Second Order Methods for Bandit Optimization and Control

Arun Suggala* ARUNSS@GOOGLE.COM
Google Research

Y. Jennifer Sun* YS7849@PRINCETON.EDU
Princeton University

Praneeth Netrapalli PNETRAPALLI@GOOGLE.COM
Google Research

Elad Hazan EHAZAN@PRINCETON.EDU
Princeton University and Google DeepMind

Editors: Shipra Agrawal and Aaron Roth

Abstract

Bandit convex optimization (BCO) is a general framework for online decision making under uncertainty. While tight regret bounds for general convex losses have been established, existing algorithms achieving these bounds have prohibitive computational costs for high dimensional data.

In this paper, we propose a simple and practical BCO algorithm inspired by the online Newton step algorithm. We show that our algorithm achieves optimal (in terms of horizon) regret bounds for a large class of convex functions that satisfy a condition we call κ-convexity. This class contains a wide range of practically relevant loss functions including linear losses, quadratic losses, and generalized linear models. In addition to optimal regret, this method is the most efficient known algorithm for several well-studied applications including bandit logistic regression.

Furthermore, we investigate the adaptation of our second-order bandit algorithm to online convex optimization with memory. We show that for loss functions with a certain affine structure, the extended algorithm attains optimal regret. This leads to an algorithm with optimal regret for bandit LQ problem under a fully adversarial noise model, thereby resolving an open question posed in (Gradu et al., 2020) and (Sun et al., 2023).

Finally, we show that the more general problem of BCO with (non-affine) memory is harder. We derive a $\tilde{\Omega}(T^{2/3})$ regret lower bound, even under the assumption of smooth and quadratic losses.

Keywords: Bandit Convex Optimization, Nonstochastic Control, Second Order Methods

1. Introduction

Bandit convex optimization (BCO) is a prominent framework for online decision-making. It can be described as an interactive game between a learner and an adversary. At time $t \in \mathbb{N}$, the learner must choose an action x_t from a convex constraint set $\mathcal{K} \subset \mathbb{R}^d$. Once x_t is chosen and played by the learner, the adversary reveals a convex loss function $f_t : \mathbb{R}^d \to \mathbb{R}$, to which the learner suffers loss $f_t(x_t)$. Unlike full-information settings where the learner observes the loss function f_t at each round, bandit feedback provides only scalar feedback—the loss associated with the chosen action, *i.e.,* the scalar $f_t(x_t)$. The learner's goal is to minimize *regret* over a time horizon T, which is

[*] Equal contribution, alphabetical order.

© 2024 A. Suggala, Y.J. Sun, P. Netrapalli & E. Hazan.

defined to be the difference between the total loss incurred by the learner and the best fixed action in $\mathcal{K}$ had the loss sequences were known ahead of the time:

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x).$$

There is a long line of work in the online learning community that aims to attain optimal regret guarantees in bandit convex optimization, see e.g. Lattimore and Szepesvári (2020). The optimal regret for this setting in terms of the number of iterations, on the order of $O(\sqrt{T})$, was obtained in Bubeck et al. (2017). However, the regret of this method has high polynomial dependence on the dimension, and similarly the running time is polynomial in both dimension and number of iterations, rendering it impractical in many applications. Motivated by the need for more efficient methods, works by Abernethy et al. (2009); Hazan and Levy (2014); Suggala et al. (2021) obtained more practical algorithms, but for restrictive classes of loss functions. Namely, these latter works apply to linear, strongly-convex and smooth, and quadratic losses, respectively.

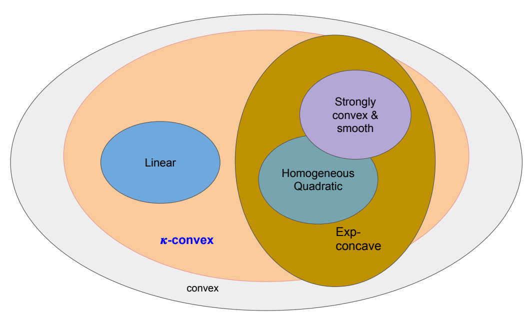
One of the interesting remaining open problems in the area is to design an online algorithm that (1) works for a rich class of functions, (2) is computationally efficient, and (3) obtains $O(\sqrt{T})$ regret without heavy dependence on the dimension. We advance this research direction with an efficient second-order method with contributions summarized below.

1.1. Our contributions to bandit convex optimization (BCO)

Our first contribution is a new algorithm called Bandit Newton Step (Algorithm 1), which is a natural adaptation of the Online Newton Step algorithm (Hazan et al., 2007) to the bandit feedback setting. Our algorithm is an improper learning technique, *i.e.,* it plays actions that lie outside of the constraint set $\mathcal{K}$, but competes with the best point inside of $\mathcal{K}$. Such improper learning algorithms are applicable to the applications we consider, such as online regression and online control. The guarantees of our algorithm apply to a broad class of convex functions that satisfy a curvature assumption called κ-convexity, which is formally defined as the following:

Definition 1 (κ-convexity) *A function f is called κ-convex over domain $\mathcal{K} \subseteq \mathbb{R}^d$ iff the following holds: f is convex and twice differentiable almost everywhere, and moreover $\exists c, C > 0$ and a PSD matrix $0 \preceq H \preceq I$ such that the Hessian of f at any $x \in \mathcal{K}$ satisfies*

$$cH \preceq \nabla^2 f(x) \preceq CH\ ,\ \frac{C}{c} \leq \kappa.$$



As shown in the diagram above, the class of κ-convex functions generalizes loss functions encountered in prior works (Abernethy et al., 2009; Hazan and Levy, 2014; Suggala et al., 2021) and finds broad applicability in online learning. Notably, this class relaxes the requirements of strong convexity and smoothness assumptions, demanding them only along directions where the Hessian H is nonzero. Consequently, it encompasses quadratic, (generalized) linear loss functions. A more detailed discussion on κ-convexity is included in Appendix A.

Contribution 1. Algorithm 1 guarantees for the class of κ-**convex** loss functions: (1) $O(d^{2.5}\sqrt{T})$ regret upper bound, and (2) per-iteration computational complexity of $O(d^2)$.

Paper	Losses	Advers.	Regret	Running time
Abernethy et al. (2009)	linear	✓	$\tilde{O}(d\sqrt{T})$	$O(d^2)$
Hazan and Levy (2014)	strongly convex, smooth	✓	$\tilde{O}(d\sqrt{T})$	$O(d)$
Suggala et al. (2021)	convex quadratic	✓	$\tilde{O}(d^{16}\sqrt{T})$	$O(d^4)$
Bubeck et al. (2021)	bounded convex	✓	$\tilde{O}(d^{10.5}\sqrt{T})$	$\mathrm{poly}(d,T)$
Lattimore (2020)	convex	✓	$\tilde{O}(d^{2.5}\sqrt{T})$	$\exp(d,T)$
Lattimore and Gyorgy (2021)	convex	×	$\tilde{O}(d^{4.5}\sqrt{T})$	$\mathrm{poly}(d)^{\dagger}$
Lattimore and György (2023)	Lipschitz convex	×	$\tilde{O}(d^{1.5}\sqrt{T})$	$O(d^3)$
Theorem 3	κ-convex	✓	$\tilde{O}(d^{2.5}\sqrt{T})$	$O(d^2)$

Table 1: Comparison with relevant prior works that achieve $\tilde{O}(\sqrt{T})$ regret for bandit convex optimization. The last column presents the per-iteration runtime excluding projections. †: this algorithm has d^4 operations that happen infrequently, thus running time over T iterations is at least $\max\{d^4, Td\}$.

The consequences of this result extend to some well studied problem settings with κ-convex losses, encompassing linear and logistic regression problems (Observation 2). For these settings, our results imply optimal in T regret, and computationally efficient bandit algorithms. A detailed discussion of these contributions is provided in Appendix B.

Contribution 1.a. Applied to bandit logistic regression problems, Algorithm 1 guarantees regret $O(d^{2.5}e^{2D}\sqrt{T})$ and running time of $O(d^2)$, where D is the diameter of the domain of the linear predictor.

Contribution 1.b. Applied to bandit linear regression problems, Algorithm 1 guarantees regret $O(d^{2.5}\sqrt{T})$ and running time of $O(d^2)$.

1.2. Our contributions to online control

Paper	Noise	Observability	Regret
Gradu et al. (2020)	adversarial	full	$\tilde{O}(T^{3/4})$
Cassel and Koren (2020)	stochastic	full	$\tilde{O}(\sqrt{T})$
Cassel and Koren (2020)	adversarial	full	$\tilde{O}(T^{2/3})$
Sun et al. (2023)	semi-adversarial	partial	$\tilde{O}(\sqrt{T})$
Theorem 5	adversarial	partial	$\tilde{O}(\sqrt{T})$

Table 2: Comparison with relevant prior works for bandit control of LQR problem.

Next, we proceed to studying the extension of our bandit algorithm to online control of linear dynamical systems. The setting of online control is more challenging and general than BCO, since each instantaneous loss depends on the system's current state, which then depends on the past controls chosen by the learner. Very often, online control algorithms are derived from standard online learning algorithms applied to loss functions with memory. In this setting (Anava et al. (2015)), at each round t, the loss function $f_t : (\mathbb{R}^d)^m \to \mathbb{R}$ depends not only on the current decision of the learner but the most recent m decisions (m is often referred to as the memory length). The learner's objective is to minimize regret, now defined as

$$\text{Memory-Regret}_T = \sum_{t=m}^{T} f_t(x_{t-m+1}, \ldots, x_t) - \min_{x \in \mathcal{K}} \sum_{t=m}^{T} f_t(x, \ldots, x).$$

The reason memory is important for control is that for stable linear dynamical systems, the effect of past states and control decays exponentially over time, allowing truncation and reduction to online learning with memory. The best known bound for bandit linear control with adversarial noise is $\tilde{O}(T^{2/3})$ (Cassel and Koren, 2020). The work of (Sun et al., 2023) gave a $O(\sqrt{T})$ regret bound for online LQ problem [1], but under a more restrictive semi-adversarial noise model. The model assumes that the noise contains a stochastic component, whose second moment admits a lower bound. It remained an open problem to resolve the optimal rates for bandit linear control and bandit convex optimization with memory (BCO-M) in general.

We address this question, and give new tight upper and lower bounds for these settings as follows.

Contribution 2. For bandit LQ problem, we propose NBPC, an efficient controller algorithm that achieves $\tilde{O}(\sqrt{T})$ regret against fully adversarial noise model.

This upper bound leverages the special memory structure, called affine memory, in linear control problems. To see the significance of the affine memory structure, we give a tight lower bound for BCO with general memory, as per below.

Contribution 3. BCO-M with general quadratic and smooth loss functions has a $\tilde{\Omega}(T^{2/3})$ regret lower bound even when some degree of improper learning is allowed.

Our analysis establishes a clear distinction between the complexities of bandit linear control and BCO-M, indicating that structural properties of linear control problems are crucial for obtaining optimal regret bounds.

1.3. Related works

Bandit convex optimization. The setting of BCO is the limited (zero-order) analogue of online convex optimization, see (Hazan, 2022; Lattimore and Szepesvári, 2020) for an in-depth treatment of OCO, BCO, and other bandit settings. The first BCO algorithms for adversarial losses were gradient based (Flaxman et al., 2004), and generally applicable, but did not attain the optimal regret bound. A flurry of research followed with more general and efficient methods, culminating in an optimal regret bound and polynomial time algorithm (Bubeck et al., 2021). That still is not the end of the picture, as the regret bounds and running time for the latter algorithm have a large (polynomial) dependence on the dimension.

1. Linear quadratic (LQ) problem refers to controlling of linear dynamical systems with quadratic costs. LQ problem is one of the most fundamental problems in control theory.

Special cases of loss functions were addressed in a sequence of works. Notable in these is the optimal regret algorithm for bandit linear optimization (Abernethy et al., 2009), which is the most important subcase. Further generalizations with improved runtimes and/or regret bounds include strongly convex + smooth (Hazan and Levy, 2014), and quadratic losses (Suggala et al., 2021). Another important research theme explores efficient bandit algorithms for stochastic losses. The second-order method of Lattimore and György (2023) is the most relevant to our work; it develops a bandit Newton approach for minimizing Lipschitz convex losses over $\mathbb{R}^d$.

Online control. Online nonstochastic control considers the problem of controlling a discrete-time linear dynamical system with an adversarially chosen perturbation and cost sequence. The goal is to compete with the best fixed policy in some benchmark policy class. Agarwal et al. (2019) is the first work that leverages the stability assumption on the system and proved reduction from nonstochastic control to online learning with memory, where the latter was studied in (Anava et al., 2015). The line of work is too extensive to include here, and we refer those who are interested to Hazan and Singh (2022) for a survey. The most relevant works to ours are perhaps Cassel and Koren (2020); Gradu et al. (2020); Sun et al. (2023), where the authors also studied nonstochastic control. A comparison of results is included in Table 2.

There are two other relevant works in the full information setting (Simchowitz et al., 2020; Simchowitz, 2020). Simchowitz et al. (2020) proposed the state-of-art benchmark policy class of disturbance response controllers (DRCs), which both (Sun et al., 2023) and we use as the benchmark policy class for regret evaluation. Simchowitz (2020) leveraged the special structure of control loss and studied the subclass of OCO-M with memory problem which they referred to as OCO with Affine Memory (OCO-AM). This insight is crucial to our analysis.

Lower bounds for bandit algorithms. Bandit convex optimization (BCO) is strictly harder than its full information counterpart, and therefore in general the regret is lower bounded by $\Omega(\sqrt{T})$. Shamir (2013) showed a tight $\Omega(\sqrt{T})$ lower bound on the regret that holds even for strongly convex and smooth loss functions.

Bandit convex optimization with memory (BCO-M) is harder than BCO. Dekel et al. (2014); Koren et al. (2017) showed a $\tilde{\Omega}(T^{2/3})$-lower bound for multi-armed bandit problems with switching/moving cost. The addition of a switching cost is a form of loss function with memory and therefore inspires the question that whether bandit optimization with memory over compact and convex decision set with natural loss functions exhibits the same restriction. We answer this question by showing $\tilde{\Omega}(T^{2/3})$ regret lower bound for quadratic and smooth loss that is adaptive at time t to the algorithm's decisions up to time $t - m$, where m is the memory length. We emphasize that most if not all existing upper bound analysis for BCO and BCO-M holds for such adaptivity assumption. Table 3 includes a summary of results in regret lower bounds for bandit algorithms.

Paper	Decision set	Regret	Note
Dekel et al. (2014)	discrete	$\tilde{\Omega}(T^{2/3})$	0-1 moving cost
Koren et al. (2017)	discrete	$\tilde{\Omega}(T^{2/3})$	metric moving cost
Cesa-Bianchi et al. (2013)	continuous	$\Omega(T^{2/3})$	0-1 moving cost
Theorem 6	continuous	$\tilde{\Omega}(T^{2/3})$	smooth, quadratic loss with memory

Table 3: Comparison with relevant prior works on lower bounds for bandit algorithms.

1.4. Paper organization

In Section 2 we give details of the setting and notations considered. In Section 3 we give the simplest version of Bandit Newton Step (BNS) and main theorem statement. Following that we give the extension of BNS to the OCO with affine memory setting in Section 4. Lower bounds for OCO with memory are given in Section 5. Details of the application to control and proof details are left to the appendix.

1.5. Notations

Denote $\mathbb{B}^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ to be the unit ball in $\mathbb{R}^d$, and $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ to be the unit sphere in $\mathbb{R}^d$. We consider the following matrix norms: $\|M\|_{\text{op}} = \sup\{\|Mx\|_2 : \|x\|_2 = 1\}$ denotes the operator norm. $\|M\| = \rho(M)$ denotes the largest singular value of M. For $u_1, \dots, u_n \in \mathbb{R}^d$, $u_{1:n} = (u_1, \dots, u_n) \in \mathbb{R}^{dn}$ denotes the concatenated vector. For matrix-vector products, $A_1, \dots, A_n \in \mathbb{R}^{m \times d}$, $A_{1:n}u_{1:n} = (A_1u_1, \dots, A_nu_n) \in \mathbb{R}^{mn}$ denotes the concatenated matrix-vector product. For $A \in \mathbb{R}^{d \times d}$ such that $A \preceq 0$, we denote the induced matrix norm on $\mathbb{R}^d$ as $\|v\|_A = \sqrt{v^\top A v}$ and its dual norm as $\|\cdot\|_A^*$. It can be checked that $\|\cdot\|_A^* = \|\cdot\|_{A^{-1}}$ if A is invertible.

2. Preliminaries

2.1. BCO

We consider the BCO problem over a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$. At each round, the learner is allowed to play a point $y_t \in \mathbb{R}^d$, after which a loss $f_t(y_t) \in \mathbb{R}$ is revealed to the learner. If y_t is restricted to lie in $\mathcal{K}$, then the learner is called a *proper* learner. If y_t can lie outside $\mathcal{K}$, then the learner is called *improper*. In this work, we provide an improper learning algorithm for κ-convex losses. The performance of the learner is measured by regret, i.e. the difference between the loss incurred by the learner and the best single point $x^* \in \mathcal{K}$ had the loss functions were known ahead of the time, i.e.

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(y_t) - \min_{x^* \in \mathcal{K}} \sum_{t=1}^{T} f_t(x^*).$$

We make the following assumptions on the sequence of loss functions $\{f_t\}_{t=1}^T$:

Assumption 1 (Oblivious Adversary) *The sequence of losses $\{f_t\}_{t=1}^T$ are fixed ahead of the game (i.e. the sequence of loss unctions does not depend on the learner's decisions).*

Assumption 2 (κ-convex losses) *The sequence of losses $\{f_t\}_{t=1}^T$ are κ-convex. That is, $\exists c, C > 0$, and PSD matrices $0 \preceq H_t \preceq I$ such that*

$$\forall x \in \mathcal{K} + \mathbb{B}^d,\ t \in [T], \quad cH_t \preceq \nabla^2 f_t(x) \preceq CH_t\ ,\ \textit{where}\ \frac{C}{c}\ \leq\ \kappa.$$

Assumption 3 (Bounded range and gradients) *$\exists B, L > 0$ such that $\{f_t\}_{t=1}^T$ satisfies*

$$\max_{1 \leq t \leq T} \sup_{x \in \mathcal{K}} f_t(x) \leq B, \quad \max_{1 \leq t \leq T} \sup_{x \in \mathcal{K}} \|\nabla f_t(x)\|_2 \leq L.$$

2.2. Control of LQ problem

An important application of BCO-M is its reduction from nonstochastic LQ control. Consider the following partially observable linear dynamical system governed by dynamics (A, B) and observation matrix C:

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

where $\mathbf{x}_t, \mathbf{u}_t$ are the state and control played at time t, respectively. $\mathbf{w}_t$ is the perturbation at time t. The system's state is not directly accessible. Instead, an observation, $\mathbf{y}_t$, which is usually a noisy low-dimension projection of the state $\mathbf{x}_t$ coupled with observation noise $\mathbf{e}_t$, is accessible. A, B, C are of appropriate dimensions, i.e. $A \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{x}}}$ $B \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{u}}}$, and $C \in \mathbb{R}^{d_{\mathbf{y}} \times d_{\mathbf{x}}}$, where $d_{\mathbf{x}}, d_{\mathbf{y}}, d_{\mathbf{u}}$ are the dimensions of $\mathbf{x}_t, \mathbf{y}_t, \mathbf{u}_t$. $\mathbf{w}_t \in \mathbb{R}^{d_{\mathbf{x}}}$, and $\mathbf{e}_t \in \mathbb{R}^{d_{\mathbf{y}}}$. In addition, we assume the system has bounded dynamics and permits a strongly stable linear policy:

Assumption 4 (Strongly stabilizable with bounded dynamics) *The matrices that govern the system dynamics and observations in Eq. (1) are bounded: $\|A\|_{\text{op}} \le \kappa_A$, $\|B\|_{\text{op}} \le \kappa_B$, $\|C\|_{\text{op}} \le \kappa_C$, and $\exists K \in \mathbb{R}^{d_{\mathbf{u}} \times d_{\mathbf{y}}}$ with $A + BKC = HLH^{-1}$ for some $H \succ 0$ and $\max\{\|K\|, \|H\|\|H^{-1}\|\} \le \kappa$, $\|L\| \le 1 - \gamma$ for some $\kappa > 0$, $0 < \gamma \le 1$.*

In the nonstochastic control setting, we make no stochastic assumption on the perturbation and noise sequence $\{\mathbf{w}_t, \mathbf{e}_t\}_{t=1}^T$, other than that they are bounded, i.e.

$$\max_{t \in [T]} \{\max\{\|\mathbf{w}_t\|_2, \|\mathbf{e}_t\|_2\}\} \le R_{\mathbf{w},\mathbf{e}}.$$

Moreover, at time t, after a control $\mathbf{u}_t$ is played by the learner, a time-varying quadratic cost

$$c_t(\mathbf{y}_t, \mathbf{u}_t) = \mathbf{y}_t^\top Q_t \mathbf{y}_t + \mathbf{u}_t^\top R_t \mathbf{u}_t \tag{2}$$

is revealed to the learner. We make the following assumption on the adversary.

Assumption 5 (Oblivious adversary) *The adversary that chooses the cost and perturbation sequences is oblivious.*

Similar to Simchowitz (2020); Sun et al. (2023), we make strong convexity assumption on c_t:

Assumption 6 (Strongly convex and smooth quadratic cost) *$\exists \beta \ge 1 \ge \alpha > 0$ such that $Q_t, R_t \succeq \alpha I$ and $Q_t, R_t \preceq \beta I$.*

In the bandit setting with partially observable systems, the learner only has access to the scalar loss $c_t(\mathbf{y}_t, \mathbf{u}_t)$ in addition to the observation $\mathbf{y}_t$ at time t. The performance in nonstochastic control is to minimize *regret*, defined by the total cost incurred by the learner $\mathcal{A}$ over a time horizon T and the would-be cost incurred by the best policy in a benchmark policy class Π, had the costs been known ahead of the game, i.e.

$$\text{Control-Regret}_T(\mathcal{A}) = \sum_{t=1}^T c_t(\mathbf{y}_t^{\mathcal{A}}, \mathbf{u}_t^{\mathcal{A}}) - \min_{\pi \in \Pi} \sum_{t=1}^T c_t(\mathbf{y}_t^{\pi}, \mathbf{u}_t^{\pi}). \tag{3}$$

Here, $(\mathbf{y}_t^{\mathcal{A}}, \mathbf{u}_t^{\mathcal{A}})$ is the observation-control pair reached by learner $\mathcal{A}$, and $(\mathbf{y}_t^{\pi}, \mathbf{u}_t^{\pi})$ is the would-be observation-control pair if the policy π was carried from the beginning of the time. It is evident from the definition of regret that the strength of regret as a performance metric strongly depends on the richness of the policy class Π. The standard benchmark policy class used in literature (Simchowitz et al. (2020); Simchowitz (2020); Sun et al. (2023)) is the disturbance response controller (DRC) policy class, formally given by the following definition:

Definition 2 (DRC policy class) *(1) A disturbance response controller (DRC) is a policy π_M of length m parametrized by a sequence of m matrices $M = (M^{[j]})_{j=0}^{m-1}$ of dimension $d_{\mathbf{u}} \times d_{\mathbf{y}}$ such that the control at t is given by $\mathbf{u}_t^{\pi_M} = K\mathbf{y}_t + \sum_{j=0}^{m-1} M^{[j]}\mathbf{y}_{t-j}^K$, where $\mathbf{y}_t^K$ is the would-be observation had the linear policy K been carried out from the beginning of the time.*

(2) A DRC policy class $\mathcal{M}(m, R_{\mathcal{M}})$ parametrized by $(m, R_{\mathcal{M}})$ is the set of all disturbance response controllers of length m that obey the norm bound:

$$\|M\|_{\ell_1,\mathrm{op}} = \sum_{j=0}^{m-1} \|M^{[j]}\|_{\mathrm{op}} \leq R_{\mathcal{M}}.$$

3. `BNS`: Bandit Newton Step for κ-Convex Losses

In this section we describe our bandit Newton algorithm `BNS` (see Algorithm 1), which achieves $\tilde{O}(\sqrt{T})$ regret guarantee for the class of κ-convex functions (Definition 1). At a high level, our algorithm tries to estimate the missing information (*i.e.,* gradient and Hessian) about the unknown loss function and then takes a Newton step to compute the next action. To estimate the gradient and Hessian at time step t, we employ the following randomized sampling scheme: We first sample a point uniformly from an ellipsoid centered at the current iterate x_t. The covariance matrix of this ellipsoid is constructed using prior Hessian estimates of the loss functions $\{f_s\}_{s=1}^{t-1}$. We then obtain one-point feedback from the adversary regarding the loss value at the sampled point. This feedback enables gradient and Hessian estimation (see lines 3-5 of Algorithm 1).

Algorithm 1 `BNS`: Bandit Newton Step

Input: convex compact set $\mathcal{K} \subset \mathbb{R}^d$, step size $\eta > 0$, condition number $\kappa' > 0$, time horizon $T \in \mathbb{N}$.

1: Initialize: $x_1 \in \mathcal{K}$, $\tilde{A}_0 = I$
2: **for** $t = 1, \ldots, T$ **do**
3: Draw $v_{t,1}, v_{t,2}$ uniformly from $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d | \|x\|_2 = 1\}$, and let $y_t = x_t + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1} + v_{t,2})$.
4: Play y_t, observe $f_t(y_t)$.
5: Create gradient and Hessian estimates $\tilde{\nabla}_t, \tilde{H}_t$:

$$\tilde{\nabla}_t = 2df_t(y_t)\tilde{A}_{t-1}^{\frac{1}{2}}v_{t,1}, \quad \tilde{H}_t = 2d^2 f_t(y_t)\tilde{A}_{t-1}^{\frac{1}{2}}(v_{t,1}v_{t,2}^\top + v_{t,2}v_{t,1}^\top)\tilde{A}_{t-1}^{\frac{1}{2}}.$$

6: Update $\tilde{A}_t = \tilde{A}_{t-1} + \frac{\eta}{\kappa'}\tilde{H}_t$. Compute $x_{t+1} = \prod_{\mathcal{K}}^{\tilde{A}_t}\left[x_t - \eta\tilde{A}_t^{-1}\tilde{\nabla}_t\right]$, where $\prod_{\mathcal{K}}^{\tilde{A}_t}$ is the projection onto $\mathcal{K}$ w.r.t the norm $\|\cdot\|_{\tilde{A}_t}$.
7: **end for**

Theorem 3 (BNS regret) *For $d, T \in \mathbb{N}$, suppose that $\{f_t\}_{t=1}^T$ and the convex compact set $\mathcal{K} \subset \mathbb{R}^d$ satisfy Assumptions 1,2,3. Let $B^* := B + \sqrt{2}(L + \sqrt{2}C)$. Then, BNS (Algorithm 1) with input $(\mathcal{K}, \eta, \kappa', T)$ with $\kappa' \geq \kappa$, and $\eta \leq \kappa'(24d^{3/2}B^*\kappa\sqrt{T\log(dT^2)})^{-1}$ satisfies the following regret guarantee for any $x \in \mathcal{K}$:*

$$\mathbb{E}\left[\sum_{t=1}^T f_t(y_t) - f_t(x)\right] \leq \frac{\mathrm{diam}(\mathcal{K})^2}{2\eta} + 3\eta d^2(B^*)^2T + \frac{2d\kappa\kappa'\log(1+\eta CT/\kappa')}{\eta}.$$

For the case of $\mathrm{diam}(\mathcal{K}) = \sqrt{d}$, by setting $\kappa' = \kappa$, and $\eta = (24d^{3/2}B^*\sqrt{T\log(dT^2)})^{-1}$, we have

$$\mathbb{E}[\textit{Regret}_T] \leq \tilde{O}(d^{2.5}\kappa^2B^*\sqrt{T}).$$

Proof (Sketch) One of the key steps in the proof is to show that the cumulative Hessian estimator $\tilde{A}_{t-1}$ concentrates well around the true cumulative Hessian w.h.p (Lemma 13). This result implies that the iterates generated by the algorithm are well defined. Another implication of this result is that the action y_t chosen by the learner is at most distance 2 (measured in ℓ_2 norm) away from $\mathcal{K}$ w.h.p. Our regret analysis relies on the following two functions: $\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x) := \mathbb{E}_{u\sim\mathbb{B},v\sim\mathbb{S}}[f_t(x + \frac{1}{2}\tilde{A}_{t-1}^{-1/2}(u+v))|\mathcal{F}_{t-1}]$, $\tilde{f}_t^{\mathbb{S},\mathbb{S}}(x) := \mathbb{E}_{u,v\sim\mathbb{S}}[f_t(x + \frac{1}{2}\tilde{A}_{t-1}^{-1/2}(u+v))|\mathcal{F}_{t-1}]$, where $\mathbb{S} = \mathbb{S}^{d-1}$ and $\mathbb{B} = \mathbb{B}^d$. These two functions are different smoothed variants of f_t. A simple application of Stokes' theorem (Flaxman et al., 2004) shows that $\mathbb{E}[\tilde{\nabla}_t|\mathcal{F}_{t-1}] = \nabla\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t)$, and $\mathbb{E}[f_t(y_t)|\mathcal{F}_{t-1}] = \tilde{f}_t^{\mathbb{S},\mathbb{S}}(x_t)$. In the rest of the proof we perform the following regret decomposition and bound each of the terms

$$\sum_{t=1}^T \mathbb{E}[f_t(y_t) - f_t(x)] = \sum_{t=1}^T \mathbb{E}[\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t) - \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x)] + \mathbb{E}[\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x) - f_t(x)] + \mathbb{E}[\tilde{f}_t^{\mathbb{S},\mathbb{S}}(x_t) - \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t)].$$

The second and third terms capture the penalty we pay for exploration. To bound the first term, we view our algorithm as performing stochastic Newton method on the smoothed losses $\tilde{f}_t^{\mathbb{B},\mathbb{S}}$ (Lemma 12). ■

Running time. Most steps of the BNS algorithm can be implemented with only matrix-vector products, that run in time $O(d^2)$ for dense matrices. In addition, its SVD can be maintained in $O(d^2)$ time using the techniques of (Arbenz and Golub, 1988; Gu and Eisenstat, 1994), since we have only rank-2 updates each iteration. This allows us to compute both the inverse and inverse square-root in linear time.

Comparison with relevant prior work. We take a moment to compare Algorithm 1 with that in (Suggala et al., 2021; Lattimore and György, 2023). While these works have also developed bandit Newton methods, our work differs from theirs in several key aspects: (a) unlike the method of Suggala et al. (2021), which is restricted to quadratic losses, our algorithm is applicable to a broader class of loss functions. Furthermore, our approach is simple and avoids focus regions and restart conditions of Bubeck et al. (2017); Suggala et al. (2021). The main reason for this is that κ-convexity helps us get an optimistic hessian estimate for the entire action space. Finally, the regret of our algorithm has a better dependence in d. (b) In contrast to Lattimore and György (2023), which focuses on stochastic losses in unconstrained domains, our algorithm is designed for the more general adversarial bandit setting in constrained domains.

Improved dependence on κ. Observe that the regret in Theorem 3 has a κ^2 dependence. In Appendix D.2, we provide an alternate analysis that improves this dependence to κ, but at the cost of worse dependence on d. In particular, this analysis leads to a regret of $\tilde{O}(d^3 \kappa B^* \sqrt{T})$, instead of $\tilde{O}(d^{2.5} \kappa^2 B^* \sqrt{T})$ stated in Theorem 3. This improved κ dependence is especially useful in problems with very large κ, *e.g.,* logistic regression where $\kappa = e^{\text{diam}(\mathcal{K})}$ (Appendix B.1). We believe a tighter concentration bound for the cumulative Hessian $\tilde{A}_t$ can improve the regret to $\tilde{O}(d^{1.5} \kappa B^* \sqrt{T})$.

4. `NBPC`: Newton Bandit Perturbation Controller for LQ problem

One extension of interest is the application of a BCO algorithm to online control of LQ problem under bandit feedback. Recall the linear time-invariant dynamical system defined in Eq. (1) and the control regret metric in Eq. (3) with quadratic costs of the structure in Eq. (2). A natural question is that whether an optimal regret guarantee can be extended to bandit nonstochastic LQ problem.

Challenges. The possibility of an optimal rate in the bandit nonstochastic LQ problem remains unresolved. Optimal rates are established only for stochastic (Cassel and Koren, 2020) or semi-adversarial (Sun et al., 2023) settings via reduction to BCO-M. The transition from semi-adversarial to fully adversarial models is significant, especially since strong convexity, which generally improves regret bounds in online learning, does not guarantee the same for loss functions in non-stochastic control problems faced with adversarial disturbances. This is because the loss function reduced from the control problem does not necessarily inherit strong convexity in the presence of fully adversarial disturbances.

Can optimal regret in bandit LQ problem be established for a fully adversarial noise model? We answer this question in two folds. On one hand, in Section 5, we will show the limitation of BCO-M for general convex loss functions by showing a lower bound of $\tilde{\Omega}(T^{2/3})$. However, in this section we will show that despite the hardness of BCO-M for general convex loss functions, an optimal $\tilde{O}(\sqrt{T})$ regret is still attainable if we leverage additional structure of the control problem. Simchowitz (2020) established that the reduction from nonstochastic LQ problem falls within a subclass of the OCO-M framework, where the the dependence on previous decisions assumes an affine structure. This discovery enabled the derivation of the first optimal logarithmic regret bounds for full-information (as opposed to bandit) LQ problem in the presence full adversarial disturbances. The challenge remains in extending the optimal regret bound to the bandit settings. With only scalar feedback, the learner must balance the bias and variance of the gradient estimator variance to obtain optimal regret. Without strong convexity and additional structural assumptions, this balance becomes harder to achieve.

This section will be organized as the following: Section 4.1 will introduce the preliminaries of the problem of bandit quadratic optimization with affine memory (BQO-AM) including important assumptions. Section 4.2 describes `BNS-AM`, an algorithm that achieves optimal regret guarantee for the class of problems described in Section 4.1. Section 4.3 will discuss the reduction from bandit LQ problem to the problem of BQO-AM and introduce the controller `NBPC` (Newton Bandit Perturbation Controller) which achieves the optimal $\tilde{O}(\sqrt{T})$ control regret.

4.1. BQO-AM preliminaries

In the BQO-AM setting, similar to the general BCO-M setting, the learner is asked to play a decision $y_t \in \mathbb{R}^d$ at time t. After the decision is made, an adversary reveals a cost function with memory length m, $f_t : (\mathbb{R}^d)^m \to \mathbb{R}$, that is evaluated the learner's m most recent decisions. For the subclass of BQO-AM problems, f_t admits the following structure:

Assumption 7 (Affine memory structure) *The loss function $f_t : (\mathbb{R}^d)^m \to \mathbb{R}$ admits the structure*

$$f_t(y_{t-m+1:t}) = q_t\left(B_t + \sum_{i=0}^{m-1} G^{[i]} Y_{t-i} y_{t-i}\right), \tag{4}$$

where $G = (G^{[i]})_{i=0}^{m-1}$ is a sequence of matrices of dimension $n \times p$, $Y_{t-m+1:t}$ are m matrices of dimension $p \times d$, and B_t is a vector of dimension n. $q_t : \mathbb{R}^n \to \mathbb{R}$ is a quadratic function with Hessian Q_t. We assume that $\alpha I \preceq Q_t \preceq \beta I$, for some $0 < \alpha \leq 1 \leq \beta$. Additionally, we assume that the learner has access to the following quantities:

$$H_t = G_t^\top G_t \in \mathbb{R}^{d\times d}, \quad G_t = \sum_{i=0}^{m-1} G^{[i]} Y_{t-i} \in \mathbb{R}^{n\times d}, \quad \forall t \in [T].$$

We note that the strong convexity assumption here does not translate to the strong convexity of f_t without further assumptions on $Y_{t-m+1:t}$. Let $\bar{f}_t$ denote the induced unary form of f_t: $\bar{f}_t(x) = f_t(x, \ldots, x)$. The Hessian $\nabla^2 \bar{f}_t$ satisfy the following property: $\alpha H_t \preceq \nabla^2 \bar{f}_t \preceq \beta H_t$ since $\nabla^2 \bar{f}_t = G_t^\top Q_t G_t \in [\alpha H_t, \beta H_t]$. We make the following regularity assumptions.

Assumption 8 (Diameter and gradient bound) *$\mathcal{K} \subset \mathbb{R}^d$ has Euclidean diameter bound D, i.e.*

$$\sup_{x,y\in\mathcal{K}} \|x - y\|_2 \leq D.$$

The value of f_t is bounded by B [2], and gradient norm is bounded L over $\mathcal{K}^m$, $\forall t$, i.e.

$$\max_{t\in[T]} \sup_{z\in\mathcal{K}^m} |f_t(z)| \leq B, \quad \max_{t\in[T]} \sup_{z\in\mathcal{K}^m} \|\nabla f_t(z)\|_2 \leq L.$$

In existing literature on BCO-M, the algorithm holds regret guarantee against m-step delayed adversaries (Gradu et al. (2020); Sun et al. (2023)). Here, we also allow the same adaptivity of the adversary:

Assumption 9 (Adaptivity of the adversary) *The adversary is allowed to be $(t-m)$-steps adaptive. In particular, if we denote as $\mathbb{F}_t$ the filtration generated by the algorithm's randomness up to time t, then the loss function f_t supplied by the adversary at time t is $\mathbb{F}_{t-m}$-measurable.*

Moreover, we make the following regularity assumptions on G, Y_t. In particular, it can be shown that when reducing from nonstochastic LQ problems satisfying Assumption 4,5,6, the following assumption is satisfied.

2. The assumption on bounded function value is non-standard. We add this assumption for simplicity. All of our results hold if substituting the value bound by LD.

Assumption 10 (Exponential decay and positive convolution invertibility-modulus) *G, Y_t satisfy that for $m = \mathrm{poly}(\log T)$,*

$$\kappa(G) = \min\left\{1, \inf_{\sum_{n\geq 0}\|u_n\|_2^2=1}\left\{\sum_{n=0}^{\infty}\left\|\sum_{i=0}^{n} G^{[i]}u_{n-i}\right\|_2^2\right\}\right\} = \Omega(1),$$

$$\|G\|_{\ell_1,\mathrm{op}} = \sum_{i=0}^{m-1}\|G^{[i]}\|_{\mathrm{op}} \leq R_G = O(1), \quad \sum_{i=m}^{\infty}\|G^{[i]}\|_{\mathrm{op}} \leq \frac{R_G}{T},$$

$$\max_{0\leq i\leq m-1}\|(G^{[i]})^\top G^{[i]}\|_2 \leq \tilde{R}_G = O(1), \quad \max_{t\in[T]}\|Y_t\|_{\mathrm{op}} \leq R_Y = O(1).$$

4.2. BNS-AM: Algorithm and Guarantees for BQO-AM problems

We introduce the following BNS-AM algorithm for BQO-AM problems. The update rule resembles that of BNS (Algorithm 1), except that the learner uses H_t to update the preconditioner matrix in place of the Hessian estimator as in BNS.

Algorithm 2 BNS-AM: Bandit Newton Step with Affine Memory

Input: convex compact set $\mathcal{K} \subset \mathbb{R}^d$, step size $\eta > 0$, memory parameter $m \in \mathbb{N}$, time horizon $T \in \mathbb{N}$, curvature parameter $\alpha > 0$.

Initialize: $x_1 = \cdots = x_m \in \mathcal{K}$, $\tilde{g}_{0:m-1} = \mathbf{0}_d$, $\hat{A}_{0:m-1} = mI$.
Sample $u_t \sim \mathbb{S}^{d-1}$ i.i.d. uniformly at random for $t = 1, \dots, m$.
Set $y_t = x_t + \hat{A}_{t-1}^{-\frac{1}{2}}u_t$, $t = 1, \dots, m$.
for $t = m, \dots, T$ **do**
  Play y_t, observe $f_t(y_{t-m+1:t})$, receive H_t.
  Update $\hat{A}_t = \hat{A}_{t-1} + \frac{\eta\alpha}{2}H_t$.
  Create gradient estimate: $\tilde{g}_t = df_t(y_{t-m+1:t})\sum_{j=0}^{m-1}\hat{A}_{t-1-j}^{\frac{1}{2}}u_{t-j} \in \mathbb{R}^d$.
  Update $x_{t+1} = \prod_{\mathcal{K}}^{\hat{A}_{t-m+1}}\left[x_t - \eta\hat{A}_{t-m+1}^{-1}\tilde{g}_{t-m+1}\right]$.
  Sample $u_{t+1} \sim \mathbb{S}^{d-1}$ uniformly at random, independent of previous steps.
  Set $y_{t+1} = x_{t+1} + \hat{A}_{t-m+1}^{-\frac{1}{2}}u_{t+1}$.
end for

The affine memory structure and access to H_t in Assumption 7 provides upper and lower bounds on the Hessian of the loss functions, addressing the challenge in creating Hessian estimators in high dimensions for loss functions with memory. This allows BNS-AM to achieve optimal regret for BQO-AM problems.

Theorem 4 (BNS-AM regret) *For $d, T \in \mathbb{N}$, suppose that the sequence of loss functions $\{f_t\}_{t=1}^T$ and the convex compact set $\mathcal{K} \subset \mathbb{R}^d$ satisfy Assumptions 7,8,9,10. Then, BNS-AM (Algorithm 2) with inputs $(\mathcal{K}, \eta, m, T, \alpha)$ with $m = \mathrm{poly}(\log T)$, $\eta = \tilde{\Theta}\left(\frac{1}{\alpha\sqrt{T}}\right)$, α given by Assumption 7 satisfies the following regret guarantee against any $x \in \mathcal{K}$:*

$$\mathbb{E}[\textit{Memory-Regret}_T(x)] \leq \tilde{O}\left(\frac{\beta}{\alpha}B^*\sqrt{T}\right),$$

with $B^* = B + (L + \beta\sqrt{m})\sqrt{m}$*.*

Proof overview. We decompose the regret into losses incurred by exploration perturbation, movement cost due to estimating the instantaneous loss with the unary form of loss evaluated at the current iterate, and the regret of bandit Newton step without memory.

$$f_t(y_{t-m+1:t}) - \bar{f}_t(x) = \underbrace{f_t(y_{t-m+1:t}) - f_t(x_{t-m+1:t})}_{\text{(perturbation loss)}} + \underbrace{f_t(x_{t-m+1:t}) - \bar{f}_t(x_t)}_{\text{(movement cost)}} + \underbrace{\bar{f}_t(x_t) - \bar{f}_t(x)}_{\text{(no-memory regret)}}$$

The affine memory structure together with regularity conditions allow us to bound each of the above terms by $\tilde{O}(\sqrt{T})$.

4.3. Reduction from the Bandit LQ problem

The `BNS-AM` algorithm and its regret guarantee for BQO-AM problems described in the previous section has significant applications in bandit LQ problems. In particular, recall the bandit LQ problem described in Section 2.2. The bandit LQ problem can be indeed formulated as a BQO-AM problem up to negligible truncation errors. In particular, $\exists f_t$ satisfying Assumptions 7,8,9,10 such that the control cost function c_t can be expressed by f_t. The reduction is formally proved in Section F.1 and F.2. This reduction makes it possible to directly adapt the `BNS-AM` algorithm to the control setting. The control algorithm, which we refer to as Newton Bandit Perturbation Controller (`NBPC`), is specified in Section F.3 with the following control regret guarantee. An extension of the result to unknown systems is included in Appendix G.

Theorem 5 (`NBPC` regret, Algorithm 4) *For* $T, d_{\mathbf{x}}, d_{\mathbf{u}}, d_{\mathbf{y}} \in \mathbb{N}$*, consider a partially observable linear dynamical system (Eq. (1)) with dynamics* (A, B, C) *and state, control, and observation dimensions* $d_{\mathbf{x}}, d_{\mathbf{u}}, d_{\mathbf{y}} \in \mathbb{N}$ *respectively, and a sequence of cost functions* $\{c_t\}_{t=1}^T$ *satisfying Assumptions 4,5,6. Suppose the controller* `NBPC` *is run with inputs* $(\mathcal{M}(m, R_{\mathcal{M}}), \eta, T, (A, B, C), K, \alpha)$ *with* $m = \text{poly}(\log T)$*, DRC policy class* $\mathcal{M}(m, R_{\mathcal{M}})$ *(Definition 2),* $\eta = \tilde{\Theta}\left(\frac{1}{\alpha\sqrt{T}}\right)$*, strongly stabilizing controller* K*, and* $\alpha > 0$ *given by Assumption 6. Then,* `NBPC` *satisfies the following regret guarantee:*

$$\mathbb{E}[\textit{Control-Regret}_T(\textit{NBPC})] = \mathbb{E}\left[\sum_{t=1}^{T} c_t(\mathbf{y}_t, \mathbf{u}_t) - \min_{M \in \mathcal{M}(m, R_{\mathcal{M}})} \sum_{t=1}^{T} c_t(\mathbf{y}_t^M, \mathbf{u}_t^M)\right] \leq \tilde{O}\left(\frac{\beta^2}{\alpha}\sqrt{T}\right).$$

5. Lower Bound for BCO-M

As mentioned in the previous section, BCO-M is hard for general convex loss functions. In this section, we turn our attention to the general problem of BCO-M. We are interested in a lower bound under Assumption 9. We show a regret lower bound of $\tilde{\Omega}(T^{2/3})$ for this setting even for quadratic and smooth loss functions and $m = 2$. This result holds even if we allow some degree of improper learning.

Inspired by the work of Dekel et al. (2014), where they showed a $\tilde{\Omega}(T^{2/3})$-regret lower bound for multi-armed bandits with switching costs, even if the switching costs part of the loss functions are given to the learner through a full information feedback model. To establish a lower bound in our setting, we note that Sun et al. (2023) showed that for strongly convex quadratic functions it is possible to attain $\tilde{O}(\sqrt{T})$ regret upper bound. The main relaxation of assumption in our setting is that we remove the strong convexity assumption, thus allowing quadratic loss of the form $(x_t - x_{t-1})^2$, which has a non-invertible Hessian. This form of quadratic losses has an intuitive connection to the switching costs analyzed by Dekel et al. (2014).

However, directly reducing from the hard case constructed by Dekel et al. (2014) is insufficient: (1) we are interested in a continuous decision space, and the switching cost at 0 is non-smooth, and (2) if we are interested in quadratic switching costs, then the switching cost decays much faster around 0, making direct reductions from the hard instance for discrete set inferior to the lower bound of $T^{2/3}$. However, with a modified instance, we are able to recover the $T^{2/3}$ lower bound, even in the setting where we allow improper learning.

Theorem 6 (BCO-M lower bound) *There exists a sequence of loss functions such that the regret incurred by any (possibly randomized) online bandit quadratic optimization with memory length $m \geq 2$ algorithm $\mathcal{A}$ is at least $\tilde{\Omega}(T^{2/3})$.*

Proof idea. We briefly sketch the idea of the loss construction here. The loss function at time t consists of three parts: a scaled linear component with a random sign, a random process indexed at t, and a moving cost between two consecutive iterates. The best strategy depends on the random sign, which is chosen by the adversary ahead of time. For a sufficiently small scale on the linear component, the learner needs to move sufficiently to learn the random sign from bandit feedback and incur the corresponding moving cost, leading to the $\tilde{\Omega}(T^{2/3})$ lower bound.

A matching upper bound for quadratic and smooth functions in the presence of $\mathbb{F}_{t-m}$-adaptive adversary can also be established. We believe that the improper leaning algorithm in (Cassel and Koren, 2020) can also be generalized to scenarios where the adversary is allowed to be $\mathbb{F}_{t-m}$-adaptive. However, since this was not spelled out explicitly, we give an alternative first-order proper learning algorithm for completeness in Section I.

6. Conclusions, discussions, and future work

In this paper, we considered a bandit version of the Online Newton Step algorithm that attains near-optimal regret bounds for a large class of convex loss functions we call κ-convex with a low dimension dependence. In addition to the application to bandit convex optimization, we showed how our methods close an open problem in online control and separated the difficulty of bandit LQR/LQG and the general BCO-M problem. Many interesting open questions remain:

- **Proper BCO for κ-convex losses.** `BNS` is improper. This prohibits applications to problems where proper learning is required (e.g. portfolio selection). It is interesting to see whether there exists efficient proper learning algorithm that achieves a $O(\sqrt{T})$ regret with reasonable dependence on the dimension for the class of κ-convex problems.
- **Extension to nonstochastic control with general cost functions.** It is worth exploring if tighter bounds are achievable for bandit linear control with general convex loss functions with adversarial perturbations.
- **Lower bound for oblivious adversary.** If the adversary is oblivious in BCO-M problems, we are interested in whether $\Omega(T^{2/3})$ regret lower bound can be achieved.

Acknowledgments

EH and JS gratefully acknowledge funding from the National Science Foundation, the Office of Naval Research, and Open Philanthropy.

References

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. 2009.

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

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

Peter Arbenz and Gene H. Golub. On the spectral decomposition of hermitian matrices modified by low rank perturbations with applications. *SIAM Journal on Matrix Analysis and Applications*, 9 (1):40–58, 1988.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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

Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *Journal of the ACM (JACM)*, 68(4):1–35, 2021.

Brian Bullins, Elad Hazan, and Tomer Koren. The limits of learning with missing data. *Advances in Neural Information Processing Systems*, 29, 2016.

Asaf Cassel and Tomer Koren. Bandit linear control. *Advances in Neural Information Processing Systems*, 33:8872–8882, 2020.

Nicolo Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Efficient learning with partially observed attributes. *Journal of Machine Learning Research*, 12(10), 2011.

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

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

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467, 2014.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160. PMLR, 2019.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 546–580. PMLR, 2022.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference On Learning Theory*, pages 167–208. PMLR, 2018.

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

Ming Gu and Stanley C Eisenstat. A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem. *SIAM journal on Matrix Analysis and Applications*, 15(4):1266–1276, 1994.

Elad Hazan. *Introduction to online convex optimization*. MIT Press, 2022.

Elad Hazan and Satyen Kale. Newtron: an efficient bandit algorithm for online multiclass prediction. *Advances in neural information processing systems*, 24, 2011.

Elad Hazan and Tomer Koren. Linear regression with limited observation. *arXiv preprint arXiv:1206.4678*, 2012.

Elad Hazan and Kfir Levy. Bandit convex optimization: Towards tight bounds. *Advances in Neural Information Processing Systems*, 27, 2014.

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

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

Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209. PMLR, 2014.

Tomer Koren, Roi Livni, and Yishay Mansour. Multi-armed bandits with metric movement costs. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Tor Lattimore and Andras Gyorgy. Improved regret for zeroth-order stochastic convex bandits. In *Conference on Learning Theory*, pages 2938–2964. PMLR, 2021.

Tor Lattimore and András György. A second-order method for stochastic bandit convex optimisation. *arXiv preprint arXiv:2302.05371*, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

H Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004. Proceedings 17*, pages 109–123. Springer, 2004.

H Brendan McMahan and Matthew Streeter. Open problem: Better bounds for online logistic regression. In *Conference on Learning Theory*, pages 44–1. JMLR Workshop and Conference Proceedings, 2012.

Shai Shalev-Shwartz and Yoram Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69:115–142, 2007.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24. PMLR, 2013.

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

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

Arun Sai Suggala, Pradeep Ravikumar, and Praneeth Netrapalli. Efficient bandit convex optimization: Beyond linear losses. In *Conference on Learning Theory*, pages 4008–4067. PMLR, 2021.

Y Jennifer Sun, Stephen Newman, and Elad Hazan. Optimal rates for bandit nonstochastic control. *arXiv preprint arXiv:2305.15352*, 2023.

Joel Tropp. Freedman's inequality for matrix martingales. 2011.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Julian Zimmert and Tor Lattimore. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Conference on Learning Theory*, pages 3285–3312. PMLR, 2022.

Contents

Appendix A. Discussion on κ-convexity

In this section, we provide a more elaborated discussion on the curvature assumption of κ-convexity (Definition 1, Section 1.1). It is easy to see that κ-convexity is implied by linearity (by taking $H = 0$ in Definition 1) and by strong convexity together with smoothness (by taking $H = I$, and c, C to be the strong convexity and smoothness parameters).

A common relaxation of strong convexity used in the literature of Newton step based algorithms to obtain optimal rates is exp-concavity, which requires strong convexity in the direction of the gradient. The class of functions that satisfy exp-concavity is formally given by the following definition.

Definition 7 (Exp-concave functions, (Hazan, 2022)) *A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is α-exp-concave over a convex compact set $\mathcal{K} \subset \mathbb{R}^n$ if $g : \mathcal{K} \to \mathbb{R}$ given by $g(x) = e^{-\alpha f(x)}$ is concave. Equivalently, if f is twice differentiable, $\nabla^2 f(x) \succeq \alpha \nabla f(x) \nabla f(x)^\top$ holds for all $x \in \mathcal{K}$.*

The motivation for considering exp-concave functions goes beyond their appealing curvature properties, which enable optimal regret guarantees. Many popular losses that do not satisfy strong convexity are exp-concave, such as the squared loss used in linear regression, logistic loss for classification, and logarithmic loss for portfolio optimization.

Unfortunately, the conditions of exp-concavity and κ-convexity are not directly comparable: linear functions do not satisfy exp-concavity, while κ-convexity requires smoothness assumption and uniform bounds on the Hessian over $\mathcal{K}$. Luckily, κ-convexity does capture a rich class of functions whose curvature properties lie between convexity and strong convexity, encompassing the most popular exp-concave loss considered in practice. To see this, first we establish Observation 1, which states κ-convexity for the class of functions that can be written as a composition of an affine function and a strongly convex and smooth function.

Observation 1 *Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex compact set. For $n, d \in \mathbb{N}$, $\alpha, \beta > 0$, consider the class of functions composed of an affine function and a strongly convex and smooth function*

$$\mathcal{L}_{\alpha,\beta,n,d}(\mathcal{K}) = \{f : \mathbb{R}^n \to \mathbb{R} \,|\, f(x) = g \circ \ell(x),\ \ell \in \mathcal{C}^2 : \mathbb{R}^n \to \mathbb{R}^d,\ \nabla^2 \ell = 0,$$
$$g \in \mathcal{C}^2 : \mathbb{R}^d \to \mathbb{R},\ \alpha I_d \preceq \nabla^2 g(z) \preceq \beta I_d,\ \forall z \in \ell(\mathcal{K})\}. \tag{5}$$

$f \in \mathcal{L}_{\alpha,\beta,n,d}(\mathcal{K})$ is κ-convex on $\mathcal{K}$ with $\kappa = \frac{\beta}{\alpha}$ and $H = \nabla \ell \nabla \ell^\top$.

In Observation 2, we further note that many popular loss functions of interests are described by the function class in Observation 1.

Observation 2 *The following loss functions belong to $\mathcal{L}_{\alpha,\beta,n,d}(\mathcal{K})$ defined in Eq. (5) for any convex compact domain $\mathcal{K} \subset \mathbb{R}^n$ and bounded $\mathcal{X} \subset \mathbb{R}^n, \mathcal{Y} \subset \mathbb{R}$:*

- *(Squared loss in linear regression) $w \mapsto (y - w^\top x)^2$, $x \in \mathcal{X}, y \in \mathcal{Y}$.*
- *(Logistic loss in classification) $w \mapsto \log(1 + \exp(-y \cdot x^\top w))$, $x \in \mathcal{X}, y \in \mathcal{Y}$.*
- *(Logarithmic loss in portfolio optimization) $w \mapsto -\log(w^\top x)$, $x \in \mathcal{X}$.*

In Appendix B, we will further discuss the implications of our improved bandit algorithm on the three examples described in Observation 2, including comparisons with previous results in these problems.

Appendix B. Applications of Bandit Newton Step

In this section, we illustrate the applicability of our bandit Newton method (Algorithm 1) to a few problems where it achieves optimal regret guarantees. First, we acknowledge that since the algorithm in Bubeck et al. (2021) works for a richer class of loss functions than ours, applying their algorithm will also give the optimal regret bounds in terms of T in these applications. However, the high polynomial dependence on the dimension ($d^{9.5}$) and the $\mathrm{poly}(d, T)$ amortized running time make the algorithm by Bubeck et al. (2021) impractical in many real-world applications. Therefore, for simplicity, we are going to exclude the comparison with Bubeck et al. (2021) in this section.

B.1. Online Logistic Regression

Logistic regression is a widely utilized and effective technique for performing classification tasks. For simplicity, we consider binary classification. In this framework, the goal is to predict the probability that a given input point belongs to one of two possible classes. These classes are typically represented by the label set $\mathcal{Y} = \{-1, 1\}$. The foundational assumption of logistic regression is that the log-odds, or the logarithm of the odds ratio between the two classes, can be linearly modeled by the input features. Mathematically, this relationship is expressed as:

$$\log\left(\frac{\mathbb{P}(Y = 1 \mid X = x)}{1 - \mathbb{P}(Y = 1 \mid X = x)}\right) = w^\top x + b,$$

where x denotes the feature vector, w represents the weight vector, and b is the bias term. This formulation provides a linear decision boundary in the feature space.

In the online setting of logistic regression, the learning process is sequential and adaptive. At each time step t, an adversary presents a new instance, consisting of a feature vector and a label $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$. Subsequently, the learning algorithm selects a linear predictor w_t from a set $\mathcal{W} = \{w \in \mathbb{R}^d : \|w\|_2 \le D\}$, where $\|w\|_2 \le D$ ensures that the chosen predictor does not have excessively large weights, thereby controlling the model's complexity. The performance of the predictor w_t is then evaluated based on the incurred loss, which, for logistic regression, is defined as:

$$f_t(w_t) = f(w_t; x_t, y_t) = \log(1 + \exp(-y_t \cdot x_t^\top w_t)).$$

In the bandit framework, where the sequence of labeled data may be confidential and not directly observable by the learner, the learner is only exposed to a scalar classification cost, denoted as $f_t(w_t)$. It is often assumed for simplicity that $\|x_t\|_2 \le 1$, $\forall t$, in which case the Hessian of f_t satisfies

$$\frac{e^{-D}}{(1+e^{-D})^2} x_t x_t^\top \preceq \nabla^2 f_t(w_t) = \frac{\exp(-y_t \cdot x_t^\top w_t)}{(1+\exp(-y_t \cdot x_t^\top w_t))^2} \cdot x_t x_t^\top \preceq \frac{1}{2} x_t x_t^\top.$$

This shows that f_t is κ-convex with $\kappa = \exp(D)$, making it suitable to apply Algorithm 1.

Corollary 8 (Regret bound for bandit logistic regression) *Suppose a learner is playing over a decision set $\mathcal{W}$ with diameter bound D against an oblivious adversary picking a labeled vector $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$ with $\|x_t\|_2 \le 1$, then Algorithm 1 guarantees*

$$\mathbb{E}[\mathit{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^T f(w_t; x_t, y_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^T f(w; x_t, y_t)\right] \le O(d^{2.5} e^{2D} \sqrt{T}).$$

Let's take a moment to compare the result of Corollary 8 with existing results. To the best of our knowledge, this is the *first* bandit logistic regression algorithm that works against adversarially chosen labeled vectors that achieves an *optimal* regret bound in T and $d^{5/2}$ dependence on d. Table 4 summarized some of the state-of-art results in the field. First, let's focus on the logistic bandits setting. Hazan and Kale (2011) and Foster et al. (2018) both studied the problem which they refer to as bandit multi-class prediction, which we refer to as the semi-bandit feedback model in Table 4. In this setting, although the learner accesses loss through bandit feedback, the vector x_t is revealed to the learner. The more general case where both the vector and the label are given to the learner through bandit feedback was studied only under stochastic assumption on the input vector, in which case $\tilde{O}(e^D\sqrt{T})$ and $\tilde{O}(\sqrt{T})$ bounds were achieved by Abeille and Lazaric (2017) and Dong et al. (2019) for frequentist and Bayesian regret bounds, respectively.

One question remaining is regarding the exponential dependence on D in the bound of Corollary 8. To shed some light on this question, we consider the full information setting. Hazan et al. (2007) showed that running online newton step gives the optimal dependence ($\log T$) in T but suffers an exponential factor in terms of the diameter D, leading to the open problem of whether a $\tilde{O}(\text{poly}(D))$ regret bound is attainable in McMahan and Streeter (2012). Later, Hazan and Levy (2014) showed a negative result that for $d \geq 2$, the regret for this problem is at least $\tilde{\Omega}(e^D \vee \sqrt{DT})$ if the algorithm is proper, i.e. $w_t \in \mathcal{W}$. Foster et al. (2018) showed that when allowing improper learning, $\tilde{O}(1)$-regret bound is attainable. It will be an interesting open problem to investigate whether the exponential dependence on D can be dropped in logistic bandits.

Paper	Feedback	Advers.	Proper	Regret	Comp.	Note
Hazan et al. (2007)	full	✓	✓	$\tilde{O}(e^D)$	$O(d^2)$	
Hazan et al. (2014)	full	×	✓	$\tilde{\Omega}(e^D \vee \sqrt{DT})$	–	
Foster et al. (2018)	full	✓	×	$\tilde{O}(1)$	$\text{poly}(d,T)$	
Hazan and Kale (2011)	semi-bandit	✓	✓	$\tilde{O}(e^D \wedge DT^{2/3})$	$O(d^2)$	
Foster et al. (2018)	semi-bandit	✓	×	$\tilde{O}(e^D \wedge \sqrt{T})$	$\text{poly}(d,T)$	
Dong et al. (2019)	bandit	×	✓	$\tilde{O}(\sqrt{T})$	$\text{poly}(d)$	Bayesian
Faury et al. (2022)	bandit	×	✓	$\tilde{O}(e^D \vee \sqrt{T})$	$O(d^2)$	frequentist
Corollary 8	bandit	✓	×	$O(e^{2D}\sqrt{T})$	$O(d^2)$	

Table 4: Comparison with relevant prior works for online logistic regression. $\tilde{O}, \tilde{\Omega}$ in the regret column hide all parameters other than D, T and logarithmic factors in D, T.

B.2. Online Linear Regression with Limited Observations

Consider the standard online learning setting for linear regression problems. At each time t, the learner is supplied with a vector $x_t \in \mathcal{X} \subset \mathbb{R}^d$ and is asked to output a weight vector $w_t \in \mathcal{W} \subset \mathbb{R}^d$ such that the learner's label prediction for x_t is $\hat{y}_t = w_t^\top x_t \in \mathbb{R}$. After the learner outputs w_t, the adversary reveals the true label y_t, to which the learner suffers loss $f(w_t; x_t) = (\hat{y}_t - y_t)^2$.

Previous works such as (Cesa-Bianchi et al., 2011; Hazan and Koren, 2012; Bullins et al., 2016) have considered this framework in the setting of limited observations. In the limited observation setting, the learner observes only a subset of the attributes in x_t. We consider a more generalized version of the problem, where the learner does not necessarily observe (x_t, y_t) at all but simply a scalar loss of $f(w_t; x_t) \in \mathbb{R}_+$. Note that the loss is κ-convex with $\kappa = 1$.

Corollary 9 (Regret bound for bandit linear regression) *Suppose a learner is playing over a decision set $\mathcal{W}$ with diameter bound $D_{\mathcal{W}}$ against an oblivious adversary picking a labeled vector (x_t, y_t), then Algorithm 1 guarantees*

$$\mathbb{E}[\text{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} f(w_t; x_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^{T} f(w; x_t)\right] \leq O(d^{2.5}\sqrt{T}).$$

Appendix C. Online Newton Step in Full Information Setting

In this section, we provide the preliminaries on the online Newton step (ONS) algorithm in the full information setting. This will serve as a building block to derive the regret guarantee of our proposed Newton step based bandit algorithm (Algorithm 1).

Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex compact set. ONS chooses each iterate through Newton-step descent with a preconditioner A_t set to be the scaled cumulative Hessian of the loss functions received so far.

Algorithm 3 Online Newton Step (ONS)

Input: convex compact set $\mathcal{K} \subset \mathbb{R}^d$, step size $\eta > 0$, Hessian multiplier $\kappa' > 0$, time horizon $T \in \mathbb{N}$.

1: Initialize: $x_1 \in \mathcal{K}$, $A_0 = I$.
2: **for** $t = 1, \ldots, T$ **do**
3: Play x_t, observe f_t.
4: Update $A_t = A_{t-1} + \frac{\eta}{\kappa'}\nabla^2 f_t(x_t)$, and compute x_{t+1}

$$x_{t+1} = \prod_{\mathcal{K}}^{A_t}\left[x_t - \eta A_t^{-1}\nabla f_t(x_t)\right]$$

5: **end for**

The following theorem bounds the regret of ONS against any single $x \in \mathcal{K}$, as a function of the step size η, the diameter of $\mathcal{K}$, and the curvature parameter of the sequence of loss functions.

Theorem 10 (ONS Full Information Regret) *Suppose that the Online Newton Step (Algorithm 3) with input $(\mathcal{K}, \eta, \kappa', T)$ applied to a sequence of loss functions $\{f_t\}_{t=1}^T$ that are twice differentiable. Moreover, suppose $\{A_t\}_{t=1\ldots T}$, the cumulative Hessians, are invertible. Then, Algorithm 3 guarantees the following regret upper bound for any $x \in \mathcal{K}$,*

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x) \;\leq\; \frac{\text{diam}(\mathcal{K})^2}{2\eta} + \sum_{t=1}^{T} \frac{\eta}{2}\|\nabla f_t(x_t)\|_{A_t^{-1}}^2 - \sum_{t=1}^{T} \Delta_t(x; \kappa').$$

Here, $\Delta_t(x; \kappa')$ is defined as

$$\Delta_t(x; \kappa') \coloneqq f_t(x) - f_t(x_t) - \langle \nabla f_t(x_t), x - x_t\rangle - \frac{1}{2\kappa'}(x - x_t)^\top \nabla^2 f_t(x_t)(x - x_t).$$

Proof For any $x \in \mathcal{K}$, we have that

$$\begin{aligned}
\|x_{t+1} - x\|_{A_t}^2 &\overset{(a)}{\leq} \|x_t - \eta A_t^{-1}\nabla f_t(x_t) - x\|_{A_t}^2 \\
&= \|x_t - x\|_{A_t}^2 - 2\eta \left\langle A_t^{-1}\nabla f_t(x_t), x_t - x\right\rangle_{A_t} + \eta^2 \|A_t^{-1}\nabla f_t(x_t)\|_{A_t}^2 \\
&= \|x_t - x\|_{A_t}^2 - 2\eta \left\langle \nabla f_t(x_t), x - x_t\right\rangle + \eta^2 \|\nabla f_t(x_t)\|_{A_t^{-1}}^2 \\
&= \|x_t - x\|_{A_{t-1}}^2 + 2\eta \left(\frac{1}{2\kappa'}\|x_t - x\|_{\nabla^2 f_t(x_t)}^2 + \left\langle \nabla f_t(x_t), x - x_t\right\rangle\right) + \eta^2 \|\nabla f_t(x_t)\|_{A_t^{-1}}^2.
\end{aligned}$$

Where (a) is due to the projection property. Now, consider the second term in the right hand side of the last equality. By definition of $\Delta_t(x;\kappa')$, we can rewrite it as

$$\frac{1}{2\kappa'}\|x_t - x\|_{\nabla^2 f_t(x_t)}^2 + \left\langle \nabla f_t(x_t), x - x_t\right\rangle = f_t(x) - f_t(x_t) - \Delta_t(x;\kappa')$$

Substituting this in the previous display, and rearranging the terms gives us

$$f_t(x_t) - f_t(x) \;\leq\; \frac{\|x_t - x\|_{A_{t-1}}^2 - \|x_t - x\|_{A_t}^2}{2\eta} + \frac{\eta}{2}\|\nabla f_t(x_t)\|_{A_t^{-1}}^2 - \Delta_t(x;\kappa').$$

Summing this over $t = 1 \dots T$ gives us

$$\sum_{t=1}^T f_t(x_t) - f_t(x) \;\leq\; \frac{\|x_1 - x\|_2^2}{2\eta} + \sum_{t=1}^T \frac{\eta}{2}\|\nabla f_t(x_t)\|_{A_t^{-1}}^2 - \sum_{t=1}^T \Delta_t(x;\kappa').$$

We now use the fact that $\|x_1 - x\|_2 \;\leq\; \text{diam}(\mathcal{K})$ to obtain the desired bound. ■

Using the regret bound established in Theorem 10, we will derive the regret inequality for the bandit Newton based algorithm through a reduction from the full information setting.

Appendix D. Regret of Bandit Newton Method for Improper Learning (Section 2.1)

In this section, we first provide a result which converts any online second order algorithm in full information setting to one that uses stochastic gradients and Hessians estimators in the bandit setting. In Section D.3, we rely on this result to bound the regret of Algorithm 1.

We first begin by formally defining the class of second order online convex optimization (OCO) algorithms - the family of regret minimization algorithms for which this reduction works - in Definition 11.

Definition 11 (Second order OCO algorithm) *Let $\mathcal{A}$ be a deterministic online convex optimization algorithm on $\mathcal{K} \subset \mathbb{R}^d$ receiving an arbitrary sequence of $T \in \mathbb{N}$ twice differentiable loss functions $f_1, \dots, f_T : \mathbb{R}^d \to \mathbb{R}$ and producing decisions $x_1 \leftarrow \mathcal{A}(\emptyset), \dots, x_t \leftarrow \mathcal{A}(f_1, \dots, f_{t-1})$. $\mathcal{A}$ is called a second order online algorithm if the following holds:*

- *Let $\hat{f}_t : \mathbb{R}^d \to \mathbb{R}$ be the quadratic function defined as*

$$\hat{f}_t(x) = \frac{1}{2}x^\top \nabla^2 f_t(x_t)x + \nabla f_t(x_t)^\top x - x_t^\top \nabla^2 f_t(x_t)x.$$

Then $\forall t \in [T]$:

$$\mathcal{A}(f_1, \dots, f_{t-1}) = \mathcal{A}(\hat{f}_1, \dots, \hat{f}_{t-1}).$$

By Definition 11, it is clear that the iterates x_t of a second order OCO algorithm $\mathcal{A}$ is completely determined by the gradients and Hessians of the loss functions of the previous iterations at the corresponding decision points. Therefore, we can rewrite

$$x_t \leftarrow \mathcal{A}(\nabla f_1(x_1), \ldots, \nabla f_{t-1}(x_{t-1}), \nabla^2 f_1(x_1), \ldots, \nabla^2 f_{t-1}(x_{t-1})).$$

We consider a formal reduction from any second order OCO algorithm to an algorithm in the bandit setting.

Lemma 12 (Bandit Reduction) *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex compact set. Let $\mathcal{A}$ be a second order online algorithm (Definition 11) on $\mathcal{K}$ that ensures a regret bound with respect any $x \in \mathcal{K}$ of the following form for any sequence of twice differentiable loss functions $\{f_t\}_{t=1}^T$ over a time horizon $T \in \mathbb{N}$:*

$$\sum_{t=1}^T f_t(x_t) - f_t(x) \ \le \ B_{\mathcal{A},x}(f_1, \ldots f_T).$$

Define the points $\{x_t\}_{t=1}^T$ as: $x_1 \leftarrow \mathcal{A}(\emptyset)$, $x_t \leftarrow \mathcal{A}(\tilde{\nabla}_1, \ldots \tilde{\nabla}_{t-1}, \tilde{H}_1, \ldots \tilde{H}_{t-1})$, i.e. x_t is given by the output of $\mathcal{A}$ given gradients $\tilde{\nabla}_1, \ldots, \tilde{\nabla}_{t-1}$ and Hessians $\tilde{H}_1, \ldots, \tilde{H}_{t-1}$, where $\tilde{\nabla}_t, \tilde{H}_t$ are (conditionally) unbiased estimators of the gradient and Hessian of f_t at x_t, respectively, i.e.

$$\mathbb{E}[\tilde{\nabla}_t | x_1, f_1, \ldots x_t, f_t] = \nabla f_t(x_t),$$
$$\mathbb{E}[\tilde{H}_t | x_1, f_1, \ldots x_t, f_t] = \nabla^2 f_t(x_t).$$

Define the stochastic approximations $h_t : \mathcal{K} \to \mathbb{R}$ of $f_t : \mathcal{K} \to \mathbb{R}$ as follows:

$$h_t(x) = \left\langle \tilde{\nabla}_t, x \right\rangle + \frac{1}{2}(x - x_t)^\top \tilde{H}_t (x - x_t)$$

Then, the following holds for any $x \in \mathcal{K}$:

$$\sum_{t=1}^T \mathbb{E}\left[f_t(x_t) - f_t(x)\right] \ \le \ \mathbb{E}\left[B_{\mathcal{A},x}(h_1, \ldots h_T)\right] - \sum_{t=1}^T \mathbb{E}[\Delta_t(x)],$$

where $\Delta_t(x)$ is the error in second order Taylor series expansion of f_t which is defined as

$$\Delta_t(x) \coloneqq f_t(x) - f_t(x_t) - \langle \nabla f_t(x_t), x - x_t \rangle - \frac{1}{2}(x - x_t)^\top \mathbb{E}[\tilde{H}_t | x_1, f_1, \ldots x_t, f_t](x - x_t).$$

Proof Observe that from the definition of h_t, we have $\nabla h_t(x_t) = \tilde{\nabla}_t, \nabla^2 h_t(x_t) = \tilde{H}_t$. So, deterministically applying a second order algorithm $\mathcal{A}$ on h_t is equivalent to stochastically applying $\mathcal{A}$ on functions f_t. So by the regret assumption on $\mathcal{A}$, we have

$$\sum_{t=1}^T h_t(x_t) - h_t(x) \ \le \ B_{\mathcal{A},x}(h_1, \ldots h_T). \tag{6}$$

Next, note that

$$\mathbb{E}[h_t(x_t)] \overset{(a)}{=} \mathbb{E}\left[\left\langle \mathbb{E}[\tilde{\nabla}_t | x_1, f_1, \ldots x_t, f_t], x_t \right\rangle\right] = \mathbb{E}\left[\langle \nabla f_t(x_t), x_t \rangle\right],$$

where we used the fact that $\tilde{\nabla}_t$ is an unbiased estimate of the true gradient in (a). A similar argument shows that

$$\mathbb{E}[h_t(x)] = \mathbb{E}\left[\langle \nabla f_t(x_t), x\rangle\right] + \frac{1}{2}\mathbb{E}\left[(x - x_t)^\top \mathbb{E}[\tilde{H}_t | x_1, f_1, \dots x_t, f_t](x - x_t)\right].$$

So, we have

$$\begin{aligned}\mathbb{E}[h_t(x_t) - h_t(x)] &= \mathbb{E}\left[\langle \nabla f_t(x_t), x_t - x\rangle\right] - \frac{1}{2}\mathbb{E}\left[(x - x_t)^\top \mathbb{E}[\tilde{H}_t | x_1, f_1, \dots x_t, f_t](x - x_t)\right] \\ &= \mathbb{E}[f_t(x_t) - f_t(x) + \Delta_t(x)].\end{aligned}$$

The Lemma now follows by taking expectations in Equation (6). ■

Together with Theorem 10, we are almost ready to establish the regret guarantee for Algorithm 1. Note that Algorithm 1 uses unbiased Hessian estimators $\tilde{H}_t$ (Line 5) that is not necessarily positive semidefinite. However, we will show that the preconditioner $\tilde{A}_t$, the cumulative Hessian, concentrates well around its mean, making the operation in Line 6 of Algorithm 1 well-defined.

D.1. Concentration of Cumulative Hessian Estimate

The following lemma shows that the cumulative Hessian estimators concentrates around its mean with high probability.

Lemma 13 (Concentration of cumulative Hessian estimate) *Consider a sequence of functions $\{f_t\}_{t=1}^T$ that satisfies Assumption 1, Assumption2, and Assumption 3. $\forall t \in [T]$, define the smoothed function $\tilde{f}_t^{\mathbb{B},\mathbb{B}}$ of f_t as*

$$\tilde{f}_t^{\mathbb{B},\mathbb{B}}(x) := \mathbb{E}_{u\sim\mathbb{B}, v\sim\mathbb{B}}\left[f_t\left(x + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(u+v)\right)\Big|\mathcal{F}_{t-1}\right],$$

where $\mathcal{F}_t$ denotes the filtration generated by the algorithm's possible randomness up to time t. Let $\tilde{A}_t = \tilde{A}_{t-1} + \frac{\eta}{\kappa'}\tilde{H}_t$, and $A_t = A_{t-1} + \frac{\eta}{\kappa'}\nabla^2 \tilde{f}_t^{\mathbb{B},\mathbb{B}}(x_t)$, where $A_0 = \tilde{A}_0 = I$. Let $B^* = B + \sqrt{2}(L + \sqrt{2}C)$. Suppose $T = \tilde{\Omega}(d)$, and $\eta \leq \kappa'(24d^{3/2}B^*\kappa\sqrt{T\log(dT^2)})^{-1}$, $\forall t$, with probability at least $1 - \frac{t}{T^2}$, for every $s \leq t$,

$$\|I - A_s^{-\frac{1}{2}}\tilde{A}_s A_s^{-\frac{1}{2}}\|_2 \leq \frac{1}{2}.$$

Proof First observe that using Stokes' theorem (Flaxman et al., 2004), we have

$$\mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}] = \nabla^2 \tilde{f}_t^{\mathbb{B},\mathbb{B}}(x_t)$$

where $\mathcal{F}_t$ denotes the filtration generated by $\{v_{s,1}, v_{s,2}\}_{s=1}^t$.

Proof by induction. Base case $t = 0$ is given by construction. Suppose the inequalities hold with probability at least $1 - \frac{t-1}{T^2}$ for any $0 \le s \le t-1$. Recall that f_t is assumed to be bounded by B and L-Lipschitz over $\mathcal{K}$ and C-smooth. Note that $|f_t(y_t)|$ is bounded by

$$\begin{aligned}|f_t(y_t)| &\le |f_t(x_t)| + \left|f_t\left(x_t + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right) - f_t(x_t)\right| \\ &\le B + \left\|\nabla f_t\left(x_t + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right)\right\|_2 \left\|\frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right\|_2 \\ &\le B + \left(L + C\left\|\frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right\|_2\right)\left\|\frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right\|_2 \\ &\le B + (L + C\|\tilde{A}_{t-1}^{-\frac{1}{2}}\|_2)\|\tilde{A}_{t-1}^{-\frac{1}{2}}\|_2.\end{aligned}$$

By induction hypothesis, with probability at least $1-\frac{t-1}{T^2}$, $\tilde{A}_{t-1} \succeq \frac{1}{2}A_{t-1} = \frac{1}{2}(I+\eta/\kappa'\sum_{s=1}^{t-1}\mathbb{E}_{s-1}[\tilde{A}_s]) \succeq \frac{1}{2}I$, and thus the above inequality implies $|f_t(y_t)| \le B + \sqrt{2}\left(L + \sqrt{2}C\right) =: B^*$. Next, define L_t and U_t as follows

$$L_t \coloneqq I + \frac{c\eta}{\kappa'}\sum_{s=1}^{t} H_s, \quad U_t \coloneqq I + \frac{C\eta}{\kappa'}\sum_{s=1}^{t} H_s.$$

Observe that from our definition of κ-convex losses, A_t can be lower and upper bounded as follows

$$L_t \preceq A_t \preceq U_t.$$

Now consider the following

$$\begin{aligned}\|I - A_t^{-\frac{1}{2}}\tilde{A}_t A_t^{-\frac{1}{2}}\|_2 &= \|A_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)A_t^{-\frac{1}{2}}\|_2 \\ &\le \|A_t^{-\frac{1}{2}}L_t^{1/2}\|_2^2\|L_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)L_t^{-\frac{1}{2}}\|_2 \\ &\overset{(a)}{\le} \|L_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)L_t^{-\frac{1}{2}}\|_2,\end{aligned}$$

where (a) follows from the fact that $L_t \preceq A_t$. This shows that it suffices to bound $\|L_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)L_t^{-\frac{1}{2}}\|_2$. Next, consider the following

$$L_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)L_t^{-\frac{1}{2}} = \frac{\eta}{\kappa'}\sum_{s=1}^{t} L_t^{-\frac{1}{2}}(\mathbb{E}[\tilde{H}_s \mid \mathcal{F}_{s-1}] - \tilde{H}_s)L_t^{-\frac{1}{2}},$$

where $Z_s = L_t^{-\frac{1}{2}}(\mathbb{E}[\tilde{H}_s \mid \mathcal{F}_{s-1}] - \tilde{H}_s)L_t^{-\frac{1}{2}}$ forms a martingale sequence with respect to the filtration $\mathcal{F}_s$. We rely on matrix Freedman inequality to bound $\|\sum_{s=1}^t Z_s\|$ (Tropp, 2011). To do this, we derive bounds for the first and second moments of Z_s.

Bounding Z_s. We first show that Z_s is a bounded random variable. To see this, note that by definition of $\tilde{H}_t$, and the fact that $-2I \preceq v_{t,1}v_{t,2}^\top + v_{t,2}v_{t,1}^\top \preceq 2I$, we have

$$\begin{aligned}&-4d^2B^*\tilde{A}_{s-1} \preceq \tilde{H}_s \preceq 4d^2B^*\tilde{A}_{s-1}, \\ &-4d^2B^*\tilde{A}_{s-1} \preceq \mathbb{E}[\tilde{H}_s \mid \mathcal{F}_{s-1}] \preceq 4d^2B^*\tilde{A}_{s-1}.\end{aligned}$$

Thus, with probability at least $1 - \frac{t-1}{T^2}$, $\forall s \leq t$,

$$\begin{aligned} Z_s &\preceq 8d^2 B^* L_t^{-\frac{1}{2}} \tilde{A}_{s-1} L_t^{-\frac{1}{2}} \\ &\preceq 8d^2 B^* L_{s-1}^{-\frac{1}{2}} \tilde{A}_{s-1} L_{s-1}^{-\frac{1}{2}} \\ &\overset{(a)}{\preceq} 8d^2 B^* \kappa A_{s-1}^{-\frac{1}{2}} \tilde{A}_{s-1} A_{s-1}^{-\frac{1}{2}} \\ &\overset{(b)}{\preceq} 12 d^2 B^* \kappa I, \end{aligned}$$

and

$$\begin{aligned} Z_s &\succeq -4d^2 B^* L_t^{-\frac{1}{2}} \tilde{A}_{s-1} L_t^{-\frac{1}{2}} \\ &\succeq -4d^2 B^* L_{s-1}^{-\frac{1}{2}} \tilde{A}_{s-1} L_{s-1}^{-\frac{1}{2}} \\ &\overset{(c)}{\succeq} -4d^2 B^* \kappa A_{s-1}^{-\frac{1}{2}} \tilde{A}_{s-1} A_{s-1}^{-\frac{1}{2}} \\ &\overset{(d)}{\succeq} -6 d^2 B^* \kappa I, \end{aligned}$$

where $(a), (c)$ follow from the definition of L_t, U_t and the fact that the losses are κ-convex, and $(b), (d)$ follow from the induction hypothesis that $\|I - A_s^{-\frac{1}{2}} \tilde{A}_s A_s^{-\frac{1}{2}}\|_2 \leq \frac{1}{2}$ and $\tilde{A}_s$ is PSD , $\forall s \leq t-1$ w.h.p.

Bounding 2^{nd} moments of Z_s. Next, we bound the second moments of Z_s

$$\mathbb{E}[Z_s^2 | \mathcal{F}_{s-1}] \overset{(a)}{\preceq} \mathbb{E}[L_t^{-\frac{1}{2}} \tilde{H}_s L_t^{-1} \tilde{H}_s L_t^{-\frac{1}{2}} | \mathcal{F}_{s-1}],$$

where (a) follows from the fact that for any random matrix X: $\mathbb{E}[(X - \mathbb{E}[X]) X - \mathbb{E}[X])^\top] \preceq \mathbb{E}[XX^\top]$. Continuing

$$\begin{aligned} &\mathbb{E}[L_t^{-\frac{1}{2}} \tilde{H}_s L_t^{-1} \tilde{H}_s L_t^{-\frac{1}{2}} | \mathcal{F}_{s-1}] \\ &= L_t^{-1/2} \tilde{A}_{s-1}^{1/2} \mathbb{E}[4d^4 f_s(y_s)^2 (v_{s,1} v_{s,2}^\top + v_{s,2} v_{s,1}^\top) \tilde{A}_{s-1}^{1/2} L_t^{-1} \tilde{A}_{s-1}^{1/2} (v_{s,1} v_{s,2}^\top + v_{s,2} v_{s,1}^\top) | \mathcal{F}_{s-1}] \tilde{A}_{s-1}^{1/2} L_t^{-1/2} \\ &\overset{(a)}{\preceq} L_t^{-1/2} \tilde{A}_{s-1}^{1/2} \mathbb{E}[6d^4 f_s(y_s)^2 (v_{s,1} v_{s,2}^\top + v_{s,2} v_{s,1}^\top)(v_{s,1} v_{s,2}^\top + v_{s,2} v_{s,1}^\top) | \mathcal{F}_{s-1}] \tilde{A}_{s-1}^{1/2} L_t^{-1/2}, \\ &\overset{(b)}{\preceq} 6d^4 (B^*)^2 \kappa L_t^{-1/2} \tilde{A}_{s-1}^{1/2} \mathbb{E}[(v_{s,1} v_{s,2}^\top + v_{s,2} v_{s,1}^\top)(v_{s,1} v_{s,2}^\top + v_{s,2} v_{s,1}^\top) | \mathcal{F}_{s-1}] \tilde{A}_{s-1}^{1/2} L_t^{-1/2}, \end{aligned}$$

where (a) follows from the fact that $\tilde{A}_{s-1}^{1/2} L_t^{-1} \tilde{A}_{s-1}^{1/2} \preceq 1.5\kappa I$ w.h.p, and (b) follows from the fact that $f_s(y_s) \leq B^*$. Next, we rely on the facts that $\mathbb{E}[v_{s,1} v_{s,2}^\top v_{s,1} v_{s,2}^\top] = d^{-2} I$, and $\mathbb{E}[v_{s,1} v_{s,2}^\top v_{s,2} v_{s,1}^\top] = d^{-1} I$ to obtain

$$\begin{aligned} &\mathbb{E}[L_t^{-\frac{1}{2}} \tilde{H}_s L_t^{-1} \tilde{H}_s L_t^{-\frac{1}{2}} | \mathcal{F}_{s-1}] \\ &\preceq 12 d^3 (B^*)^2 (1 + d^{-1}) \kappa L_t^{-1/2} \tilde{A}_{s-1} L_t^{-1/2} \\ &\overset{(a)}{\preceq} 36 d^3 (B^*)^2 \kappa^2, \end{aligned}$$

where (a) follows from the fact that $L_t^{-1/2} \tilde{A}_{s-1} L_t^{-1/2} \preceq 1.5\kappa I$ w.h.p.

Matrix Freedman. With the above bounds on the martingale sequence, we can bound $L_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)L_t^{-\frac{1}{2}}$ by matrix Freedman inequality (Lemma 14) [3]

$$\begin{aligned} \|I - \tilde{A}_t^{-\frac{1}{2}} A_t \tilde{A}_t^{-\frac{1}{2}}\|_2 &\leq \|L_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)L_t^{-\frac{1}{2}}\|_2 \\ &\leq \frac{12\eta\kappa d^{3/2} B^* \sqrt{t\log(dT^2)}}{\kappa'} + \frac{8\eta\kappa d^2 B^* \log(dT^2)}{\kappa'}. \end{aligned}$$

The RHS of the above inequality is less than $1/2$ for our choice of η. ■

Lemma 14 (Matrix Freedman, Theorem 1.2 in (Tropp, 2011)) *Consider a self-adjoint matrix martingale $\{Y_k\}_{k\geq 0}$ of dimension d. Let $\{X_k\}_{k\geq 1}$ be the difference sequence of $\{Y_k\}_{k\geq 0}$. Assume that the difference sequence is uniformly bounded almost surely: $\exists R$ such that*

$$\lambda_{\max}(X_k) \leq R \quad \textit{almost surely} \quad \forall k.$$

Define the predictable quadratic variation process

$$W_k = \sum_{j=1}^{k} \mathbb{E}[X_j^2 \mid \mathbb{F}_{j-1}], \quad \forall k.$$

Then $\forall \varepsilon \geq 0, \sigma^2 > 0$,

$$\mathbb{P}\left(\exists k \geq 0 : \lambda_{\max}(Y_k) \geq \varepsilon \textit{ and } \|W_k\|_2 \leq \sigma^2\right) \leq d \cdot \exp\left(-\frac{\varepsilon^2/2}{\sigma^2 + R\varepsilon/3}\right).$$

Corollary 15 *Let $B^* = B + \sqrt{2}(L + \sqrt{2}C)$. With probability at least $1 - \frac{1}{T}$, the y_t played by Algorithm 1 satisfies*

$$|f_t(y_t)| \leq B^*, \; \forall t \in [T].$$

Proof Note that in the proof of Lemma 13, we have shown that $|f_t(y_t)| \leq B^*$ if $\|I - A_{t-1}^{-\frac{1}{2}} \tilde{A}_{t-1} A_{t-1}^{-\frac{1}{2}}\|_2 \leq \frac{1}{2}$. Thus,

$$\mathbb{P}\left(|f_t(y_t)| \leq B^*, \; \forall t \in [T]\right) \geq \mathbb{P}\left(\|I - A_t^{-\frac{1}{2}} \tilde{A}_t A_t^{-\frac{1}{2}}\|_2 \leq \frac{1}{2}, \; \forall t \in [T]\right) \geq 1 - \frac{1}{T}.$$

■

3. Freedman's inequality requires the first and second moment bounds to hold almost surely. Observe that we only have these bounds in high probability. However, there is a standard workaround to this (see for instance Suggala et al. (2021)), where one can create an alternate martingale that satisfies these bounds a.s., and is exactly equal to the original martingale w.h.p.

D.2. Concentration of Cumulative Hessian Estimate [Improved κ dependence]

In this section, we provide an alternate analysis for Hessian concentration that improves the dependence of regret on κ, but at the cost of worse dependence on d. In particular, this analysis leads to a regret of $\tilde{O}(d^3\kappa B^*\sqrt{T})$, instead of $\tilde{O}(d^{2.5}\kappa^2 B^*\sqrt{T})$ stated in Theorem 3. This analysis relies on empirical Freedman's inequality (Zimmert and Lattimore, 2022) instead of the matrix Freedman used in the previous section.

Lemma 16 (Concentration of cumulative Hessian estimate) *Consider a sequence of functions $\{f_t\}_{t=1}^T$ that satisfies Assumption 1, Assumption 2, and Assumption 3. $\forall t \in [T]$, define the smoothed function $\tilde{f}_t^{\mathbb{B},\mathbb{B}}$ of f_t as*

$$\tilde{f}_t^{\mathbb{B},\mathbb{B}}(x) \coloneqq \mathbb{E}_{u\sim\mathbb{B},v\sim\mathbb{B}}\left[f_t\left(x+\frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(u+v)\right)\Big|\mathcal{F}_{t-1}\right],$$

where $\mathcal{F}_t$ denotes the filtration generated by the algorithm's possible randomness up to time t. Let $\tilde{A}_t = \tilde{A}_{t-1} + \frac{\eta}{\kappa'}\tilde{H}_t$, and $A_t = A_{t-1} + \frac{\eta}{\kappa'}\nabla^2\tilde{f}_t^{\mathbb{B},\mathbb{B}}(x_t)$, where $A_0 = \tilde{A}_0 = I$. Let $B^* = B+\sqrt{2}(L+\sqrt{2}C)$. Suppose $T = \Omega(d^2)$, and $\eta \le \kappa'\left(200d^2\sqrt{T\log dTB^*}\right)^{-1}$, $\forall t$, with probability at least $1-\frac{t}{T^2}$, for every $s \le t$,

$$\|I - A_s^{-\frac{1}{2}}\tilde{A}_s A_s^{-\frac{1}{2}}\|_2 \le \frac{1}{2}.$$

Proof Similar to the proof of Lemma 13, we rely on induction to prove the result.

Proof by induction. Base case $t = 0$ is given by construction. Suppose the inequalities hold with probability at least $1-\frac{t-1}{T^2}$ for any $0 \le s \le t-1$. Recall that f_t is assumed to be bounded by B and L-Lipschitz over $\mathcal{K}$ and C-smooth. Note that $|f_t(y_t)|$ is bounded by

$$\begin{aligned}
|f_t(y_t)| &\le |f_t(x_t)| + \left|f_t\left(x_t + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right) - f_t(x_t)\right| \\
&\le B + \left\|\nabla f_t\left(x_t + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right)\right\|_2 \left\|\frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right\|_2 \\
&\le B + \left(L + C\left\|\frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right\|_2\right)\left\|\frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(v_{t,1}+v_{t,2})\right\|_2 \\
&\le B + (L + C\|\tilde{A}_{t-1}^{-\frac{1}{2}}\|_2)\|\tilde{A}_{t-1}^{-\frac{1}{2}}\|_2.
\end{aligned}$$

By induction hypothesis, with probability at least $1-\frac{t-1}{T^2}$, $\tilde{A}_{t-1} \succeq \frac{1}{2}A_{t-1} \succeq \frac{1}{2}I$, and thus the above inequality implies $|f_t(y_t)| \le B + \sqrt{2}\left(L+\sqrt{2}C\right) =: B^*$.

Next, consider the following

$$\tilde{A}_t = \tilde{A}_{t-1} + \frac{\eta}{\kappa'}\tilde{H}_t = \tilde{A}_{t-1}^{1/2}\left(I + \frac{2d^2 f_t(y_t)\eta}{\kappa'}(v_{t,1}v_{t,2}^\top + v_{t,2}v_{t,1}^\top)\right)\tilde{A}_{t-1}^{1/2}$$

For our choice of η, it is easy to verify that $\frac{1}{2} \preceq \left(I + \frac{2d^2 f_t(y_t)\eta}{\kappa'}(v_{t,1}v_{t,2}^\top + v_{t,2}v_{t,1}^\top)\right) \preceq \frac{3}{2}$. So $\frac{1}{2}\tilde{A}_{t-1} \preceq \tilde{A}_t \preceq \frac{3}{2}\tilde{A}_{t-1}$. This shows that $\tilde{A}_t$ is invertible, and is very close to $\tilde{A}_{t-1}$, with high probability. We use this inequality often in the subsequent analysis. We now show that the following holds with probability at least $1-\delta$

$$\sum_{s=1}^{t}\mathbb{E}_{s-1}[\tilde{H}_s] - gA_{t-1} \preceq \sum_{s=1}^{t}\tilde{H}_s \preceq \sum_{s=1}^{t}\mathbb{E}_{s-1}[\tilde{H}_s] + gA_{t-1},$$

where $g = \left(96d^2\sqrt{t\log\frac{dTB^*}{\delta}} + 32d^3\log\frac{dTB^*}{\delta}\right)B^*$. Let $\mathcal{C}_\epsilon$ be a finite cover of $\mathbb{S} = \{x \in \mathbb{R}^d | \|x\|_2 = 1\}$ such that for any $v \in \mathbb{S}$ there exists a $u \in \mathcal{C}_\epsilon$ such that $\|v-u\|_2 \leq \epsilon$. By Vershynin (2018), there exists a $\mathcal{C}_\epsilon$ such that $|\mathcal{C}_\epsilon| \leq \left(\frac{2}{\epsilon}+1\right)^d$. Consider any $u \in \mathcal{C}_\epsilon$, and the corresponding sequence of random variables $\{u^T\tilde{H}_s u\}_{s=1}^t$. We use Freedman's inequality to bound $|\sum_{s=1}^t u^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])u|$. To do this, we derive bounds for the first and second moments of $u^T\tilde{H}_s u$

Bounding 1^{st} moment of $u^T\tilde{H}_s u$. From the definition of $\tilde{H}_s$, and the fact that $-2I \preceq v_{t,1}v_{t,2}^\top + v_{t,2}v_{t,1}^\top \preceq 2I$, we have the following, which holds with high probability

$$|u^T\tilde{H}_s u| \leq 4d^2B^*\|u\|^2_{\tilde{A}_{s-1}} \overset{(a)}{\leq} 8d^2B^*\|u\|^2_{A_{s-1}} \overset{(b)}{<} \infty.$$

Here, inequality (a) follows from the fact that $\|I - A_s^{-\frac{1}{2}}\tilde{A}_sA_s^{-\frac{1}{2}}\|_2 \leq \frac{1}{2}$, $\forall s \leq t-1$ w.h.p. And inequality (b) follows from the fact that the loss functions f_t are smooth.

Bounding 2^{nd} moment of $u^T\tilde{H}_s u$. Next, we bound the second moments of $u^T\tilde{H}_s u$

$$\begin{aligned}
&\mathbb{E}[(u^T\tilde{H}_s u)^2|\mathcal{F}_{s-1}] = \mathbb{E}[u^T\tilde{H}_s uu^T\tilde{H}_s u|\mathcal{F}_{s-1}]\\
&= u^T\tilde{A}_{s-1}^{1/2}\mathbb{E}[4d^4f_s(y_s)^2(v_{s,1}v_{s,2}^\top + v_{s,2}v_{s,1}^\top)\tilde{A}_{s-1}^{1/2}uu^T\tilde{A}_{s-1}^{1/2}(v_{s,1}v_{s,2}^\top + v_{s,2}v_{s,1}^\top)|\mathcal{F}_{s-1}]\tilde{A}_{s-1}^{1/2}u\\
&\overset{(a)}{\leq} \|u\|^2_{\tilde{A}_{s-1}}u^T\tilde{A}_{s-1}^{1/2}\mathbb{E}[4d^4f_s(y_s)^2(v_{s,1}v_{s,2}^\top + v_{s,2}v_{s,1}^\top)(v_{s,1}v_{s,2}^\top + v_{s,2}v_{s,1}^\top)|\mathcal{F}_{s-1}]\tilde{A}_{s-1}^{1/2}u,\\
&\overset{(b)}{\leq} 4d^4(B^*)^2\|u\|^2_{\tilde{A}_{s-1}}u^T\tilde{A}_{s-1}^{1/2}\mathbb{E}[(v_{s,1}v_{s,2}^\top + v_{s,2}v_{s,1}^\top)(v_{s,1}v_{s,2}^\top + v_{s,2}v_{s,1}^\top)|\mathcal{F}_{s-1}]\tilde{A}_{s-1}^{1/2}u,
\end{aligned}$$

where (a) follows from the fact $\tilde{A}_{s-1}^{1/2}uu^T\tilde{A}_{s-1}^{1/2} \preceq \|u\|^2_{\tilde{A}_{s-1}}I$, and (b) follows from the fact that $f_s(y_s) \leq B^*$ w.h.p. Next, we rely on the facts that $\mathbb{E}[v_{s,1}v_{s,2}^\top v_{s,1}v_{s,2}^\top] = d^{-2}I$, and $\mathbb{E}[v_{s,1}v_{s,2}^\top v_{s,2}v_{s,1}^\top] = d^{-1}I$ to obtain

$$\mathbb{E}[(u^T\tilde{H}_s u)^2|\mathcal{F}_{s-1}] \leq 8d^3(1+d^{-1})(B^*)^2\|u\|^4_{\tilde{A}_{s-1}} \leq 16d^3(B^*)^2\|u\|^4_{\tilde{A}_{s-1}} \leq 64d^3(B^*)^2\|u\|^4_{A_{s-1}}.$$

Empirical Freedman's Inequality. Applying the empirical Freedman's inequality from Lemma 17 on $\{u^T\tilde{H}_s u\}_{s=1}^t$ gives us the following bound, which holds with probability at least $1-\delta$

$$\Big|\sum_{s=1}^{t}u^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])u\Big| \leq \left(48d^{3/2}\sqrt{t\log\frac{dTB^*}{\delta}} + 16d^2\log\frac{dTB^*}{\delta}\right)B^*\|u\|^2_{A_{t-1}}.$$

Taking a union bound over all $u \in \mathcal{C}_\epsilon$, we get the following which holds with probability at least $1-\delta$, for any $u \in \mathcal{C}_\epsilon$

$$\Big|\sum_{s=1}^{t}u^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])u\Big| \leq \left(48d^{3/2}\sqrt{t\log\frac{dTB^*|\mathcal{C}_\epsilon|}{\delta}} + 16d^2\log\frac{dTB^*|\mathcal{C}_\epsilon|}{\delta}\right)B^*\|u\|^2_{A_{t-1}}.$$

We now show that the above bound also holds for any $u \in \mathbb{S}$, for appropriate choice of ϵ. Consider any $u \in \mathbb{S}$ and let v be the closest point to u that lies in $\mathcal{C}_\epsilon$.

$$\begin{aligned}
&\sum_{s=1}^{t} \left[u^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])u - v^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])v\right] \\
&\quad = \sum_{s=1}^{t}(v-u)^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])(v-u) + \sum_{s=1}^{t} 2(u-v)^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])v \\
&\quad \leq \sum_{s=1}^{t}\left|(v-u)^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])(v-u)\right| + \sum_{s=1}^{t} 2\left|(u-v)^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])v\right|.
\end{aligned}$$

Using similar arguments as used for bounding the first moment of $u^T\tilde{H}_s u$, where we showed that $|u^T\tilde{H}_s u| \leq 8d^2B^*\|u\|_{A_{t-1}}^2$, we get

$$\begin{aligned}
&\sum_{s=1}^{t} \left[u^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])u - v^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])v\right] \\
&\quad \leq 16td^2B^*\|v-u\|_{A_{t-1}}^2 + 32td^2B^*\|v-u\|_{A_{t-1}}\|v\|_{A_{t-1}} \\
&\quad \overset{(a)}{=} O\left(t^{3/2}d^2B^*\epsilon^2 + t^{3/2}d^2B^*\epsilon\right),
\end{aligned}$$

where (a) follows from the fact that $\|A_{t-1}\|_2 = O\left(1 + \frac{\eta t}{\kappa'}\right)$. Choosing $\epsilon = \frac{g}{t^2d^2}$, for some appropriate constant g, we have the following which holds with probability at least $1-\delta$

$$\forall u \in \mathbb{S}, \quad \Big|\sum_{s=1}^{t} u^T(\tilde{H}_s - \mathbb{E}_{s-1}[\tilde{H}_s])u\Big| \leq \left(96d^2\sqrt{t\log\frac{dTB^*}{\delta}} + 32d^3\log\frac{dTB^*}{\delta}\right)B^*\|u\|_{A_{t-1}}^2.$$

This shows that, with probability at least $1-\delta$

$$\sum_{s=1}^{t}\mathbb{E}_{s-1}[\tilde{H}_s] - gA_{t-1} \preceq \sum_{s=1}^{t}\tilde{H}_s \preceq \sum_{s=1}^{t}\mathbb{E}_{s-1}[\tilde{H}_s] + gA_{t-1},$$

where $g = \left(96d^2\sqrt{t\log\frac{dTB^*}{\delta}} + 32d^3\log\frac{dTB^*}{\delta}\right)B^*$. Rewriting this equation gives us

$$A_t - \frac{\eta g}{\kappa'}A_{t-1} \preceq \tilde{A}_t \preceq A_t + \frac{\eta g}{\kappa'}A_{t-1},$$

Since $A_{t-1} \preceq A_t$, we get

$$\left(1 - \frac{\eta g}{\kappa'}\right)A_t \preceq \tilde{A}_t \preceq \left(1 + \frac{\eta g}{\kappa'}\right)A_t.$$

For our choice of $\eta \leq \kappa'\left(200d^2\sqrt{T\log\frac{dTB^*}{\delta}}\right)^{-1}, T = \Omega(d^2)$, we get

$$\frac{1}{2}A_t \preceq \tilde{A}_t \preceq \frac{3}{2}A_t.$$

This finishes the proof of the Lemma.

■

Lemma 17 (Strengthened Freedman's inequality (Zimmert and Lattimore, 2022)) *Let $\{X_t\}_{t=1,2\dots}$ be a martingale difference sequence w.r.t a filtration $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ such that $\mathbb{E}[X_t|\mathcal{F}_t] = 0$ and assume $\mathbb{E}[|X_t||\mathcal{F}_t] < \infty$ a.s. Then with probability at least $1-\delta$*

$$\Big|\sum_{t=1}^T X_t\Big| \le 3\sqrt{V_T \log\left(\frac{2\max\{U_T, \sqrt{V_T}\}}{\delta}\right)} + 2U_T \log\left(\frac{2\max\{U_T, \sqrt{V_T}\}}{\delta}\right),$$

where $V_T = \sum_{t=1}^T \mathbb{E}_{t-1}[X_t^2], U_T = \max\{1, \max_{t\in[T]} X_t\}$.

D.3. Proof of Theorem 3

Let $\mathbb{B} = \{x : \|x\|_2 \le 1\}, \mathbb{S} = \{x : \|x\|_2 = 1\}$ be the unit ball and unit sphere in $\mathbb{R}^d$. We define smoothed functions $\tilde{f}_t^{\mathbb{B},\mathbb{B}}, \tilde{f}_t^{\mathbb{B},\mathbb{S}}, \tilde{f}_t^{\mathbb{S},\mathbb{S}}$ as follows

$$\begin{aligned}
\tilde{f}_t^{\mathbb{B},\mathbb{B}}(x) &:= \mathbb{E}_{u\sim\mathbb{B}, v\sim\mathbb{B}}\left[f_t\left(x + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(u+v)\right)\Big|\mathcal{F}_{t-1}\right] \\
\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x) &:= \mathbb{E}_{u\sim\mathbb{B}, v\sim\mathbb{S}}\left[f_t\left(x + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(u+v)\right)\Big|\mathcal{F}_{t-1}\right] \\
\tilde{f}_t^{\mathbb{S},\mathbb{S}}(x) &:= \mathbb{E}_{u\sim\mathbb{S}, v\sim\mathbb{S}}\left[f_t\left(x + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(u+v)\right)\Big|\mathcal{F}_{t-1}\right].
\end{aligned}$$

Observe that

$$\mathbb{E}[f_t(y_t) \mid \mathcal{F}_{t-1}] = \tilde{f}_t^{\mathbb{S},\mathbb{S}}(x_t),$$

where $\mathcal{F}_t$ denotes the filtration generated by $\{v_{s,1}, v_{s,2}\}_{s=1}^t$. By Stokes' theorem (Flaxman et al., 2004), we have that the gradient and Hessian estimators constructed in Algorithm 1 satisfy

$$\mathbb{E}[\tilde{\nabla}_t \mid \mathcal{F}_{t-1}] = \nabla\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t), \quad \mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}] = \nabla^2\tilde{f}_t^{\mathbb{B},\mathbb{B}}(x_t).$$

Observe that the Hessian and gradients estimated in Algorithm 1 are not of the same function. Interestingly, despite this mismatch, we can derive $\sqrt{T}$ regret of the algorithm. Define $A_t := A_{t-1} + \frac{\eta}{\kappa'}\nabla^2\tilde{f}_t^{\mathbb{B},\mathbb{B}}(x_t)$, with $A_0 = I$. Throughout the proof, we assume that the cumulative Hessian concentrates well; that is, the following holds

$$\|A_t^{-\frac{1}{2}}(A_t - \tilde{A}_t)A_t^{-\frac{1}{2}}\|_2 \le \frac{1}{2}.$$

This assumption is formally proved in Lemma 13, which says that the above stated inequality holds simultaneously for all $t \in [T]$ with probability at least $1 - \frac{1}{T}$. Therefore, we can without loss of generality assume that the above inequality holds deterministically by suffering an additional constant in the regret bound.

For any $x \in \mathcal{K}$, we can decompose the regret as the following:

$$\begin{aligned}
&\sum_{s=1}^{T} \mathbb{E}[f_s(y_s) - f_s(x)] \\
&= \sum_{s=1}^{T} \mathbb{E}[\tilde{f}_s^{\mathbb{S},\mathbb{S}}(x_s) - f_s(x)] \\
&= \sum_{s=1}^{T} \mathbb{E}[\tilde{f}_s^{\mathbb{B},\mathbb{S}}(x_s) - f_s(x)] + \mathbb{E}[\tilde{f}_s^{\mathbb{S},\mathbb{S}}(x_s) - \tilde{f}_s^{\mathbb{B},\mathbb{S}}(x_s)] \\
&= \underbrace{\sum_{s=1}^{T} \mathbb{E}[\tilde{f}_s^{\mathbb{B},\mathbb{S}}(x_s) - \tilde{f}_s^{\mathbb{B},\mathbb{S}}(x)]}_{T_1} + \underbrace{\sum_{s=1}^{T} \mathbb{E}[\tilde{f}_s^{\mathbb{B},\mathbb{S}}(x) - f_s(x)]}_{T_2} + \underbrace{\sum_{s=1}^{T} \mathbb{E}[\tilde{f}_s^{\mathbb{S},\mathbb{S}}(x_s) - \tilde{f}_s^{\mathbb{B},\mathbb{S}}(x_s)]}_{T_3},
\end{aligned}$$

where we will bound T_1, T_2, and T_3 separately.

Bounding T_1. To bound this term, we rely on Lemma 12. In particular, we instantiate it with the ONS algorithm described in Algorithm 3, and use $\tilde{\Delta}_t, \tilde{H}_t$ as the stochastic estimates of $\nabla \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t), \nabla^2 \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t)$. This gives us the following bound

$$T_1 \;\leq\; \frac{\text{diam}(\mathcal{K})^2}{2\eta} + \sum_{t=1}^{T} \frac{\eta}{2} \mathbb{E}\left[\|\tilde{\nabla}_t\|^2_{\tilde{A}_t^{-1}}\right] - \sum_{t=1}^{T} \mathbb{E}\left[\Delta_t(x)\right],$$

where $\Delta_t(x)$ is defined as

$$\Delta_t(x) \coloneqq \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x) - \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t) - \left\langle \nabla \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t), x - x_t \right\rangle - \frac{1}{2\kappa'}(x - x_t)^\top \mathbb{E}[\tilde{H}_t | \mathcal{F}_{t-1}](x - x_t).$$

For our choice of $\kappa' (\geq \kappa)$, and our assumption on the Hessian of f_t, $\Delta_t(x)$ is always greater than 0. This is because

$$\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x) - \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t) - \left\langle \nabla \tilde{f}_t^{\mathbb{B},\mathbb{S}}(x_t), x - x_t \right\rangle \geq \frac{c}{2}\|x - x_t\|^2_{H_t} \geq \frac{1}{2\kappa'}(x - x_t)^\top \mathbb{E}[\tilde{H}_t | \mathcal{F}_{t-1}](x - x_t).$$

Next, we bound $\|\tilde{\nabla}_t\|_{\tilde{A}_t^{-1}}$:

$$\begin{aligned}
\|\tilde{\nabla}_t\|^2_{\tilde{A}_t^{-1}} &= 2d^2 f_t(y_t)^2 v_{t,1}^\top \tilde{A}_{t-1}^{\frac{1}{2}} \tilde{A}_t^{-1} \tilde{A}_{t-1}^{\frac{1}{2}} v_{t,1} \\
&\overset{(a)}{\leq} 6d^2 f_t(y_t)^2 v_{t,1}^\top A_{t-1}^{\frac{1}{2}} A_t^{-1} A_{t-1}^{\frac{1}{2}} v_{t,1} \\
&\overset{(b)}{\leq} 6d^2 (B^*)^2,
\end{aligned}$$

where (a) follows from the fact that $\frac{1}{2}A_t \preceq \tilde{A}_t \preceq \frac{3}{2}A_t$, and (b) follows from Corollary 15. Substituting this in the above upper bound for T_1 gives us

$$T_1 \;\leq\; \frac{\text{diam}(\mathcal{K})^2}{2\eta} + 3\eta d^2 (B^*)^2 T.$$

Bounding T_2. We now upper bound T_2 by $\tilde{O}(\frac{1}{\eta})$. To see this, consider the following

$$
\begin{aligned}
T_2 &= \sum_{t=1}^{T} \mathbb{E}[\tilde{f}_t^{\mathbb{B},\mathbb{S}}(x) - f_t(x)] \\
&= \sum_{t=1}^{T} \mathbb{E}_{u_t\sim\mathbb{B}, v_t\sim\mathbb{S}} \left[f_t\left(x + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(u_t+v_t) \right) - f_t(x) \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{u_t\sim\mathbb{B}, v_t\sim\mathbb{S}} \left[\frac{1}{2}\nabla f_t(x)^\top \tilde{A}_{t-1}^{-\frac{1}{2}}(u_t+v_t) + \frac{1}{8}(u_t+v_t)^\top \tilde{A}_{t-1}^{-\frac{1}{2}} \nabla^2 f_t(x(u_t,v_t)) \tilde{A}_{t-1}^{-\frac{1}{2}}(u_t+v_t) \right] \\
&= \frac{1}{8}\sum_{t=1}^{T} \mathbb{E}_{u_t\sim\mathbb{B}, v_t\sim\mathbb{S}}[(u_t+v_t)^T \tilde{A}_{t-1}^{-\frac{1}{2}} \nabla^2 f_t(x(u_t,v_t)) \tilde{A}_{t-1}^{-\frac{1}{2}}(u_t+v_t)],
\end{aligned}
$$

where the last equality follows from the fact that conditioning on $\mathcal{F}_{t-1}$, the first-order term vanishes. Here, $x(u_t, v_t)$ is a point on the line connecting x and $x + \frac{1}{2}\tilde{A}_{t-1}^{-\frac{1}{2}}(u_t+v_t)$. Next, observe that $\nabla^2 f_t(x(u_t,v_t)) \preceq CH_t \preceq \frac{C}{c}\mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}] \preceq \kappa\mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}]$. Substituting this in the previous display gives us

$$
T_2 \;\le\; \frac{\kappa}{8}\sum_{t=1}^{T} \mathbb{E}[(u_t+v_t)^T \tilde{A}_{t-1}^{-\frac{1}{2}} \mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}] \tilde{A}_{t-1}^{-\frac{1}{2}}(u_t+v_t)].
$$

Next, since by concentration of Hessian estimate accumulates, we have $\tilde{A}_t \succeq \frac{1}{2}A_t$ and the fact that

$$
A_{t-1} = A_t - \frac{\eta}{\kappa'}\mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}] \succeq A_t - c\eta I \succeq \frac{1}{2}A_t,
$$

we bound T_2 as

$$
\begin{aligned}
T_2 \;&\le\; \frac{\kappa}{4}\sum_{t=1}^{T} \mathbb{E}[(u_t+v_t)^\top A_{t-1}^{-\frac{1}{2}} \mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}] A_{t-1}^{-\frac{1}{2}}(u_t+v_t)] \\
&\le\; \frac{\kappa}{2}\sum_{t=1}^{T} \mathbb{E}[(u_t+v_t)^\top A_{t}^{-\frac{1}{2}} \mathbb{E}[\tilde{H}_t \mid \mathcal{F}_{t-1}] A_{t}^{-\frac{1}{2}}(u_t+v_t)].
\end{aligned}
$$

Continuing, we have with $|\cdot|$ denoting the determinant of a square matrix,

$$
\begin{aligned}
T_2 &\le \frac{\kappa\kappa'}{2\eta}\sum_{t=1}^{T} \mathbb{E}[(u_t+v_t)^\top A_t^{-\frac{1}{2}}(A_t - A_{t-1})A_t^{-\frac{1}{2}}(u_t+v_t)] \\
&= \frac{\kappa\kappa'}{\eta}\sum_{s=1}^{T} \mathbb{E}[\mathrm{tr}(A_t^{-\frac{1}{2}}(A_t - A_{t-1})A_t^{-\frac{1}{2}})] \\
&\overset{(c)}{\le} \frac{\kappa\kappa'}{\eta}\mathbb{E}\left[\log\frac{|A_T|}{|A_0|}\right] \\
&\overset{(d)}{\le} \frac{\kappa\kappa' d\log(1+\eta CT/\kappa')}{\eta},
\end{aligned}
$$

where inequality (c) follows from the fact that for $A, B \succeq 0$,

$$A^{-1} \cdot (A - B) \leq \log \frac{|A|}{|B|},$$

and inequality (d) follows from the fact that the since the largest eigenvalue $\lambda_{\max}(A_T)$ of A_T satisfies $\lambda_{\max}(A_T) \leq 1 + \eta CT/\kappa'$,

$$\log |A_T| \leq \log \left(\lambda_{\max}(A_T)^d \right) \leq d \log(1 + \eta CT/\kappa').$$

Bounding T_3.

$$\begin{aligned} T_3 &= \sum_{s=1}^{T} \mathbb{E}[\tilde{f}_s^{\mathbb{S},\mathbb{S}}(x_s) - \tilde{f}_s^{\mathbb{B},\mathbb{S}}(x_s)] \\ &\overset{(a)}{\leq} \sum_{s=1}^{T} \mathbb{E}[\tilde{f}_s^{\mathbb{S},\mathbb{S}}(x_s) - f_s(x_s)], \end{aligned}$$

where (a) follows from the convexity of f_s:

$$\begin{aligned} \tilde{f}_s^{\mathbb{B},\mathbb{S}}(x_s) &= \mathbb{E}_{u_s \sim \mathbb{B}, v_s \sim \mathbb{S}} \left[f_s \left(x_s + \frac{1}{2} \tilde{A}_{t-1}^{-\frac{1}{2}} (u_s + v_s) \right) \right] \\ &\geq f_s \left(x_s + \frac{1}{2} \mathbb{E}_{u_s \sim \mathbb{B}, v_s \sim \mathbb{S}} \left[\tilde{A}_{t-1}^{-\frac{1}{2}} (u_s + v_s) \right] \right) \\ &= f_s(x_s) \end{aligned}$$

Using similar arguments as T_2 to bound $\mathbb{E}[\tilde{f}_s^{\mathbb{S},\mathbb{S}}(x_s) - f_s(x_s)]$, we obtain

$$T_3 \;\leq\; \frac{\kappa\kappa' d \log(1 + \eta CT/\kappa')}{\eta}.$$

Combining the bounds for T_1, T_2, T_3 gives us the required regret bound.

Appendix E. Regret of BNS−AM for BQO-AM problems (Section 4.2)

In this section, we prove Theorem 4. We first make two observations on BNS−AM. BNS−AM does improper learning, as the decisions y_t's do not necessarily lie within $\mathcal{K}$ (Line 10 in Algorithm 2). We bound the function value and gradient evaluated at y_t's in Remark 18. The dependence of the iterates are explained in Remark 19.

Remark 18 (Value and gradient bound of y_t's) *Since Algorithm 2 does improper learning, it is essential to show that the loss and gradient at each of the y_t played by the algorithm is bounded. Note that $\forall t$,*

$$\begin{aligned} \|\nabla f_t(y_{t-m+1:t})\|_2 &\leq \|\nabla f_t(x_{t-m+1:t})\|_2 + \beta \|y_{t-m+1:t} - x_{t-m+1:t}\|_2 \leq L + \beta\sqrt{m}, \\ |f_t(y_{t-m+1:t})| &\leq |f_t(x_{t-m+1:t})| + |\nabla f_t(y_{t-m+1:t})^\top (y_{t-m+1:t} - x_{t-m+1:t})| \\ &\leq B + (L + \beta\sqrt{m})\sqrt{m}. \end{aligned}$$

We denote $B^* = B + (L + \beta\sqrt{m})\sqrt{m}$.

Remark 19 (Filtration and independence from delay) *Denote $\mathbb{F}_t = \sigma(\{u_s\}_{s\leq t})$ to be the filtration generated by Algorithm 2 random sampling step. Then by Assumption 9, $f_t, H_t, \hat{A}_t$ are $\mathbb{F}_{t-m}$-measurable. By the delayed updates in Algorithm 2, x_t is $\mathbb{F}_{t-m}$-measurable. $u_t, y_t, \tilde{g}_t$ are $\mathbb{F}_t$-measurable.*

We now prove Theorem 4. Following usual procedure of bounding with-memory regret in bandit setting, we note that the regret can be decomposed into three terms, which we will bound separately:

$$\mathbb{E}[\text{Regret}_T(x)] = \underbrace{\mathbb{E}\left[\sum_{t=m}^{T} f_t(y_{t-m+1:t}) - f_t(x_{t-m+1:t})\right]}_{\text{(perturbation loss)}} + \underbrace{\mathbb{E}\left[\sum_{t=m}^{T} f_t(x_{t-m+1:t}) - \bar{f}_t(x_t)\right]}_{\text{(movement cost)}} + \underbrace{\mathbb{E}\left[\sum_{t=m}^{T} \bar{f}_t(x_t) - \bar{f}_t(x)\right]}_{\text{(underlying regret)}}.$$

We will decompose the proof as the following: First, we will show some useful properties of the gradient estimator $\tilde{g}_t$. Then, we will bound the three terms above separately.

E.1. Properties of the gradient estimator

The gradient estimator constructed in Line 7, Algorithm 2 is a conditionally unbiased estimator of the divergence of the loss function f_t evaluated at $(x_{t-m+1}, \ldots, x_t)$. Together with the smoothness assumption on f_t and stability of the algorithm, this implies that the bias incurred by using $\tilde{g}_t$ as an estimator of $\nabla \bar{f}_t(x_t)$ is small.

Lemma 20 (Gradient estimator) *$\forall t \geq m$, the conditional expectation of the gradient estimator $\tilde{g}_t$ constructed in Line 7, Algorithm 2 is given by*

$$\mathbb{E}[\tilde{g}_t \mid \mathbb{F}_{t-m}] = \sum_{i=1}^{m} [\nabla f_t(x_{t-m+1:t})]_i,$$

where $[\nabla f_t(x_{t-m+1:t})]_i$ denotes the i-th d-vector in the dm-vector $\nabla f_t(x_{t-m+1:t})$.

Proof The proof follows similarly to the proof of Lemma C.3 in (Sun et al., 2023). In particular, we note that for a real-valued quadratic function on $\mathbb{R}^n$ $q(x) = \frac{1}{2}x^\top Ax + b^\top x + c$, a symmetric positive-definite matrix $M \in \mathbb{R}^{d\times d}$, a filtration $\mathbb{F}$, and a random unit vector $u \in \mathbb{R}^n$ satisfying (1) u is symmetric with zero first and third moments conditioning on $\mathbb{F}$, (2) u satisfies $\mathbb{E}[uu^\top \mid \mathbb{F}] = \frac{r}{n} I_{n\times n}$, and (3) A, b, c, x are $\mathbb{F}$-measurable, then

$$\begin{aligned}
\mathbb{E}\left[q(x+Mu)M^{-1}u \mid \mathbb{F}\right] &= \frac{1}{2}\mathbb{E}\left[(x+Mu)^\top A(x+Mu)M^{-1}u \mid \mathbb{F}\right] + \mathbb{E}\left[b^\top (x+Mu)M^{-1}u \mid \mathbb{F}\right] \\
&\quad + \mathbb{E}[cM^{-1}u \mid \mathbb{F}] \\
&= \frac{1}{2}\mathbb{E}[x^\top AMuM^{-1}u \mid \mathbb{F}] + \frac{1}{2}\mathbb{E}[u^\top MAxM^{-1}u \mid \mathbb{F}] + \mathbb{E}[b^\top MuM^{-1}u \mid \mathbb{F}] \\
&= \frac{1}{2}M^{-1}\mathbb{E}[uu^\top \mid \mathbb{F}]MA^\top x + \frac{1}{2}M^{-1}\mathbb{E}[uu^\top \mid \mathbb{F}]MAx + M^{-1}\mathbb{E}[uu^\top \mid \mathbb{F}]Mb \\
&= \frac{r}{2n}(A+A^\top)x + \frac{r}{n}b \\
&= \frac{r}{n}\nabla q(x).
\end{aligned}$$

Consider the following matrix $M_t \in \mathbb{R}^{dm\times dm}$:

$$M_t = \begin{bmatrix} \hat{A}_{t-m}^{-\frac{1}{2}} & 0 & \dots & 0 \\ 0 & \hat{A}_{t-m+1}^{-\frac{1}{2}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{A}_{t-1}^{-\frac{1}{2}} \end{bmatrix}.$$

By Remark 19, f_t, M_t is $\mathbb{F}_{t-m}$-measurable, and the concatenated vector $v_t = [u_{t-m+1}, \dots u_t]$ satisfies that the first and third moments of v_t is 0 conditioning on $\mathbb{F}_{t-m}$, and since $u_{t-m+1}, \dots, u_t$ are independent conditioning on $\mathbb{F}_{t-m}$,

$$\mathbb{E}[v_t(v_t)^\top \mid \mathbb{F}_{t-m}] = \frac{1}{d}I_{dm\times dm}.$$

By taking $r = m$, we have that

$$\nabla f_t(x_{t-m+1:t}) = \mathbb{E}[df_t(y_{t-m+1:t})M_t^{-1}v_t \mid \mathbb{F}_{t-m}].$$

We note that by definition of the gradient estimator $\tilde{g}_t$,

$$\begin{aligned}
\sum_{i=1}^{m}[\nabla f_t(x_{t-m+1:t})]_i &= \mathbb{E}\left[df_t(y_{t-m+1:t})\sum_{i=1}^{m}[M_t^{-1}v_t]_i \mid \mathbb{F}_{t-m}\right] \\
&= \mathbb{E}\left[df_t(y_{t-m+1:t})\sum_{s=t-m+1}^{t}\hat{A}_s^{\frac{1}{2}}u_s \mid \mathbb{F}_{t-m}\right] \\
&= \mathbb{E}\left[\tilde{g}_t \mid \mathbb{F}_{t-m}\right].
\end{aligned}$$

■

E.2. Bounding perturbation loss

We start with the perturbation loss and prove the following proposition:

Proposition 21 *The perturbation loss is bounded by*

$$\mathbb{E}\left[\sum_{t=m}^{T} f_t(y_{t-m+1:t}) - f_t(x_{t-m+1:t})\right] \le \beta R_G \left[\frac{10mR_G R_Y}{\kappa(G)}\sqrt{T} + \frac{4md\log(\sigma T)}{\eta\alpha\kappa(G)} + dR_Y^2\sqrt{T}\right].$$

Proof Note that by the quadratic assumption on f_t, we have by second-order Taylor expansion:

$$\sum_{t=m}^{T}\mathbb{E}[f_t(y_{t-m+1:t}) - f_t(x_{t-m+1:t})] = \sum_{t=m}^{T}\mathbb{E}[\nabla f_t(x_{t-m+1:t})^\top \hat{A}_{t-m:t-1}^{-\frac{1}{2}} u_{t-m+1:t}] + \frac{1}{2}\sum_{t=m}^{T}\mathbb{E}[u_{t-m+1:t}^\top \hat{A}_{t-m:t-1}^{-\frac{1}{2}}\nabla^2 f_t \hat{A}_{t-m:t-1}^{-\frac{1}{2}} u_{t-m:1:t}].$$

The first order term is 0 since $\forall t,\ f_t, x_{t-m+1}, \dots, x_t, \hat{A}_{t-m}, \dots, \hat{A}_{t-1}$ are $\mathbb{F}_{t-m}$-measurable, and thus

$$\mathbb{E}[\nabla f_t(x_{t-m+1:t})^\top \hat{A}_{t-m:t-1}^{-\frac{1}{2}} u_{t-m+1:t}] = \mathbb{E}[\nabla f_t(x_{t-m+1:t})^\top \hat{A}_{t-m:t-1}^{-\frac{1}{2}}\mathbb{E}[u_{t-m+1:t} \mid \mathbb{F}_{t-m}]] = 0.$$

The second term can be bounded as following. Note that by denoting $[\nabla^2 f_t]_{ij}$ as the ij-th $d\times d$ block of $\nabla^2 f_t$, we have

$$\begin{aligned}
&u_{t-m+1:t}^\top \hat{A}_{t-m:t-1}^{-\frac{1}{2}}\nabla^2 f_t \hat{A}_{t-m:t-1}^{-\frac{1}{2}} u_{t-m+1:t}\\
&= [\hat{A}_{t-m}^{-\frac{1}{2}} u_{t-m+1}, \dots, \hat{A}_{t-1}^{-\frac{1}{2}} u_t]^\top \nabla^2 f_t [\hat{A}_{t-m}^{-\frac{1}{2}} u_{t-m+1}, \dots, \hat{A}_{t-1}^{-\frac{1}{2}} u_t]\\
&= \sum_{i,j=1}^{m} u_{t-m+i}^\top \hat{A}_{t-m+i-1}^{-\frac{1}{2}}[\nabla^2 f_t]_{ij}\hat{A}_{t-m+j-1}^{-\frac{1}{2}} u_{t-m+j}.
\end{aligned}$$

By independence between u_s, u_t for $s \neq t$ and taking the expectation, we have

$$\begin{aligned}
&\mathbb{E}[u_{t-m+1:t}^\top \hat{A}_{t-m:t-1}^{-\frac{1}{2}}\nabla^2 f_t \hat{A}_{t-m:t-1}^{-\frac{1}{2}} u_{t-m:1:t}]\\
&= \sum_{i=1}^{m}\mathbb{E}\left[u_{t-m+i}^\top \hat{A}_{t-m+i-1}^{-\frac{1}{2}}[\nabla^2 f_t]_{ii}\hat{A}_{t-m+i-1}^{-\frac{1}{2}} u_{t-m+i}\right]\\
&\le \sum_{i=1}^{m}\mathbb{E}\left[\operatorname{tr}(\hat{A}_{t-m+i-1}^{-\frac{1}{2}}[\nabla^2 f_t]_{ii}\hat{A}_{t-m+i-1}^{-\frac{1}{2}})\right]\\
&\le \mathbb{E}\left[\operatorname{tr}\left(\hat{A}_{t-m}^{-\frac{1}{2}}\sum_{i=1}^{m}[\nabla^2 f_t]_{ii}\hat{A}_{t-m}^{-\frac{1}{2}}\right)\right],
\end{aligned}$$

where the last steps follow from that $\hat{A}_s \preceq \hat{A}_t$, $\forall s \le t$, and $\operatorname{tr}(\cdot)$ is linear. Note that since f_t is quadratic, $[\nabla^2 f_t]_{ii} = \nabla^2_{y_{t-m+i}} f_t$, where $\nabla^2_{y_{t-m+i}} f_t$ is the Hessian of f_t w.r.t. y_{t-m+i}. $\forall 1 \le i \le m-1$, by smoothness assumption on Q_t,

$$[\nabla^2 f_t]_{ii} = \nabla^2_{y_{t-m+i}} f_t = Y_{t-m+i}^\top (G^{[m-i]})^\top Q_t G^{[m-i]} Y_{t-m+i} \preceq \beta\tilde{R}_G Y_{t-m+i}^\top Y_{t-m+i}. \tag{7}$$

Denote $\sigma = \max_{t\in[T]} \lambda_{\max}(H_t) \leq R_G^2 R_Y^2$. For simplicity, assume that $\sigma \geq 1$ and $\eta\alpha\sigma m \leq 2$, which will be satisfied by our choice of η and m for large enough T. Summing over all iterations, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{t=m}^{T} \mathbb{E}[u_{t-m+1:t}^\top \hat{A}_{t-m:t-1}^{-\frac{1}{2}} \nabla^2 f_t \hat{A}_{t-m:t-1}^{-\frac{1}{2}} u_{t-m:1:t}] \\
&\leq \frac{1}{2}\sum_{t=m}^{T} \mathbb{E}\left[\operatorname{tr}\left(\hat{A}_{t-m}^{-\frac{1}{2}} \sum_{i=1}^{m} [\nabla^2 f_t]_{ii} \hat{A}_{t-m}^{-\frac{1}{2}}\right)\right] \\
&\leq \frac{\beta \tilde{R}_G}{2}\sum_{t=m}^{T} \mathbb{E}\left[\operatorname{tr}\left(\hat{A}_{t-m}^{-\frac{1}{2}}\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\hat{A}_{t-m}^{-\frac{1}{2}}\right)\right] \\
&\leq \beta \tilde{R}_G \sum_{t=m}^{T} \mathbb{E}\left[\operatorname{tr}\left(\hat{A}_{t}^{-\frac{1}{2}}\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\hat{A}_{t}^{-\frac{1}{2}}\right)\right],
\end{aligned}$$

where th second inequality follows from Eq. (7), and the third inequality follows from that since $\hat{A}_t \succeq mI$, $\forall t$, we have that $\forall s < t$,

$$\hat{A}_t = \hat{A}_s + \frac{\eta\alpha}{2}\sum_{r=s+1}^{t} H_r \preceq \hat{A}_s + \frac{\eta\alpha\sigma m(t-s)}{2} I \preceq \left(1 + \frac{\eta\alpha\sigma(t-s)}{2}\right)\hat{A}_s, \tag{8}$$

which implies that assuming $\eta\alpha\sigma m \leq 2$, we have $\hat{A}_t \preceq \max\{2, \eta\alpha\sigma m\}\hat{A}_{t-m} \preceq 2\hat{A}_{t-m}$.

Here, we use a similar "blocking" technique used in Simchowitz (2020) to bound this term. In particular, for some $\tau \in \mathbb{Z}_{++}$ to be determined later (we will take $\tau = \lfloor\sqrt{T}\rfloor$), consider endpoints $k_j = \tau(j-1) + m$, $j = 1, \ldots, J$, where $J = \lfloor\frac{T-m}{\tau}\rfloor$. Then, we can rewrite the sum on the right hand side as

$$\begin{aligned}
\sum_{t=m}^{T} \operatorname{tr}\left(\hat{A}_t^{-\frac{1}{2}}\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\hat{A}_t^{-\frac{1}{2}}\right) &= \sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1} \operatorname{tr}\left(\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\cdot \hat{A}_t^{-1}\right) \\
&\quad + \sum_{t=J\tau+m}^{T} \operatorname{tr}\left(\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\cdot \hat{A}_t^{-1}\right) \\
&\leq_{(1)} \sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1} \operatorname{tr}\left(\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\cdot \hat{A}_t^{-1}\right) \\
&\quad + \tau d R_Y^2,
\end{aligned}$$

where (1) follows from that $T - m - J\tau < \tau$ and since $\hat{A}_t \succeq mI$ and $\operatorname{tr}(A) \leq d\|A\|_2$ for $A \in \mathbb{R}^{d\times d}$ such that $A \succeq 0$ is symmetric,

$$\operatorname{tr}\left(\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\cdot \hat{A}_t^{-1}\right) \leq \frac{d}{m}\sum_{s=t-m+1}^{t} \|Y_s^\top Y_s\|_2 \leq d R_Y^2.$$

To bound the first term on the right hand side, since $\hat{A}_t \succeq \hat{A}_{k_j}$ for $t \geq k_j$, we have

$$\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1} \operatorname{tr}\left(\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\cdot \hat{A}_t^{-1}\right) \leq \frac{4m}{\kappa(G)} \underbrace{\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\left(\frac{\kappa(G)}{4m}\operatorname{tr}\left(\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\cdot \hat{A}_{k_j}^{-1}\right)}_{(a)},$$

To bound (a), we further decompose (a) as

$$(a) = \underbrace{\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1} \operatorname{tr}\left(\left(-\frac{H_t}{2} + \frac{\kappa(G)}{4m}\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right)\cdot \hat{A}_{k_j}^{-1}\right)}_{(i)} + \underbrace{\frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}\left(\left(\sum_{t=k_j}^{k_{j+1}-1} H_t\right)\cdot \hat{A}_{k_j}^{-1}\right)}_{(ii)}.$$

To see the bound on the first term, we use Proposition 4.8 from Simchowitz (2020). We include the proof of this result in Section E.6.1 for completion.

Lemma 22 (Proposition 4.8 in Simchowitz (2020)) *$\forall Y_1, \ldots, Y_T$, we have that:*

$$\sum_{t=m}^{T} H_t \succeq \frac{\kappa(G)}{2}\sum_{t=1}^{T} Y_t^\top Y_t - 5mR_G R_Y I.$$

If we let $\tilde{Y}_t = Y_{t+k_j-m}$, and $\tilde{H}_t = H_{t+k_j-m}$, then Lemma 22 implies that

$$\begin{aligned}
\sum_{t=k_j}^{k_{j+1}-1}\left(\frac{\kappa(G)}{4m}\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right) &\preceq \frac{\kappa(G)}{4}\sum_{t=k_j-m+1}^{k_{j+1}-1} Y_t^\top Y_t \\
&\preceq \frac{\kappa(G)}{4}\sum_{t=1}^{\tau+m-1} \tilde{Y}_t^\top \tilde{Y}_t \\
&\preceq \sum_{t=m}^{\tau+m-1} \frac{\tilde{H}_t}{2} + \frac{5mR_G R_Y}{2} I \\
&= \sum_{t=k_j}^{k_{j+1}-1} \frac{H_t}{2} + \frac{5mR_G R_Y}{2} I.
\end{aligned}$$

Thus, (i) is bounded by

$$\begin{aligned}
(i) &= \sum_{j=1}^{J} \underbrace{\left[\left(\sum_{t=k_j}^{k_{j+1}-1} -\frac{H_t}{2} + \frac{\kappa(G)}{4m}\sum_{s=t-m+1}^{t} Y_s^\top Y_s\right) - \frac{5mR_G R_Y}{2} I\right]}_{\preceq 0} \cdot \hat{A}_{k_j}^{-1} + \frac{5mR_G R_Y}{2}\sum_{j=1}^{J}\operatorname{tr}(\hat{A}_{k_j}^{-1}) \\
&\leq \frac{5dR_G R_Y}{2}\left\lfloor \frac{T}{\tau} \right\rfloor.
\end{aligned}$$

Since by Eq. 8, $\hat{A}_{k_{j+1}-1} \preceq \max\{2, \eta\alpha\sigma\tau\}\hat{A}_{k_j}$, (ii) is bounded by

$$\begin{aligned}
(ii) &= \frac{1}{2}\sum_{j=1}^{J}\left(\sum_{t=k_j}^{k_{j+1}-1} H_t\right)\cdot \hat{A}_{k_j}^{-1} \\
&= \frac{1}{2\eta\alpha}\sum_{j=1}^{J}\left(\hat{A}_{k_{j+1}-1} - \hat{A}_{k_j-1}\right)\cdot \hat{A}_{k_j}^{-1} \\
&\leq \frac{\max\{2,\eta\alpha\sigma\tau\}}{2\eta\alpha}\sum_{j=1}^{J}\left(\hat{A}_{k_{j+1}-1} - \hat{A}_{k_j-1}\right)\cdot \hat{A}_{k_{j+1}-1}^{-1} \\
&\leq \frac{\max\{2,\eta\alpha\sigma\tau\}}{2\eta\alpha}\sum_{j=1}^{J}\log\left(\frac{|\hat{A}_{k_{j+1}-1}|}{|\hat{A}_{k_j-1}|}\right) \\
&\leq \frac{\max\{2,\eta\alpha\sigma\tau\}}{2\eta\alpha}\log(|\hat{A}_T|) \\
&\leq_{(2)} \frac{\max\{2,\eta\alpha\sigma\tau\}}{2\eta\alpha} d\log(\sigma T),
\end{aligned}$$

where (2) follows from $|\hat{A}_T| \leq \|\hat{A}_T\|_2^d$.

Combining, we have

$$\text{(perturbation loss)} \leq \beta R_G\left[\frac{4m}{\kappa(G)}\left(\frac{5dR_GR_Y}{2}\left\lfloor\frac{T}{\tau}\right\rfloor + \frac{\max\{2,\eta\alpha\sigma\tau\}}{2\eta\alpha}d\log(\sigma T)\right) + \tau d R_Y^2\right],$$

by setting $\tau = \lfloor\sqrt{T}\rfloor$ and assuming $\eta \leq \frac{2}{\alpha\sigma\sqrt{T}} \leq \frac{2}{\alpha\sigma\tau}$, we have

$$\text{(perturbation loss)} \leq \beta R_G\left[\frac{10mR_GR_Y}{\kappa(G)}\sqrt{T} + \frac{4md\log(\sigma T)}{\eta\alpha\kappa(G)} + dR_Y^2\sqrt{T}\right].$$

■

E.3. Bounding movement cost

The movement cost depends on the stability of the algorithm and is bounded by the following lemma:

Lemma 23 *The movement cost is bounded by*

$$\mathbb{E}\left[\sum_{t=m}^{T} f_t(x_{t-m+1:t}) - \bar{f}_t(x_t)\right] \leq \eta d B^* L m^2 T.$$

Proof The movement cost is bounded by the Lipschitz constant L of f_t's and the Euclidean distances between neighboring iterates.

Lemma 24 (Generalized Pythagorean Theorem) *For any positive definite, symmetric matrix A and its induced norm $\|\cdot\|_A$ over $\mathbb{R}^d$, let C be a convex, closed, nonempty subset of $\mathbb{R}^d$, $y \in \mathbb{R}^d$, and $x = \Pi_C(y)$ w.r.t. $\|\cdot\|_A$. Then $\forall z \in C$,*

$$\|x - z\|_A \leq \|y - z\|_A.$$

Note that by the above lemma, take $y = x_{t-1} - \eta \hat{A}_{t-m}^{-1} \tilde{g}_{t-m}$, $x = x_t$, $z = x_{t-1}$, and $\|\cdot\| = \|\cdot\|_{\hat{A}_{t-m}}$, we have that $\forall t$,

$$\|x_t - x_{t-1}\|_{\hat{A}_{t-m}}^2 \le \|\eta \hat{A}_{t-m}^{-1} \tilde{g}_{t-m}\|_{\hat{A}_{t-m}}^2 = \eta^2 \|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}^2,$$

where since $\hat{A}_{t-m} \succeq \hat{A}_{t-m-1}$,

$$\begin{aligned}
\|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}^2 &\le d^2 (B^*)^2 \sum_{i,j=0}^{m-1} u_{t-i}^\top \hat{A}_{t-m-1-i}^{\frac{1}{2}} \hat{A}_{t-m}^{-1} \hat{A}_{t-m-1-j}^{\frac{1}{2}} u_{t-j} \\
&\le d^2 (B^*)^2 m^2 \left\| \hat{A}_{t-m-1}^{\frac{1}{2}} \hat{A}_{t-m}^{-1} \hat{A}_{t-m-1}^{\frac{1}{2}} \right\|_{\mathrm{op}} \\
&\le d^2 (B^*)^2 m^2,
\end{aligned}$$

Thus,

$$\|x_t - x_{t-1}\|_2 \le \frac{1}{\sqrt{m}} \|x_t - x_{t-1}\|_{\hat{A}_{t-m}} \le \sqrt{m} \eta d B^*.$$

By the Lipschitz assumption on f_t over $\mathcal{K}$, the movement cost is bounded by

$$\begin{aligned}
\sum_{t=m}^{T} \mathbb{E}[f_t(x_{t-m+1:t}) - \bar{f}_t(x_t)] &\le L \sum_{t=m}^{T} \mathbb{E}[\|(x_{t-m+1}, \dots, x_t) - (x_t, \dots, x_t)\|_2] \\
&= L \sum_{t=m}^{T} \mathbb{E}\left[\left(\sum_{s=t-m+1}^{t-1} \|x_s - x_t\|_2^2 \right)^{\frac{1}{2}} \right] \\
&\le L \sum_{t=m}^{T} \mathbb{E}\left[\left(\sum_{s=t-m+1}^{t-1} (t-s) \sum_{r=s+1}^{t} \|x_r - x_{r-1}\|_2^2 \right)^{\frac{1}{2}} \right] \\
&\le \eta d B^* L m^2 T.
\end{aligned}$$

■

E.4. Bounding underlying regret

The underlying regret can be bounded by the following lemma:

Lemma 25 *The underlying regret is bounded by*

$$\mathbb{E}\left[\sum_{t=m}^{T} \bar{f}_t(x_t) - \bar{f}_t(x) \right] \le \frac{mD^2}{2\eta} + 2\eta\sigma \max\{\alpha, 1\} d^2 (B^*)^2 m^3 T + \eta \beta \tilde{R}_G R_Y^2 d^2 B^* D m^{\frac{9}{2}} T + mB.$$

Proof By projection onto convex set, we have that $\forall x \in \mathcal{K}$,

$$\begin{aligned}
\|x_t - x\|_{\hat{A}_{t-m}}^2 &\le \|x_{t-1} - x - \eta \hat{A}_{t-m}^{-1} \tilde{g}_{t-m}\|_{\hat{A}_{t-m}}^2 \\
&= \|x_{t-1} - x\|_{\hat{A}_{t-m-1}}^2 + \frac{1}{2} \|x_{t-1} - x\|_{\eta\alpha H_{t-m}}^2 - 2\eta \tilde{g}_{t-m}^\top (x_{t-1} - x) + \eta^2 \|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}^2 \\
&\le \|x_{t-1} - x\|_{\hat{A}_{t-m-1}}^2 + \|x_{t-m} - x\|_{\eta\alpha H_{t-m}}^2 + \|x_{t-m} - x_{t-1}\|_{\eta\alpha H_{t-m}}^2 - 2\eta \tilde{g}_{t-m}^\top (x_{t-m} - x) \\
&\quad + 2\eta \tilde{g}_{t-m}^\top (x_{t-m} - x_{t-1}) + \eta^2 \|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}^2,
\end{aligned}$$

where the last inequality follows from that for a square matrix $H \succeq 0$, $\|x+y\|_H^2 \leq 2(\|x\|_H^2+\|y\|_H^2)$. Rearranging, we have

$$\begin{aligned}
&\tilde{g}_{t-m}^\top (x_{t-m} - x) - \frac{1}{2}\|x_{t-m} - x\|_{\alpha H_{t-m}}^2 \\
&\leq \frac{1}{2\eta}(\|x_{t-1} - x\|_{\hat{A}_{t-m-1}}^2 - \|x_t - x\|_{\hat{A}_{t-m}}^2) + \frac{\alpha}{2}\|x_{t-m} - x_{t-1}\|_{H_{t-m}}^2 + \tilde{g}_{t-m}^\top (x_{t-m} - x_{t-1}) \\
&\quad + \frac{\eta}{2}\|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}^2.
\end{aligned}$$

We can bound the terms on the right hand side as the following: by using the bounds on $\|x_t - x_{t-1}\|_2$ and $\|\tilde{g}_t\|_{\hat{A}_t^{-1}}$ in Section E.3, we have

$$\begin{aligned}
\frac{\alpha}{2}\|x_{t-m} - x_{t-1}\|_{H_{t-m}}^2 &\leq \frac{\alpha\sigma m}{2}\sum_{s=t-m+1}^{t-1}\|x_s - x_{s-1}\|_2^2 \leq \frac{1}{2}\eta^2\alpha\sigma d^2(B^*)^2 m^3, \\
\tilde{g}_{t-m}^\top (x_{t-m} - x_{t-1}) &\leq \|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}\|x_{t-m} - x_{t-1}\|_{\hat{A}_{t-m}} \\
&\leq \|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}\sum_{s=t-m+1}^{t-1}\|x_s - x_{s-1}\|_{\hat{A}_{t-m}} \\
&\leq 2\|\tilde{g}_{t-m}\|_{\hat{A}_{t-m}^{-1}}\sum_{s=t-m+1}^{t-1}\|x_s - x_{s-1}\|_{\hat{A}_{s-m}} \\
&\leq 2\eta d^2(B^*)^2 m^3.
\end{aligned}$$

Therefore, substituting these bounds into the right hand side of the inequality, we have

$$\begin{aligned}
\tilde{g}_{t-m}^\top (x_{t-m} - x) - \frac{1}{2}\|x_{t-m} - x\|_{\alpha H_{t-m}}^2 &\leq \frac{1}{2\eta}(\|x_{t-1} - x\|_{\hat{A}_{t-m-1}}^2 - \|x_t - x\|_{\hat{A}_{t-m}}^2) + \frac{1}{2}\eta^2\alpha\sigma d^2(B^*)^2 m^3 \\
&\quad + 2\eta d^2(B^*)^2 m^3 + \frac{1}{2}\eta d^2(B^*)^2 m^2 \\
&\leq \frac{1}{2\eta}(\|x_{t-1} - x\|_{\hat{A}_{t-m-1}}^2 - \|x_t - x\|_{\hat{A}_{t-m}}^2) \\
&\quad + 2\eta\sigma \max\{\alpha, 1\} d^2(B^*)^2 m^3.
\end{aligned}$$

Note that since by Lemma 20,

$$\mathbb{E}[\tilde{g}_{t-m} \mid \mathbb{F}_{t-2m}] = \sum_{i=1}^m \nabla_i f_{t-m}(x_{t-2m+1:t-m}).$$

When bounding perturbation loss, we showed that $\forall t \in [T]$, $\forall i \in [m]$,

$$[\nabla^2 f_t]_{ii} \preceq \beta\tilde{R}_G Y_{t-m+i}^\top Y_{t-m+i} \preceq \beta\tilde{R}_G R_Y^2 I_{d\times d}.$$

Thus, since $\nabla^2 f_t \succeq 0$, we have

$$\|\nabla^2 f_t\|_2 \leq \mathrm{tr}(\nabla^2 f_t) = \sum_{i=1}^m \mathrm{tr}([\nabla^2 f_t]_{ii}) \leq \sum_{i=1}^m d\|[\nabla^2 f_t]_{ii}\|_2 \leq dm\beta\tilde{R}_G R_Y^2.$$

By Lemma A.2 in Gradu et al. (2020), $\forall x$,

$$\nabla \bar{f}_t(x) = \sum_{i=0}^{m-1} \nabla_i f_t(x, \ldots, x).$$

By smoothness of f_{t-m},

$$\begin{aligned}
\|\mathbb{E}[\tilde{g}_{t-m} \mid \mathbb{F}_{t-2m}] - \nabla \bar{f}_{t-m}(x_{t-m})\|_2 &\leq \sqrt{m}\|\nabla f_{t-m}(x_{t-2m+1:t-m}) - \nabla f_{t-m}(x_{t-m}, \ldots, x_{t-m})\|_2 \\
&\leq d\beta \tilde{R}_G R_Y^2 m^2 \sum_{s=t-2m+1}^{t-m-1} \|x_s - x_{t-m}\|_2 \\
&\leq d\beta \tilde{R}_G R_Y^2 m^2 \sum_{s=t-2m+1}^{t-m-1} \sum_{r=s+1}^{t-m} \|x_s - x_{s-1}\|_2 \\
&\leq \eta\beta \tilde{R}_G R_Y^2 d^2 B^* m^{\frac{9}{2}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\nabla \bar{f}_{t-m}(x_{t-m})^\top (x_{t-m} - x) - \frac{1}{2}\|x_{t-m} - x\|^2_{\alpha H_{t-m}}\right] &\leq \frac{1}{2\eta}\mathbb{E}\left[\|x_{t-1} - x\|^2_{\hat{A}_{t-m-1}} - \|x_t - x\|^2_{\hat{A}_{t-m}}\right] \\
&\quad + \mathbb{E}[(\nabla \bar{f}_{t-m}(x_{t-m}) - \tilde{g}_{t-m})^\top (x_{t-m} - x)] \\
&\quad + 2\eta\sigma \max\{\alpha, 1\} d^2 (B^*)^2 m^3 \\
&\leq \frac{1}{2\eta}\mathbb{E}\left[\|x_{t-1} - x\|^2_{\hat{A}_{t-m-1}} - \|x_t - x\|^2_{\hat{A}_{t-m}}\right] \\
&\quad + 2\eta\sigma \max\{\alpha, 1\} d^2 (B^*)^2 m^3 \\
&\quad + \eta\beta \tilde{R}_G R_Y^2 d^2 B^* D m^{\frac{9}{2}}.
\end{aligned}$$

Since $\nabla^2 \bar{f}_t = G_t^\top Q_t G_t \succeq \alpha H_t$, summing over all iterations, we have that $\forall x \in \mathcal{K}$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=m}^{T} \bar{f}_t(x_t) - \bar{f}_t(x)\right] &\leq \sum_{t=2m}^{T} \mathbb{E}\left[\bar{f}_{t-m}(x_{t-m}) - \bar{f}_{t-m}(x)\right] + mB \\
&\leq \sum_{t=2m}^{T} \mathbb{E}\left[\nabla \bar{f}_{t-m}(x_{t-m})^\top (x_{t-m} - x) - \frac{1}{2}\|x_{t-m} - x\|^2_{\alpha H_{t-m}}\right] + mB \\
&\leq \frac{1}{2\eta}\sum_{t=2m}^{T} \mathbb{E}\left[\|x_{t-1} - x\|^2_{\hat{A}_{t-m-1}} - \|x_t - x\|^2_{\hat{A}_{t-m}}\right] + 2\eta\sigma \max\{\alpha, 1\} d^2 (B^*)^2 m^3 T \\
&\quad + \eta\beta \tilde{R}_G R_Y^2 d^2 B^* m^{\frac{9}{2}} T + mB \\
&\leq \frac{1}{2\eta}\mathbb{E}\left[\|x_{2m-1} - x\|^2_{\hat{A}_{m-1}}\right] + 2\eta\sigma \max\{\alpha, 1\} d^2 (B^*)^2 m^3 T \\
&\quad + \eta\beta \tilde{R}_G R_Y^2 d^2 B^* m^{\frac{9}{2}} T + mB \\
&\leq \frac{mD^2}{2\eta} + 2\eta\sigma \max\{\alpha, 1\} d^2 (B^*)^2 m^3 T + \eta\beta \tilde{R}_G R_Y^2 d^2 B^* D m^{\frac{9}{2}} T + mB.
\end{aligned}$$

■

E.5. Assembling regret bound

Combining bounds in Section E.2, Section E.3, Section E.4 and using that $\sigma \leq R_G^2 R_Y^2$, the overall expected regret w.r.t. any $x \in \mathcal{K}$ is given by

$$\mathbb{E}[\text{Regret}_T(x)] \leq \beta R_G \left[\frac{10mR_G R_Y}{\kappa(G)} \sqrt{T} + \frac{4md \log(\sigma T)}{\eta \alpha \kappa(G)} + dR_Y^2 \sqrt{T} \right] + \eta d B^* L m^2 T$$
$$+ \frac{mD^2}{2\eta} + 2\eta\sigma \max\{\alpha, 1\} d^2 (B^*)^2 m^3 T + \eta\beta \tilde{R}_G R_Y^2 d^2 B^* D m^{\frac{9}{2}} T + mB.$$

By taking $\eta = \tilde{\Theta}\left(\frac{1}{\alpha\sqrt{T}}\right)$, we have

$$\mathbb{E}[\text{Regret}_T(x)] = \tilde{O}\left(\frac{\beta}{\alpha} B^* \sqrt{T}\right),$$

where $\tilde{O}(\cdot)$ hides logarithmic factors in T and all natural parameters $R_G, \tilde{R}_G, R_Y, \kappa(G), B, L, D, d$.

E.6. Proofs of Technical Lemmas

E.6.1. PROOF OF LEMMA 22

Fix a sequence of $Y_1, \ldots, Y_T$ and G. Let u be any unit vector, and by definition of H_t. Denote $u_t = Y_t v$, and let $Y_t = 0, \forall t \leq 0$ and $t > T$. Then,

$$\begin{aligned}
v^\top \left(\sum_{t=m}^{T} H_t \right) v &= \sum_{t=m}^{T} \left\| \sum_{i=0}^{m-1} G^{[i]} u_{t-i} \right\|_2^2 \\
&\geq \sum_{t=1}^{T+m-1} \left\| \sum_{i=0}^{m-1} G^{[i]} u_{t-i} \right\|_2^2 - 2mR_G^2 R_Y^2 \\
&\geq_{(1)} \frac{1}{2} \sum_{t=1}^{T+m-1} \left\| \sum_{i=0}^{\infty} G^{[i]} u_{t-i} \right\|_2^2 - \sum_{t=1}^{T+m-1} \left\| \sum_{i=m}^{\infty} G^{[i]} u_{t-i} \right\|_2^2 - 2mR_G^2 R_Y^2 \\
&\geq_{(2)} \frac{1}{2} \sum_{t=1}^{T+m-1} \left\| \sum_{i=0}^{\infty} G^{[i]} u_{t-i} \right\|_2^2 - \frac{R_G^2 R_Y^2}{T} - 4mR_G^2 R_Y^2 \\
&=_{(3)} \frac{1}{2} \sum_{t=1}^{\infty} \left\| \sum_{i=0}^{t} G^{[i]} u_{t-i} \right\|_2^2 - 5mR_G^2 R_Y^2 \\
&\geq_{(4)} \frac{\kappa(G)}{2} \sum_{t=1}^{T} \|Y_t v\|_2^2 - 5mR_G^2 R_Y^2 \\
&= \frac{\kappa(G)}{2} v^\top \left(\sum_{t=1}^{T} Y_t^\top Y_t \right) - 5mR_G^2 R_Y^2.
\end{aligned}$$

where we use the inequality $\|x-y\|_2^2 \geq \frac{1}{2}\|x\|_2^2 - \|y\|_2^2$ in (1), the decaying assumption on G such that $\sum_{i=m}^{\infty} \|G^{[i]}\|_{\text{op}} \leq \frac{R_G}{T}$ in (2), $Y_t = 0\ \forall t \leq 0$ and $t > T$ in (3), and definition of $\kappa(G)$ in Assumption 10 in (4).

E.6.2. PROOF OF LEMMA 24

Consider the functions $f_y(v) = \|v-y\|_A^2$ on C and $g_{x,z}(t) = f(tz+(1-t)x)$ defined on $t \in [0,1]$. Since C is convex, we have $tz+(1-t)x \in C, \forall t \in [0,1]$. By assumption, $x = \Pi_C(y)$, $g_{x,z}$ attains minimum at $t=0$. $g'_{x,z}(0) = 2(z-x)^\top A(x-y) \geq 0$. Then,

$$\|y-z\|_A^2 = \|y-x+x-z\|_A^2 = \|y-x\|_A^2 + \|x-z\|_A^2 + 2(x-y)^\top A(z-x) \geq \|x-z\|_A^2.$$

Appendix F. Regret of NBPC (Section 4.3)

F.1. Reduction from LQ control to BQO-AM

Note that by letting $G^{[i]} = \begin{bmatrix} C \\ KC \end{bmatrix} (A+BKC)^{i-1}B \in \mathbb{R}^{(d_{\mathbf{y}}+d_{\mathbf{u}})\times d_{\mathbf{u}}}$, we have that the observation-control pair $(\mathbf{y}_t, \mathbf{u}_t)$ reached by playing DRC matrices $(M_1, \dots, M_{t-1})$ can be expressed as

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{u}_t \end{bmatrix} = \begin{bmatrix} \mathbf{y}_t^K \\ K\mathbf{y}_t^K \end{bmatrix} + \sum_{i=1}^{t} G^{[i]}(\mathbf{u}_{t-i} - K\mathbf{y}_{t-i}) = \begin{bmatrix} \mathbf{y}_t^K \\ K\mathbf{y}_t^K \end{bmatrix} + \sum_{i=1}^{t} G^{[i]} \left(\sum_{j=0}^{m-1} M_{t-i}^{[j]} \mathbf{y}_{t-i-j}^K \right).$$

The sequence of matrices $G = \{G^{[i]}\}_{i\geq 1}$ is called the Markov operator. For simplicity, we assume that the system dynamics and the stabilizing linear policy K is known in this section. With these knowledge, the above equation implies that the signals $\mathbf{y}_t^K$ can be directly computed by the learner from the observations.

Moreover, let $d = md_{\mathbf{y}}d_{\mathbf{u}}$, and let $\mathfrak{e} : (\mathbb{R}^{d_{\mathbf{u}}\times d_{\mathbf{y}}})^m \to \mathbb{R}^d$ denote the natural embedding of a DRC controller M in $\mathbb{R}^d$, with inverse $\mathfrak{e}^{-1}$, i.e. for $M^{[0]}, \dots, M^{[m-1]} \in \mathbb{R}^{d_{\mathbf{u}}\times d_{\mathbf{y}}}$, $k \in [m-1]$, $i \in [d_{\mathbf{u}}]$, $j \in [d_{\mathbf{y}}]$,

$$\mathfrak{e}(M^{[0]}, \dots, M^{[m-1]})_{kd_{\mathbf{u}}d_{\mathbf{y}}+(i-1)d_{\mathbf{y}}+j} = M_{ij}^{[k]}.$$

Denote $\mathfrak{e}_{\mathbf{y}} : (\mathbb{R}^{d_{\mathbf{y}}})^m \to \mathbb{R}^{d_{\mathbf{u}}\times d}$ such that $\forall M^{[0]}, \dots, M^{[m-1]} \in \mathbb{R}^{d_{\mathbf{u}}\times d_{\mathbf{y}}}$,

$$\mathfrak{e}_{\mathbf{y}}(\mathbf{y}_{t-m+1}^K, \dots, \mathbf{y}_t^K)\mathfrak{e}(M^{[0]}, \dots, M^{[m-1]}) = \sum_{j=0}^{m-1} M^{[j]}\mathbf{y}_{t-j}^K.$$

We shorthand $Y_t = \mathfrak{e}_{\mathbf{y}}(\mathbf{y}_{t-m+1}^K, \dots, \mathbf{y}_t^K)$, and $\mathfrak{e}(M) = \mathfrak{e}(M^{[0]}, \dots, M^{[m-1]})$. An explicit formulation of $\mathfrak{e}_{\mathbf{y}}$ is given by

$$Y_{i,j+1:j+d_{\mathbf{y}}} = \mathbf{y}_{t-k+1}^K \quad \text{if} \quad j = (k-1)d_{\mathbf{u}}d_{\mathbf{y}} + (i-1)d_{\mathbf{y}}, \quad \forall i \in [d_{\mathbf{u}}], k \in [m].$$

Therefore, for any controller that plays DRC policies, the cost function can be re-written as

$$\begin{aligned} c_t(\mathbf{y}_t, \mathbf{u}_t) = &\left(B_t + \sum_{i=0}^{m-1} G^{[i]} Y_{t-i}\mathfrak{e}(M_{t-i}) \right)^\top \begin{bmatrix} Q_t & \mathbf{0}_{d_{\mathbf{y}}\times d_{\mathbf{u}}} \\ \mathbf{0}_{d_{\mathbf{u}}\times d_{\mathbf{y}}} & R_t \end{bmatrix} \left(B_t + \sum_{i=0}^{m-1} G^{[i]} Y_{t-i}\mathfrak{e}(M_{t-i}) \right) \\ &=: f_t(\mathfrak{e}(M_{t-m+1}), \dots, \mathfrak{e}(M_t)), \end{aligned}$$

where $B_t = (\mathbf{y}_t^K, K\mathbf{y}_t^K) + \sum_{i=m}^{t} G^{[i]} Y_{t-i}\mathfrak{e}(M_{t-i})$. Let $\bar{f}_t$ denote its induced unary form, i.e. $\bar{f}_t(\mathfrak{e}(M)) = f_t(\mathfrak{e}(M), \dots, \mathfrak{e}(M))$. Note that f_t is an adaptive function of the learners decision before the time $t-m+1$ through B_t. B_t is independent of the algorithm's decisions $M_{t-m+1:t}$, and Y_t's are independent of the algorithm's decisions. Therefore, the adversarial model in Assumption 9 is satisfied.

F.2. Regularity conditions

Lemma 26 (Bounds on LQR/LQG induced with-memory loss functions) $\forall M_1, \dots, M_T \in \mathcal{M}(m, R_{\mathcal{M}})$,

- *For* $m \geq 1 + \left(\log \frac{1}{1-\gamma}\right)^{-1} \log T$,

$$\sum_{i=0}^{\infty} \|G^{[i]}\|_{\text{op}} \leq 1 + \frac{c\kappa\sqrt{1+\kappa^2}\kappa_C\kappa_B}{\gamma} =: R_G, \quad \sum_{i=m}^{\infty} \|G^{[i]}\|_{\text{op}} \leq \frac{R_G}{T},$$
$$\sum_{i=1}^{\infty} \|C(A+BKC)^{i-1}B\|_{\text{op}} \leq \frac{c\kappa_C\kappa_B\kappa}{\gamma} =: R_{G_{\mathbf{y}}}, \quad \sum_{i=1}^{\infty} \|C(A+BKC)^{i-1}\|_{\text{op}} \leq \frac{c\kappa_C\kappa}{\gamma} =: R_{G_{\mathbf{y}},K},$$
$$\max_{1\leq i\leq m-1} \|(G^{[i]})^\top G^{[i]}\|_2 \leq (1+\kappa^2)\kappa_C^2\kappa_B^2\kappa^2 =: \tilde{R}_G.$$

- $\max_{t\in[T]} \|\mathbf{y}_t^K\|_2 \leq (R_{G,K}+1)R_{\mathbf{w},\mathbf{e}} =: R_{\text{nat}}$.
- $\max_{t\in[T]} |c_t(\mathbf{y}_t^{M_{1:t-1}}, \mathbf{u}_t^{M_{1:t}})| = \max_{t\in[T]} |f_t(\mathfrak{e}(M_{t-m+1}), \dots, \mathfrak{e}(M_t))| \leq \beta R_{\text{nat}}^2((1+R_{G_{\mathbf{y}}}R_{\mathcal{M}})^2(1+2\kappa^2) + 2R_{\mathcal{M}}^2) =: B$.
- $\max_{M^1,M^2\in\mathcal{M}(m,R_{\mathcal{M}})} \|\mathfrak{e}(M^1) - \mathfrak{e}(M^2)\|_2 \leq \sqrt{m\max\{d_{\mathbf{u}}, d_{\mathbf{y}}\}}R_{\mathcal{M}} =: D$.
- $\max_{t\in[T]} \|Y_t\|_2 \leq \sqrt{md_{\mathbf{y}}d_{\mathbf{u}}^2}R_{\text{nat}} =: R_Y$.
- $\max_{t\in[T]} \|\nabla f_t(\mathfrak{e}(M_{t-m+1}), \dots, \mathfrak{e}(M_t))\| \leq 2\beta\sqrt{m}R_Y R_G(R_Y R_G D + 2R_{\text{nat}}) =: L$.
- $\kappa(G) \geq \frac{1}{4}\min\{1, \kappa^{-2}\}$.

Proof By Assumption 4, we know that

$$\begin{aligned}
\sum_{i=0}^{\infty} \|G^{[i]}\|_{\text{op}} &\leq 1 + \sum_{i=1}^{\infty} \sqrt{\|C(A+BKC)^{i-1}B\|_{\text{op}}^2 + \|KC(A+BKC)^{i-1}B\|_{\text{op}}^2} \\
&\leq 1 + c\sqrt{1+\kappa^2}\kappa_C\kappa_B \sum_{i=1}^{\infty} \|HL^{i-1}H^{-1}\|_2 \\
&\leq 1 + c\kappa\sqrt{1+\kappa^2}\kappa_C\kappa_B \sum_{i=0}^{\infty} (1-\gamma)^i \\
&= 1 + \frac{c\kappa\sqrt{1+\kappa^2}\kappa_C\kappa_B}{\gamma} \\
&= R_G,
\end{aligned}$$

where c is some constant possibly depending on the dimension of H, L. By similar argument and the choice of m, we can establish the other inequalities. Moreover,

$$\max_{1\leq i\leq m-1} \|(G^{[i]})^\top G^{[i]}\|_2 \leq (1+\kappa^2)\kappa_C^2\kappa_B^2\kappa^2(1-\gamma)^{2(i-i)} \leq (1+\kappa^2)\kappa_C^2\kappa_B^2\kappa^2 = \tilde{R}_G.$$

Therefore, denote $\mathbf{x}_t^K$ as the would-be state if the linear policy K was played from the beginning of the time, we have

$$\max_{t\in[T]} \|\mathbf{y}_t^K\|_2 = \max_{t\in[T]} \left\|C\mathbf{x}_t^K + \mathbf{e}_t\right\| = \max_{t\in[T]} \left\|C\sum_{i=0}^{t-1}(A+BKC)^i \mathbf{w}_{t-i} + \mathbf{e}_t\right\|_2 \le (R_{G,K}+1)R_{\mathbf{w},\mathbf{e}},$$

and

$$\max_{t\in[T]} \left\|\sum_{j=0}^{m-1} M_t^{[j]} \mathbf{y}_{t-j}^K\right\|_2 \le \max_{t\in[T]} \|M_t\|_{\ell_1,\mathrm{op}} \max_{t\in[T]} \|\mathbf{y}_t^K\|_2 \le R_{\mathcal{M}} R_{\mathrm{nat}},$$

which implies that $\max_{t\in[T]} \left\|\mathbf{y}_t^{M_{1:t-1}}\right\| \le R_{\mathrm{nat}}(1+R_{G_{\mathbf{y}}}R_{\mathcal{M}})$. Thus, c_t's are bounded by

$$\begin{aligned}
\max_{t\in[T]} c_t(\mathbf{y}_t^{M_{1:t-1}}, \mathbf{u}_t^{M_{1:t}}) &\le \beta\left(\|\mathbf{y}_t^{M_{1:t-1}}\|_2^2 + \left\|K\mathbf{y}_t^{M_{1:t-1}} + \sum_{j=0}^{m-1} M_t^{[j]}\mathbf{y}_{t-j}^K\right\|_2^2\right)\\
&\le \beta R_{\mathrm{nat}}^2((1+R_{G_{\mathbf{y}}}R_{\mathcal{M}})^2(1+2\kappa^2)+2R_{\mathcal{M}}^2)\\
&= B.
\end{aligned}$$

The diameter of $\mathcal{M}(m, R_{\mathcal{M}})$ is given by

$$\begin{aligned}
\max_{M_1,M_2\in\mathcal{M}(m,R_{\mathcal{M}})} \|\mathfrak{e}(M_1)-\mathfrak{e}(M_2)\|_2 &\le \sqrt{m} \max_{M_1,M_2\in\mathcal{M}(m,R_{\mathcal{M}})} \max_{0\le j\le m-1} \|M_1^{[j]} - M_2^{[j]}\|_F\\
&\le \sqrt{m\max\{d_{\mathbf{u}}, d_{\mathbf{y}}\}} \max_{M_1,M_2\in\mathcal{M}(m,R_{\mathcal{M}})} \max_{0\le j\le m-1} \|M_1^{[j]} - M_2^{[j]}\|_{\mathrm{op}}\\
&\le \sqrt{m\max\{d_{\mathbf{u}}, d_{\mathbf{y}}\}} R_{\mathcal{M}},
\end{aligned}$$

where the second inequality follows from that $\|A\|_F \le \sqrt{\mathrm{rank}(A)}\sigma_{\max}(A) \le \sqrt{\mathrm{rank}(A)}\|A\|_{\mathrm{op}}$.

To bound the gradient of f_t, denote $\nabla_i f_t$ as the gradient of f_t w.r.t. $\mathfrak{e}(M_{t-m+i})$, and then we have $\forall t\in[T], i\in[m]$,

$$\begin{aligned}
\|\nabla_i f_t(\mathfrak{e}(M_{t-m+1}),\ldots,\mathfrak{e}(M_t))\|_2 &\le 2\beta\left\|Y_{t-m+i}^\top (G^{[m-i]})^\top \left(B_t + \sum_{i=0}^{m-1} G^{[i]} Y_{t-i}\mathfrak{e}(M_{t-i})\right)\right\|_2\\
&\le 2\beta \max_{t\in[T]} \|Y_t\|_{\mathrm{op}} R_G \left(\max_{t\in[T]} \|B_t\|_2 + R_G \max_{t\in[T]} \|Y_t\|_{\mathrm{op}} D\right)
\end{aligned}$$

We can bound B_t and Y_t as follows:

$$\begin{aligned}
&\max_{t\in[T]} \|B_t\|_2 = \max_{t\in[T]} \left\|\mathbf{y}_t^K + \sum_{i=m}^{t} G^{[i]} \left(\sum_{j=0}^{m-1} M_{t-i}^{[j]}\mathbf{y}_{t-j}^K\right)\right\|_2 \le R_{\mathrm{nat}}\left(1+\frac{R_G R_{\mathcal{M}}}{T}\right),\\
&\max_{t\in[T]} \|Y_t\|_2 \le \sqrt{m d_{\mathbf{y}} d_{\mathbf{u}}^2} \max_{t\in[T]} \|\mathbf{y}_t^K\|_2 \le \sqrt{m d_{\mathbf{y}} d_{\mathbf{u}}^2} R_{\mathrm{nat}} =: R_Y,
\end{aligned}$$

where the second inequality is given by the definition of $\mathfrak{e}_{\mathbf{y}}$. Thus, assuming $\frac{R_G R_{\mathcal{M}}}{T} \leq 1$,

$$\begin{aligned}\max_{t\in[T]} \|\nabla f_t(\mathfrak{e}(M_{t-m+1}), \ldots, \mathfrak{e}(M_t))\|_2 &\leq \sqrt{m} \max_{t\in[T],1\leq i\leq m} \|\nabla_i f_t(\mathfrak{e}(M_{t-m+1}), \ldots, \mathfrak{e}(M_t))\|_2 \\ &\leq 2\beta\sqrt{m} R_Y R_G (R_Y R_G D + 2R_{\text{nat}}) \\ &=: L.\end{aligned}$$

The last inequality on $\kappa(G) \geq \frac{1}{4}\min\{1, \kappa^{-2}\}$ is given by Lemma 3.1 in Simchowitz (2020). ■

F.3. `NBPC`: Newton Bandit Perturbation Controller Algorithm

Algorithm 4 Newton Bandit Perturbation Controller (`NBPC`)

Input: DRC policy class $\mathcal{M}(m, R_{\mathcal{M}})$, step size $\eta > 0$, time horizon T, system Markov operator G from system parameters (A, B, C), (κ, γ)-strongly stable linear policy K, strong convexity parameter $\alpha > 0$.

Initialize: $M_1^{[j]} = \cdots = M_m^{[j]} = \mathbf{0}_{d_{\mathbf{u}}\times d_{\mathbf{y}}}$, $\forall j \in [m]$, $\tilde{g}_{0:m-1} = \mathbf{0}_{md_{\mathbf{u}}d_{\mathbf{y}}}$, $\hat{A}_{0:m-1} = mI_{md_{\mathbf{u}}d_{\mathbf{y}}\times md_{\mathbf{u}}d_{\mathbf{y}}}$.
Sample $\varepsilon_t \sim \mathbb{S}^{md_{\mathbf{u}}d_{\mathbf{y}}-1}$ i.i.d. uniformly at random for $t = 1, \ldots, m$.
Set $\widetilde{M}_t = \mathfrak{e}^{-1}(\mathfrak{e}(M_t) + \hat{A}_{t-1}^{-\frac{1}{2}}\varepsilon_t)$, $t = 1, \ldots, m$.
Play control $\mathbf{u}_t = K\mathbf{y}_t$, incur cost $c_t(\mathbf{y}_t, \mathbf{u}_t)$ for $t = 1, \ldots, m$.
for $t = m, \ldots, T$ **do**
  Play control $\mathbf{u}_t = \mathbf{u}_t^{\widetilde{M}_t} = K\mathbf{y}_t + \sum_{j=0}^{m-1} \widetilde{M}_t^{[j]}\mathbf{y}_{t-j}^K$, incur cost $c_t(\mathbf{y}_t, \mathbf{u}_t) = f_t(\mathfrak{e}(\widetilde{M}_{t-m+1:t}))$.
  System involves as $\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \mathbf{w}_t$ and $\mathbf{y}_{t+1} = C\mathbf{x}_{t+1} + \mathbf{e}_{t+1}$. Receive new observation $\mathbf{y}_{t+1}$, compute signal

$$\mathbf{y}_{t+1}^K = \mathbf{y}_{t+1} - \sum_{i=1}^{t+1} G^{[i]} \left(\sum_{j=0}^{m-1} \widetilde{M}_{t+1-i}^{[j]} \mathbf{y}_{t+1-i-j}^K \right),$$

  where $\forall t \leq 0$, $\mathbf{y}_t^K \stackrel{\text{def}}{=} 0$ and $\widetilde{M}_t \stackrel{\text{def}}{=} 0$.
  Compute Hessian information matrix $H_t \in \mathbb{R}^{md_{\mathbf{y}}d_{\mathbf{u}}\times md_{\mathbf{y}}d_{\mathbf{u}}}$:

$$H_t = G_t^\top G_t, \quad G_t = \sum_{i=0}^{m-1} G^{[i]} Y_{t-i}, \quad Y_t = \mathfrak{e}_{\mathbf{y}}(\mathbf{y}_{t-m+1}^K, \ldots, \mathbf{y}_t^K).$$

  Update $\hat{A}_t = \hat{A}_{t-1} + \frac{\eta\alpha}{2} H_t$.
  Create gradient estimate: $\tilde{g}_t = md_{\mathbf{u}}d_{\mathbf{y}}c_t(\mathbf{y}_t, \mathbf{u}_t)\sum_{j=0}^{m-1} \hat{A}_{t-1-j}^{\frac{1}{2}}\varepsilon_{t-j}$.
  Update $M_{t+1} = \mathfrak{e}^{-1}\left(\prod_{\mathfrak{e}(\mathcal{M}(m,R_{\mathcal{M}})}^{\hat{A}_{t-m+1}} \left[\mathfrak{e}(M_t) - \eta\hat{A}_{t-m+1}^{-1}\tilde{g}_{t-m+1}\right]\right)$.
  Sample $\varepsilon_{t+1} \sim \mathbb{S}^{md_{\mathbf{u}}d_{\mathbf{y}}-1}$ uniformly at random, independent of previous steps.
  Set $\widetilde{M}_{t+1} = \mathfrak{e}^{-1}(\mathfrak{e}(M_{t+1}) + \hat{A}_{t-m+1}^{-\frac{1}{2}}\varepsilon_{t+1})$.
end for

F.4. Proof of Theorem 5

Let R_G, R_Y, B, D, L be consistent with the notations in Lemma 26, and

$$M^* = \operatorname*{arg\,min}_{M \in \mathcal{M}(m, R_{\mathcal{M}})} \sum_{t=1}^{T} c_t(\mathbf{y}_t^M, \mathbf{u}_t^M).$$

Note that treating the first $2m-2$ steps as burn-in loss, we have

$$\begin{aligned}\text{Regret}_T(\texttt{NBPC}) \leq 2mB &+ \left(\sum_{t=2m-1}^{T} f_t(\mathfrak{e}(\widetilde{M}_{t-m+1:t})) - \bar{f}_t(\mathfrak{e}(M^*))\right) \\ &+ \left(\sum_{t=2m-1}^{T} \bar{f}_t(\mathfrak{e}(M^*)) - c_t(\mathbf{y}_t^{M^*}, \mathbf{u}_t^{M^*})\right).\end{aligned}$$

By Theorem 4 and Lemma 26, by taking $\eta = \tilde{\Theta}\left(\frac{1}{\alpha\sqrt{T}}\right)$, the second term is bounded by $\tilde{O}\left(\frac{\beta^2}{\alpha}\sqrt{T}\right)$, where $\tilde{O}$ hides logarithmic terms in T and natural parameters as those derived in Lemma 26. The third term can be bounded as the following:

$$\sum_{t=2m-1}^{T} \bar{f}_t(\mathfrak{e}(M^*)) - c_t(\mathbf{y}_t^{M^*}, \mathbf{u}_t^{M^*}) \leq L \sum_{t=2m-1}^{T} \left\| \sum_{i=m}^{t} G^{[i]} Y_{t-i}(\mathfrak{e}(M^*) - \mathfrak{e}(M_{t-i})) \right\|_2 \leq L R_G R_Y D. \tag{9}$$

Appendix G. Extension to control of unknown systems (Section 4.3)

Consider the setting the dynamics of the system are unknown to the learner. In this case, the learner needs to first estimate the system dynamics and obtain an estimated Markov operator $\hat{G}$, and then use $\hat{G}$ in place of G in the control algorithm.

This section will be organized as the following: Section G.1 will analyze the regret guarantee if instead of the true Markov operator G, the learner receives an estimate $\hat{G}$ satisfying $\|\hat{G} - G\|_{\ell_1, \text{op}} \leq \varepsilon_G$.

G.1. Known Stabilizing Controller

First, we consider the case where a (κ, γ)-stabilizing linear policy K is known to the learner, although the learner has no information to the system's dynamics (A, B, C) and therefore cannot compute G. Simchowitz (2020) has studied this problem in the full information setting, where they proved a quadratic sensitivity of the control regret to the estimation error ε_G of $\hat{G}$. The main difference between our setting and and that of Simchowitz (2020) comes from the bandit gradient estimation step and the delayed updates in Algorithm 2 to establish conditional independence.

We start with the assumption that there exists an estimation algorithm that returns a $\hat{G}$ sufficiently close to the true Markov operator G:

Assumption 11 (Accurate system estimator) *Assume that the learner has access to an estimate $\hat{G}$ of the Markov operator G satisfying*

$$\|G - \hat{G}\|_{\ell_1, \text{op}} \leq \varepsilon_G, \quad \hat{G}^{[i]} = \mathbf{0}_{(d_{\mathbf{y}} + d_{\mathbf{u}}) \times d_{\mathbf{u}}}, \forall i \geq m.$$

Remark 27 *The system estimation via least-square algorithm used in (Simchowitz et al., 2020; Simchowitz, 2020; Sun et al., 2023) satisfies the conditions in Assumption 11 after $N = O(1/\varepsilon_G^2)$ iterations with high probability (see, for instance, Theorem 6b in (Simchowitz et al., 2020)). Since we pay at most BN-regret during the estimation step, $\tilde{O}(\sqrt{T})$ is preserved if ε_G propagates quadratically in the regret bound (i.e. the learner's regret suffers an additional $\tilde{O}(T\varepsilon_G^2)$ due to estimation error).*

Preliminaries. During the course of the algorithm, the learner uses the Markov operator to compute $\hat{\mathbf{y}}_t^K$. Therefore, with an estimated Markov operator, the $\mathbf{y}_t^K$'s computed by the learner are no longer accurate. When reducing to BQO-AM, the loss functions that the learner sees become

$$\hat{F}_t(\mathfrak{e}(M_{t-m+1:t})) = \left(\hat{B}_t + \sum_{i=0}^{m-1} \hat{G}^{[i]}\hat{Y}_{t-i}\mathfrak{e}(M_{t-i})\right)^\top \begin{bmatrix} Q_t & \mathbf{0}_{d_\mathbf{y}\times d_\mathbf{u}} \\ \mathbf{0}_{d_\mathbf{u}\times d_\mathbf{y}} & R_t \end{bmatrix} \left(\hat{B}_t + \sum_{i=0}^{m-1} \hat{G}^{[i]}\hat{Y}_{t-i}\mathfrak{e}(M_{t-i})\right),$$

where $\hat{B}_t = (\hat{\mathbf{y}}_t^K, K\hat{\mathbf{y}}_t^K), \hat{Y}_t$ are the counterparts of B_t, Y_t defined in Section 4.3 using estimates $\hat{G}$ and $\hat{\mathbf{y}}_t^K$. Note that we do not have the additional terms in $\hat{B}_t$ that depends on the learner's past controls since $\hat{G}^{[i]} = \mathbf{0}_{(d_\mathbf{y}+d_\mathbf{u})\times d_\mathbf{u}}$ for $i \geq m$. First, we consider a pseudo-loss function:

$$\tilde{F}_t(\mathfrak{e}(M_{t-m+1:t})) = \left(\tilde{B}_t + \sum_{i=0}^{m-1} G^{[i]}\hat{Y}_{t-i}\mathfrak{e}(M_{t-i})\right)^\top \begin{bmatrix} Q_t & \mathbf{0}_{d_\mathbf{y}\times d_\mathbf{u}} \\ \mathbf{0}_{d_\mathbf{u}\times d_\mathbf{y}} & R_t \end{bmatrix} \left(\tilde{B}_t + \sum_{i=0}^{m-1} G^{[i]}\hat{Y}_{t-i}\mathfrak{e}(M_{t-i})\right),$$

where $\tilde{B}_t = (\mathbf{y}_t^K, K\mathbf{y}_t^K)$. Let $\hat{f}_t, \tilde{f}_t$ denote their induced unary form, respectively. Although the learner has no access to $\tilde{F}_t$ (which requires the knowledge of G), Proposition D.8 in Simchowitz (2020) suggests that the control regret for unknown system is implied by the bound for some $\nu > 0$ on the following quantity:

$$\mathbb{E}\left[\sum_{t=m}^{T} \tilde{F}_t(\mathfrak{e}(\widetilde{M}_{t-m+1:t})) - \tilde{f}_t(\mathfrak{e}(M)) + \nu\sum_{t=m}^{T} \|Y_t(\mathfrak{e}(M_{t-m+1}) - \mathfrak{e}(M))\|_2^2\right], \tag{10}$$

where $\widetilde{M}_t$ is the DRC policy played by the learner at time t, and M_t is the projected DRC policy by Algorithm 4 at time t. Therefore, in the rest of this section, we focus on bounding the quantity in Eq. (10).

G.2. High-level proof

Throughout this section, we assume that Assumption 11 holds, i.e. we have access to a sufficiently accurate estimator $\hat{G}$ of the Makrov operator G. For a comparator $x \in \mathfrak{e}(\mathcal{M}(m, R_\mathcal{M}))$, we denote

$$\widetilde{\text{Regret}}_T(x) = \sum_{t=m}^{T} \tilde{F}_t(y_{t-m+1:t}) - \tilde{f}_t(x), \quad \widehat{\text{Regret}}_T(x) = \sum_{t=m}^{T} \hat{F}_t(y_{t-m+1:t}) - \hat{f}_t(x),$$

where $y_t = \mathfrak{e}(\widetilde{M}_t)$. Similarly, we denote $x_t = \mathfrak{e}(M_t)$. Suppose Algorithm 2 is run with $\hat{F}_t$ with a slightly modified cumulative Hessian $\hat{A}_t = mI + \frac{\eta\alpha}{6}\sum_{s=m}^{t}\hat{H}_s$, where $\hat{H}_t = \hat{G}_t^\top\hat{G}_t$, $\hat{G}_t = \sum_{i=0}^{m-1}\hat{G}^{[i]}\hat{Y}_{t-i}$. We start by noting that $\widehat{\text{Regret}}_T(x)$ can be bounded via the regret bound for known systems by treating $\hat{G}$ as the true Markov operator.

Lemma 28 $\forall x \in \mathfrak{e}(\mathcal{M}(m, R_{\mathcal{M}}))$, $\mathbb{E}\left[\widehat{\textit{Regret}}_T(x)\right] \leq \tilde{O}(\sqrt{T})$.

Lemma 29 and Lemma 30 relates the regret of controlling an unknown system to $\widehat{\text{Regret}}_T(x)$.

Lemma 29 $\forall \nu > 0$, $\exists x^* \in \mathfrak{e}(\mathcal{M}(\frac{m}{3}, \frac{R_{\mathcal{M}}}{2}))$ *such that*

$$\mathbb{E}\left[\text{ControlRegret}_T\right] \leq \mathbb{E}[\widetilde{\textit{Regret}}_T(x^*)] + \tilde{O}(1) \cdot T\varepsilon_G^2 \left(1 + \frac{1}{\nu}\right) + \nu \sum_{t=m}^{T} \mathbb{E}[\|\hat{Y}_t(x_t - x^*)\|_2^2] + \tilde{O}(1),$$

where $\text{ControlRegret}_T := \sum_{t=1}^T c_t(\mathbf{y}_t, \mathbf{u}_t) - \min_{M \in \mathcal{M}(\frac{m}{3}, \frac{R_{\mathcal{M}}}{2})} \sum_{t=1}^T c_t(\mathbf{y}_t^M, \mathbf{u}_t^M)$.

Lemma 30 is an analogous result to Lemma 5.1-5.4 in Simchowitz (2020), providing a relationship between the two quantities defined above.

Lemma 30 $\forall \nu > 0$, *the regret for any* $x \in \mathfrak{e}(\mathcal{M}(m, R_{\mathcal{M}}))$ *for* $\tilde{F}_t$ *is bounded by*

$$\mathbb{E}[\widetilde{\textit{Regret}}_T(x)] \leq 2B(\mathbb{E}[\widehat{\textit{Regret}}_T(x)]) + \frac{\nu}{m} \sum_{t=m}^{T} \sum_{i=0}^{m-1} \mathbb{E}\left[\|\hat{Y}_{t-i}(x_t - x)\|_2^2\right] - \frac{\alpha}{12} \sum_{t=m}^{T} \mathbb{E}\left[\|x_t - x\|_{\tilde{H}_t}^2\right]$$
$$+ \left(\beta R_{\hat{Y}}^2 + \frac{\alpha R_{\hat{Y}}^2 D^2}{3} + \frac{12}{\alpha}\beta^2 D^2 + \frac{(\beta D R_G)^2 m}{\nu}\right) \varepsilon_G^2 T,$$

where $B(\cdot)$ *denotes any upper bound on a given quantity, and* $\tilde{H}_t$ *is defined analogously to* $\hat{H}_t$ *for* $\tilde{F}_t$*'s.*

Lemma 31 *For* $\nu = \frac{\alpha\kappa(G)}{192}$, *we have*

$$\frac{2\nu}{m} \sum_{t=m}^{T} \sum_{i=0}^{m-1} \mathbb{E}\left[\|\hat{Y}_{t-i}(x_t - x)\|_2^2\right] - \frac{\alpha}{12} \sum_{t=m}^{T} \mathbb{E}\left[\|x_t - x\|_{\tilde{H}_t}^2\right] \leq \tilde{O}(\sqrt{T}),$$

and therefore by Lemma 30,

$$\mathbb{E}[\widetilde{\textit{Regret}}_T(x)] \leq 2B(\mathbb{E}[\widehat{\textit{Regret}}_T(x)]) - \frac{\alpha\kappa(G)}{192m} \sum_{t=m}^{T} \sum_{i=0}^{m-1} \mathbb{E}\left[\|\hat{Y}_{t-i}(x_t - x)\|_2^2\right] + \tilde{O}(1) \cdot \varepsilon_G^2 T.$$

Combining Lemma 28-31, we conclude that

$$\mathbb{E}\left[\text{ControlRegret}_T\right] \leq \tilde{O}(\sqrt{T}) + O(T\varepsilon_G^2).$$

G.3. Omitted proofs in Section G.1

G.3.1. Proof of Lemma 28

To see the desired regret bound, it suffices to derive analogous bounds to Lemma 26 to bound the natural parameters necessary for analysis in Section 4. First, we have by Assumption 11,

$$R_{\hat{G}} \leq R_G + \varepsilon_G, \quad \tilde{R}_{\hat{G}} \leq \tilde{R}_G \varepsilon_G (R_{\hat{G}} + R_G), \quad \kappa(\hat{G}) \geq \frac{1}{4}\min\{1, \kappa^{-2}\}.$$

It is left to establish bounds on $\hat{B}, \hat{L}, R_{\hat{Y}}$. Note that assuming $\varepsilon_G \le \frac{1}{2R_{\mathcal{M}}}$,

$$\begin{aligned}
\max_{t\in[T]} \left\| \sum_{j=0}^{m-1} M_t^{[j]} \hat{\mathbf{y}}_{t-j}^K \right\|_2 &\le R_{\mathcal{M}} \max_{t\in[T]} \max_{t-m+1\le s\le t} \|\hat{\mathbf{y}}_s^K\|_2 \\
&\le R_{\mathcal{M}} \left(R_{\text{nat}} + \max_{t\in[T]} \max_{t-m+1\le s\le t} \left\| \sum_{i=0}^{s} (G_{\mathbf{y}}^{[i]} - \hat{G}_{\mathbf{y}}^{[i]}) \sum_{j=0}^{m-1} M_{s-i}^{[j]} \hat{\mathbf{y}}_{s-i-j}^K \right\|_2 \right) \\
&\le R_{\mathcal{M}} \left(R_{\text{nat}} + \varepsilon_G \max_{t\in[T]} \left\| \sum_{j=0}^{m-1} M_t^{[j]} \hat{\mathbf{y}}_{t-j}^K \right\|_2 \right),
\end{aligned}$$

and thus $\max_{t\in[T]} \left\|\sum_{j=0}^{m-1} M_t^{[j]} \hat{\mathbf{y}}_{t-j}^K\right\|_2 \le 2R_{\mathcal{M}} R_{\text{nat}}$, which implies $\max_{t\in[T]} \|\hat{\mathbf{y}}_t^K\|_2 \le 2R_{\text{nat}}$. Therefore, following the analysis of Lemma 26, $\hat{B} \le 4B$, $R_{\hat{Y}} \le 2R_Y$, $\hat{L} \le 8L$.

G.3.2. PROOF OF LEMMA 29

Denote

$$M^* = \operatorname*{arg\,min}_{M\in\mathcal{M}(m,R_{\mathcal{M}})} \sum_{t=1}^{T} c_t(\mathbf{y}_t^M, \mathbf{u}_t^M), \quad x^* = \mathfrak{e}(M^*).$$

When the system is unknown and the control algorithm is run with an estimated Markov operator $\hat{G}$, we have that

$$c_t(\mathbf{y}_t, \mathbf{u}_t) = \left\| \tilde{B}_t + \sum_{i=0}^{t} G^{[i]} \hat{Y}_{t-i} \mathfrak{e}(\widetilde{M}_{t-i}) \right\|_{P_t}^2 .$$

The control regret for unknown system with an estimated Markov operator $\hat{G}$ can be decomposed as

$$\begin{aligned}
\text{ControlRegret}_T = &\underbrace{\sum_{t=1}^{2m-2} c_t(\mathbf{y}_t, \mathbf{u}_t)}_{\le 2m\hat{B}} + \underbrace{\sum_{t=2m-1}^{T} c_t(\mathbf{y}_t, \mathbf{u}_t) - \tilde{F}_t(y_{t-m+1:t})}_{\text{(loss approximation error)}} \\
&+ \underbrace{\sum_{t=2m-1}^{T} \tilde{F}_t(y_{t-m+1:t}) - \tilde{f}_t(x^*)}_{\widetilde{\text{Regret}}_T(x^*)} + \underbrace{\sum_{t=2m-1}^{T} \tilde{f}_t(x^*) - \bar{f}_t(x^*)}_{\text{(comparator approximation error)}} \\
&+ \underbrace{\sum_{t=2m-1}^{T} \bar{f}_t(x^*) - \sum_{t=2m-1}^{T} c_t(\mathbf{y}_t^{M^*}, \mathbf{u}_t^{M^*})}_{\text{(control approximation error)}} .
\end{aligned}$$

We bound each of the terms as follows: since

$$\|\Delta_t\|_2 := \left\| \sum_{i=m}^{t} G^{[i]} \hat{Y}_{t-i} \mathfrak{e}(\widetilde{M}_{t-i}) \right\|_2 \le \frac{R_G R_{\hat{Y}} \hat{D}}{T},$$

thus

$$\begin{aligned}(\text{loss approximation error}) &\le \beta \sum_{t=2m-1}^{T} \|\Delta_t\|_2 \left(2 \left\| \tilde{B}_t + \sum_{i=0}^{t} G^{[i]} \hat{Y}_{t-i} x_{t-i} \right\|_2 + \|\Delta_t\|_2 \right) \\ &= \tilde{O}(1).\end{aligned}$$

The control approximation error is bounded by $\tilde{O}(1)$ by Eq. 9. By Lemma D.10 in Simchowitz (2020),

$$(\text{comparator approximation error}) \le \nu \sum_{t=m}^{T} \|\hat{Y}_t(x_t - x^*)\|_2^2 + \tilde{O}(1) \cdot T\varepsilon_G^2 \left(1 + \frac{1}{\nu}\right) + \tilde{O}(1).$$

G.3.3. PROOF OF LEMMA 30

Lemma 32 (Gradient error) *Let* $c(\beta, \kappa, G, \hat{G}, D, \hat{Y}) = 2\beta R_{\hat{Y}}^2 \left(\sqrt{1+\kappa^2} + D(R_G + R_{\hat{G}})\right)$. *Then, there holds*

$$\max_{t\in[T]} \max_{M_{t-m+1:t}\in\mathcal{M}(m,R_{\mathcal{M}})} \|\nabla\hat{F}_t(\mathfrak{e}(M_{t-m+1:t})) - \nabla\tilde{F}_t(\mathfrak{e}(M_{t-m+1:t}))\|_2 \le c(\beta, \kappa, G, \hat{G}, D, \hat{Y})\varepsilon_G\sqrt{m},$$

$$\nabla\tilde{f}_t(\mathfrak{e}(M_t)) - \nabla\hat{f}_t(\mathfrak{e}(M_t)) = 2\tilde{G}_t^\top P_t(\tilde{G}_t - \hat{G}_t)\mathfrak{e}(M_t) + 2(\tilde{G}_t - \hat{G}_t)^\top P_t\hat{G}_t\mathfrak{e}(M_t),$$

where $P_t = \begin{bmatrix} Q_t & \mathbf{0}_{d_{\mathbf{y}}\times d_{\mathbf{u}}} \\ \mathbf{0}_{d_{\mathbf{u}}\times d_{\mathbf{y}}} & R_t \end{bmatrix}$.

Proof The second equality follows directly by taking the gradient of the functions. To bound the difference of the gradients of $\hat{F}_t$ and $\tilde{F}_t$, Note that $\forall t$ and $\forall M_{t-m+1:t}$,

$$\begin{aligned}\|\nabla_i\hat{F}_t(\mathfrak{e}(M_{t-m+1:t})) - \nabla_i\tilde{F}_t(\mathfrak{e}(M_{t-m+1:t}))\|_2 &\le 2\beta R_{\hat{Y}} \|(\hat{G}^{[m-i]} - G^{[m-i]})\hat{B}_t\|_2 \\ &\quad + 2\beta R_{\hat{Y}}^2 D \left\| \sum_{j=0}^{m-1} (\hat{G}^{[m-i]})^\top \hat{G}^{[j]} - (G^{[m-i]})^\top \hat{G}^{[j]} \right\| \\ &\le 2\beta R_{\hat{Y}}^2 (\sqrt{1+\kappa^2} + D(R_G + R_{\hat{G}}))\varepsilon_G.\end{aligned}$$

■

Let $y_t = \mathfrak{e}(\widetilde{M}_t)$. We again decompose the regret into perturbation loss, movement cost, and underlying regret and bound each separately. We define the following quantities: denote

$$\widehat{\text{PerturbLoss}} := \mathbb{E}\left[\sum_{t=m}^{T} \hat{F}_t(y_{t-m+1:t}) - \hat{F}_t(x_{t-m+1:t})\right],$$
$$\widetilde{\text{PerturbLoss}} := \mathbb{E}\left[\sum_{t=m}^{T} \tilde{F}_t(y_{t-m+1:t}) - \tilde{F}_t(x_{t-m+1:t})\right],$$
$$\widehat{\text{MoveCost}} := \mathbb{E}\left[\sum_{t=m}^{T} \hat{F}_t(x_{t-m+1:t}) - \hat{f}_t(x_t)\right],\ \widetilde{\text{MoveCost}} := \mathbb{E}\left[\sum_{t=m}^{T} \tilde{F}_t(x_{t-m+1:t}) - \tilde{f}_t(x_t)\right],$$
$$\widehat{\text{ONSRegret}}(x) := \mathbb{E}\left[\sum_{t=m}^{T} \hat{f}_t(x_t) - \hat{f}_t(x)\right],\ \widetilde{\text{ONSRegret}}(x) := \mathbb{E}\left[\sum_{t=m}^{T} \tilde{f}_t(x_t) - \tilde{f}_t(x)\right].$$

Bounding perturbation loss. Similar to the analysis in Section E.2, we have

$$\begin{aligned}
\widetilde{\text{PerturbLoss}} &= \frac{1}{2}\sum_{t=m}^{T}\sum_{i=1}^{m}\mathbb{E}\left[u_{t-m+i}^\top \hat{A}_{t-m+i-1}^{-\frac{1}{2}}[\nabla^2 \tilde{F}_t]_{ii}\hat{A}_{t-m+i-1}^{-\frac{1}{2}}u_{t-m+i}\right]\\
&\leq \frac{\beta}{2}\sum_{t=m}^{T}\sum_{i=1}^{m}\mathbb{E}\left[\left\|G^{[m-i]}\hat{Y}_{t-m+i}\hat{A}_{t-m+i-1}^{-\frac{1}{2}}u_{t-m+i}\right\|_2^2\right]\\
&\leq \underbrace{\beta\sum_{t=m}^{T}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\hat{G}^{[m-i]}\hat{Y}_{t-m+i}\hat{A}_{t-m+i-1}^{-\frac{1}{2}}u_{t-m+i}\right\|_2^2\right]}_{\leq 2B(\widehat{\text{PerturbLoss}})}\\
&\quad + \beta\sum_{t=m}^{T}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\left(\hat{G}^{[m-i]} - G^{[m-i]}\right)\hat{Y}_{t-m+i}\hat{A}_{t-m+i-1}^{-\frac{1}{2}}u_{t-m+i}\right\|_2^2\right].
\end{aligned}$$

The bound on the first term can be derived identically as in Proposition 21. The second term can be bounded as

$$\beta\sum_{t=m}^{T}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\left(\hat{G}^{[m-i]} - G^{[m-i]}\right)\hat{Y}_{t-m+i}\hat{A}_{t-m+i-1}^{-\frac{1}{2}}u_{t-m+i}\right\|_2^2\right] \leq \varepsilon_G^2\beta R_{\hat{Y}}^2 T.$$

Bounding movement cost. The bound for movement cost is simply the Lipschitz constant multiplying the Euclidean distance between iterates. Therefore, let $\hat{L}$ denote the Lipschitz constant for $\hat{F}_t$, respectively, we have

$$\widetilde{\text{MoveCost}} \leq \frac{\max_{t\in[T],M_{t-m+1:t}}\|\nabla\tilde{F}_t(M_{t-m+1:t})\|_2}{\hat{L}}B(\widehat{\text{MoveCost}}) \leq 2B(\widehat{\text{MoveCost}}),$$

where the last inequality follows from Lemma 32 and assuming $\varepsilon_G \leq (c(\beta,\kappa,G,\hat{G},D,\hat{Y})\sqrt{m})^{-1}L$.

Bounding underlying regret. By construction, we have $\alpha\tilde{H}_t \preceq \nabla^2\tilde{f}_t \preceq \beta\tilde{H}_t$ and $\alpha\hat{H}_t \preceq \nabla^2\hat{f}_t \preceq \beta\hat{H}_t$, and thus $\forall x \in \mathcal{K}$,

$$\begin{aligned}
\widetilde{\text{ONSRegret}}(x) &\leq \sum_{t=m}^{T} \mathbb{E}\left[\nabla\tilde{f}_t(x_t)^\top(x_t - x) - \frac{1}{2}\|x_t - x\|^2_{\alpha\tilde{H}_t}\right] \\
&= \sum_{t=m}^{T} \mathbb{E}\left[\nabla\hat{f}_t(x_t)^\top(x_t - x) - \frac{1}{6}\|x_t - x\|^2_{\alpha\hat{H}_t}\right] \\
&\quad + \sum_{t=m}^{T} \mathbb{E}\left[(\nabla\tilde{f}_t(x_t) - \nabla\hat{f}_t(x_t))^\top(x_t - x)\right] + \frac{\alpha}{6}\sum_{t=m}^{T}\mathbb{E}\left[\|x_t - x\|^2_{\hat{H}_t - 3\tilde{H}_t}\right] \\
&\leq B(\widehat{\text{ONSRegret}}(x)) + \sum_{t=m}^{T} \mathbb{E}\left[(\nabla\tilde{f}_t(x_t) - \nabla\hat{f}_t(x_t))^\top(x_t - x)\right] - \frac{\alpha}{6}\sum_{t=m}^{T}\mathbb{E}\left[\|x_t - x\|^2_{\tilde{H}_t}\right] \\
&\quad + \frac{\alpha}{3}\sum_{t=m}^{T}\mathbb{E}\left[\|(\hat{G}_t - \tilde{G}_t)^\top(x_t - x)\|_2^2\right] \\
&\leq B(\widehat{\text{ONSRegret}}(x)) + \sum_{t=m}^{T} \mathbb{E}\left[(\nabla\tilde{f}_t(x_t) - \nabla\hat{f}_t(x_t))^\top(x_t - x)\right] - \frac{\alpha}{6}\sum_{t=m}^{T}\mathbb{E}\left[\|x_t - x\|^2_{\tilde{H}_t}\right] \\
&\quad + \frac{\alpha\varepsilon_G^2 R_{\hat{Y}}^2 D^2 T}{3}.
\end{aligned}$$

By Lemma 32, we have $\forall t$,

$$\begin{aligned}
\frac{1}{2}(\nabla\tilde{f}_t(x_t) - \nabla\hat{f}_t(x_t))^\top(x_t - x) &= \underbrace{x_t^\top(\tilde{G}_t - \hat{G}_t)^\top P_t^\top \tilde{G}_t(x_t - x)}_{(1)} + \underbrace{x_t^\top \hat{G}_t^\top P_t^\top(\tilde{G}_t - \hat{G}_t)(x_t - x)}_{(2)} \\
&\leq \underbrace{\frac{6}{\alpha}\beta^2 D^2\varepsilon_G^2 + \frac{\alpha}{24}\|x_t - x\|^2_{\tilde{H}_t}}_{(1)} + \underbrace{\frac{(\beta D R_G)^2 m}{2\nu}\varepsilon_G^2 + \frac{\nu}{2m}\sum_{i=0}^{m-1}\|\hat{Y}_{t-i}(x_t - x)\|_2^2}_{(2)},
\end{aligned}$$

where the inequality follows by applying Cauchy-Schwarz $u^\top v \leq \frac{1}{2\lambda}\|u\|_2^2 + \frac{\lambda}{2}\|v\|_2^2$ with $\lambda = \frac{\alpha}{12}$ for (1) and $\lambda = \frac{\nu}{m}$ for (2).

The desired inequality follows from combining the bounds.

G.3.4. PROOF OF LEMMA 31

Using the same blocking method and definitions of k_j's as in Section E.2, we have

$$\begin{aligned}\frac{2\nu}{m}\sum_{t=m}^{T}\sum_{i=0}^{m-1}\mathbb{E}\left[\|\hat{Y}_{t-i}(x_t-x)\|_2^2\right] &\leq \frac{2\nu}{m}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\sum_{i=0}^{m-1}\mathbb{E}\left[\|\hat{Y}_{t-i}(x_t-x)\|_2^2\right]+2\nu R_{\hat{Y}}^2D^2\tau\\ &\leq \frac{4\nu}{m}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\sum_{i=0}^{m-1}\mathbb{E}\left[\|\hat{Y}_{t-i}(x_{k_j}-x)\|_2^2\right]\\ &\quad+\frac{4\nu}{m}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\sum_{i=0}^{m-1}\mathbb{E}\left[\|\hat{Y}_{t-i}(x_t-x_{k_j})\|_2^2\right]+2\nu R_{\hat{Y}}^2D^2\tau.\end{aligned}$$

Applying similar analysis as in Section E.2,

$$\begin{aligned}\frac{4\nu}{m}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\sum_{i=0}^{m-1}\mathbb{E}\left[\|\hat{Y}_{t-i}(x_{k_j}-x)\|_2^2\right] &\leq \frac{8\nu}{\kappa(G)}\sum_{j=1}^{J}\left[\left(\sum_{t=k_j}^{k_{j+1}-1}\mathbb{E}\left[\|x_{k_j}-x\|_{\tilde{H}_t}^2\right]\right)+5mR_GR_{\hat{Y}}D^2\right]\\ &\leq \frac{8\nu}{\kappa(G)}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\mathbb{E}\left[\|x_{k_j}-x\|_{\tilde{H}_t}^2\right]+\frac{40\nu mR_GR_{\hat{Y}}D^2}{\kappa(G)}\left\lfloor\frac{T}{\tau}\right\rfloor.\end{aligned}$$

On the other hand,

$$\frac{\alpha}{12}\sum_{t=m}^{T}\mathbb{E}\left[\|x_t-x\|_{\tilde{H}_t}^2\right]\geq\frac{\alpha}{24}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\mathbb{E}\left[\|x_{k_j}-x\|_{\tilde{H}_t}^2\right]-\frac{\alpha}{12}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\mathbb{E}\left[\|x_t-x_{k_j}\|_{\tilde{H}_t}^2\right].$$

By choice of $\nu=\frac{\alpha\kappa(G)}{192}$, the left hand side of the desired inequality is upper bounded by

$$\underbrace{\frac{\alpha\kappa(G)}{48m}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\sum_{i=0}^{m-1}\mathbb{E}\left[\|\hat{Y}_{t-i}(x_t-x_{k_j})\|_2^2\right]}_{(1)}+\underbrace{\frac{\alpha}{12}\sum_{j=1}^{J}\sum_{t=k_j}^{k_{j+1}-1}\mathbb{E}\left[\|x_t-x_{k_j}\|_{\tilde{H}_t}^2\right]}_{(2)}+2\nu R_{\hat{Y}}^2D^2\tau+\frac{40\nu mR_GR_{\hat{Y}}D^2}{\kappa(G)}\left\lfloor\frac{T}{\tau}\right\rfloor.$$

We bound the first two terms. First, note that assuming $\frac{\eta\alpha\varepsilon_GR_{\hat{Y}}T}{6}\leq 1$, we have

$$\begin{aligned}\hat{H}_t &\preceq \tilde{H}_t+\sum_{i=0}^{m-1}(\hat{G}^{[i]}-G^{[i]})\hat{Y}_{t-i}\preceq\tilde{H}_t+\varepsilon_GR_{\hat{Y}}I,\\ \hat{A}_t &= mI+\frac{\eta\alpha}{6}\sum_{s=m}^{t}\hat{H}_t\preceq mI+\frac{\eta\alpha}{6}\sum_{s=m}^{t}\tilde{H}_t+\frac{\eta\alpha}{6}\cdot\varepsilon_GR_{\hat{Y}}tI\preceq 2\tilde{A}_t.\end{aligned}$$

First, we bound (2). Note that

$$\begin{aligned}
\|x_t - x_{k_j}\|_{\tilde{H}_t}^2 &\le 2\|\tilde{G}_t \tilde{A}_t^{-\frac{1}{2}} \hat{A}_t^{\frac{1}{2}} (x_t - x_{k_j})\|_2^2 \\
&\le 2\mathrm{tr}(\tilde{A}_t^{-\frac{1}{2}} \tilde{H}_t \tilde{A}_t^{-\frac{1}{2}}) \|x_t - x_{k_j}\|_{\hat{A}_t}^2 \\
&\le \frac{12}{\eta\alpha} \frac{\log(|\tilde{A}_t|)}{\log(|\tilde{A}_{t-1}|)} \cdot \tau \sum_{s=k_j+1}^{t} \|x_s - x_{s-1}\|_{\hat{A}_t}^2 \\
&\le \frac{12}{\eta\alpha} \frac{\log(|\tilde{A}_t|)}{\log(|\tilde{A}_{t-1}|)} \cdot \tau \max\{2, \eta\alpha\hat{\sigma}(\tau+m)\} \sum_{s=k_j+1}^{t} \|x_s - x_{s-1}\|_{\hat{A}_{s-m}}^2 \\
&\le \frac{12}{\eta\alpha} \frac{\log(|\tilde{A}_t|)}{\log(|\tilde{A}_{t-1}|)} \cdot \tau^2 \max\{2, \eta\alpha\hat{\sigma}(\tau+m)\} \eta^2 d^2 (\hat{B}^*)^2 m^2.
\end{aligned}$$

Now, by letting $\tau = \lfloor \sqrt{T} \rfloor$, and assuming that $\eta \le \frac{1}{\alpha\hat{\sigma}\sqrt{T}}$, we have that

$$\|x_t - x_{k_j}\|_{\tilde{H}_t}^2 \le \frac{24\eta\tau^2 d^2 (\hat{B}^*)^2 m^2}{\alpha} \cdot \frac{\log(|\tilde{A}_t|)}{\log(|\tilde{A}_{t-1}|)}.$$

Thus,

$$(2) \le \sum_{t=m}^{T} \mathbb{E}\left[\|x_t - x_{k_j}\|_{\tilde{H}_t}^2\right] \le \frac{24\eta\tau^2 d^2 (\hat{B}^*)^2 m^2}{\alpha} \sum_{t=m}^{T} \frac{\log(|\tilde{A}_t|)}{\log(|\tilde{A}_{t-1}|)} \le \frac{24\eta d^3 (\hat{B}^*)^2 m^2 T}{\alpha} \log(\tilde{\sigma} T).$$

To see the bound on (1): $\forall 0 \le i \le m-1$,

$$\begin{aligned}
\|\hat{Y}_{t-i}(x_t - x_{k_j})\|_2^2 &= \|\hat{Y}_{t-i} \hat{A}_t^{-\frac{1}{2}} \hat{A}_t^{\frac{1}{2}} (x_t - x_{k_j})\|_2^2 \\
&\le \mathrm{tr}(\hat{A}_t^{-\frac{1}{2}} \hat{Y}_{t-i}^\top \hat{Y}_{t-i} \hat{A}_t^{-\frac{1}{2}}) \|x_t - x_{k_j}\|_{\hat{A}_t}^2 \\
&\le \mathrm{tr}(\hat{A}_{t-i}^{-\frac{1}{2}} \hat{Y}_{t-i}^\top \hat{Y}_{t-i} \hat{A}_{t-i}^{-\frac{1}{2}}) \cdot 2\eta^2 \tau^2 d^2 (\hat{B}^*)^2 m^2,
\end{aligned}$$

where the last inequality follows similarly from before. Summing over, we have that

$$\begin{aligned}
(1) &\le 2\eta^2 \tau^2 d^2 (\hat{B}^*)^2 m^2 \sum_{t=m}^{T} \sum_{i=0}^{m-1} \mathrm{tr}(\hat{A}_{t-i}^{-\frac{1}{2}} \hat{Y}_{t-i}^\top \hat{Y}_{t-i} \hat{A}_{t-i}^{-\frac{1}{2}}) \\
&\le 4\eta^2 \tau^2 d^2 (\hat{B}^*)^2 m \sum_{t=m}^{T} \mathrm{tr}\left(\hat{A}_t^{-\frac{1}{2}} \left(\sum_{i=0}^{m-1} \hat{Y}_{t-i}^\top \hat{Y}_{t-i}\right) \hat{A}_t^{-\frac{1}{2}}\right).
\end{aligned}$$

By the same analysis as in Section E.2,

$$\sum_{t=m}^{T} \mathrm{tr}\left(\hat{A}_t^{-\frac{1}{2}} \left(\sum_{i=0}^{m-1} \hat{Y}_t^\top \hat{Y}_t\right) \hat{A}_t^{-\frac{1}{2}}\right) \le \frac{10 m R_{\hat{G}} R_{\hat{Y}}}{\kappa(\hat{G})} \sqrt{T} + \frac{4md \log(\hat{\sigma} T))}{\eta\alpha\kappa(\hat{G})} + d R_{\hat{Y}}^2 \sqrt{T}.$$

The result follows from taking $\eta = \tilde{\Theta}(\frac{1}{\alpha\sqrt{T}})$, $\tau = \lfloor \sqrt{T} \rfloor$, in which case

$$(1) = \tilde{O}(\sqrt{T}), \quad (2) = \tilde{O}(\sqrt{T}), \quad 2\nu R_{\hat{Y}}^2 D^2 \tau + \frac{40\nu m R_G R_{\hat{Y}} D^2}{\kappa(G)} \left\lfloor \frac{T}{\tau} \right\rfloor = \tilde{O}(\sqrt{T}).$$

G.4. Discussion: unknown stabilizing controllers

To further generalize the results in the previous sections, we believe that the same regret bound can be attained for unknown system with unknown stabilizing controller K. Here, we sketch the idea to arrive at the same regret bound but omit the proof. In this case, we further assume that the system is strongly controllable and observable.

Definition 33 (Strong controllability and observability) *A system (A, B) is (k, κ)-strongly controllable if the matrix*

$$C_k = \begin{bmatrix} B & AB & \dots & A^{k-1}B \end{bmatrix} \in \mathbb{R}^{d_{\mathbf{x}} \times k d_{\mathbf{u}}}$$

has full row rank, and $\|(C_k C_k^\top)^{-1}\|_2 \leq \kappa$. A partially observable system (A, B, C) is observable if the matrix

$$O = \begin{bmatrix} C^\top & (CA)^\top & \dots & (CA^{d_{\mathbf{x}}-1})^\top \end{bmatrix}^\top \in \mathbb{R}^{d_{\mathbf{x}} d_{\mathbf{y}} \times d_{\mathbf{x}}}$$

has full column rank.

We first need a Markov operator estimator, which we can run a variant of the system estimation algorithm 1 in Chen and Hazan (2021) to obtain a good estimator for $G_{\mathbf{y}}$. Then, we run the Ho-Kalman algorithm to extract $\hat{A}, \hat{B}, \hat{C}$ that are close to A, B, C up to linear transformations by unitary matrices. With the estimated $\hat{A}, \hat{B}, \hat{C}$, we can run a variant of the controller recovery algorithm in Chen and Hazan (2021) to obtain a a controller $\hat{K}$. With the $\hat{G}, \hat{K}$, we can run Algorithm 4.

Appendix H. BCO-M lower bound: Proof of Theorem 6 (Section 5)

In this section, we prove Theorem 6. We start by constructing the sequence of loss functions. On a high level, the loss function f_t has memory length of 2 and has three different components: a scaled linear loss of the difference between the current iterate and the iterate two steps before, a random variable from a random process indexed at time t, and a quadratic moving cost between the current and previous iterates. Concretely, the loss function takes the following form:

$$f_t(x_{t-1}, x_t) = \varepsilon \ell^*(x_t - x_{t-2}) + \bar{n}_t + (x_t - x_{t-1})^2, \tag{11}$$

where ℓ^* takes values in $\{-1, 1\}$ with equal probability and is fixed ahead of time but unknown to the learner; ε is some scaling factor; $\bar{n}_t$ is a multi-scaled random walk indexed at time t. We view f_t as an adaptive loss function to the learner's decision at time $t-2$.

One important distinction we make is the notion of regret in the presence of an adaptive adversary. In the presence of an adaptive adversary described in Assumption 9, a function f_t of memory length m is itself a function of the decision $x_{1:t-m}$. Therefore, we can write $f_t(x_{t-m+1:t}) = F_t(x_{1:t})$. We are interested in the following regret definition: for $x \in \mathcal{K}$, the regret w.r.t. x is given by

$$\sum_{t=1}^{T} F_t(x_{1:t}) - \sum_{t=1}^{T} F_t(x_{1:t-m}, x, \dots, x). \tag{12}$$

This is the common regret notion in literature (e.g. Auer et al. (2002); McMahan and Blum (2004)). Cesa-Bianchi et al. (2013) compares this regret notion with a stronger *policy regret* defined as

$$\sum_{t=1}^{T} F_t(x_{1:t}) - \sum_{t=1}^{T} F_t(x,\ldots,x),$$

and showed that the latter has a lower bound of $\Omega(T)$. Therefore, we focus on establishing a lower bound with respect to the regret defined in Eq. (12).

With the notion of regret in Eq. (12), the regret of loss functions in Eq. (11) is given by $\sum_{t=1}^{T} f_t(x_{t-1}, x_t) - \min_x \sum_{t=1}^{T} f_t(x,x)$. The optimal strategy is thus determined by the value of ℓ^*. In the bandit setting, the learner has no access to ℓ^* and for sufficiently small ε, the information regarding the sign of ℓ^* through the scaler feedback is limited. The tradeoff between moving decisions to learn ℓ^* and incurring moving cost results in the suboptimal bounds derived in Theorem 6.

This section will be organized as the following: Appendix H.1 spells out the construction of the multi-scaled random walk used in the construction of f_t in Eq. (11) and its useful properties; Appendix H.2 establishes the regularity of the constructed loss function (in particular it is bounded with high probability); Appendix H.3 proves the regret lower bound in Theorem 6.

H.1. Construction of multi-scaled random walk and loss functions

We consider the multi-scaled random walk constructed in Dekel et al. (2014).

Definition 34 (Multi-scaled random walk) *Given a time horizon $T \in \mathbb{Z}_{++}$, let $\xi_1,\ldots,\xi_T$ be i.i.d. $N(0,\sigma^2)$ random variables for some $\sigma \in \mathbb{R}_{++}$. The multi-scaled random walk is a random process $\{\bar{n}_t\}_{t=1}^T$ defined by*

$$\bar{n}_t = \bar{n}_{\rho(t)} + \xi_t,$$
$$\rho(t) = t - \max\{2^i : i \geq 0,\ 2^i \text{ divides } t\}.$$

Given the definition of the process $\{\bar{n}_t\}_{t=1}^T$, consider the sequence of loss functions that is adaptive to the learner's decisions at time $t-2$ for every t.

Definition 35 (Loss instance) *Consider the $(t-2)$-adaptive loss function $L_t(\cdot;\varepsilon) : [0,1] \to \mathbb{R}$ parameterized by $\varepsilon > 0$ at time t:*

$$L_t(x;\varepsilon) = \varepsilon \ell^*(x - x_{t-2}) + \bar{n}_t,$$

where ℓ^* is a Rademacher random variable chosen before the game starts unknown to the learner, i.e. $\ell^* = \pm 1$ with probability $\frac{1}{2}$ equally.

We define the with-memory loss functions as the sum of the loss instance defined in Definition 35 and the quadratic moving cost between x_t and x_{t-1}.

Definition 36 (Regret, moving cost, and performance metric) *The additional moving cost takes the form*

$$M = \sum_{t=1}^{T} (x_t - x_{t-1})^2.$$

We assume that the learner has full access to M. Note that this only makes the problem easier. Define the with-memory quadratic loss functions over $[0,1] \times [0,1]$ as

$$f_t(x_{t-1}, x_t) = L_t(x_t, \varepsilon) + (x_t - x_{t-1})^2.$$

Let $\delta > 0$ be the improper learning parameter. The regret w.r.t. any $x \in [\delta, 1-\delta]$ is measured by

$$\begin{aligned}\mathbb{E}[\textit{Regret}_T(x)] &= \mathbb{E}\left[\sum_{t=1}^{T} f_t(x_{t-1}, x_t) - \sum_{t=1}^{T} f_t(x, x)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} L_t(x_t; \varepsilon) + M - \sum_{t=1}^{T} L_t(x; \varepsilon)\right].\end{aligned}$$

Remark 37 *Note that the f_t defined in Definition 36 is quadratic, smooth, and adaptive to the learner's decision at time $t-2$. In particular, this fits the assumption on the adversary model in Section 4.*

The following property of the multi-scaled random walk (MRW) constructed in Definition 34 is proved in Lemma 2 in Dekel et al. (2014). To summarize the proof, we note that the depth and width defined in Lemma 38 is bounded by the number of 1's and 0's in the binary representation of t respectively, which are bounded by $\lfloor \log_2 T \rfloor + 1$.

Lemma 38 (Depth and width of MRW, Lemma 2 in Dekel et al. (2014)) *The parent set of t is defined by $\rho^*(t) = \{\rho(t)\} \cup \rho^*(\rho(t))$, with $\rho^*(1) = \emptyset$. For a fixed t, the cut of t is defined by the number of integer s such that t lies between s and its parent, i.e. $\mathrm{cut}(t) = \{s \in [T] : \rho(s) < t \le s\}$. The depth $d(\rho)$ and width $w(\rho)$ of the multi-scaled random walk defined in Definition 34 are defined and bounded by*

$$\begin{aligned} d(\rho) &= \max_{t \in [T]} |\rho^*(t)| \le \lfloor \log_2 T \rfloor + 1, \\ w(\rho) &= \max_{t \in [T]} |\mathrm{cut}(t)| \le \lfloor \log_2 T \rfloor + 1. \end{aligned}$$

H.2. Bound on loss functions

To establish a meaningful lower bound, we need to have f_t be bounded. It is unclear that f_t's are bounded due to the component of $\bar{n}_t$ in $L_t(x_t, \varepsilon)$. However, by the bound on the depth of $\{\bar{n}_t\}_{t=1}^T$, we argue in this section that by setting $\sigma \sim \frac{1}{\log T}$, we can without loss of generality assume that $f_t \in [-3, 3]$. In particular, Lemma 38 implies a tail bound on process $\{\bar{n}_t\}_{t=1}^T$, which directly follows from tail bounds for the sum of i.i.d. Gaussian random variables.

Lemma 39 *The random process $\{\bar{n}_t\}_{t=1}^T$ obeys the tail bound that $\forall \gamma \in (0, 1)$,*

$$\mathbb{P}\left(\max_{1 \le t \le T} |\bar{n}_t| > \sigma \cdot 2 \log\left(\frac{T}{\gamma}\right)\right) \le \gamma.$$

Proof By Gaussian tail bound we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le t\le T}|\bar{n}_t| > \sigma\cdot 2\log\left(\frac{T}{\gamma}\right)\right)\\
&\le \sum_{t=1}^{T}\mathbb{P}\left(\left|\sum_{s\in\{\rho^*(t)\cup\{t\}\}}\xi_s\right| > \sigma\cdot 2\log\left(\frac{T}{\gamma}\right)\right)\\
&\le \sum_{t=1}^{T}\exp\left(-\log\frac{T}{\gamma}\right)\\
&=\gamma,
\end{aligned}$$

since $\sum_{s\in\{\rho^*(t)\cup\{t\}\}}\xi_s$ is $N(0,(|\rho^*(t)|+1)\sigma^2)$ distributed. ■

Note that by definition of L_t, Lemma 39 implies that by setting

$$\sigma = \frac{1}{2\log T},$$

and $\varepsilon \le 1$, $\gamma = \frac{1}{T}$, we have with probability at least $1-\frac{1}{T}$,

$$\max_{t\in[T]}|L_t(x_t;\varepsilon)| \le 1 + \max_{t\in[T]}|\bar{n}_t| \le 2.$$

Thus, we can without loss of generality assume that $L_t(x_t;\varepsilon)$ is indeed bounded between $[-3,3]$ for all t: note that since the addition $\bar{n}_t$ to $L_t(x;\varepsilon)$ cancels for the learner and the comparator, the regret is always bounded by

$$\text{Regret}_T = \sum_{t=1}^{T}\varepsilon\ell^*(x_t - x^*) + M \le (\varepsilon+1)T \le 2T.$$

Let A denote the event when $L_t(x_t;\varepsilon)\in[-3,3], \forall t\in[T]$, then if $\mathbb{E}[\text{Regret}_T] = \tilde{\Omega}(T^{\frac{2}{3}})$, since

$$2\gamma T + \mathbb{E}[\text{Regret}_T \mid A] \ge \mathbb{E}[\text{Regret}_T] = \tilde{\Omega}(T^{\frac{2}{3}}),$$

we have that $\mathbb{E}[\text{Regret}_T \mid A] = \tilde{\Omega}(T^{\frac{2}{3}})$.

H.3. $T^{\frac{2}{3}}$-lower bound

We prove a lower bound against any randomized algorithm. For any randomized algorithm, the decision random variable X_t played at time t is a random function of the observations $Y_1,\dots,Y_{t-1}$, where $Y_t = L_t(x_t;\varepsilon)$. Note that Y_t decomposes as

$$Y_t = Y_{\rho(t)} + \varepsilon\ell^*(X_t - X_{t-2}) - \varepsilon\ell^*(X_{\rho(t)} - X_{\rho(t)-2}) + \xi_t.$$

Then, conditioning on $Y_{1:t-1}$ and $X_{1:t}$,

$$\begin{aligned}
Y_t \mid (Y_{1:t-1}, X_{1:t}, \ell^*=1) &\stackrel{D}{=} N\left(Y_{\rho(t)} + \varepsilon(X_t - X_{t-2}) - \varepsilon(X_{\rho(t)} - X_{\rho(t)-2}), \sigma^2\right),\\
Y_t \mid (Y_{1:t-1}, X_{1:t}, \ell^*=-1) &\stackrel{D}{=} N\left(Y_{\rho(t)} - \varepsilon(x_t - x_{t-2}) + \varepsilon(x_{\rho(t)} - x_{\rho(t)-2}), \sigma^2\right).
\end{aligned}$$

We are interested in the KL divergence of the distribution $\mathbb{Q}_0^{t|t-1}$ and $\mathbb{Q}_1^{t|t-1}$ of the random variable Y_t. The KL divergence between two uni-variate Gaussian distributions with the same variance σ^2 and different expectation μ_1, μ_2 is given by

$$D_{\mathrm{KL}}(N(\mu_1, \sigma^2) \parallel N(\mu_2, \sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

For finite sequence of random variables $\{X_s\}_{s\in S_X}, \{Y_s\}_{s\in S_Y}$ mapping from $(\Omega, \Sigma, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathrm{Leb}_{\mathbb{R}})$, we slightly abuse notation and denote as $\mathbb{P}(\{X_s\}_{s\in S_X} \mid \{Y_s\}_{s\in S_Y})$ the joint distribution of $\{X_s\}_{s\in S_X}$ conditioning on the sub product σ-algebra of Σ generated by the sequence of random variables $\{Y_s\}_{s\in S_Y}$. We have then

$$\begin{aligned}
&D_{\mathrm{KL}}(\mathbb{P}(Y_t \mid (Y_{1:t-1}, X_{1:t}, \ell^* = 1)) \parallel \mathbb{P}(Y_t \mid (Y_{1:t-1}, X_{1:t}, \ell^* = -1)))\\
&= \frac{(2\varepsilon(X_t - X_{t-2}) - 2\varepsilon(X_{\rho(t)} - X_{\rho(t)-2}))^2}{2\sigma^2}\\
&\le \frac{8\varepsilon^2(X_t - X_{t-2})^2 + 8\varepsilon^2(X_{\rho(t)} - X_{\rho(t)-2})^2}{2\sigma^2}\\
&\le \frac{8\varepsilon^2}{\sigma^2}[(X_t - X_{t-1})^2 + (X_{t-1} - X_{t-2})^2 + (X_{\rho(t)} - X_{\rho(t)-1})^2 + (X_{\rho(t)-1} - X_{\rho(t)-2})^2], \quad (13)
\end{aligned}$$

where we use the fact that $(a+b)^2 \le 2(a^2+b^2)$. The KL divergence satisfies the chain rule stated in the following lemma.

Lemma 40 (Chain rule for KL divergence) *Let X, Y be two random variables, and let $\mathbb{P}$, $\mathbb{Q}$ be two joint distribution over X, Y. Let $\mathbb{P}_X$, $\mathbb{Q}_X$ be the marginal distribution of X under $\mathbb{P}$, $\mathbb{Q}$, respectively, and let $\mathbb{P}_{Y|X}$, $\mathbb{Q}_{Y|X}$ be the conditional distribution of Y given X under $\mathbb{P}$, $\mathbb{Q}$. Then,*

$$D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{Q}) = D_{\mathrm{KL}}(\mathbb{P}_X \parallel \mathbb{Q}_X) + \mathbb{E}_{X\sim\mathbb{P}_X}[D_{\mathrm{KL}}(\mathbb{P}_{Y|X} \parallel \mathbb{Q}_{Y|X})].$$

By Lemma 40, we can bound the KL divergence of the joint distribution of $Y_{1:T}, X_{1:T}$.

$$\begin{aligned}
&D_{\mathrm{KL}}(\mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = 1) \parallel \mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = -1)) && (14)\\
&= \underbrace{D_{\mathrm{KL}}(\mathbb{P}(Y_{1:T-1}, X_{1:T} \mid \ell^* = 1) \parallel \mathbb{P}(Y_{1:T-1}, X_{1:T} \mid \ell^* = -1))}_{(1)} && (15)\\
&\quad + \mathbb{E}_{Y_{1:T-1}, X_{1:T}|\ell^*=1}[D_{\mathrm{KL}}(\mathbb{P}(Y_T \mid Y_{1:T-1},, X_{1:T}, \ell^* = 1) \parallel \mathbb{P}(Y_T \mid Y_{1:T-1},, X_{1:T}, \ell^* = -1)],
\end{aligned}$$

where (1) can be further decomposed as

$$\begin{aligned}
&(1)\\
&\quad = D_{\mathrm{KL}}(\mathbb{P}(Y_{1:T-1}, X_{1:T-1} \mid \ell^* = 1) \parallel \mathbb{P}(Y_{1:T-1}, X_{1:T-1} \mid \ell^* = -1))\\
&\qquad + \mathbb{E}_{Y_{1:T-1}, X_{1:T-1}|\ell^*=1}\underbrace{[D_{\mathrm{KL}}(\mathbb{P}(X_T \mid Y_{1:T-1}, X_{1:T-1}, \ell^* = 1) \parallel \mathbb{P}(X_T \mid Y_{1:T-1}, X_{1:T-1}, \ell^* = -1))]}_{(2)}.
\end{aligned}$$

Note that the distribution of $X_T \mid Y_{1:T-1}, X_{1:T-1}$ is independent of ℓ^* since X_T depends on ℓ^* only through $Y_{1:T-1}$, thus we have $(2) = 0$. Recursively apply this argument, we have that

$$\begin{aligned}
&D_{\mathrm{KL}}(\mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = 1) \parallel \mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = -1)) \qquad (16)\\
&= \sum_{t=1}^{T} \mathbb{E}_{Y_{1:t-1}, X_{1:t} \mid \ell^*=1}[D_{\mathrm{KL}}(\mathbb{P}(Y_t \mid Y_{1:t-1,}, X_{1:t}, \ell^* = 1) \parallel \mathbb{P}(Y_t \mid Y_{1:t-1,}, X_{1:t}, \ell^* = -1)]\\
&\leq \frac{8\varepsilon^2}{\sigma^2} \sum_{t=1}^{T} \mathbb{E}_{X_{1:t}}[(X_t - X_{t-1})^2 + (X_{t-1} - X_{t-2})^2 + (X_{\rho(t)} - X_{\rho(t)-1})^2\\
&\qquad + (X_{\rho(t)-1} - X_{\rho(t)-2})^2 \mid \ell^* = 1]\\
&\leq \frac{16\varepsilon^2}{\sigma^2} \sum_{t=1}^{T} \mathbb{E}_{X_{1:t}}[(X_t - X_{t-1})^2 + (X_{\rho(t)} - X_{\rho(t)-1})^2 \mid \ell^* = 1]\\
&\leq \frac{16(\lfloor \log_2 T \rfloor + 2)\varepsilon^2}{\sigma^2} \mathbb{E}_{X_{1:T}}[M \mid \ell^* = 1], \qquad (17)
\end{aligned}$$

where the first inequality follows from Eq. (13). To see the last inequality, note that we can bound

$$\begin{aligned}
&\mathbb{E}_{X_{1:T}}\left[\sum_{t=1}^{T} (X_{\rho(t)} - X_{\rho(t)-1})^2 \mid \ell^* = 1\right]\\
&\leq \left(\max_{t \in [T]} |\{s \in [T] : \rho(s) = t\}|\right) \mathbb{E}_{X_{1:T}}\left[\sum_{t=1}^{T} (X_t - X_{t-1})^2 \mid \ell^* = 1\right]\\
&\leq w(\rho) \mathbb{E}_{X_{1:T}}\left[\sum_{t=1}^{T} (X_t - X_{t-1})^2 \mid \ell^* = 1\right],
\end{aligned}$$

where the last inequality follows since if for some t, $|\{s \in [T] : \rho(s) = t\}| = n \geq 1$, then let $\{s_1, \ldots, s_n\}$ be set of all $s \in [T]$ satisfying $\rho(s) = t$ in increasing order. We note that $\mathrm{cut}(s_1) \geq n$ since $\forall i \in \{1, \ldots, n\}$, $t = \rho(s_i) = \rho(s_1) < s_1 \leq s_i$. Therefore, we have $w(\rho) \geq \max_{t \in [T]} |\{s \in [T] : \rho(s) = t\}|$.

The KL divergence gives a bound on the probability measure of any measurable event A through Pinsker's inequality, stated below:

Lemma 41 (Pinsker's inequality) *Let $\mathbb{P}, \mathbb{Q}$ be two probability distributions on a measurable space $(Y, \mathbb{F})$, then we have*

$$\sup_{A \in \mathbb{F}} \{|\mathbb{P}(A) - \mathbb{Q}(A)|\} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{Q})}.$$

Let $\mathbb{F}$ be the σ-algebra generated by the random variables $Y_{1:T}, X_{1:T}$. Then, Pinsker's inequality applied to (17),

$$\begin{aligned}
&\sup_{A \in \mathbb{F}} \{|\mathbb{P}(A \mid \ell^* = 1) - \mathbb{P}(A \mid \ell^* = -1)|\}\\
&\leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = 1) \parallel \mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = -1))}\\
&\leq \frac{\sqrt{8(\lfloor \log_2 T \rfloor + 2)}\varepsilon}{\sigma} \sqrt{\mathbb{E}[M \mid \ell^* = 1]}.
\end{aligned}$$

Similar to the above analysis, if we switch $\mathbb{Q}_0$ and $\mathbb{Q}_1$, we will have a symmetrical inequality:

$$\begin{aligned}&\sup_{A\in\mathbb{F}}\{|\mathbb{P}(A \mid \ell^* = -1) - \mathbb{P}(A \mid \ell^* = 1)|\}\\ &\le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = -1) \parallel \mathbb{P}(Y_{1:T}, X_{1:T} \mid \ell^* = 1))}\\ &\le \frac{\sqrt{8(\lfloor \log_2 T \rfloor + 2)}\varepsilon}{\sigma}\sqrt{\mathbb{E}[M \mid \ell^* = -1]}.\end{aligned}$$

Combining the two inequalities and by concavity of the function $f(x) = \sqrt{x}$ for $x \ge 0$,

$$\begin{aligned}&\sup_{A\in\mathbb{F}}\{|\mathbb{P}(A \mid \ell^* = 1) - \mathbb{P}(A \mid \ell^* = -1)|\}\\ &\le \frac{\sqrt{8(\lfloor \log_2 T \rfloor + 2)}\varepsilon}{\sigma}\left(\frac{1}{2}\sqrt{\mathbb{E}[M \mid \ell^* = 1]} + \frac{1}{2}\sqrt{\mathbb{E}[M \mid \ell^* = -1]}\right)\\ &\le \frac{\sqrt{8(\lfloor \log_2 T \rfloor + 2)}\varepsilon}{\sigma}\sqrt{\mathbb{E}[M]}.\end{aligned}$$

Since we view the dependence on x_{t-2} in L_t as an adaptive adversary, and therefore the optimal fixed point $x^* \in [\delta, 1-\delta]$ would be $x^* = \delta$ if $\ell^* = 1$ and $x^* = 1-\delta$ if $\ell^* = -1$, we can express the regret as

$$\begin{aligned}\mathbb{E}[\mathrm{Regret}_T] &\ge \left[\sum_{t=1}^{T}\frac{1}{2}\mathbb{E}[\varepsilon(X_t - \delta) \mid \ell^* = 1] + \frac{1}{2}\mathbb{E}[\varepsilon(1-\delta-X_t) \mid \ell^* = -1]\right] + \mathbb{E}[M]\\ &= \frac{\varepsilon T}{2} - \varepsilon\delta T - \frac{\varepsilon}{2}\left[\sum_{t=1}^{T}\mathbb{E}[X_t \mid \ell^* = 1] - \mathbb{E}[X_t \mid \ell^* = -1]\right] + \mathbb{E}[M].\end{aligned}$$

Denote $Z = \sum_{t=1}^{T} X_t$. Z is by definition measurable with respect to $\mathbb{F}$. Since $X_t \in [0,1]$, $\forall t$, $Z \in [0, T]$. We can further express the sum of differences between the expectation of x_t under $\ell^* = 1$ and $\ell^* = -1$ by

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[X_t \mid \ell^* = 1] - \mathbb{E}[X_t \mid \ell^* = -1] &= \mathbb{E}[Z \mid \ell^* = 1] - \mathbb{E}[Z \mid \ell^* = -1]\\ &= \int_0^T z(\mathbb{P}(Z \in dz \mid \ell^* = 1) - \mathbb{P}(Z \in dz \mid \ell^* = -1))\\ &\le \frac{\sqrt{8(\lfloor \log_2 T \rfloor + 2)}\varepsilon}{\sigma}\sqrt{\mathbb{E}[M]}T,\end{aligned}$$

Therefore, the regret is lower bounded by

$$\mathbb{E}[\mathrm{Regret}_T] \ge \frac{\varepsilon(1-2\delta)T}{2} - \frac{\varepsilon^2 T}{2}\frac{\sqrt{8(\lfloor \log_2 T \rfloor + 2)}}{\sigma}\sqrt{\mathbb{E}[M]} + \mathbb{E}[M],$$

where the right hand side is a quadratic function of $\sqrt{\mathbb{E}[M]}$ minimized at

$$\sqrt{\mathbb{E}[M]} = \frac{\varepsilon^2 T}{4}\frac{\sqrt{8(\lfloor \log_2 T \rfloor + 2)}}{\sigma}$$

with the minimal value equal to

$$\frac{\varepsilon(1-2\delta)T}{2}-\frac{\varepsilon^4T^2(\lfloor\log_2 T\rfloor+2)}{\sigma^2} \geq \frac{\varepsilon(1-2\delta)T}{2}-\frac{2\varepsilon^4T^2\log T}{\sigma^2}.$$

Recall that σ is chosen to be $\sigma=\frac{1}{2\log T}$, and choosing $\varepsilon=\Theta(T^{-1/3}/\log T)$ we have

$$\mathbb{E}[\text{Regret}_T]\geq \frac{\varepsilon(1-2\delta)T}{2}-\frac{2\varepsilon^4T^2\log T}{\sigma^2}=\tilde{\Omega}(T^{2/3}).$$

Appendix I. BCO-M upper bound

In this section, we provide a first-order, proper learning algorithm (Algorithm 5) for BCO-M problems that achieves $\tilde{O}(T^{2/3})$ regret upper bound for convex quadratic, smooth functions that can have adaptivity described in Assumption 9, which matches the lower bound established in Theorem 6. The algorithm is based on the FTRL (Follow-the-Regularized-Leader, Shalev-Shwartz and Singer (2007)) algorithm and makes use of self-concordant barriers over compact convex sets to ensure proper learning.

I.1. Preliminaries on Self-Concordant Barriers

First, we introduce self-concordant barriers, which we will make use of in Algorithm 5.

Definition 42 (Self-concordant barrier) *A $\mathcal{C}^3$ function $R(\cdot)$ over a closed convex set $\mathcal{K}\subset\mathbb{R}^d$ with non-empty interior is a ν-self-concordant barrier of $\mathcal{K}$ if it satisfies the following two properties:*

1. *(Boundary property) For any sequence $\{x_n\}_{n\in\mathbb{N}}\subset \text{int}(\mathcal{K})$ such that $\lim_{n\to\infty}x_n=x\in\partial\mathcal{K}$, $\lim_{n\to\infty}R(x_n)=\infty$.*
2. *(Self-concordant) $\forall x\in\text{int}(\mathcal{K})$, $h\in\mathbb{R}^d$,*
   (a) $|\nabla^3R(x)[h,h,h]|\leq 2|\nabla^2R(x)[h,h]|^{3/2}$.
   (b) $|\langle\nabla R(x),h\rangle|\leq\sqrt{\nu}|\nabla^2R(x)[h,h]|^{1/2}$.

It is well known that self-concordant functions satisfy the following properties:

Proposition 43 (Properties of self-concordant functions, Suggala et al. (2021)) *ν-self-concordant barriers over $\mathcal{K}$ satisfy the following properties:*

1. *Sum of two self-concordant functions is self-concordant. Linear and quadratic functions are self-concordant.*
2. *If $x,y\in\mathcal{K}$ satisfies $\|x-y\|_{\nabla^2R(x)}<1$, then the following inequality holds:*

$$(1-\|x-y\|_{\nabla^2R(x)})^2\nabla^2R(x)\preceq\nabla^2R(y)\preceq\frac{1}{(1-\|x-y\|_{\nabla^2R(x)})^2}\nabla^2R(x). \tag{18}$$

3. *The Dikin ellipsoid centered at any point in the interior of* $\mathcal{K}$ *w.r.t. a self-concordant barrier* $R(\cdot)$ *over* $\mathcal{K}$ *is completely contained in* $\mathcal{K}$*. Namely,*

$$\{y \in \mathbb{R}^d \mid \|y - x\|_{\nabla^2 R(x)} \le 1\} \subset \mathcal{K}, \;\; \forall x \in \textit{int}(\mathcal{K}). \tag{19}$$

where

$$\|v\|_{\nabla^2 R(x)} \overset{def}{=} \sqrt{v^\top \nabla^2 R(x) v}$$

4. $\forall x, y \in \textit{int}(\mathcal{K})$*:*

$$R(y) - R(x) \le \nu \log \frac{1}{1 - \pi_x(y)},$$

where $\pi_x(y) \overset{def}{=} \inf\{t \ge 0 : x + t^{-1}(y - x) \in \mathcal{K}\}$.

I.2. BQO-M algorithm and analysis

We introduce Algorithm 5, the assumptions, and guarantees.

Assumption 12 *We assume the convex compact constraint set* $\mathcal{K} \subset \mathbb{R}^d$ *and the convex, quadratic loss function* $f_t : \mathcal{K}^m \to \mathbb{R}$ *satisfies Assumption 8 and 9 in Section 4.1. Moreover, the Hessian is bounded above and satisfies* $\nabla^2 f_t \preceq \beta I_{dm}$.

Algorithm 5 Bandit Quadratic Optimization with Memory

Input: convex compact set $\mathcal{K}$, step size $\eta > 0$, perturbation parameter $\delta \in [0, 1]$, α-strongly convex ν-self-concordant barrier R over $\mathcal{K}$, memory length $m \in \mathbb{N}$, time horizon $T \in \mathbb{N}$.

1: Initialize: $x_1 = \cdots = x_m = \arg\min_{x \in \mathcal{K}} R(x)$, $\tilde{g}_{0:m-1} = \mathbf{0}_d$, $A_1 = \cdots = A_m = \nabla^2 R(x_1)$.
2: Sample $u_t \sim \mathbb{S}^{d-1}$ i.i.d. uniformly at random for $t = 1, \ldots, m$.
3: Set $y_t = x_t + \delta A_t^{-\frac{1}{2}} u_t$, $t = 1, \ldots, m$.
4: **for** $t = m, \ldots, T$ **do**
5:     Play y_t, observe $f_t(y_{t-m+1:t})$.
6:     Create gradient estimator:

$$\tilde{g}_t = \frac{d}{\delta} f_t(y_{t-m+1:t}) \sum_{j=0}^{m-1} A_{t-j}^{\frac{1}{2}} u_{t-j} \in \mathbb{R}^d.$$

7:     Update $x_{t+1} = \arg\min_{x \in \mathcal{K}} \eta \sum_{s=m}^{t-m+1} \langle \tilde{g}_s, x \rangle + R(x)$, $A_{t+1} = \nabla^2 R(x_{t+1})$.
8:     Independently from previous steps, sample $u_{t+1} \sim \mathbb{S}^{d-1}$ uniformly at random.
9:     Set $y_{t+1} = x_{t+1} + \delta A_{t+1}^{-\frac{1}{2}} u_{t+1}$.
10: **end for**

Before stating the regret guarantee, we note that Algorithm 5 satisfies two properties: proper learning and delayed dependence, given by the following two remarks.

Remark 44 (Correctness) *All y_t's played by Algorithm 5 satisfy that $y_t \in \mathcal{K}$ since $\|y_t - x_t\|^2_{\nabla^2 R(x_t)} = \delta^2 u_t^\top A_t^{-\frac{1}{2}} A_t A_t^{-\frac{1}{2}} u_t = \delta^2 \leq 1$. By Eq. (19) in Proposition 43, $y_t \in \mathcal{K}$.*

Remark 45 (Delayed dependence) *In Algorithm 5, $y_t, \tilde{g}_t$ are $\mathbb{F}_t$-measurable, and x_t, A_t are $\mathbb{F}_{t-m}$-measurable.*

Similar to Theorem 20, the constructed gradient estimators $\tilde{g}_t$ is a conditionally unbiased estimator of the divergence of f_t evaluated at $(x_{t-m+1}, \dots, x_t)$.

Lemma 46 (Unbiased gradient estimator) *$\tilde{g}_t$ is a conditionally unbiased estimator of the following statistics:*

$$\mathbb{E}[\tilde{g}_t \mid \mathbb{F}_{t-m}] = \nabla \cdot f_t(x_{t-m+1:t}),$$

where $\mathbb{F}_t = \sigma(\{u_s\}_{s \leq t})$ is the filtration generated by the algorithm's random sampling up to time t, and $\nabla \cdot f_t(x_{t-m+1:t})$ denotes the divergence of f_t evaluated at $x_{t-m+1:t}$.

Proof Similar as in the proof of Lemma 20. ■

Finally, the regret guarantee of Algorithm 5 is given by the following proposition.

Proposition 47 *For $d, m, T \in \mathbb{N}$, convex compact set $\mathcal{K} \subset \mathbb{R}^d$ and $\{f_t\}_{t=m}^T$ satisfying Assumption 12, α-strongly convex ν-self-concordant barrier function R over $\mathcal{K}$, suppose Algorithm 5 is run with input $(\mathcal{K}, \eta, \delta, \alpha, R, m, T)$. The regret can be decomposed as*

$$\mathbb{E}[\textit{Regret}_T(x)] = \underbrace{\sum_{t=m}^T \mathbb{E}[f_t(y_{t-m+1:t}) - f_t(x_{t-m+1:t})]}_{(1:\textit{exploration loss})} + \underbrace{\sum_{t=m}^T \mathbb{E}[f_t(x_{t-m+1:t}) - \bar{f}_t(x_t)]}_{(2:\textit{movement loss})}$$
$$+ \underbrace{\sum_{t=m}^T \mathbb{E}[\bar{f}_t(x_t) - \bar{f}_t(x)]}_{(3:\textit{underlying regret})},$$

and each of the above terms can be bounded as following

$$(1) \leq C_1 \delta^2 T, \quad (2) \leq C_2 \frac{\eta^2 T}{\delta^2}, \quad (3) \leq C_3 \frac{\eta T}{\delta^2} + C_4 \frac{1}{\eta} + C_5,$$

where $C_1, \dots, C_5$ are given by

$$C_1 = \frac{\beta m}{2\alpha}, \; C_2 = \frac{8\beta d^2 B^2 m^5}{\alpha}, \; C_3 = 16 d^2 B^2 m^3 + \frac{4 d \beta D B m^4}{\sqrt{\alpha}}, \; C_4 = \nu \log T, \; C_5 = mB.$$

In particular, by setting $\delta = T^{-\frac{1}{6}}$, $\eta = T^{-\frac{2}{3}}$, and $m = \mathrm{poly}(\log T)$, we have $\forall x \in \mathcal{K}$, $\mathbb{E}[\textit{Regret}_T(x)] \leq \tilde{O}(T^{\frac{2}{3}})$.

Proof Proposition 47 can be established similarly as Theorem 4, where we bound each of the decomposed loss as follows.

Perturbation loss. The perturbation loss can be bounded by

$$\begin{aligned}\sum_{t=m}^{T}\mathbb{E}[f_t(y_{t-m+1:t})-f_t(x_{t-m+1:t})] &\leq \sum_{t=m}^{T}\delta\mathbb{E}[\nabla f_t(x_{t-m+1:t})^\top A_{t-m+1:t}^{-\frac{1}{2}}u_{t-m+1:t}] \\ &\quad+\frac{\delta^2}{2}\mathbb{E}[u_{t-m+1:t}^\top A_{t-m+1:t}^{-\frac{1}{2}}(\beta I)A_{t-m+1:t}^{-\frac{1}{2}}u_{t-m+1:t}] \\ &=_{(1)}\frac{\delta^2\beta}{2}\sum_{t=m}^{T}\sum_{s=t-m+1}^{t}\mathbb{E}[u_s^\top(\nabla^2 R(x_s))^{-1}u_s] \\ &\leq_{(2)}\frac{\delta^2\beta m}{2\alpha}T \\ &=C_1\delta^2 T,\end{aligned}$$

where (1) follows from the m-step delayed dependence and independence of the perturbations, and (2) follows from the α-strong convexity assumption on the self-concordant barrier $R(\cdot)$.

Movement loss. To bound the movement loss, it is necessary and sufficient to bound the ℓ_2-distance between neighboring iterates, i.e. $\|x_{t+1}-x_t\|_2$.

Denote $\Phi_t(x)=\eta\sum_{s=m}^{t-m+1}\langle\tilde{g}_s,x\rangle+R(x)$. By specification of Algorithm 5, we have that $x_{t+1}=\arg\min_{x\in\mathcal{K}}\Phi_t(x)$. By Taylor's theorem, $\exists\alpha\in[0,1]$ such that $z=\alpha x_t+(1-\alpha)x_{t+1}$ satisfies

$$\begin{aligned}\frac{1}{2}\|x_t-x_{t+1}\|_{\nabla^2 R(z)}^2 &=_{(3)}\Phi_t(x_t)-\Phi_t(x_{t+1})-\nabla\Phi_t(x_{t+1})^\top(x_t-x_{t+1}) \\ &\leq_{(4)}(\Phi_{t-1}(x_t)-\Phi_{t-1}(x_{t+1}))+\eta\langle\tilde{g}_{t-m+1},x_t-x_{t+1}\rangle \\ &\leq_{(5)}\eta\|\tilde{g}_{t-m+1}\|_{\nabla^2 R(x_t)}^*\|x_t-x_{t+1}\|_{\nabla^2 R(x_t)},\end{aligned}$$

where (3) follows from $\nabla^2\Phi_t(x)=\nabla^2 R(x),\forall x\in\mathcal{K}$; (4) follows from $\nabla\Phi_t(x_{t+1})^\top(x_t-x_{t+1})\geq 0$ by optimality condition; (5) follows from $x_t=\arg\min_{x\in\mathcal{K}}\Phi_{t-1}(x)$ and generalized Cauchy-Schwarz. It is clear from this expression that it suffices to bound $\|\tilde{g}_{t-m+1}\|_{\nabla^2 R(x_t)}^*$ since

$$\|x_t-x_{t+1}\|_2\leq\frac{2\eta}{\sqrt{\alpha}}\|\tilde{g}_{t-m+1}\|_{\nabla^2 R(x_t)}^*.$$

Lemma 48 *Provided that* $\eta\leq\frac{\delta}{4dBm}\left(1-\exp\left(-\frac{\log 2}{m}\right)\right)$, $\forall t\geq m$, *it holds that*

$$\|\tilde{g}_{t-m+1}\|_{\nabla^2 R(x_t)}^*\leq\frac{2dBm}{\delta}.$$

Proof The base case follows by construction since $\tilde{g}_1=\cdots=\tilde{g}_{m-1}=0$, the claimed inequality holds for any $m\leq t\leq 2m-2$.

Induction hypothesis. Suppose for some $t\geq 2m-1$, $\|\tilde{g}_{s-m+1}\|_{\nabla^2 R(x_s)}^*\leq\frac{2dBm}{\delta}$ holds $\forall s\leq t-1$.

Induction. Consider $\|\tilde{g}_{t-m+1}\|^*_{\nabla^2 R(x_t)}$:

$$\left(\|\tilde{g}_{t-m+1}\|^*_{\nabla^2 R(x_t)}\right)^2 \le \frac{d^2B^2}{\delta^2} \sum_{j,k=0}^{m-1} u_{t-j}^\top A_{t-j}^{\frac{1}{2}} A_t^{-1} A_{t-k}^{\frac{1}{2}} u_{t-k}$$
$$\le \frac{d^2B^2m^2}{\delta^2} \max_{0\le j\le m-1} \left\| A_{t-j}^{\frac{1}{2}} A_t^{-1} A_{t-j}^{\frac{1}{2}} \right\|_2. \tag{20}$$

By induction hypothesis and choice of η, $\forall s \le t$, we have

$$\|x_s - x_{s-1}\|_{\nabla^2 R(x_{s-1})} \le 2\eta \|\tilde{g}_{s-m}\|^*_{\nabla^2 R(x_{s-1})} \le \frac{4\eta dBm}{\delta} \le 1.$$

Then, by Eq. (18) in Proposition 43, we have that $\forall s \le t$,

$$A_s \succeq \left(1 - \frac{4\eta dBm}{\delta}\right)^2 A_{s-1} \;\Rightarrow\; A_t \succeq \left(1 - \frac{4\eta dBm}{\delta}\right)^{2m} A_{t-i},\ \forall i \in \{0, \dots, m-1\}.$$

By assumption on η, we have that

$$A_t \succeq \left(\exp\left(-\frac{\log 2}{m}\right)\right)^{2m} A_{t-i} = \frac{1}{4} A_{t-i}, \quad \forall i \in \{0, \dots, m-1\}.$$

Thus, Eq. (20) implies that

$$\|\tilde{g}_{t-m+1}\|^*_{\nabla^2 R(x_t)} \le \frac{2dBm}{\delta}.$$

■

Lemma 48 directly implies that $\|x_t - x_{t+1}\|_2 \le \frac{4\eta dBm}{\delta\sqrt{\alpha}}$, $\forall t$, which together with the assumption that f_t is β-smooth, gives that

$$\begin{aligned}
\sum_{t=m}^{T} \mathbb{E}[f_t(x_{t-m+1:t}) - \bar{f}_t(x_t)] &\le \frac{\beta}{2} \sum_{t=m}^{T} \sum_{s=t-m+1}^{t} \mathbb{E}\|x_s - x_t\|_2^2] \\
&\le \frac{\beta m}{2} \sum_{t=m}^{T} \sum_{s=t-m+1}^{t} \sum_{r=s+1}^{t} \mathbb{E}[\|x_r - x_{r-1}\|_2^2] \\
&\le \frac{\beta m^3 T}{2} \cdot \frac{16\eta^2 d^2 B^2 m^2}{\delta^2 \alpha} \\
&= \frac{8\eta^2 \beta d^2 B^2 m^5}{\delta^2 \alpha} \\
&= C_2 \frac{\eta^2}{\delta^2} T.
\end{aligned}$$

Underlying regret. To see the bound on the underlying regret, note that by Lemma 46 and analysis in Section E.4, we have that

$$\left\|\mathbb{E}[\tilde{g}_t \mid \mathbb{F}_{t-m}] - \nabla \bar{f}_t(x_t)\right\|_2 \leq m\beta \sum_{s=t-m+1}^{t} \sum_{r=s+1}^{t} \|x_r - x_{r-1}\|_2 \leq \frac{4\eta d\beta B m^4}{\delta\sqrt{\alpha}}.$$

By the above bias bound,

$$\begin{aligned}
&\sum_{t=m}^{T} \mathbb{E}[\bar{f}_t(x_t) - \bar{f}_t(x)] \\
&\leq \sum_{t=m}^{T-m+1} \left(\mathbb{E}\left[\tilde{g}_t^\top (x_t - x)\right] + \mathbb{E}[(\nabla \bar{f}_t(x_t) - \mathbb{E}[\tilde{g}_t \mid \mathbb{F}_{t-m}])^\top (x_t - x)]\right) + mB \\
&\leq \underbrace{\sum_{t=m}^{T-m+1} \mathbb{E}\left[\tilde{g}_t^\top (x_{t+m} - x)\right]}_{(a)} + \underbrace{\sum_{t=m}^{T-m+1} \mathbb{E}\left[\tilde{g}_t^\top (x_t - x_{t+m})\right]}_{(b)} + \frac{4\eta d\beta D B m^4 T}{\delta\sqrt{\alpha}} + mB.
\end{aligned}$$

We will bound (a) and (b). (a) is given by a variant of the FTL-BTL lemma (Lemma 5.2.2, Hazan (2022)).

Lemma 49 (FTL-BTL, with memory) *$\forall t \geq 2m-2$, $t \in \mathbb{N}$, $\forall x \in \mathcal{K}$, the $x_t, \tilde{g}_t$ given by Algorithm 5 with step size $\eta > 0$ and self-concordant barrier $R(\cdot)$ satisfies the following inequality:*

$$\sum_{s=m}^{t-m+1} \eta\langle \tilde{g}_s, x\rangle + R(x) \geq \sum_{s=m}^{t-m+1} \eta\langle \tilde{g}_s, x_{s+m}\rangle + R(x_m). \tag{21}$$

Proof We prove Eq. (21) by induction. For $t = 2m-2$, the right hand side equals to $R(x_m)$, and $R(x_m) \leq R(x)$, $\forall x \in \mathcal{K}$ since $x_m = \arg\min_{x\in\mathcal{K}} R(x)$ (Line 1, Algorithm 5). Suppose Eq. (21) holds for some $t \geq 2m-2$. Consider $t+1$. We have $\forall x \in \mathcal{K}$, by Line 7,

$$\sum_{s=m}^{t-m+2} \eta\langle \tilde{g}_s, x_{t+2}\rangle + R(x_{t+2}) \leq \sum_{s=m}^{t-m+2} \eta\langle \tilde{g}_s, x\rangle + R(x).$$

By induction hypothesis, we have $\forall x \in \mathcal{K}$,

$$\begin{aligned}
\sum_{s=m}^{t-m+2} \eta\langle \tilde{g}_s, x\rangle + R(x) &\geq \sum_{s=m}^{t-m+1} \eta\langle \tilde{g}_s, x_{s+m}\rangle + R(x_m) + \eta\langle \tilde{g}_{t-m+2}, x_{t+2}\rangle \\
&= \sum_{s=m}^{t-m+2} \eta\langle \tilde{g}_s, x_{s+m}\rangle + R(x_m).
\end{aligned}$$

■

Lemma 49 implies a bound on (a): we can without loss of generality assume that the comparator x satisfies $\pi_{2m-1}(x) > 1 - \frac{1}{T}$ at a cost of $O(1)$, and then by Proposition 43,

$$(a) \leq \frac{1}{\eta}(R(x) - R(x_m)) \leq \frac{\nu \log T}{\eta}.$$

By bounds established for the local norm of $(x_{t+1} - x_t)$ and $\tilde{g}_t$, we have

$$(b) \leq \sum_{t=m}^{T-m+1} \mathbb{E}\left[\|\tilde{g}_t\|^*_{\nabla^2 R(x_t)} \left(\sum_{i=0}^{m-1} \|x_{t+i} - x_{t+i+1}\|_{\nabla^2 R(x_t)}\right)\right] \leq \frac{16\eta d^2 B^2 m^3 T}{\delta^2},$$

since $\nabla^2 R(x_t) \preceq 4\nabla^2 R(x_{t+i})$ for all $0 \leq i \leq m-1$. Combining the bounds, we have that

$$\begin{aligned}\sum_{t=m}^{T} \mathbb{E}[\bar{f}_t(x_t) - \bar{f}_t(x)] &\leq \frac{\nu \log T}{\eta} + \frac{16\eta d^2 B^2 m^3 T}{\delta^2} + \frac{4\eta d\beta DBm^4 T}{\delta\sqrt{\alpha}} + mB \\ &= C_3 \frac{\eta T}{\delta^2} + C_4 \frac{1}{\eta} + C_5.\end{aligned}$$

■