_1^\star,\ldots,\lambda_T^\star}\left\{\sqrt{T}+\lambda_{1:T-1}^\star+\sum_{t=1}^{T-1}(\sigma_{1:t}+\lambda_{0:t}^\star)^{-1/2}\right\}\right).$$

Notably, the optimal regularization coefficients $\{\lambda_t^\star\}_{t=1}^T$ only exist in analysis. Therefore, we can plug in any feasible $\{\lambda_t^\dagger\}_{t=1}^T$ to achieve different guarantees in various cases. Specifically, by doing this, our results can recover existing regret guarantees, be robust to corrupted quadratic properties, and consider decaying quadraticity in bandit LQR control. The details of these implications will be shown in Section 3.3. At the end of this part, we provide several remarks on the results and techniques in this work.

Remark 1. We have proved that Algorithm 1, designed for BCO with heterogeneous curvatures, can also handle switching cost terms, as shown in Theorem 2. Our contribution lies mainly in the technical aspect — achieving a simple analysis for the stability term of $\|\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$, which enables us to control the switching costs and provide a simplified analysis for BCO with heterogeneous curvatures. ◁

Remark 2. We clarify that the techniques in this work (with-history reduction and stability analysis) are general beyond the specific problem studied here. Specifically, the with-history reduction (Sun et al., 2023) can be used in other bandit non-stochastic control problems to avoid the undesired truncation error emerging in previous reduction methods. Our stability analysis is also not restricted to this specific context, as it can be used in other problems requiring FTRL with time-varying step sizes and self-concordant barriers (or normal barriers by domain lifting). ◁

Remark 3. In the control-reduced BCO-SC problem, we need to handle a non-oblivious setup due to the with-history losses. We emphasize again that the obliviousness inherited from the control part makes our inner online learning algorithm still feasible, as explained in Section 3.1. ◁

3.3. Back to Bandit LQR Control

In this part, we apply the results to the bandit LQR control problem, and obtain interpolated theoretical guarantees that are adaptive to the true curvatures of cost functions. Our results can thus recover existing results and imply new bounds in certain cases, as summarized in Table 1.

We adapt the method and analysis in Section 3.2 and obtain our method, which is summarized in Algorithm 2. In summary, we adopt DRP policy (Definition 2) and define with-history functions (Definition 3). We utilize the lazy-update method of Cassel and Koren (2020) to realize effective gradient estimation on a smaller time horizon. When the algorithm is scheduled to update, we follow the same method in the online learning context, that is, using Algorithm 1 as a subroutine. The algorithms for Lipschitz and smooth functions vary only in the setups of regularization coefficients $\{\lambda_t\}_{t=1}^T$ and step sizes $\{\eta_t\}_{t=1}^T$. We defer these configurations to (3.5) and (3.6) due to page constraints.

Algorithm 2 Heterogeneous Bandit LQR Control

Input: Initial regularization coefficient λ_0, lifted domain $\boldsymbol{\mathcal{M}}$, normal barrier $\Psi(\cdot)$ on $\boldsymbol{\mathcal{M}}$
Initialize: $\bar{\boldsymbol{M}}_1=\arg\min_{\boldsymbol{M}\in\boldsymbol{\mathcal{M}}}\Psi(\boldsymbol{M})$, $d_c=md_{\mathbf{y}}d_{\mathbf{u}}$, $\mathcal{S}=\emptyset$ (record the time steps when the algorithm updates)
for $t=H',\ldots,T$ **do**
  Receive an observation $\mathbf{y}_t$
  Compute Nature's $\mathbf{y}_t^{\mathrm{nat}}=\mathbf{y}_t-\sum_{i=1}^{t-1}G^{[i]}\mathbf{u}_{t-i}$
  Submit $\mathbf{u}_t(M_t)=\sum_{i=0}^{m-1}M_t^{[i]}\mathbf{y}_{t-i}^{\mathrm{nat}}$ via DRP policy
  Receive $c_t(\mathbf{y}_t,\mathbf{u}_t)$ and its curvature α_t
  Compute curvature σ_t of the with-history loss
  Draw a random bit $b_t\sim\mathrm{Bern}(1/H)$
  if $b_t\prod_{i=1}^{H-1}(1-b_{t-i})=1$ **then**
    Update $\mathcal{S}=\mathcal{S}\cup\{t\}$
    Compute λ_t and η_{t+1} using (3.5) for Lipschitz functions or (3.6) for smooth functions
    ▷ $\sigma_{1:t}$ and $\lambda_{0:t}$ are computed only on $\mathcal{S}$
    Send $(c_t(\mathbf{y}_t,\mathbf{u}_t),\sigma_t,\bar{\boldsymbol{M}}_t,\eta_{t+1},\lambda_t)$ to Algorithm 1 and obtain $\boldsymbol{M}_{t+1}$
  else
    Maintain the current policy $\boldsymbol{M}_{t+1}=\boldsymbol{M}_t$
  end
end

We provide our regret bounds for bandit LQR control in Theorem 3, which is analogous to the online learning context (Theorem 2). The proof can be found in Appendix D.1.

Theorem 3. *Denoting by $\{\lambda_t^\star\}_{t=1}^T$ the optimal regularization coefficients, for Lipschitz and α_t-quadratic costs, under Assumptions 1-2, by setting the regularization coefficients $\{\lambda_t\}_{t=0}^T$ and step sizes $\{\eta_t\}_{t=1}^T$ as (3.5), Algorithm 2 can bound the expected policy regret (i.e., $\mathbb{E}[\mathrm{REG}_T^{\mathrm{C}}]$) as*

$$\widetilde{\mathcal{O}}\left(\inf_{\lambda_1^\star,\ldots,\lambda_T^\star}\left\{T^{1/3}+\lambda_{1:T-1}^\star+\sum_{t=1}^{T-1}(\alpha_{1:t}+\lambda_{0:t}^\star)^{-1/3}\right\}\right).$$

If the cost functions are additionally smooth, by choosing the regularization coefficients $\{\lambda_t\}_{t=0}^T$ and step sizes $\{\eta_t\}_{t=1}^T$ as (3.6), Algorithm 2 bounds the expected policy regret as

$$\widetilde{\mathcal{O}}\left(\inf_{\lambda_1^\star,\ldots,\lambda_T^\star}\left\{\sqrt{T}+\lambda_{1:T-1}^\star+\sum_{t=1}^{T-1}(\alpha_{1:t}+\lambda_{0:t}^\star)^{-1/2}\right\}\right).$$

As noted in Section 3.2.2, the optimal regularization coefficients $\{\lambda_t^\star\}_{t=1}^T$ only exist in analysis. Therefore, we can plug in any feasible $\{\lambda_t^\dagger\}_{t=1}^T$ to achieve different guarantees

in various cases. Specifically, below we provide direct corollaries considering the cases of *pure convexity / quadraticity*, *corrupted quadraticity*, and *decaying quadraticity*. The corresponding proofs are deferred to Appendix D.2.

① **Pure Convexity/Quadraticity.** Corollary 1 recovers the $\widetilde{\mathcal{O}}(T^{3/4})$ regret for convex costs (Gradu et al., 2020; Cassel and Koren, 2020). For quadratic costs, we obtain a *new* $\widetilde{\mathcal{O}}(\alpha^{-1/3}T^{2/3})$ bound. For smooth costs, Corollary 2 recovers the $\widetilde{\mathcal{O}}(T^{2/3})$ regret for convex costs (Cassel and Koren, 2020). Our $\widetilde{\mathcal{O}}(\alpha^{-1/2}\sqrt{T})$ for quadratic functions matches $\widetilde{\mathcal{O}}(\alpha^{-1}\sqrt{T})$ of Sun et al. (2023) in the dependence of T, with a tighter dependence on the curvature α.

Corollary 1. *For Lipschitz costs, under Assumptions 1, 2, when $\{c_t(\cdot,\cdot)\}_{t=1}^T$ are convex, Algorithm 2 with configuration (3.5) achieves an $\widetilde{\mathcal{O}}(T^{3/4})$ regret. When $\{c_t(\cdot,\cdot)\}_{t=1}^T$ are α-quadratic, Algorithm 2 achieves $\widetilde{\mathcal{O}}(\alpha^{-1/3}T^{2/3})$.*

Corollary 2. *For Lipschitz and smooth costs, under Assumptions 1, 2, when $\{c_t(\cdot,\cdot)\}_{t=1}^T$ are convex, Algorithm 2 with (3.6) achieves an $\widetilde{\mathcal{O}}(T^{2/3})$ regret. When $\{c_t(\cdot,\cdot)\}_{t=1}^T$ are α-quadratic, Algorithm 2 achieves $\widetilde{\mathcal{O}}(\alpha^{-1/2}\sqrt{T})$.*

We note that an $\widetilde{\mathcal{O}}(\sqrt{T})$ bound is also obtained in Theorem 8 of Cassel and Koren (2020) for strongly convex and smooth functions. However, a caveat is that they allow an *improper learning*, i.e., allowing policies in a larger domain outside $\mathcal{M}$ with the compared policy still in $\mathcal{M}$. This eliminates the use of self-concordant barriers, hence avoiding the associated hard-to-control truncation errors.

Our $\widetilde{\mathcal{O}}(\sqrt{T})$ bound is near-optimal compared to the $\Omega(\sqrt{T})$ lower bound (Shamir, 2013), as also stated by Sun et al. (2023). Although the other results are sub-optimal in this sense, they have matched the state-of-the-art bounds attainable by *efficient* algorithms (Flaxman et al., 2005; Saha and Tewari, 2011). While some recent breakthroughs (Bubeck et al., 2017; Lattimore, 2020) achieve the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$, they are computationally expensive and not efficiently implementable in practice. Achieving the optimal regret with efficient methods in BCO is extremely challenging and is still open in the community.

② **Corrupted Quadraticity.** A naïve solution is to discard the non-quadratic (i.e., convex) ones, where the optimal $\widetilde{\mathcal{O}}(T^{2/3})$ can be maintained when $M = \mathcal{O}(T^{2/3})$ for Lipschitz costs. Corollary 3 shows that in the worst case (the first bound), when $M = \mathcal{O}(T^{8/9})$ functions are convex, the algorithm can still maintain the same regret bound. In the best case (the second bound), the desired bound is attainable even when $M = \Theta(T)$. Similarly, for smooth costs, a naïve solution can maintain the $\widetilde{\mathcal{O}}(\sqrt{T})$ bound when $M = \mathcal{O}(\sqrt{T})$. Corollary 4 shows that the same bound is attainable when $M = \mathcal{O}(T^{3/4})$ in the worst case (the first bound), and even when $M = \Theta(T)$ in the best case (the second bound).

Corollary 3. *For Lipschitz costs, under Assumptions 1, 2, when M functions in $\{c_t(\cdot,\cdot)\}_{t=1}^T$ are convex and the rest are α-quadratic, Algorithm 2 with configuration (3.5) guarantees $\widetilde{\mathcal{O}}(M^{3/4} + \alpha^{-1/3}(T-M)^{2/3})$. If the quadratic functions appear in the first $(T-M)$ rounds, the bound can be further improved to $\widetilde{\mathcal{O}}(\alpha^{-1/3}T(T-M)^{-1/3})$.*

Corollary 4. *For Lipschitz and smooth costs, under Assumptions 1, 2, when M functions in $\{c_t(\cdot,\cdot)\}_{t=1}^T$ are convex and the rest are α-quadratic, Algorithm 2 with configuration (3.6) guarantees $\widetilde{\mathcal{O}}(\sqrt{T} + M^{2/3} + \alpha^{-1/2}(T-M)^{1/2})$. If the quadratic functions appear in the first $(T-M)$ rounds, it can be further improved to $\widetilde{\mathcal{O}}(\sqrt{T} + \alpha^{-1/2}T(T-M)^{-1/2})$.*

③ **Decaying Quadraticity.** Corollaries 5-6 study cost functions with decaying quadraticity. The results show that when $\alpha_t = t^{-\gamma}$ with $\gamma > 0$, the bounds will degenerate. And when γ exceeds a specific threshold, the results become the same as optimizing on purely convex costs, even if the costs still exhibit quadratic properties.

Corollary 5. *For Lipschitz costs, under Assumptions 1, 2, when $c_t(\cdot,\cdot)$ is α_t-quadratic with $\alpha_t = t^{-\gamma}$ for some $\gamma \in [0,1]$, Algorithm 2 with configuration (3.5) achieves an $\widetilde{\mathcal{O}}(T^{2/3+\gamma/3})$ regret if $\gamma \in [0, 1/4]$, and $\widetilde{\mathcal{O}}(T^{3/4})$ otherwise.*

Corollary 6. *For Lipschitz and smooth costs, under Assumptions 1, 2, when $c_t(\cdot,\cdot)$ is α_t-quadratic with $\alpha_t = t^{-\gamma}$ for some $\gamma \in [0,1]$, Algorithm 2 with (3.6) achieves an $\widetilde{\mathcal{O}}(T^{1/2+\gamma/2})$ regret if $\gamma \in [0, 1/3]$, and $\widetilde{\mathcal{O}}(T^{2/3})$ otherwise.*

Remark 4. As opposed to the online setting, the Lipschitzness assumption is still required for smooth costs in bandit control. This is mainly caused by the reduction from bandit control to BCO-M (and BCO-SC). The results for BCO-SC do not require Lipschitzness for smooth loss functions. ◁

4. Conclusion and Future Directions

In this paper, we study bandit LQR control with heterogeneous curvatures. We obtain interpolated guarantees that are adaptive to the true curvatures of costs. This is done via a lossless with-history reduction scheme and a simple analysis of a local-norm stability term essential in both regret and switching cost analysis, using Newton decrement.

As discusses at the end of Section 2, studies in universal online learning with bandit feedback remains largely unexplored. This work, along with that of Luo et al. (2022), has made a first step in addressing this problem by considering known but heterogeneous curvatures. An important future direction would be to investigate homogeneous but unknown curvatures, a more challenging setup in universal online learning, with bandit feedback. This problem is difficult even within bandit convex optimization, considering only convex or strongly convex online functions.

Acknowledgements

This research was supported by NSFC (62361146852, 62206125, 61921006) and JiangsuSF (BK20220776). The authors thank Jennifer Sun for the discussions on non-stochastic control, especially about the with-history reduction to BCO with memory. Peng Zhao thanks Tomer Koren for helpful discussions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference Computational Learning Theory (COLT)*, pages 263–274, 2008.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M. Kakade, and Karan Singh. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 111–119, 2019.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: Price of past mistakes. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 784–792, 2015.

Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 65–72, 2007.

Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6): 503–515, 1954.

Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing (STOC)*, pages 72–85, 2017.

Asaf Cassel and Tomer Koren. Bandit linear control. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 8872–8882, 2020.

Asaf Cassel, Alon Cohen, and Tomer Koren. Rate-optimal online convex optimization in adaptive linear control. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 7410–7422, 2022.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1028–1037, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 4192–4201, 2018.

Ofer Dekel, Ambuj Tewari, and Raman Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1747–1754, 2012.

Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, pages 726–736, 2009.

Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 385–394, 2005.

Paula Gradu, John Hallman, and Elad Hazan. Non-stochastic control with bandit feedback. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 10764–10774, 2020.

Elad Hazan. Introduction to Online Convex Optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan and Kfir Y. Levy. Bandit convex optimization: Towards tight bounds. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 784–792, 2014.

Elad Hazan and Karan Singh. Introduction to online non-stochastic control. *ArXiv preprint*, arXiv:2211.09619, 2022.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Tor Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3):311–334, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: High-probability data-dependent regret bounds for adversarial bandits and MDPs. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 15522–15533, 2020.

Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. *Optimal Control*. John Wiley & Sons, 2012.

Haipeng Luo, Mengxiao Zhang, and Peng Zhao. Adaptive bandit convex optimization with heterogeneous curvature. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, pages 1576–1612, 2022.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 10154–10164, 2019.

Deepan Muthirayan, Jianjun Yuan, and Pramod P. Khargonekar. Adaptive gradient online control. In *Proceedings of the 2022 American Control Conference (ACC)*, pages 2136–2141, 2022.

Yurii E. Nesterov and Arkadii Nemirovskii. *Interior-point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

Francesco Orabona. A modern introduction to online learning. *ArXiv preprint*, arXiv:1912.13213, 2019.

Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 636–642, 2011.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the 26th Annual Conference Computational Learning Theory (COLT)*, pages 3–24, 2013.

Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 18318–18329, 2020.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Proceedings of 33rd Conference on Learning Theory (COLT)*, pages 3320–3436, 2020.

Y. Jennifer Sun, Stephen Newman, and Elad Hazan. Optimal rates for bandit nonstochastic control. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 21908–21919, 2023.

Tim van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3666–3674, 2016.

Yu-Hu Yan, Peng Zhao, and Zhi-Hua Zhou. Universal online learning with gradual variations: A multi-layer online ensemble approach. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 37682–37715, 2023.

Lijun Zhang, Guanghui Wang, Jinfeng Yi, and Tianbao Yang. A simple yet universal strategy for online convex optimization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26605–26623, 2022.

Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26865–26894, 2022.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, 2003.

A. Self-concordant Barrier and Normal Barrier

In this section, we provide some background knowledge about self-concordant barriers and normal barriers.

Definition 5. *A function $\psi : \mathrm{int}(\mathcal{W}) \to \mathbb{R}$ is a self-concordant function if: (i) ψ is three times continuously differentiable and convex, and becomes infinity along any sequence of points approaching the boundary of the domain; and (ii) For every $x \in \mathbb{R}^d$ and $w \in \mathrm{int}(\mathcal{W})$, it holds that $|\nabla^3\psi(w)[x,x,x]| \le 2|\nabla^2\psi(w)[x,x]|^{3/2}$.*

Definition 6. *If $|\nabla\psi(w)[x]| \le \nu^{1/2}(|\nabla^2\psi(w)[x,x]|)^{1/2}$, a self-concordant function ψ is a ν-self-concordant barrier.*

Lemma 2 (Theorem 2.5.1 of Nesterov and Nemirovskii (1994))**.** *For each closed convex body $\mathcal{W} \subseteq \mathbb{R}^d$, there always exists an $\mathcal{O}(d)$-self-concordant barrier on $\mathcal{W}$.*

Lemma 3 (Proposition 2.3.4 of Nesterov and Nemirovskii (1994))**.** *If Ψ is a ν-normal barrier on $\mathcal{W} \subseteq \mathbb{R}^d$, then for any $w_1, w_2 \in \mathrm{int}(\mathcal{W})$, it holds that: (i) $\|w\|^2_{\nabla^2\Psi(w)} = w^\top\nabla^2\Psi(w)w = \nu$; (ii) $\nabla^2\Psi(w)w = -\nabla\Psi(w)$; (iii) $\Psi(w_1) \ge \Psi(w_2) - \nu\ln\frac{-\langle\nabla\Psi(w_2), w_1\rangle}{\nu}$; and (iv) $\|\nabla\Psi(w)\|^2_{\nabla^{-2}\Psi(w)} = \nu$.*

Lemma 4 (Proposition 5.1.4 of Nesterov and Nemirovskii (1994))**.** *If ψ is a ν-self-concordant barrier on $\mathcal{W} \subseteq \mathbb{R}^d$, then the function $\psi(w,b) \triangleq 400(\psi(w/b) - 2\nu\ln b)$ is a $\bar{\nu}$-normal barrier on $\mathrm{con}(\mathcal{W})$ with $\bar{\nu} = 800\nu$, where $\mathrm{con}(\mathcal{W}) = \{\mathbf{0}\} \cup \{(w,b) \mid w/b \in \mathcal{W}, w \in \mathbb{R}^d, b > 0\}$ is the conic hull of $\mathcal{W}$ lifted to $\mathbb{R}^{d+1}$.*

Lemma 5 (Proposition 2.3.2 of Nesterov and Nemirovskii (1994))**.** *Let $\psi : \mathrm{int}(\mathcal{W}) \mapsto \mathbb{R}$ be a ν-self-concordant barrier over the closed convex set $\mathcal{W}$, then for any $w, w' \in \mathrm{int}(\mathcal{W})$, we have $\psi(w') - \psi(w) \le \nu\log\frac{1}{1-\pi_w(w')}$, where $\pi_w(w') \triangleq \arg\inf_{t\ge 0}\{w + t^{-1}(w'-w) \in \mathcal{W}\}$ is the Minkowski function of $\mathcal{W}$ whose pole is on w, which is always in $[0,1]$.*

B. Omitted Details of Section 3.1 (Reduction to Online Learning)

In this section, we give some omitted details of Section 3.1, including the truncation lemma (Lemma 6), the relationship of regret bounds between the whole time horizon and the lazy-update horizon (Lemma 7), and the preservation of function properties between the control cost and the truncated function (Lemma 8).

Lemma 6 (Appendix D.3 of Sun et al. (2023))**.** *For Lipschitz costs, choosing memory length $H = \Theta(\log T)$ and $m = \Theta(\log T)$ in the definition of DRP (Definition 2), for any fixed DRP policy $M \in \mathcal{M}$, the cumulative gap between the with-history loss function $\widetilde{f}_t(M)$ (Definition 3) and the costs $c_t(\mathbf{y}_t(M), \mathbf{u}_t(M))$ can be bounded by*

$$\sum_{t=1}^{T}\widetilde{f}_t(M) - \sum_{t=1}^{T}c_t(\mathbf{y}_t(M), \mathbf{u}_t(M)) \le \mathcal{O}(1).$$

Interested readers can refer to Appendix D.3 of Sun et al. (2023) for the proof.

Lemma 7 (Lemma 11 of Cassel and Koren (2020))**.** *Suppose the lazy-update method chooses random bits $b_{1:T}$ from Bern($1/H$) independently, then for any decision sequence $M_{H':T}$ and fixed comparator M, it holds that*

$$\mathbb{E}\left[\sum_{t=H'}^{T}\widetilde{f}_t(M_t) - \sum_{t=H'}^{T}\widetilde{f}_t(M)\right] \le 3H\cdot\mathbb{E}\left[\sum_{t\in\mathcal{S}}\widetilde{f}_t(M_t) - \sum_{t\in\mathcal{S}}\widetilde{f}_t(M)\right],$$

where H' denote the effective memory length.

The Lipschitzness, strong convexity, and smoothness of the cost function can be preserved to the with-history loss functions, as summarized in the following lemma.

Lemma 8 (Function Properties Connection (Sun et al., 2023, Lemma D.6))**.** *The functional relationship between the cost function c_t and the with-history loss function f_t (in Definition 3) or its unary version $\widetilde{f}_t$ is as follows:*

(i) For L_c-Lipschitz costs, f_t is L_f-Lipschitz with $L_f \triangleq 2L_c\sqrt{(1+RR_G)^2 + R^2}R_G R^2_{\mathrm{nat}}\sqrt{H}$, where R appears in the definition of DRP policy class $\mathcal{M}$ (Definition 2).

(ii) For α_t-quadratic costs, $\mathbb{E}_{t-H-1}[\widetilde{f}_t]$ is σ_t-strongly convex with $\sigma_t \triangleq \alpha_t\left(\mathrm{Var}_{\mathbf{e}} + \mathrm{Var}_{\boldsymbol{\xi}}\frac{\sigma_{\min}(C)}{1+\|A\|_2^2}\right)$, where the expectation $\mathbb{E}_{t-H-1}[\cdot]$ is taken on the randomness up to the $(t-H-1)$-th round, due to Definition 3.

(iii) For β_c-smooth costs, $\widetilde{f}_t$ is β_f-smooth with $\beta_f \triangleq 4\beta_c R^2_{\mathrm{nat}}R^2_G H$.

In above results, R_{nat} denotes the upper bound for $\|\mathbf{y}_t^{\mathrm{nat}}\|_2$ and R_G is the upper bound for $\|G\|_{\ell_1,\mathrm{op}} = \sum_{i=1}^{\infty}\|G^{[i]}\|_{\mathrm{op}}$.

C. Omitted Proofs of Section 3.2 (BCO with Switching Cost)

In this section, we provide the proofs for Section 3.2, including simplified regret analyses for BCO with heterogeneous curvatures, Lemma 1 and Theorem 2.

C.1. Simplified Regret Analysis for BCO with Heterogeneous Curvatures

In this part, we provide the regret analysis for BCO with heterogeneous curvatures (Luo et al., 2022) in Lemma 9 and Lemma 10, with *simplified* proofs using our Theorem 1 for the crucial term of $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_{\boldsymbol{H}_t}$.

Lemma 9 (Theorem 21 of Luo et al. (2022))**.** *If the loss functions $\{h_t\}_{t=1}^T$ is L_f-Lipschitz continuous and σ_t-strongly convex, Algorithm 2 of Luo et al. (2022) ensures with any regularization coefficients $\{\lambda_t\}_{t=1}^T \in (0,1)$:*

$$\mathbb{E}\left[\sum_{t=1}^T h_t(w_t) - \min_{w\in\mathcal{W}}\sum_{t=1}^T h_t(w)\right] \le \widetilde{\mathcal{O}}\left(T^{1/3} + \lambda_{1:T} + \sum_{t=1}^T(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3}\right).$$

Lemma 10 (Lemma 19 of Luo et al. (2022))**.** *If the loss functions $\{h_t\}_{t=1}^T$ is β_f-smooth and σ_t-strongly convex, Algorithm 1 of Luo et al. (2022) ensures with any regularization coefficients $\{\lambda_t\}_{t=1}^T \in (0,1)$:*

$$\mathbb{E}\left[\sum_{t=1}^T h_t(w_t) - \min_{w\in\mathcal{W}}\sum_{t=1}^T h_t(w)\right] \le \widetilde{\mathcal{O}}\left(\sqrt{T} + \lambda_{1:T} + \sum_{t=1}^T(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2}\right).$$

Proof of Lemma 9. The proof mainly follows the Theorem 12 and Lemma 20 of Luo et al. (2022), and with our novel analysis (i.e., Theorem 1) for the crucial stability term $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_{\boldsymbol{H}_t}$. To begin with, the regret can be decomposed into the following parts:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T h_t(w_t) - \sum_{t=1}^T h_t(w^\star)\right] = \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}_t) - \sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}^\star)\right] \\
&= \underbrace{\sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}_t) - \sum_{t=1}^T \boldsymbol{h}_t(\bar{\boldsymbol{w}}_t)}_{\text{① Exploration}} + \underbrace{\sum_{t=1}^T \boldsymbol{h}_t(\bar{\boldsymbol{w}}_t) - \sum_{t=1}^T \widetilde{\boldsymbol{h}}_t(\bar{\boldsymbol{w}}_t)}_{\text{② Regularization-I}} + \underbrace{\sum_{t=1}^T \widetilde{\boldsymbol{h}}_t(\bar{\boldsymbol{w}}_t) - \sum_{t=1}^T \widehat{\boldsymbol{h}}_t(\bar{\boldsymbol{w}}_t)}_{\text{③ Smooth-I}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \widehat{\boldsymbol{h}}_t(\bar{\boldsymbol{w}}_t) - \sum_{t=1}^T \widehat{\boldsymbol{h}}_t(\widetilde{\boldsymbol{w}})\right]}_{\text{⑤ Reg-Term}} \\
&+ \underbrace{\sum_{t=1}^T \widehat{\boldsymbol{h}}_t(\widetilde{\boldsymbol{w}}) - \sum_{t=1}^T \widetilde{\boldsymbol{h}}_t(\widetilde{\boldsymbol{w}})}_{\text{③ Smooth-II}} + \underbrace{\sum_{t=1}^T \widetilde{\boldsymbol{h}}_t(\widetilde{\boldsymbol{w}}) - \sum_{t=1}^T \boldsymbol{h}_t(\widetilde{\boldsymbol{w}})}_{\text{② Regularization-II}} + \underbrace{\sum_{t=1}^T \boldsymbol{h}_t(\widetilde{\boldsymbol{w}}) - \sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}^\star)}_{\text{④ Comparator-Bias}},
\end{aligned} \tag{C.1}$$

where $w^\star \in \arg\min_{w\in\mathcal{W}}\sum_{t=1}^T h_t(w)$, $\boldsymbol{w}^\star = (w^\star, 1)$. $\boldsymbol{h}_t$ is the lifted version of $h_t(\cdot)$. $\widetilde{\boldsymbol{h}}_t(\boldsymbol{w}) \triangleq \boldsymbol{h}_t(\boldsymbol{w}) + \frac{\lambda_t}{2}\|\boldsymbol{w}_{(1:d)}\|_2^2$ is a regularized version of $\boldsymbol{h}_t(\cdot)$. $\widehat{\boldsymbol{h}}_t(\boldsymbol{w}) \triangleq \mathbb{E}_{\boldsymbol{b}\in\mathbb{B}^{d+1}}[\widetilde{\boldsymbol{h}}_t(\boldsymbol{w} + \boldsymbol{H}_t^{-1/2}\boldsymbol{b})]$ is the smoothed version of $\widetilde{\boldsymbol{h}}_t(\cdot)$. $\widetilde{\boldsymbol{w}} \triangleq (1-1/T)\cdot \boldsymbol{w}^\star + 1/T \cdot \bar{\boldsymbol{w}}_1$, where $\bar{\boldsymbol{w}}_1$ is the minimizer of $\Psi(\cdot)$, i.e., $\bar{\boldsymbol{w}}_1 = \arg\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}}\Psi(\boldsymbol{w})$.

① Exploration term, using the L_f-Lipschitzness of $\boldsymbol{h}_t(\cdot)$, can be bounded by

$$\sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}_t) - \sum_{t=1}^T \boldsymbol{h}_t(\bar{\boldsymbol{w}}_t) \le L_f\sum_{t=1}^T\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2 = L_f\sum_{t=1}^T\|\boldsymbol{H}_t^{-1/2}\boldsymbol{s}_t\|_2 \le \left(\sum_{t=1}^T(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3}\right).$$

② Regularization-I and Regularization-II can be bounded using the definition of $\widetilde{\boldsymbol{h}}_t(\cdot)$. For any $\boldsymbol{w}\in\boldsymbol{\mathcal{W}}$,

$$\sum_{t=1}^T \widetilde{\boldsymbol{h}}_t(\boldsymbol{w}) - \sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}) \le \mathcal{O}\left(\sum_{t=1}^T\lambda_t\right) = \mathcal{O}(\lambda_{1:T}). \tag{C.2}$$

③ SMOOTH-I and SMOOTH-II, using Lipschitzness of $\widetilde{\boldsymbol{h}}_t(\cdot)$ and the definition of $\widehat{\boldsymbol{h}}_t(\cdot)$, can be bounded as

$$\sum_{t=1}^{T}\widehat{\boldsymbol{h}}_t(\boldsymbol{w})-\sum_{t=1}^{T}\widetilde{\boldsymbol{h}}_t(\boldsymbol{w})=\mathbb{E}_{\boldsymbol{b}\in\mathbb{B}^{d+1}}\left[\sum_{t=1}^{T}\widetilde{\boldsymbol{h}}_t(\boldsymbol{w}+\boldsymbol{H}_t^{-1/2}\boldsymbol{b})-\sum_{t=1}^{T}\widetilde{\boldsymbol{h}}_t(\boldsymbol{w})\right]\le\mathcal{O}\left(\sum_{t=1}^{T}(L_f+\lambda_t)\|\boldsymbol{H}_t^{-1/2}\boldsymbol{b}\|_2\right)$$
$$\le\mathcal{O}\left(\sum_{t=1}^{T}\frac{1}{\sqrt{\eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})}}\right)\le\mathcal{O}\left(\sum_{t=1}^{T}(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3}\right),$$

for any $\boldsymbol{w}\in\boldsymbol{\mathcal{W}}$, where the last step is due to the step size setup (3.5).

④ COMPARATOR-BIAS term, using the definition of $\widetilde{\boldsymbol{w}}$, i.e., $\widetilde{\boldsymbol{w}}\triangleq(1-1/T)\cdot\boldsymbol{w}^\star+1/T\cdot\bar{\boldsymbol{w}}_1$, can be bounded as

$$\sum_{t=1}^{T}\boldsymbol{h}_t(\widetilde{\boldsymbol{w}})-\sum_{t=1}^{T}\boldsymbol{h}_t(\boldsymbol{w}^\star)=\sum_{t=1}^{T}\boldsymbol{h}_t\left(\left(1-\frac{1}{T}\right)\cdot\boldsymbol{w}^\star+\frac{1}{T}\cdot\bar{\boldsymbol{w}}_1\right)-\sum_{t=1}^{T}\boldsymbol{h}_t(\boldsymbol{w}^\star)$$
$$\le\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{h}_t(\bar{\boldsymbol{w}}_1)-\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{h}_t(\boldsymbol{w}^\star)\le\mathcal{O}(1),\tag{C.3}$$

which can be ignored. In the end, we analyze the most important ⑤ REG-TERM. Since $\widehat{\boldsymbol{h}}_t(\cdot)$ is $(\sigma_t+\lambda_t)$-strongly convex,

$$\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\boldsymbol{h}}_t(\bar{\boldsymbol{w}}_t)-\sum_{t=1}^{T}\widehat{\boldsymbol{h}}_t(\widetilde{\boldsymbol{w}})\right]\le\mathbb{E}\left[\sum_{t=1}^{T}\left(\nabla\widehat{\boldsymbol{h}}_t(\bar{\boldsymbol{w}}_t)^\top(\bar{\boldsymbol{w}}_t-\widetilde{\boldsymbol{w}})-\frac{\sigma_t+\lambda_t}{2}\|\bar{\boldsymbol{w}}_t-\widetilde{\boldsymbol{w}}\|_2^2\right)\right]$$
$$=\mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{g}_t^\top(\bar{\boldsymbol{w}}_t-\widetilde{\boldsymbol{w}})-\frac{\sigma_t+\lambda_t}{2}\|\bar{\boldsymbol{w}}_t-\widetilde{\boldsymbol{w}}\|_2^2\right)\right]=\mathbb{E}\left[\sum_{t=1}^{T}\boldsymbol{\ell}_t(\bar{\boldsymbol{w}}_t)-\sum_{t=1}^{T}\boldsymbol{\ell}_t(\widetilde{\boldsymbol{w}})\right],$$

where the second step is due to the unbiased gradient estimator, and the last step defines $\boldsymbol{\ell}_t(\boldsymbol{w})\triangleq\boldsymbol{g}_t^\top\boldsymbol{w}+\frac{\sigma_t+\lambda_t}{2}\|\boldsymbol{w}-\bar{\boldsymbol{w}}_t\|_2^2$. Consequently, the FTRL update rule $\bar{\boldsymbol{w}}_{t+1}=\arg\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}}\boldsymbol{F}_{t+1}(\boldsymbol{w})$ can be rewritten as

$$\bar{\boldsymbol{w}}_{t+1}=\arg\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}}\boldsymbol{F}_{t+1}(\boldsymbol{w})=\arg\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}}\left\{\sum_{s=1}^{t}\boldsymbol{\ell}_s(\boldsymbol{w})+\boldsymbol{R}_{t+1}(\boldsymbol{w})\right\}=\arg\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}}\left\{\sum_{s=1}^{t}\boldsymbol{\ell}_s(\boldsymbol{w})+\boldsymbol{R}'_{t+1}(\boldsymbol{w})\right\},$$

where $\boldsymbol{R}_{t+1}(\boldsymbol{w})\triangleq\frac{\lambda_0}{2}\|\boldsymbol{w}\|_2^2+\frac{1}{\eta_{t+1}}\Psi(\boldsymbol{w})$, and $\boldsymbol{R}'_{t+1}(\boldsymbol{w})\triangleq\boldsymbol{R}_{t+1}(\boldsymbol{w})-\frac{1}{\eta_{t+1}}\Psi(\bar{\boldsymbol{w}}_1)=\frac{\lambda_0}{2}\|\boldsymbol{w}\|_2^2+\frac{1}{\eta_{t+1}}(\Psi(\boldsymbol{w})-\Psi(\bar{\boldsymbol{w}}_1))$. Using the regret guarantee of FTRL (Lemma 12), it holds that

$$\sum_{t=1}^{T}\boldsymbol{\ell}_t(\bar{\boldsymbol{w}}_t)-\sum_{t=1}^{T}\boldsymbol{\ell}_t(\widetilde{\boldsymbol{w}})\le\boldsymbol{R}'_{T+1}(\widetilde{\boldsymbol{w}})-\boldsymbol{R}'_1(\bar{\boldsymbol{w}}_1)+\sum_{t=1}^{T}\nabla\boldsymbol{\ell}_t(\bar{\boldsymbol{w}}_t)^\top(\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1})+\sum_{t=1}^{T}(\boldsymbol{R}'_t(\bar{\boldsymbol{w}}_{t+1})-\boldsymbol{R}'_{t+1}(\bar{\boldsymbol{w}}_{t+1}))$$
$$\le\boldsymbol{R}'_{T+1}(\widetilde{\boldsymbol{w}})-\boldsymbol{R}'_1(\bar{\boldsymbol{w}}_1)+\sum_{t=1}^{T}\boldsymbol{g}_t^\top(\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1})\le\frac{\Psi(\widetilde{\boldsymbol{w}})-\Psi(\bar{\boldsymbol{w}}_1)}{\eta_{T+1}}+\sum_{t=1}^{T}\boldsymbol{g}_t^\top(\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1})+\mathcal{O}(1)$$
$$\le\frac{\bar{\nu}\log T}{\eta_{T+1}}+\sum_{t=1}^{T}\boldsymbol{g}_t^\top(\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1})\le\frac{\bar{\nu}\log T}{\eta_{T+1}}+2d\sum_{t=1}^{T}\|\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t},\tag{C.4}$$

where the second step is due to $\boldsymbol{\ell}_t(\bar{\boldsymbol{w}}_t)=\boldsymbol{g}_t$ and $\boldsymbol{R}'_t(\boldsymbol{w})\le\boldsymbol{R}'_{t+1}(\boldsymbol{w})$ for any $\boldsymbol{w}\in\boldsymbol{\mathcal{W}}$ since the step size sequence $\{\eta_t\}_{t=1}^T$ is non-increasing. The third step is because of $\boldsymbol{R}'_1(\bar{\boldsymbol{w}}_1)=\frac{\lambda_0}{2}\|\bar{\boldsymbol{w}}_1\|_2^2\ge0$ and $\frac{\lambda_0}{2}\|\widetilde{\boldsymbol{w}}\|_2^2\le\mathcal{O}(1)$. The fourth step is due to Lemma 5 with the fact that $\widetilde{\boldsymbol{w}}\triangleq(1-1/T)\cdot\boldsymbol{w}^\star+1/T\cdot\bar{\boldsymbol{w}}_1$, and omits the constant term $\mathcal{O}(1)$. The fifth step is due to the upper bound of the gradient estimator:

$$\|\boldsymbol{g}_t\|_{\boldsymbol{H}_t^{-1}}^2=d^2\left(h_t(w_t)+\frac{\lambda_t}{2}\|w_t\|_2^2\right)^2\left\|\boldsymbol{H}_t^{1/2}\boldsymbol{s}_t\right\|_{\boldsymbol{H}_t^{-1}}^2\le d^2\left(1+\frac{\lambda_t}{2}\right)^2\le4d^2,\tag{C.5}$$

which implies $\|\boldsymbol{g}_t\|_{\boldsymbol{H}_t^{-1}}\le2d$. The second step requires $h_t(\cdot)\le1$, which can be assumed without loss of generality. Consequently, the $\sum_{t\in[T]}\|\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$ term, using Theorem 1, can be bounded as

$$\sum_{t=1}^{T}\|\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}\le4d\sum_{t=1}^{T}\eta_t+2\eta_1\sqrt{\bar{\nu}}\sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\le\mathcal{O}\left(\frac{1}{\eta_{T+1}}+\sum_{t=1}^{T}\eta_t\right).$$

Plugging the above upper bound back into (C.4), we can upper-bound the REG-TERM as

$$\text{REG-TERM} \leq \mathcal{O}\left(\frac{\log T}{\eta_{T+1}} + \sum_{t=1}^{T}\eta_t\right) \leq \widetilde{\mathcal{O}}\left(T^{1/3} + \sum_{t=1}^{T}(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3}\right).$$

Combining all parts finishes the proof. □

Proof of Lemma 10. We decompose the regret using the same way as (C.1) in Lemma 9. To begin with, ① EXPLORATION, using the β_f-smoothness of $\boldsymbol{h}_t(\cdot)$, can be bounded by

$$\sum_{t=1}^{T}\boldsymbol{h}_t(\boldsymbol{w}_t) - \sum_{t=1}^{T}\boldsymbol{h}_t(\bar{\boldsymbol{w}}_t) \leq \frac{\beta_f}{2}\sum_{t=1}^{T}\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2^2 = \frac{\beta_f}{2}\sum_{t=1}^{T}\|\boldsymbol{H}_t^{-1/2}\boldsymbol{s}_t\|_2^2 \leq \left(\sum_{t=1}^{T}(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2}\right).$$

② REGULARIZATION-I and REGULARIZATION-II can bounded by $\mathcal{O}(\lambda_{1:T})$ as (C.2). For ③ SMOOTH-I and SMOOTH-II, using smoothness of $\widetilde{\boldsymbol{h}}_t(\cdot)$ and the definition of $\widehat{\boldsymbol{h}}_t(\cdot)$, for any $\boldsymbol{w}\in\boldsymbol{\mathcal{W}}$, it holds that

$$\begin{aligned}\sum_{t=1}^{T}\widehat{\boldsymbol{h}}_t(\boldsymbol{w}) - \sum_{t=1}^{T}\widetilde{\boldsymbol{h}}_t(\boldsymbol{w}) &= \mathbb{E}_{\boldsymbol{b}\in\mathbb{B}^{d+1}}\left[\sum_{t=1}^{T}\widetilde{\boldsymbol{h}}_t(\boldsymbol{w}+\boldsymbol{H}_t^{-1/2}\boldsymbol{b}) - \sum_{t=1}^{T}\widetilde{\boldsymbol{h}}_t(\boldsymbol{w})\right] \leq \mathcal{O}\left(\sum_{t=1}^{T}(\beta_f+\lambda_t)\|\boldsymbol{H}_t^{-1/2}\boldsymbol{b}\|_2^2\right)\\ &\leq \mathcal{O}\left(\sum_{t=1}^{T}\frac{1}{\eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})}\right) \leq \mathcal{O}\left(\sum_{t=1}^{T}(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2}\right),\end{aligned}$$

where the last step is due to the step size setup (3.6). ④ COMPARATOR-BIAS can be bounded by $\mathcal{O}(1)$ as (C.3). It remains to handle the most important ⑤ REG-TERM. Specifically, following the same proof as Lemma 9, we obtain

$$\text{REG-TERM} \leq \mathcal{O}\left(\frac{\log T}{\eta_{T+1}} + \sum_{t=1}^{T}\eta_t\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{T} + \sum_{t=1}^{T}(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2}\right).$$

Combining all parts finishes the proof. □

C.2. Proof of Lemma 1

Proof. From the proof of Theorem 1, we can first upper-bound the concerned $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1})$ as follows:

$$\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 2d\sqrt{\eta_t} + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\sqrt{\eta_1\bar{\nu}}.$$

In the following, we discuss the Lipschitzness and smoothness cases respectively.

Lipschitzness Case. Due to the step size configurations (3.5), it holds that

$$\begin{aligned}\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} &\leq d^{4/3}(L_f+1)^{-2/3}\left(\left(\frac{1}{\sigma_{1:t}+\lambda_{0:t}}+\frac{1}{T}\right)^{-1/3} - \left(\frac{1}{\sigma_{1:t-1}+\lambda_{0:t-1}}+\frac{1}{T}\right)^{-1/3}\right)\\ &\leq d^{4/3}(L_f+1)^{-2/3}\left((\sigma_{1:t}+\lambda_{0:t})^{1/3} - (\sigma_{1:t-1}+\lambda_{0:t-1})^{1/3}\right) \leq d^{4/3}(L_f+1)^{-2/3}(\sigma_t+\lambda_t)^{1/3}\\ &\leq d^{4/3}(L_f+1)^{-2/3}(4L_f+1)^{1/3} \leq 2^{2/3}d^{4/3}\end{aligned}$$

where the second last step is due to $\sigma_t \leq 4L_f$ by Lemma 11. Thus it suffices to choose η_1 such that $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 2d^{4/3}\sqrt{\eta_1}(1+\sqrt{\bar{\nu}}) \leq 1/2$, i.e., $\eta_1 \leq \frac{1}{16d^{8/3}(1+\sqrt{\bar{\nu}})^2}$. Consequently, requiring $T \geq \lambda_0$, it suffices to require

$$\eta_1 = d^{-4/3}(L_f+1)^{2/3}\left(\frac{1}{\lambda_0}+\frac{1}{T}\right)^{1/3} \leq 2^{1/3}d^{-4/3}(L_f+1)^{2/3}\lambda_0^{-1/3} \leq \frac{1}{16d^{8/3}(1+\sqrt{\bar{\nu}})^2},$$

which can be satisfied by setting $\lambda_0 = 2^{13}d^4(L_f+1)^2(1+\sqrt{\bar{\nu}})^6$ and assuming $T \geq \lambda_0$.

Smoothness Case. Due to the step size configurations (3.6), it holds that

$$\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \leq 2d \left(\frac{1}{\frac{1}{2d}\sqrt{\frac{\beta_f+1}{\sigma_{1:t}+\lambda_{0:t}} + \frac{1}{T}}} - \frac{1}{\frac{1}{2d}\sqrt{\frac{\beta_f+1}{\sigma_{1:t-1}+\lambda_{0:t-1}} + \frac{1}{T}}} \right)$$
$$\leq \frac{2d}{\sqrt{\beta_f+1}} \left(\sqrt{\sigma_{1:t}+\lambda_{0:t}} - \sqrt{\sigma_{1:t-1}+\lambda_{0:t-1}} \right) \leq \frac{2d\sqrt{\sigma_t+\lambda_t}}{\sqrt{\beta_f+1}} \leq 2d,$$

where the last step is due to $\lambda_t \in (0,1)$ and $\sigma_t \leq \beta_f$. Thus it suffices to choose η_1 such that $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 2d\sqrt{\eta_1}(1+\sqrt{\bar{\nu}}) \leq 1/2$, i.e., $\eta_1 \leq \frac{1}{16d^2(1+\sqrt{\bar{\nu}})^2}$. Consequently, requiring $1/T \leq (\beta_f+1)/\lambda_0$, it suffices to require

$$\eta_1 = \frac{1}{2d}\sqrt{\frac{\beta_f+1}{\lambda_0} + \frac{1}{T}} \leq \frac{1}{2d}\sqrt{\frac{2(\beta_f+1)}{\lambda_0}} \leq \frac{1}{16d^2(1+\sqrt{\bar{\nu}})^2},$$

which can be satisfied by setting $\lambda_0 = 128d^2(\beta_f+1)(1+\sqrt{\bar{\nu}})^4$ and assuming $T \geq 128d^2(1+\sqrt{\bar{\nu}})^4$. Notably, given $\lambda(\bar{\boldsymbol{w}}_t, \boldsymbol{F}_{t+1}) \leq 1/2$, we directly obtain $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t} \leq \sqrt{\eta_1} = \mathcal{O}(1)$ from (3.8), which is useful for the switching cost analysis. The proof is finished. □

C.3. Proof of Theorem 2

Proof. In this theorem, we analyze the regret of BCO with switching cost. To begin with, we define a sequence of functions in the lifted domain $\boldsymbol{h}_t : \boldsymbol{\mathcal{W}} \mapsto \mathbb{R}$ such that $\boldsymbol{h}_t(\boldsymbol{w}) \triangleq h_t(\boldsymbol{w}_{(1:d)})$ for any $\boldsymbol{w} \in \boldsymbol{\mathcal{W}}$, where $\boldsymbol{w}_{(1:d)}$ denotes the first d entries of $\boldsymbol{w}$. It is easy to verify that $\boldsymbol{h}_t$ can still inherit the σ_t-strong convexity, L_f-Lipschitzness, and β_f-smoothness from the original function h_t. We investigate the following regrets with switching costs in the Lipschitzness and smoothness cases respectively, where some problem-dependent constants are omitted. Specifically, for Lipschitz functions,

$$\text{REG}_T^{\text{SC-LIP}} \triangleq \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}_t) - \min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}} \sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w})\right] + \sum_{t=2}^T \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2 + \sum_{t=2}^T \|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_2. \tag{C.6}$$

And for smooth functions, we study

$$\text{REG}_T^{\text{SC-SM}} \triangleq \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w}_t) - \min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}} \sum_{t=1}^T \boldsymbol{h}_t(\boldsymbol{w})\right] + \sum_{t=2}^T \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2^2 + \sum_{t=2}^T \|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_2 + \sum_{t=2}^T \|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_2^2. \tag{C.7}$$

In the following, we discuss the Lipschitzness and smoothness cases respectively.

Lipschitzness Case. To begin with, the unary regret part can be upper-bounded using the guarantee of Luo et al. (2022) for Lipschitz functions. We restate it in Lemma 9 with the corresponding proof in Appendix C.1 for self-containedness.

In the following, we focus on the switching cost terms. Since $\boldsymbol{w}_t = \bar{\boldsymbol{w}}_t + \boldsymbol{H}_t^{-1/2}\boldsymbol{b}$, $\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2$ can be bounded as

$$\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2 = \|\boldsymbol{H}_t^{-1/2}\boldsymbol{b}\|_2 \leq \mathcal{O}\left((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3}\right),$$

where the last step is because of $\boldsymbol{H}_t = \nabla^2\Psi(\bar{\boldsymbol{w}}_t) + \eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})\boldsymbol{I}$ and the step size setup (3.5). Since $\boldsymbol{H}_t \succeq \eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})\boldsymbol{I}$, we obtain

$$\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_2 \leq \frac{\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_{\boldsymbol{H}_t}}{\eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})} \leq \frac{\eta_1}{\eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})} \leq \mathcal{O}\left((\sigma_{1:t-1}+\lambda_{0:t-1})^{-2/3}\right),$$

where the second step is due to Lemma 1 and the last step is due to the step size setup (3.5) and $\eta_1 = \mathcal{O}(1)$. In the following, we show that the above term can be absorbed by $\mathcal{O}((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3})$. Specifically, due to (3.5) and $\lambda_t \in (0,1)$, it holds that $d^{2/3}(L_f+1)^{2/3}/(\sigma_{1:t}+\lambda_{0:t})^{1/3} < 1$, which implies $(\sigma_{1:t}+\lambda_{0:t})^{1/3} > d^{2/3}(L_f+1)^{2/3} > 1$. As a result, $\sigma_{1:t}+\lambda_{0:t} > 1$, and thus $\mathcal{O}((\sigma_{1:t-1}+\lambda_{0:t-1})^{-2/3}) \leq \mathcal{O}((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3})$. Combining all terms, the regret with switching cost (C.6) can be bounded by

$$\text{REG}_T^{\text{SC-LIP}} \leq \widetilde{\mathcal{O}}\left(T^{1/3} + \lambda_{1:T} + \sum_{t=1}^T (\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/3}\right) \leq \widetilde{\mathcal{O}}\left(T^{1/3} + \lambda_{1:T-1} + \sum_{t=1}^{T-1} (\sigma_{1:t}+\lambda_{0:t})^{-1/3}\right),$$

for any regularization coefficients $\{\lambda_t\}_{t=1}^T \in (0,1)$, where the last step is due to $\lambda_T \in (0,1)$ and $(\sigma_{1:0}+\lambda_0)^{-1/3} \leq \lambda_0^{-1/3} \leq 1$ ($\lambda_0 \geq 1$ because of Lemma 1).

Later, following Lemma 22 of Luo et al. (2022), by setting the regularization coefficient as $\lambda_t \approx (\sigma_{1:t}+\lambda_{0:t})^{-1/3}$, the regret of BCO with switching cost for Lipschitz functions can be bounded by the optimal regularization coefficients $\{\lambda_t^\star\}_{t=1}^T$:

$$\mathrm{REG}_T^{\text{SC-LIP}} \leq \widetilde{\mathcal{O}}\left(\inf_{\lambda_1^\star,\ldots,\lambda_T^\star}\left\{T^{1/3}+\lambda_{1:T-1}^\star+\sum_{t=1}^{T-1}(\sigma_{1:t}+\lambda_{0:t}^\star)^{-1/3}\right\}\right),$$

which finishes the proof.

Smoothness Case. To begin with, the unary regret part can be upper-bounded using the guarantee of Luo et al. (2022) for smooth functions. We restate it in Lemma 10 with the corresponding proof in Appendix C.1 for self-containedness.

In the following, we focus on the switching cost terms. Since $\boldsymbol{w}_t = \bar{\boldsymbol{w}}_t + \boldsymbol{H}_t^{-1/2}\boldsymbol{b}$, $\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2^2$ can be bounded as

$$\|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_t\|_2^2 = \|\boldsymbol{H}_t^{-1/2}\boldsymbol{b}\|_2^2 \leq \mathcal{O}\left((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2}\right),$$

where the last step is because of $\boldsymbol{H}_t \triangleq \nabla^2\Psi(\bar{\boldsymbol{w}}_t) + \eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})\boldsymbol{I}$ and the step size setup (3.6). Since $\boldsymbol{H}_t \succeq \eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})\boldsymbol{I}$, we obtain

$$\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_2 \leq \frac{\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_{\boldsymbol{H}_t}}{\eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})} \leq \frac{\eta_1}{\eta_t(\sigma_{1:t-1}+\lambda_{0:t-1})} \leq \mathcal{O}\left((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2}\right),$$

where the second step is due to Lemma 1 and the last step is due to the step size setup (3.6) and $\eta_1 = \mathcal{O}(1)$. Accordingly,

$$\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t-1}\|_2^2 \leq \mathcal{O}\left((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1}\right).$$

In the following, we show that the above term can be absorbed by $\mathcal{O}((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2})$. Specifically, due to (3.6) and $\lambda_t \in (0,1)$, it holds that $d\sqrt{\beta_f+1}/\sqrt{\sigma_{1:t}+\lambda_{0:t}} < 1$, which implies $\sqrt{\sigma_{1:t}+\lambda_{0:t}} > d\sqrt{\beta_f+1} > 1$. As a result, $\sigma_{1:t}+\lambda_{0:t} > 1$, and thus $\mathcal{O}((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1}) \leq \mathcal{O}((\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2})$. Combining all terms, the regret with switching cost (C.7) can be bounded by

$$\mathrm{REG}_T^{\text{SC-SM}} \leq \widetilde{\mathcal{O}}\left(\sqrt{T}+\lambda_{1:T}+\sum_{t=1}^{T}(\sigma_{1:t-1}+\lambda_{0:t-1})^{-1/2}\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{T}+\lambda_{1:T-1}+\sum_{t=1}^{T-1}(\sigma_{1:t}+\lambda_{0:t})^{-1/2}\right),$$

for any regularization coefficients $\{\lambda_t\}_{t=1}^T \in (0,1)$, where the last step is due to $\lambda_T \in (0,1)$ and $(\sigma_{1:0}+\lambda_0)^{-1/2} \leq \lambda_0^{-1/2} \leq 1$ ($\lambda_0 \geq 1$ because of Lemma 1).

Later, following Lemma 6 of Luo et al. (2022), by setting the regularization coefficient as $\lambda_t \approx (\sigma_{1:t}+\lambda_{0:t})^{-1/2}$, the regret of BCO with switching cost for smooth functions can be bounded by the optimal regularization coefficients $\{\lambda_t^\star\}_{t=1}^T$:

$$\mathrm{REG}_T^{\text{SC-SM}} \leq \widetilde{\mathcal{O}}\left(\inf_{\lambda_1^\star,\ldots,\lambda_T^\star}\left\{\sqrt{T}+\lambda_{1:T-1}^\star+\sum_{t=1}^{T-1}(\sigma_{1:t}+\lambda_{0:t}^\star)^{-1/2}\right\}\right),$$

which finishes the proof. □

D. Omitted Proofs of Section 3.3 (Bandit LQR Control)

In this section, we provide the omitted proofs of Section 3.3, including Theorem 3, and corollaries considering the cost functions with pure convexity and quadraticity (Corollary 1 and 2), corrupted quadraticity (Corollary 3 and 4), and decaying quadraticity (Corollary 5 and 6).

D.1. Proof of Theorem 3

Proof. To begin with, we decompose the regret in the following way.

$$\begin{aligned}\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] &= \mathbb{E}\left[\sum_{t=1}^{T} c_t(\mathbf{y}_t,\mathbf{u}_t) - \min_{\pi\in\Pi}\sum_{t=1}^{T} c_t(\mathbf{y}_t^{\pi},\mathbf{u}_t^{\pi})\right] = \mathbb{E}\left[\sum_{t=1}^{T} c_t(\mathbf{y}_t,\mathbf{u}_t) - \sum_{t=1}^{T} c_t(\mathbf{y}_t(M^\star),\mathbf{u}_t(M^\star))\right] \\ &\le \underbrace{\sum_{t=1}^{H'-1} c_t(\mathbf{y}_t,\mathbf{u}_t)}_{\textsc{Burn-In}} + \underbrace{\sum_{t=H'}^{T} c_t(\mathbf{y}_t,\mathbf{u}_t) - \sum_{t=H'}^{T} f_t(M_{[t-H:t]})}_{\textsc{term (i)}} \\ &\quad + \underbrace{\sum_{t=H'}^{T} \widetilde{f}_t(M^\star) - \sum_{t=H'}^{T} c_t(\mathbf{y}_t(M^\star),\mathbf{u}_t(M^\star))}_{\textsc{term (ii)}} + \underbrace{\mathbb{E}\left[\sum_{t=H'}^{T} f_t(M_{[t-H:t]}) - \sum_{t=H'}^{T} \widetilde{f}_t(M^\star)\right]}_{\textsc{Regret}},\end{aligned}$$

where the first line is due to the definition of the optimal fixed policy $M^\star \in \arg\min_{M\in\mathcal{M}} \sum_{t=1}^{T} c_t(\mathbf{y}_t(M),\mathbf{u}_t(M))$, which is fixed due to the cost functions $\{c_t(\cdot,\cdot)\}_{t=1}^{T}$ are chosen by an oblivious adversary. For Lipschitz costs, by choosing the memory length $H = \Theta(\log T)$, the burn-in cost can be bounded by $\sum_{t=1}^{m+H}\mathbb{E}[c_t(\mathbf{y}_t,\mathbf{u}_t)] \le \widetilde{\mathcal{O}}(1)$. TERM (I) is exactly zero due to the property of with-history loss functions, i.e., $f_t(M_{[t-H:t]}) = c_t(\mathbf{y}_t,\mathbf{u}_t)$. And TERM (II) is the truncation error, which is at most $\widetilde{\mathcal{O}}(1)$ due to Lemma 6.

Finally, it remains to deal with the REGRET term. We first give it a further decomposition:

$$\textsc{Regret} = \underbrace{\sum_{t=H'}^{T} \boldsymbol{f}_t(\boldsymbol{M}_{[t-H:t]}) - \sum_{t=H'}^{T} \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t)}_{\textsc{Memory}} + \underbrace{\mathbb{E}\left[\sum_{t=H'}^{T} \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \sum_{t=H'}^{T} \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}^\star)\right]}_{\textsc{Unary-Reg}},$$

where $\boldsymbol{f}_t : \boldsymbol{\mathcal{M}}^{H+1} \mapsto \mathbb{R}$ denotes the loss function in lifted domain such that $\boldsymbol{f}_t(\boldsymbol{M}_1,\dots,\boldsymbol{M}_{H+1}) \triangleq f_t(M_1,\dots,M_{H+1})$ for any $\boldsymbol{M}_i = (M_i, 1)$. Its unary version is correspondingly defined as $\widetilde{\boldsymbol{f}}_t(\boldsymbol{M}) \triangleq \boldsymbol{f}_t(\boldsymbol{M},\dots,\boldsymbol{M})$ for any $\boldsymbol{M} \in \boldsymbol{\mathcal{M}}$. It can be verified that $\widetilde{\boldsymbol{f}}_t(\cdot)$ can still inherit the σ_t-strong convexity, L_f-Lipschitzness, and β_f-smoothness from the original $f_t(\cdot)$.

Denoting by $\mathcal{S}$ the time horizon when the algorithm updates, the unary regret will only become $3H$ times larger (Lemma 7):

$$\textsc{Unary-Reg} \le 3H \cdot \mathbb{E}\left[\sum_{t\in\mathcal{S}} \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \sum_{t\in\mathcal{S}} \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}^\star)\right].$$

We note that since the comparator $M^\star$ (and thus $\boldsymbol{M}^\star$) is fixed, we can easily extend Lemma 7 for non-oblivious adversary. To further deal with the unary regret, we denote by $f_{t;H}(M) \triangleq \mathbb{E}_{t-H-1}[\widetilde{f}_t(M)]$ the expected version of $\widetilde{f}_t(\cdot)$ for any $M \in \mathcal{M}$. Due to Lemma 8, $f_{t;H}(\cdot)$ is σ_t-strongly convex. It is easy to verify that $\boldsymbol{f}_{t;H}(\boldsymbol{M}) \triangleq \mathbb{E}_{t-H-1}[\widetilde{\boldsymbol{f}}_t(\boldsymbol{M})]$ is also σ_t-strongly convex for any $\boldsymbol{M} = (M,1) \in \boldsymbol{\mathcal{M}}$. As a result, the $\mathbb{E}[\widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}^\star)]$ term in the unary regret can be further transformed to

$$\begin{aligned}\mathbb{E}\left[\widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}^\star)\right] &= \mathbb{E}\left[\mathbb{E}_{t-H-1}\left[\widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}^\star)\right]\right] = \mathbb{E}\left[\boldsymbol{f}_{t;H}(\boldsymbol{M}_t) - \boldsymbol{f}_{t;H}(\boldsymbol{M}^\star)\right] \\ &\le \mathbb{E}\left[\langle\nabla\boldsymbol{f}_{t;H}(\boldsymbol{M}_t), \boldsymbol{M}_t - \boldsymbol{M}^\star\rangle - \frac{\sigma_t}{2}\|\boldsymbol{M}_t - \boldsymbol{M}^\star\|_{\mathrm{F}}^2\right] \\ &= \underbrace{\mathbb{E}\left[\langle\boldsymbol{g}_t, \boldsymbol{M}_t - \boldsymbol{M}^\star\rangle - \frac{\sigma_t}{2}\|\boldsymbol{M}_t - \boldsymbol{M}^\star\|_{\mathrm{F}}^2\right]}_{\textsc{Opt-Term}} + \underbrace{\mathbb{E}\left[\langle\nabla\boldsymbol{f}_{t;H}(\boldsymbol{M}_t) - \boldsymbol{g}_t, \boldsymbol{M}_t - \boldsymbol{M}^\star\rangle\right]}_{\textsc{Bias-Term}},\end{aligned}$$

where the first line is due to the definition of $\boldsymbol{f}_{t;H}(\cdot)$, the second line is because of the strong convexity of $\boldsymbol{f}_{t;H}(\cdot)$. The OPT-TERM in the third line can be optimized in a deterministic way following the same analysis as that in Lemma 9 and Lemma 10. To optimize the BIAS-TERM, we give it a further decomposition:

$$\textsc{Bias-Term} = \mathbb{E}\left[\langle\nabla\boldsymbol{f}_{t;H}(\boldsymbol{M}_t) - \nabla\widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t), \boldsymbol{M}_t - \boldsymbol{M}^\star\rangle\right] + \mathbb{E}\left[\langle\nabla\widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \boldsymbol{g}_t, \boldsymbol{M}_t - \boldsymbol{M}^\star\rangle\right],$$

where the first term

$$\mathbb{E}\left[\langle \nabla \boldsymbol{f}_{t;H}(\boldsymbol{M}_t) - \nabla \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t), \boldsymbol{M}_t - \boldsymbol{M}^\star \rangle\right] = \mathbb{E}\left[\langle \mathbb{E}_t\left[\nabla \boldsymbol{f}_{t;H}(\boldsymbol{M}_t) - \nabla \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t)\right], \boldsymbol{M}_t - \boldsymbol{M}^\star \rangle\right] = 0 \tag{D.1}$$

because of the definition of $\boldsymbol{f}_{t;H}(\cdot)$, where $\mathbb{E}_t[\cdot]$ is taken on the randomness up to the t-th round. The second term

$$\mathbb{E}\left[\langle \nabla \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \boldsymbol{g}_t, \boldsymbol{M}_t - \boldsymbol{M}^\star \rangle\right] = \mathbb{E}\left[\langle \mathbb{E}_t\left[\nabla \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t) - \boldsymbol{g}_t\right], \boldsymbol{M}_t - \boldsymbol{M}^\star \rangle\right] = 0 \tag{D.2}$$

because the gradient estimator $\boldsymbol{g}_t$ is unbiased for the true gradient $\nabla \widetilde{\boldsymbol{f}}_t(\boldsymbol{M}_t)$ (actually $\boldsymbol{g}_t$ is an unbiased estimator of the smoothed version of $\widetilde{\boldsymbol{f}}_t$ at $\boldsymbol{M}_t$). Note that in the first step of (D.1) and (D.2), given randomness up to the t-th round, the variable $\boldsymbol{M}_t - \boldsymbol{M}^\star$ is deterministic mainly due to the fact that $\boldsymbol{M}^\star$ is fixed as discussed before. Otherwise, when facing a general non-oblivious adversary, this step will not hold due to the randomness on the comparator.

As for the memory part, since $\boldsymbol{M}_t = \bar{\boldsymbol{M}}_t + \boldsymbol{H}_t^{-1/2}\boldsymbol{s}_t$, i.e., $\mathbb{R}[\boldsymbol{M}_t] = \bar{\boldsymbol{M}}_t$, if the costs are Lipschitz, which implies L_f-Lipschitzness of $\boldsymbol{f}_t$, the memory part can be bounded by

$$\textsc{Memory} \le \frac{1}{2} L_f H^2 \sum_{t \in \mathcal{S}} (\|\bar{\boldsymbol{M}}_{t+1} - \bar{\boldsymbol{M}}_t\|_{\mathrm{F}} + 2\|\boldsymbol{M}_t - \bar{\boldsymbol{M}}_t\|_{\mathrm{F}}).$$

If the costs are smooth, which implies β_f-smoothness of $\widetilde{\boldsymbol{f}}_t$, we obtain

$$\textsc{Memory} \le \frac{1}{2} H^2 \sum_{t \in \mathcal{S}} (L_f \|\bar{\boldsymbol{M}}_{t+1} - \bar{\boldsymbol{M}}_t\|_{\mathrm{F}} + \beta_f \|\bar{\boldsymbol{M}}_{t+1} - \bar{\boldsymbol{M}}_t\|_{\mathrm{F}}^2 + 6\beta_f \|\boldsymbol{M}_t - \bar{\boldsymbol{M}}_t\|_{\mathrm{F}}^2).$$

Thus we successfully reduce the problem to bandit convex optimization with switching cost (BCO-SC). Thus we can directly use the parameter configurations of Theorem 2 and obtain the same regret guarantees therein. Finally, noticing that the strong convexity parameter α_t of the cost function $c_t(\cdot, \cdot)$ is linear in the strong convexity parameter σ_t of the with-history loss function $\widetilde{f}_t(\cdot)$, using the second part of Lemma 8 finishes the proof. □

D.2. Proofs of Corollaries 1-6

Proof of Corollary 1. Since Theorem 3 holds for any non-negative sequence of $\{\lambda_t^\star\}_{t=1}^T$, when $\{c_t(\cdot)\}_{t=1}^T$ are convex, if we choose $\lambda_1^\star = T^{3/4}$ and $\lambda_t^\star = 0$ for $t \ge 2$, then it holds that $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \le \widetilde{\mathcal{O}}(T^{3/4})$. If $\{c_t(\cdot)\}_{t=1}^T$ are α-quadratic, if we choose $\lambda_t^\star = 0$ for $t \in [T]$, it holds that $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \le \widetilde{\mathcal{O}}(\alpha^{-1/3}T^{2/3})$. The proof is finished. □

Proof of Corollary 2. Since Theorem 3 holds for any non-negative sequence of $\{\lambda_t^\star\}_{t=1}^T$, when $\{c_t(\cdot)\}_{t=1}^T$ are convex, if we choose $\lambda_1^\star = T^{2/3}$ and $\lambda_t^\star = 0$ for $t \ge 2$, then it holds that $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \le \widetilde{\mathcal{O}}(T^{2/3})$. If $\{c_t(\cdot)\}_{t=1}^T$ are α-quadratic, if we choose $\lambda_t^\star = 0$ for $t \in [T]$, it holds that $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \le \widetilde{\mathcal{O}}(\sqrt{T/\alpha})$. The proof is finished. □

Proof of Corollary 3. When the first M online functions are convex and the rest ones are α-quadratic, the regret upper-bound in Theorem 3 is the largest. By choosing $\lambda_t^\star = 0$ for $t \ge 2$, it holds that

$$\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \le \widetilde{\mathcal{O}}\left(T^{1/3} + \lambda_{1:T-1}^\star + \sum_{t=1}^{T-1} (\alpha_{1:t} + \lambda_{0:t}^\star)^{-1/3}\right) \le \widetilde{\mathcal{O}}\left(\lambda_1^\star + M\lambda_1^{\star -1/3} + \alpha^{-1/3}(T-M)^{2/3}\right),$$

where the last step omits the low-order term of $\widetilde{\mathcal{O}}(T^{1/3})$. Choosing $\lambda_1^\star = M^{3/4}$ finishes the proof of the first part. When the σ-quadratic functions appear in the first $(T-M)$ rounds, the above guarantee can be strengthened. Specifically, by choosing $\lambda_t^\star = 0$ for $t \in [T]$, we obtain

$$\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \le \widetilde{\mathcal{O}}\left(T^{1/3} + \lambda_{1:T-1}^\star + \sum_{t=1}^{T-1} (\alpha_{1:t} + \lambda_{0:t}^\star)^{-1/3}\right) \le \widetilde{\mathcal{O}}\left(\alpha^{-1/3}(T-M)^{2/3} + \alpha^{-1/3}M(T-M)^{-1/3}\right),$$

where the first term represent the regret bound of the first $(T-M)$ rounds for quadratic functions and the second term is the regret bound of the last M rounds for convex functions, finishing the proof. □

Proof of Corollary 4. When the first M online functions are convex and the rest ones are α-quadratic, the regret upper-bound in Theorem 3 is the largest. By choosing $\lambda_t^\star = 0$ for $t \geq 2$, it holds that

$$\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}\left(\sqrt{T} + \lambda_{1:T-1}^\star + \sum_{t=1}^{T-1}(\alpha_{1:t} + \lambda_{0:t}^\star)^{-1/2}\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{T} + \lambda_1^\star + M{\lambda_1^\star}^{-1/2} + \alpha^{-1/2}(T-M)^{1/2}\right).$$

Choosing $\lambda_1^\star = M^{2/3}$ finishes the proof of the first part. When the α-quadratic functions appear in the first $(T-M)$ rounds, the above guarantee can be strengthened. Specifically, by choosing $\lambda_t^\star = 0$ for $t \in [T]$, we obtain

$$\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}\left(\sqrt{T} + \lambda_{1:T-1}^\star + \sum_{t=1}^{T-1}(\alpha_{1:t} + \lambda_{0:t}^\star)^{-1/2}\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{T} + \alpha^{-1/2}(T-M)^{1/2} + \alpha^{-1/2}M(T-M)^{-1/2}\right),$$

where the first term represent the regret bound of the first $(T-M)$ rounds for quadratic functions and the second term is the regret bound of the last M rounds for convex functions, finishing the proof. □

Proof of Corollary 5. To begin with, choosing $\lambda_1^\star = T^b$ and $\lambda_t^\star = 0$ for $t \geq 2$, from Theorem 3, we obtain

$$\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}\left(T^b + \sum_{t=1}^{T-1}(t^{1-\gamma} + \lambda_1^\star)^{-1/3}\right) \leq \widetilde{\mathcal{O}}\left(T^b + \min\left\{T^{2/3+\gamma/3}, T^{1-b/3}\right\}\right),$$

where the first step omits the low-order term of $\widetilde{\mathcal{O}}(T^{1/3})$ and uses $\alpha_{1:t} = \sum_{s=1}^t s^{-\gamma} = \mathcal{O}(t^{1-\gamma})$. In the following, we discuss the above upper bound case by case. If $2/3+\gamma/3 \leq 1-b/3$, i.e., $\gamma+b \leq 1$, it holds that $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}(T^b+T^{2/3+\gamma/3})$. To minimize the upper bounds, we set $b = 2/3+\gamma/3$ and achieve $\widetilde{\mathcal{O}}(T^{2/3+\gamma/3})$. Combining $\gamma+b \leq 1$ and $b = 2/3+\gamma/3$ gives the constraint of $\gamma \leq 1/4$. Otherwise, if $2/3+\gamma/3 > 1-b/3$, i.e., $\gamma+b > 1$, we obtain $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}(T^b+T^{1-b/3})$. Choosing $b = 3/4$ gives an $\widetilde{\mathcal{O}}(T^{3/4})$ regret bound. Combining $\gamma+b > 1$ and $b = 3/4$ gives the constraint of $\gamma > 1/4$. □

Proof of Corollary 6. To begin with, choosing $\lambda_1^\star = T^b$ and $\lambda_t^\star = 0$ for $t \geq 2$, from Theorem 3, we obtain

$$\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}\left(\sqrt{T} + T^b + \sum_{t=1}^{T-1}(t^{1-\gamma} + \lambda_1^\star)^{-1/2}\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{T} + T^b + \min\left\{T^{1/2+\gamma/2}, T^{1-b/2}\right\}\right),$$

where the first step is due to $\alpha_{1:t} = \sum_{s=1}^t s^{-\gamma} = \mathcal{O}(t^{1-\gamma})$. In the following, we discuss the above upper bound case by case. If $1/2+\gamma/2 \leq 1-b/2$, i.e., $\gamma+b \leq 1$, it holds that $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}(T^b+T^{1/2+\gamma/2})$. To minimize the upper bounds, we set $b = 1/2+\gamma/2$ and achieve $\widetilde{\mathcal{O}}(T^{1/2+\gamma/2})$. Combining $\gamma+b \leq 1$ and $b = 1/2+\gamma/2$ gives the constraint of $\gamma \leq 1/3$. Otherwise, if $1/2+\gamma/2 > 1-b/2$, i.e., $\gamma+b > 1$, we obtain $\mathbb{E}[\textsc{Reg}_T^{\mathrm{C}}] \leq \widetilde{\mathcal{O}}(\sqrt{T}+T^b+T^{1-b/2})$. Choosing $b = 2/3$ gives an $\widetilde{\mathcal{O}}(T^{2/3})$ regret bound. Combining $\gamma+b > 1$ and $b = 2/3$ gives the constraint of $\gamma > 1/3$. □

E. Technical Lemmas

In this section, we provide technical lemmas about the relationship between Lipschitzness and strong convexity (Lemma 11) and a basic lemma about FTRL (Lemma 12).

Lemma 11 (Lemma 31 of Luo et al. (2022))**.** *If a convex function $f : \mathcal{W} \mapsto \mathbb{R}$ is L-Lipschitz and σ-strongly convex, and has bounded domain diameter $\max_{w_1,w_2\in\mathcal{W}} \|w_1 - w_2\|_2 \leq D$, then it holds that $\sigma \leq 4L/D$.*

Lemma 12. *Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a closed and convex feasible set, and denote by $\psi_t : \mathcal{W} \mapsto \mathbb{R}$ the convex regularizer and $h_t : \mathcal{W} \mapsto \mathbb{R}$ the convex online functions. Denoting by $F_t(w) = \sum_{s=1}^{t-1} h_s(w) + \psi_t(w)$, if the FTRL update rule is specified as $w_t \in \arg\min_{w\in\mathcal{W}} F_t(w)$, then for any $w \in \mathcal{W}$, we have*

$$\sum_{t=1}^T h_t(w_t) - \sum_{t=1}^T h_t(w) \leq \psi_{T+1}(w) - \psi_1(w_1) + \sum_{t=1}^T \nabla h_t(w_t)^\top (w_t - w_{t+1}) + \sum_{t=1}^T (\psi_t(w_{t+1}) - \psi_{t+1}(w_{t+1})).$$