Journal of Machine Learning Research 24 (2023) 1-70 Submitted 3/22; Published 6/23

Non-stationary Online Learning with Memory and Non-stochastic Control

Peng Zhao ZHAOP@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Yu-Hu Yan YANYH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Yu-Xiang Wang YUXIANGW@CS.UCSB.EDU
Department of Computer Science
University of California, Santa Barbara, CA 93106, USA

Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Editor: Shipra Agrawal

Abstract

We study the problem of Online Convex Optimization (OCO) with memory, which allows loss functions to depend on past decisions and thus captures temporal effects of learning problems. In this paper, we introduce *dynamic policy regret* as the performance measure to design algorithms robust to non-stationary environments, which competes algorithms' decisions with a sequence of changing comparators. We propose a novel algorithm for OCO with memory that provably enjoys an *optimal* dynamic policy regret in terms of time horizon, non-stationarity measure, and memory length. The key technical challenge is how to control the *switching cost*, the cumulative movements of player's decisions, which is neatly addressed by a novel switching-cost-aware online ensemble approach equipped with a new meta-base decomposition of dynamic policy regret and a careful design of meta-learner and base-learner that explicitly regularizes the switching cost. The results are further applied to tackle non-stationarity in *online non-stochastic control* (Agarwal et al., 2019), i.e., controlling a linear dynamical system with adversarial disturbance and convex cost functions. We derive a novel gradient-based controller with dynamic policy regret guarantees, which is the first controller provably competitive to a sequence of changing policies for online non-stochastic control.

Keywords: online learning, online convex optimization with memory, online non-stochastic control, non-stationary environments, dynamic policy regret, online ensemble

1. Introduction

Online Convex Optimization (OCO) (Shalev-Shwartz, 2012; Hazan, 2016) is a versatile model of learning in adversarial environments, which can be regarded as a sequential game between a player and an adversary (environments). At each round, the player makes a prediction from a convex set $\mathbf{w}_t \in \mathcal{W} \subseteq \mathbb{R}^d$, the adversary simultaneously selects a convex loss $f_t : \mathcal{W} \mapsto \mathbb{R}$, and the player incurs a loss $f_t(\mathbf{w}_t)$. The goal of the player is to minimize the cumulative loss. The framework is found useful in a variety of disciplines including learning theory, game theory, and optimization, etc (Cesa-Bianchi and Lugosi, 2006).

The standard OCO framework considers only *memoryless* adversary, in the sense that the resulting loss is only determined by the player's current prediction without involving past ones. In real-world applications, particularly those related to online decision making, it is often the case that past predictions/decisions would also contribute to the current loss, which makes the standard OCO framework not viable. To remedy this issue, Online Convex Optimization with Memory (OCO with Memory) was proposed as a simplified and elegant model to capture the temporal effects of learning problems (Merhav et al., 2002; Anava et al., 2015). Specifically, at each round, the player makes a prediction $\mathbf{w}_t \in \mathcal{W}$, the adversary chooses a loss function $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$, and the player will then suffer a loss $f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t)$. Notably, now the loss function depends on both current and past predictions. The parameter m is the memory length, and evidently the OCO with memory model reduces to the standard memoryless OCO when memory length $m = 0$. The performance measure for OCO with memory is *policy regret* (Dekel et al., 2012), defined as

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \min_{\mathbf{v} \in \mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{v}, \ldots, \mathbf{v}), \tag{1}$$

where throughout the paper we adopt the notation $\mathbf{a}_{i:j}$ to denote the vector sequence $\mathbf{a}_i, \ldots, \mathbf{a}_j$. We start the index from 1 for convenience. Recent studies apply online learners with provable low policy regret to a variety of related problems (Chen et al., 2018; Agarwal et al., 2019; Daniely and Mansour, 2019; Chen et al., 2020). However, the policy regret (1) only measures the performance versus a *fixed* comparator and is thus not suitable for learning in non-stationary and open environments (Sugiyama and Kawanabe, 2012; Zhou, 2022). For instance, in the recommendation system, the users' interest may change when looking through the product pages; in the traffic flow scheduling, the traffic network pattern changes throughout the day. Therefore, it is necessary to design online decision-making algorithms with robustness to non-stationary environments. To this purpose, we introduce the *dynamic policy regret* to guide algorithm design, measuring the competitive performance against an arbitrary sequence of *time-varying* comparators $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$, defined as

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) = \sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \sum_{t=1}^{T} f_t(\mathbf{v}_{t-m:t}). \tag{2}$$

The upper bound of $\text{D-Regret}_T(\mathbf{v}_{1:T})$ should be a function of the comparator sequence $\mathbf{v}_{1:T}$, while the algorithm is agnostic to the choice of comparators. The proposed measure is very general—it subsumes static policy regret (1) as a special case when comparators become the best predictor in hindsight, i.e., $\mathbf{v}_{1:T} = \mathbf{v}^* \in \arg\min_{\mathbf{v} \in \mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{v}, \ldots, \mathbf{v})$. Therefore, dynamic policy regret is a more stringent measure than standard policy regret and algorithms that optimize it are more robust to non-stationary environments.

The fundamental challenge of dynamic policy regret optimization is how to simultaneously compete with all comparator sequences with vastly different levels of non-stationarity. Our approach builds upon recent advance of non-stationary online learning (Zhang et al., 2018a; Zhao et al., 2020, 2021b) to hedge the uncertainty via the meta-base online ensemble structure, along with several new ingredients specifically designed for the OCO with memory setting. In particular, it is essential to control the *switching cost* for OCO with memory, the cumulative movement of player's predictions. The amount is relatively easy to control in static policy regret (Anava et al., 2015), yet becomes much harder in dynamic policy regret and could even scale linearly due to the meta-base online ensembles structure. Intuitively, online algorithms minimizing dynamic regret necessitate maintaining a certain probability of aggressive movement to catch up with potential changes within non-stationary environments, which results in tensions between dynamic regret and switching cost. We elegantly address the difficulty by proposing a *switching-cost-aware online ensemble* approach. Our approach features a novel meta-base decomposition of dynamic policy regret and a switching-cost-regularized surrogate loss, which avoids directly handling switching cost altogether but regularizes the switching cost to meta-learner and base-learner instead. Our proposed online-ensemble algorithm provably enjoys an *optimal* $\mathcal{O}(\sqrt{T(1+P_T)})$ dynamic policy regret, where $P_T = \sum_{t=2}^{T} \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2$ denotes the unknown path length of comparators. As a byproduct, our result can serve as a solution for minimizing dynamic regret of *online convex optimization with switching cost*, a variant of classic OCO setting by penalizing switching cost of returned decisions (Blum and Kalai, 1999; Gofer, 2014; Chen et al., 2018). Specifically, consider the OCO problem with online functions $h_1, \ldots, h_T$ with $h_t : \mathcal{W} \mapsto \mathbb{R}$. Denote by $\mathbf{w}_1, \ldots, \mathbf{w}_T$ the returned decisions by our algorithm. Then, we have $\sum_{t=1}^{T} h_t(\mathbf{w}_t) - \sum_{t=1}^{T} h_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \leq \mathcal{O}(\sqrt{\lambda T(1+P_T)})$, where P_T is the path length as defined above. We also establish the lower bound to show the *minimax optimality* in terms of switching-cost coefficient λ, time horizon T, and path length P_T. Compared to our conference paper (Zhao et al., 2022b), the current result improves the dependence in the memory parameter λ to be optimal, which is achieved via a novel usage of the laze update mechanism.

The results of OCO with memory yield an important application in online decision-making problems. Specifically, we investigate the problem of *online non-stochastic control* (Agarwal et al., 2019), i.e., controlling a linear dynamical system with adversarial (non-stochastic) disturbance and adversarial convex cost functions. Online non-stochastic control has attracted much recent research attention due to its relaxed assumptions on disturbances and flexibility of cost functions. Existing studies mainly focus on optimizing static policy regret, whereas the optimal controller of each round would naturally change over iterations since the disturbances and cost functions both change adversarially. Therefore, it is necessary to investigate *dynamic policy regret*, which competes controller's performance with time-varying benchmark controllers. By adopting the "disturbance-action" policy parameterization (Agarwal et al., 2019), online non-stochastic control is reduced to OCO with memory, and thus its dynamic policy regret can be optimized by a similar meta-base online ensemble structure as developed before. Our designed controller attains an $\widetilde{\mathcal{O}}(\sqrt{T(1+P_T)})$ dynamic policy regret, where P_T measures the fluctuation of compared controllers. To the best of our knowledge, this is the first controller competitive to a sequence of changing "disturbance-action" policies. Given that our techniques for OCO with memory provide a provable way to handle the memory effects of past decisions, we anticipate that they would have broader applications in online decision-making problems.

The main contributions of this paper are summarized as follows.

- We introduce *dynamic policy regret* as the performance measure to guide the algorithm design of OCO with memory and online non-stochastic control to enhance the robustness of online algorithms to non-stationary environments.
- We propose a novel algorithm for OCO with memory, which enjoys an *optimal* dynamic policy regret of order $\mathcal{O}(\sqrt{T(1+P_T)})$. To achieve this, several key algorithmic ingredients are designed to handle unknown environments and control switching cost.
- The results are further applied to the problem of online non-stochastic control, yielding an online controller with $\widetilde{\mathcal{O}}(\sqrt{T(1+P_T)})$ dynamic policy regret, which is the first online controller competitive with a sequence of *time-varying* policies.

In the following, we first review related works in Section 2 and then introduce some preliminaries in Section 3. Next, we present the main results for OCO with memory and online non-stochastic control in Section 4 and Section 5. Section 6 reports the experiments. We finally conclude the paper in Section 7. All the proofs are included in appendices.

2. Related Work

In this section, we briefly discuss related works on OCO with memory, online non-stochastic control, and dynamic regret minimization for online learning.

OCO with Memory. OCO with memory is initiated by Merhav et al. (2002), who prove an $\mathcal{O}(T^{2/3})$ policy regret by a blocking technique. Later, Anava et al. (2015) propose a simple gradient-based algorithm that provably achieves $\mathcal{O}(\sqrt{T})$ and $\mathcal{O}(\log T)$ policy regret for convex and strongly convex functions, respectively. Recent study discloses that the policy regret of OCO with memory over exp-concave functions is at least $\Omega(T^{1/3})$ (Simchowit, 2020, Theorem 2.3). One of the key concepts of OCO with memory is *switching cost*, the cumulative movement of decisions, which is also concerned in smoothed online learning (Chen et al., 2018; Goel et al., 2019; Goel and Wierman, 2019), online learning with switching budget (Altschuler and Talwar, 2018; Chen et al., 2020; Sherman and Koren, 2021; Wang et al., 2021). Online learning with memory is also studied in the prediction with expert advice setting (Geulen et al., 2010; György and Neu, 2014; Cesa-Bianchi et al., 2013; Altschuler and Talwar, 2018) and bandit settings (Dekel et al., 2012, 2014; Altschuler and Talwar, 2018; Arora et al., 2019).

Online Non-stochastic Control. Recently, there is a surge of interest to apply modern statistical and algorithmic techniques to the control problem. Online non-stochastic control is proposed by Agarwal et al. (2019), where the regret is chosen as the performance measure and the disturbance is allowed to be adversarially chosen. When online cost functions are convex and Lipschitz, Agarwal et al. (2019) obtain an $\mathcal{O}(\sqrt{T})$ policy regret for known linear dynamical system by introducing the DAC parameterization and reducing the problem to OCO with memory. Hazan et al. (2020) show an $\mathcal{O}(T^{2/3})$ policy regret for unknown system via system identification. In addition, Foster and Simchowitz (2020) propose the online learning with advantages technique and obtain logarithmic regret for known system with quadratic cost and adversarial disturbance, whose results are strengthened by Simchowit (2020) to accommodate arbitrary changing costs. All mentioned results are developed for fully observed system, and Simchowitz et al. (2020) present a clear picture for non-stochastic control with partially observed systems. We are still witnessing a variety of recent advances, for example, non-stochastic control with bandit feedback (Gradu et al., 2020a; Cassel and Koren, 2020), adaptive regret minimization (Gradu et al., 2020b; Zhang et al., 2022b,c), etc. We will present more discussions on the relationship between these works for adaptive regret minimization and our work (for dynamic regret minimization) at the end of this section. There are other related works studying non-stationary online control from the lens of competitive ratio (Shi et al., 2020; Goel and Hassibi, 2022b) and robust control (Goel and Hassibi, 2020, 2022a). In addition, there have been considerable efforts dedicated to the broader field of online (stochastic) control over the past several decades. While only a handful can be mentioned here (Guo and Ljung, 1995; Fiechter, 1997; Abbasi-Yadkori and Szepesvári, 2011; Cohen et al., 2018; Dean et al., 2020; Cassel et al., 2022a,b), interested readers can refer to the references therein to explore more recent developments in this area.

Dynamic Regret. Benchmarking the regret in term of changing comparators dates back to early development of prediction with expert advice (Herbster and Warmuth, 1998, 2001), in which they studied a special form of dynamic regret that supports the comparators change for at most S times (often referred to as S-tracking/shifting/switching regret) (Herbster and Warmuth, 1998, 2001; Bousquet and Warmuth, 2002; Cesa-Bianchi et al., 2012; György and Szepesvári, 2016; Wei et al., 2016; Zheng et al., 2019; Luo et al., 2022). For online convex optimization, Zinkevich (2003) pioneers the study of dynamic regret and shows that OGD can attain an $\mathcal{O}(\sqrt{T}(1+P_T))$ dynamic regret. Zhang et al. (2018b) show that the minimax lower bound is $\Omega(\sqrt{T(1+P_T)})$ and close the gap by proposing an algorithm with an $\mathcal{O}(\sqrt{T(1+P_T)})$ regret. Recent works achieve problem-dependent guarantees by exploiting smoothness and incorporating the optimistic online learning techniques (Zhao et al., 2020, 2021b), and other works obtain an improved rate by exploiting exp-concavity or strong convexity (Baby and Wang, 2021, 2022). More results for dynamic regret minimization have been developed in bandit convex optimization (Zhao et al., 2021a), Markov decision processes (Zhao et al., 2022a), online label shift problems (Bai et al., 2022; Baby et al., 2023), time-varying games (Zhang et al., 2022a; Yan et al., 2023), etc. We note that the dynamic regret measure studied in this paper is also called the *universal* dynamic regret, in the sense that the regret guarantee holds universally against any comparator sequence in the domain. Another special variant called the *worst-case* dynamic regret is frequently studied in the literature (Besbes et al., 2015; Jadbabaie et al., 2015; Mokhtari et al., 2016; Zhang et al., 2017; Baby and Wang, 2019; Zhang et al., 2020; Zhao and Zhang, 2021), which specifies comparators as the optimizers of online functions. The worst-case dynamic regret is less general than the universal one. Indeed, both worst-case dynamic regret and static regret are special cases of the universal dynamic regret with different choices of comparators, and we refer the reader to (Zhao et al., 2021b) for more elaborations.

More Discussions. Online non-stochastic control in non-stationary environments is also recently studied via the measure of *adaptive regret* (Hazan and Seshadhri, 2009; Daniely et al., 2015)—the regret compared to the best policy on any interval in the time horizon. Gradu et al. (2020b) propose the first controller with an $\widetilde{\mathcal{O}}(\sqrt{T})$ expected adaptive regret on any interval in the total horizon. The result is strengthened in a recent work (concurrent to our paper) (Zhang et al., 2022b), which presents a strongly adaptive controller with an $\widetilde{\mathcal{O}}(\sqrt{|\mathcal{I}|})$ deterministic adaptive regret on any interval $\mathcal{I} \subseteq [T]$. The two papers and our work all study non-stationary online control, however, the concerned measures and used techniques are completely different. **(1)** Measures: dynamic regret examines the global behavior to ensure a competitive performance with time-varying compared polices, whereas adaptive regret focuses on the local behavior with respect to a fixed strategy. Even though a black-box reduction from dynamic regret to adaptive regret has been known in the simpler setting of prediction with expert advice (i.e., online linear optimization over the simplex) (Luo and Schapire, 2015, Theorem 4), the relationship between strongly adaptive regret and universal dynamic regret for online convex optimization over the general setup (Zhang, 2020, Section 5) remains highly *unclear*, which is even more vague when further taking the switching cost into account. **(2)** Techniques: optimizing either dynamic regret or adaptive regret requires the meta-base online ensemble structure to deal with uncertainty of the non-stationary environments. However, the specific techniques, especially the way to control switching cost, exhibit significant difference. Gradu et al. (2020b) leverage the Follow-the-Leading-History framework (Hazan and Seshadhri, 2009) with a shrinking technique (Geulen et al., 2010) to keep previous experts unchanged with a certain probability to reduce the switching cost, so their result holds in expectation only. The improved result of $\mathcal{O}(\sqrt{|\mathcal{I}|})$ deterministic strongly adaptive regret bound (Zhang et al., 2022b) is achieved by a very different framework drawn inspirations from parameter-free online learning (Cutkosky, 2020). By contrast, the key ingredients of our approach are the novel meta-base decomposition and the switching-cost-regularized loss, which avoid explicitly handling the switching cost of final decisions but directly control the switching cost of meta-algorithm and individual base-algorithm. These mechanisms finally lead to a deterministic dynamic policy regret guarantee for our methods.

3. Preliminaries

This section introduces preliminaries for online convex optimization (OCO) with memory.

Problem Setup. OCO with memory is a variant of standard OCO framework to capture the long-term effects of past decisions, whose protocol is shown below.

```
for $t = m+1, \dots, T$ do
    the player chooses a decision $\mathbf{w}_t \in \mathcal{W}$;
    the adversary reveals the loss $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$ that applies to last $m+1$ decisions;
    the player suffers a loss of $f_t(\mathbf{w}_{t-m}, \dots, \mathbf{w}_t)$;
end for
```

In above, m is the memory length, and $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$ is convex in memory, which means its unary function $\widetilde{f}_t(\mathbf{w}) = f_t(\mathbf{w}, \dots, \mathbf{w})$ is convex in $\mathbf{w}$. Clearly, OCO with memory recovers the standard memoryless OCO when $m = 0$. The standard measure is policy regret (Dekel et al., 2012) as defined in (1). We introduce a strengthened measure called *dynamic policy regret* to compete with changing comparators as defined in (2). The dynamic policy regret upper bound usually involves the path length $P_T = \sum_{t=2}^{T} \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2$, which measures the variation of comparators and thus captures the environmental non-stationarity. Throughout the paper, $\mathcal{O}(\cdot)$-notation is used to express regret upper bound as a function of T and P_T, and $\widetilde{\mathcal{O}}(\cdot)$-notation omits logarithmic factors in T. To make it clear, we mention that the $\mathcal{O}(\cdot)$-notation does not hide $\log \log P_T$ or $\log \log T$ terms, even though they are indeed small.

Assumptions. Next, we introduce several standard assumptions (Anava et al., 2015). For simplicity we focus on the ℓ_2-norm and the extension to general primal-dual norms is straightforward.

Assumption 1 (coordinate-wise Lipschitzness)**.** The online function $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$ is L-coordinate-wise Lipschitz, i.e., $|f_t(\mathbf{x}_0, \ldots, \mathbf{x}_m) - f_t(\mathbf{y}_0, \ldots, \mathbf{y}_m)| \leq L \sum_{i=0}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|_2$.

Assumption 2 (bounded gradient)**.** The gradient norm of the unary loss is at most G, i.e., for all $\mathbf{w} \in \mathcal{W}$ and $t \in [T]$, $\|\nabla \widetilde{f}_t(\mathbf{w})\|_2 \leq G$.

Assumption 3 (bounded domain)**.** The domain $\mathcal{W}$ is convex, closed, and satisfies $\|\mathbf{w} - \mathbf{w}'\|_2 \leq D$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$. For convenience, we also assume $\mathbf{0} \in \mathcal{W}$.

Static Regret of OCO with Memory. This part briefly reviews the result of static policy regret. Anava et al. (2015) propose a simple approach based on the gradient descent based on the observation that when online functions are coordinate-wise Lipschitz, the policy regret can be upper bounded by the switching cost and the vanilla regret over the unary loss, formally,

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \min_{\mathbf{v} \in \mathcal{W}} \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}) \leq \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 + \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \min_{\mathbf{v} \in \mathcal{W}} \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}),$$

where $\lambda = m^2 L$. The first term is the *switching cost* measuring the cumulative movement of decisions $\mathbf{w}_{1:T}$ and the remaining term is the standard regret of memoryless OCO. Consequently, it is natural to perform Online Gradient Descent (OGD) (Zinkevich, 2003) over the unary loss $\widetilde{f}_t$, i.e., $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta \nabla \widetilde{f}_t(\mathbf{w}_t)]$, where $\eta > 0$ is the step size and $\Pi_{\mathcal{W}}[\cdot]$ denotes the projection onto the nearest point in $\mathcal{W}$. It is well-known that with an appropriate step size OGD enjoys an $\mathcal{O}(\sqrt{T})$ regret in memoryless OCO. Further, Anava et al. (2015) show that the produced decisions move sufficiently slowly. Indeed, switching cost satisfies $\sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \leq \mathcal{O}(\eta T)$, which will not affect the final regret order by choosing $\eta = \mathcal{O}(1/\sqrt{T})$. Combining both facts yields an $\mathcal{O}(\sqrt{T})$ static policy regret (Anava et al., 2015, Theorem 3.1).

4. OCO with Memory

This section presents dynamic policy regret of OCO with memory. We begin with the gentle case when the path length is known, and then handle the general case when it is unknown and present the overall result.

4.1 A Gentle Start: known path length

Similar to the static regret analysis mentioned in the last section, we first upper-bound the dynamic policy regret (2) in the following way:

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) \leq \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 + \lambda \sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2. \quad (3)$$

There are three terms in the upper bound: dynamic regret of unary functions, switching cost of final decisions, and switching cost of comparators. Therefore, it is natural to deploy OGD over unary functions, and we can prove the following dynamic policy regret guarantee. The proof can be found in Appendix B.1.

Theorem 1. *Under Assumptions 1–3, running OGD over unary losses $\widetilde{f}_1, \ldots, \widetilde{f}_T$ ensures*

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) = \sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \sum_{t=1}^{T} f_t(\mathbf{v}_{t-m:t}) \leq \mathcal{O}\Big(\eta T + \frac{1+P_T}{\eta} + P_T\Big) \tag{4}$$

for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$, where $P_T = \sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2$ is the path length measuring fluctuation of the comparator sequence.

Suppose the value of path length P_T were known a priori, Theorem 1 indicates an optimal $\mathcal{O}(\sqrt{T(1+P_T)})$ dynamic policy regret by setting step size as $\eta = \mathcal{O}(\sqrt{(1+P_T)/T})$, matching the $\Omega(\sqrt{T(1+P_T)})$ lower bound of memoryless OCO (Zhang et al., 2018a). However, this step size tuning is not realistic because we cannot attain the prior information of path length $P_T = \sum_{t=2}^{T} \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2$. Indeed, since the dynamic policy regret measure holds for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T$ that can be arbitrarily selected in the feasible domain $\mathcal{W}$, the path length P_T essentially captures the environmental non-stationarity and is *unknown* to the player. In Section 4.2, we will further elucidate the challenge of designing online algorithms that enjoy optimal dynamic policy regret and meanwhile do not require prior knowledge of environmental non-stationarity, especially due to the switching cost arising in OCO with memory. In Section 4.3, we will present our solution by introducing several novel algorithmic ingredients. Finally, in Section 4.4 we further improve the algorithm to achieve an optimal memory dependence along with the corresponding lower bound argument to show the minimax optimality of our results.

4.2 Challenge: unknown path length and switching cost of OCO with memory

As mentioned in the last paragraph, the fundamental difficulty of attaining optimal dynamic policy regret lies in the infeasible step size tuning that depends on the unknown comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T$. We emphasize that such an unpleasant dependence cannot be removed by the well-known doubling trick (Cesa-Bianchi et al., 1997), because we cannot monitor the empirical value of path length, $P_t = \sum_{s=2}^{t} \|\mathbf{v}_s - \mathbf{v}_{s-1}\|_2$, as comparators $\mathbf{v}_1, \ldots, \mathbf{v}_T$ can be arbitrarily chosen in the feasible domain $\mathcal{W}$ and are entirely unknown to the learner. Similar challenge also emerges in recent studies of memoryless non-stationary online learning (Zhang et al., 2018a; Zhao et al., 2020), inspired by which we employ the meta-base online ensemble framework to design a two-layer approach to optimize the dynamic policy regret. Below, we will first briefly review the framework and then elucidate the challenge of its application in OCO with memory, mainly due to the tension between dynamic regret and switching cost, which necessitates additional new ideas.

Meta-base Online Ensemble Framework. The framework admits a two-layer structure and is essentially an online ensemble method (Zhou, 2012; Zhao, 2021). We first need to design an appropriate pool of candidate step sizes $\mathcal{H} = \{\eta_1, \ldots, \eta_N\}$ to ensure the existence of a step size η_{i^*} that approximates optimal step size η_* well. Then, multiple base-learners $\mathcal{B}_1, \ldots, \mathcal{B}_N$ are maintained, and each performs base-algorithm (for example, OGD) with a step size $\eta_i \in \mathcal{H}$ and generates the decision sequence $\mathbf{w}_{1,i}, \mathbf{w}_{2,i}, \ldots, \mathbf{w}_{T,i}$. Finally, a meta-learner, supposed to be able to track the best base-learner, is used to combine all intermediate results of base learners to produce final output $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T$, where $\mathbf{w}_t = \sum_{i=1}^N p_{t,i}\mathbf{w}_{t,i}$. The final output of meta-base algorithm can well approximate the decision sequence of the best base-learner (the one with near-optimal step size η_{i^*}) and thus ensure a good dynamic regret bound.

Indeed, by employing OGD over unary functions $\widetilde{f}_1, \ldots, \widetilde{f}_T$ and designing a proper step size pool $\mathcal{H}$, it is not hard to prove a dynamic regret bound over unary functions, that is, $\sum_{t=1}^T \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^T \widetilde{f}_t(\mathbf{v}_t) \leq \mathcal{O}(\sqrt{T(1+P_T)})$. Then, by (9) we have

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) \leq \mathcal{O}(\sqrt{T(1+P_T)}) + \mathcal{O}(P_T) + \sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2.$$

So we are in the position to control *switching cost.* Below, we demonstrate that a vanilla deployment of the meta-base method may move too fast to achieve a sublinear switching cost and will ruin the overall policy regret bound, which necessitates additional novel algorithmic ingredients to better balance the dynamic regret and switching cost.

Switching Cost. The switching cost is the pivot of the analysis for OCO with memory. Anava et al. (2015) demonstrate that many popular OCO algorithms for static regret minimization naturally produce slow-moving decisions, however, it becomes more difficult in dynamic regret. Intuitively, for dynamic online algorithms, it is necessary to keep some probability of aggressive movement in order to catch up with the potential changes of non-stationary environments, which results in *tensions between dynamic regret and switching cost.* Formally, denote by $\mathbf{w}_t = \sum_{i=1}^N p_{t,i}\mathbf{w}_{t,i}$ the final decision returned by the two-layer approach, then the switching cost can be bounded by

$$\sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \leq D\sum_{t=2}^T \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \sum_{t=2}^T \sum_{i=1}^N p_{t,i}\|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2. \tag{5}$$

A formal proof is presented in Appendix B.2. In the upper bound, the first term $\sum_{t=2}^T \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1$ is the switching cost of meta-learner, which is at most $\mathcal{O}(\sqrt{T})$. However, the second term $\sum_{t=2}^T \sum_{i=1}^N p_{t,i}\|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$, the weighted sum of switching cost of all base-learners, becomes the major barrier, which could be very large and even grow linearly over iterations. Specifically, for each base-learner $\mathcal{B}_i$ (OGD with step size η_i), its switching cost is at most $\mathcal{O}(\eta_i T)$; additionally, to ensure a coverage of the optimal step size, the pool of candidate step sizes is usually set as $\mathcal{H} = \{\eta_i = \mathcal{O}(2^i \cdot T^{-1/2}), i \in [N]\}$ such that $\eta_1 = \mathcal{O}(T^{-1/2})$ and $\eta_N = \mathcal{O}(1)$. Therefore, the base-learner with larger step sizes would incur unacceptable switching cost, for instance, the switching cost of base-learner $\mathcal{B}_N$ could grow linearly, of order $\mathcal{O}(T)$. As a result, the term $\sum_{t=2}^T \sum_{i=1}^N p_{t,i}\|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$ could be enlarged by base-learners whose step sizes are too large and therefore is difficult to control.

4.3 Algorithmically Enforcing Low Switching Cost: a new meta-base decomposition

To resolve the challenge of switching cost in dynamic policy regret minimization, we propose a novel switching-cost-aware online ensemble approach. Specifically, we start with proposing the following new meta-base regret decomposition to avoid directly controlling switching cost of final predictions or controlling switching cost of every base-learner:

$$\begin{aligned}
&\sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \qquad (6)\\
&\leq \sum_{t=1}^{T} \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}_t \rangle + \lambda D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \lambda \sum_{t=2}^{T} \sum_{i=1}^{N} p_{t,i} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2\\
&= \underbrace{\sum_{t=1}^{T} \big(\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i}\big) + \lambda D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1}_{\texttt{meta-regret}} + \underbrace{\sum_{t=1}^{T} \big(g_t(\mathbf{w}_{t,i}) - g_t(\mathbf{v}_t)\big) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2}_{\texttt{base-regret}} .
\end{aligned}$$

The first inequality follows from the convexity of unary functions and switching cost decomposition (5), and for convenience we introduce the notation of linearized loss $g_t(\mathbf{w}) = \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w} \rangle$. The second equation is crucial, in which the key ingredient is the introduced *switching-cost-regularized surrogate loss* $\boldsymbol{\ell}_t \in \mathbb{R}^N$ for the meta-algorithm, defined as

$$\ell_{t,i} \triangleq g_t(\mathbf{w}_{t,i}) + \lambda \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2. \qquad (7)$$

Intuitively, the base-learner's switching cost is now taken into account when evaluating its performance—the meta-learner will impose more penalty on base-learners with larger switching cost. Technically, the key improvement upon previous analysis in (5) lies in the switching cost term of the base-learner: we now only need to bound switching cost of a single base-learner $\sum_{t=2}^{T} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$, which is to be contrasted to the switching cost of all the base-learners $\sum_{t=2}^{T} \sum_{i=1}^{N} p_{t,i} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$.

Furthermore, noting that the new meta-base decomposition (6) holds simultaneously for *any* index $i \in [N]$, we can therefore choose the compared index as i^* (the one with near-optimal step size) and the switching cost of this base-learner $\mathcal{B}_{i^*}$ is at most $\mathcal{O}(\eta_{i^*} T) = \mathcal{O}(\sqrt{T(1+P_T)})$. In other words, we successfully escape from those base-learners with unacceptably large step sizes, whose switching cost is too large to tolerate.

Consequently, we can tackle switching cost in the meta-base methods with the help of the switching-cost-regularized technique. The rest is more or less standard. Specifically, the meta-base regret decomposition indicates the following requirements on the base-algorithm and meta-algorithm:

- base-algorithm needs to achieve low dynamic regret over unary functions and tolerate its own switching cost $\sum_{t=2}^{T} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$;
- meta-algorithm needs to optimize the switching-cost-regularized loss to impose more penalty on base-learners with larger switching cost, and tolerate its own switching cost $\sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1$.

Below, we outline the specific configurations of our switching-cost-aware online ensemble approach (including settings of step size pool, base-algorithm, and meta-algorithm) to fulfill above requirements.

Step Size Pool. We initiate $N = \lceil \frac{1}{2}\log_2(1+T) \rceil + 1 = \mathcal{O}(\log T)$ base-learners, with step size pool set as

$$\mathcal{H} = \left\{ \eta_i \;\middle|\; \eta_i = 2^{i-1} \cdot \sqrt{\frac{D^2}{(\lambda G + G^2)T}},\ i \in [N] \right\}. \tag{8}$$

Base-algorithm. The base-algorithm is chosen as OGD running over the linearized loss $\{g_t\}_{t=1:T}$. The switching cost of each base-learner can be safely controlled, as indicated by Theorem 1. More specifically, there are N base-learners denoted by $\mathcal{B}_1, \ldots, \mathcal{B}_N$ and the base-learner $\mathcal{B}_i$ (with step size $\eta_i \in \mathcal{H}$) performs

$$\mathbf{w}_{t+1,i} = \Pi_{\mathcal{W}}[\mathbf{w}_{t,i} - \eta_i \nabla g_t(\mathbf{w}_{t,i})] = \Pi_{\mathcal{W}}[\mathbf{w}_{t,i} - \eta_i \nabla \widetilde{f}_t(\mathbf{w}_t)].$$

The second equation is from $g_t(\mathbf{w}) = \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w} \rangle$ and the update exhibits the computational advantage due to linearization: although multiple base-learners are performed, they share the same gradient and thus the algorithm only calculates one gradient per iteration, rather than N gradients as was anticipated.

Meta-algorithm. The meta-algorithm is set as the well-known Hedge algorithm (Freund and Schapire, 1997) *running over the switching-cost-regularized loss.* The weight $\boldsymbol{p}_{t+1} \in \Delta_N$ is updated by $p_{t+1,i} \propto p_{t,i} \exp(-\varepsilon \ell_{t,i})$, where $\boldsymbol{\ell}_t \in \mathbb{R}^N$ is the switching-cost-regularized surrogate loss defined in (7) and $\varepsilon > 0$ is the learning rate. Then, the meta-regret $\sum_{t=1}^T \left(\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i} \right) + \lambda D \sum_{t=2}^T \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1$, essentially the static regret with switching cost, can be well controlled with $\varepsilon = \mathcal{O}(\sqrt{1/T})$. For technical reasons, we adopt a non-uniform initialization by setting $\boldsymbol{p}_1 \in \Delta_N$ with $p_{1,i} \propto 1/(i^2+i)$. The dependence of learning rate on T can be removed by either a time-varying tuning or doubling trick.

We finally remark that base-algorithm (OGD) and meta-algorithm (Hedge) can be understood in a unified view from the aspect of Online Mirror Descent (OMD) (Nemirovsky and Yudin, 1983; Shalev-Shwartz, 2012; Srebro et al., 2011). OMD is a powerful online method accommodating general geometries and both OGD and Hedge are its special instances. We can generalize the dynamic policy regret of Theorem 1 from OGD to OMD, and this can be used to extend all the results in this paper from ℓ_2-norm to general primal-dual norms. More descriptions are supplied in Appendix B.3.

Overall Algorithm. Combining all above ingredients, we propose the Switching-Cost-Regularized Ensemble Algorithm for OCO with Memory (**Scream**) algorithm, which is based on online mirror descent and admits a two-layer meta-base online ensemble structure. Algorithm 1 presents overall procedures: each base-learner performs OGD with its step size as shown in Line 10; the meta-learner combines local decisions and updates the weight according to the switching-cost-regularized loss as described in Lines 4–9. The following theorem demonstrates that our algorithm can attain a favorable dynamic policy regret, striking a good balance between regret and switching cost.

Algorithm 1 Scream

```
Input: step size pool $\mathcal{H} = \{\eta_1, \ldots, \eta_N\}$, learning rate of meta-algorithm $\varepsilon$
Initialization: $\mathbf{w}_{1:m} \in \mathcal{W}$, $\mathbf{w}_{m,i} \in \mathcal{W}$, $\forall i \in [N]$; $\boldsymbol{p}_m \in \Delta_N$ with $p_{m,i} \propto 1/(i^2+i)$, $\forall i \in [N]$
for $t = m+1$ to $T$ do
    Receive $\mathbf{w}_{t,i}$ from base-learner $\mathcal{B}_i$ for $i \in [N]$
    Submit the decision $\mathbf{w}_t = \sum_{i=1}^N p_{t,i}\mathbf{w}_{t,i}$
    Suffer a loss of $f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t)$
    Observe the online function $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$ that applies to last $m+1$ decisions
    Construct the linearized loss by $g_t(\mathbf{w}) = \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w} \rangle$
    Construct the switching-cost-regularized loss $\boldsymbol{\ell}_t \in \mathbb{R}^N$ with $\ell_{t,i} = g_t(\mathbf{w}_{t,i}) + \lambda\|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$ for $i \in [N]$
    Update the weight $\boldsymbol{p}_{t+1} \in \Delta_N$ according to $p_{t+1,i} \propto p_{t,i}\exp(-\varepsilon\ell_{t,i})$
    Base-learner $\mathcal{B}_i$ updates the local decision by $\mathbf{w}_{t+1,i} = \Pi_{\mathcal{W}}[\mathbf{w}_{t,i} - \eta_i \nabla \widetilde{f}_t(\mathbf{w}_t)]$, $\forall i \in [N]$
end for
```

Theorem 2. *Under Assumptions 1–3, by setting the learning rate optimally of meta-algorithm as $\varepsilon = \sqrt{2/((2\lambda+G)(\lambda+G)D^2T)}$ and the step size pool $\mathcal{H}$ as* (8)*, our proposed Scream algorithm ensures that for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$, we have*

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) \leq \mathcal{O}\big(\sqrt{\lambda T(1+P_T)} + \lambda^{\frac{3}{4}}\sqrt{T}(1 + \log\log P_T) + \lambda P_T\big),$$

where $\lambda = m^2L$ and $P_T = \sum_{t=2}^T \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2$. So dynamic policy regret is $\mathcal{O}(\sqrt{T(1+P_T)})$.

The proof of Theorem 2 is presented in Appendix B.4.

Remark 1. Since the dynamic policy regret holds for any comparator sequence, by simply setting comparators as the fixed best decision in hindsight (now $P_T = 0$), our dynamic policy regret implies the $\mathcal{O}(\sqrt{T})$ static policy regret (Anava et al., 2015). Second, when omitting the consideration of the λ-dependence, the dynamic regret bound simplifies to $\mathcal{O}(\sqrt{T(1+P_T)})$, which is *minimax optimal* in terms of T and P_T, as an $\Omega(\sqrt{T(1+P_T)})$ lower bound has been established for the dynamic regret of memoryless OCO (Zhang et al., 2018a), which is a special case of OCO with memory when setting $m = 0$.

Remark 2. We further examine the *memory dependence* of the attained bounds. The dynamic policy regret in Theorem 2 exhibits a quadratic dependence on the memory length m (i.e., linear dependence on $\lambda = m^2L$). Recall that the dynamic policy regret is upper bounded by the dynamic regret of unary functions and switching cost of decisions (i.e., $\sum_{t=1}^T \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^T \widetilde{f}_t(\mathbf{v}_t) + \lambda\sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2$) as well as the switching cost/ path length of comparators (i.e., $\lambda\sum_{t=2}^T \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2 = \lambda P_T$), namely,

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) \leq \underbrace{\sum_{t=1}^T \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^T \widetilde{f}_t(\mathbf{v}_t) + \lambda\sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2}_{\texttt{dynamic regret of OCO with switching cost}} + \underbrace{\lambda\sum_{t=2}^T \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2}_{\texttt{path length } (=\lambda P_T)}. \quad (9)$$

Notably, the last path length term is the variation of comparators and thus irrelevant to the algorithm, which already exhibits a quadratic memory dependence. As a result, in the following we will focus the memory dependence of the first two terms, which is essentially the *dynamic regret of OCO with switching cost.* Indeed, our conference version (Zhao et al., 2022b) gives an $\mathcal{O}(\sqrt{\lambda T(1+P_T)}+\lambda\sqrt{T}(1+\log\log P_T)) \leq \mathcal{O}(\lambda\sqrt{T(1+P_T)})$ regret bound,[1] whereas Theorem 2 of this paper improves the result to $\mathcal{O}(\sqrt{\lambda T(1+P_T)}+\lambda^{3/4}\sqrt{T}(1+\log\log P_T)) \leq \mathcal{O}(\lambda^{3/4}\sqrt{T(1+P_T)})$ through a refined *analysis* (there is no modification on the algorithm), achieving an $\lambda^{1/4}$ improvement.

As a benefit, when choosing a fixed comparator, Theorem 2 implies an $\mathcal{O}(\lambda^{3/4}\sqrt{T})$ static regret, improving upon the $\mathcal{O}(\lambda\sqrt{T})$ static regret implication based on the dynamic policy regret in the conference version (Zhao et al., 2022b), where $\lambda = \mathcal{O}(m^2)$ is the squared memory length. Nevertheless, the best static policy regret for OCO with switching cost is $\mathcal{O}(\sqrt{\lambda T}) = \mathcal{O}(m\sqrt{T})$, which enjoys a linear dependence on the memory length (Anava et al., 2015) (see discussions in Appendix B.8 for details), and our result still exhibits a gap here. Therefore, we are wondering what the optimal memory dependence of dynamic regret for OCO with switching cost is. We answer this question in the next subsection.

4.4 Improved Algorithm with an Optimal Memory Dependence

In this part, we resolve the question raised at the end of the last subsection. Specifically, we first illustrate the failure of Scream algorithm in achieving optimal memory dependence; and then we propose an improved algorithm building upon Scream (Algorithm 1) that enjoys an $\mathcal{O}(\sqrt{\lambda T(1+P_T)})$ dynamic regret for OCO with switching cost, hence matching the $\mathcal{O}(\sqrt{\lambda T})$ static regret (Anava et al., 2015) when choosing a fixed comparator such that $P_T = 0$. We finally supply the lower bound to demonstrate the minimax optimality of our attained upper bound in terms of the memory dependence.

Failure of Scream Algorithm. Inspecting the proof of Theorem 2, we can observe that the sub-optimality of memory dependence mainly comes from the meta-regret $\sum_{t=1}^T \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle - \sum_{t=1}^T \ell_{t,i} + \lambda D \sum_{t=2}^T \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1$ (see the decomposition in (6) for more details). Specifically, consider the switching cost of meta-algorithm, which can be upper bounded as follows:

$$\lambda \sum_{t=2}^T \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 \leq \lambda \sum_{t=2}^T \varepsilon \|\ell_t\|_\infty \leq \lambda\varepsilon G_{\mathrm{meta}} T \leq \mathcal{O}(\lambda^{\frac{3}{4}}\sqrt{T}), \tag{10}$$

where the first inequality holds by the standard analysis on the meta-algorithm (see (31) for more details). The second inequality is by definition of $G_{\mathrm{meta}} = \sup_{t\in[T],i\in[N]} |\ell_{t,i}|$, that is, the maximum scale of the loss of meta-algorithm. The last inequality is due to the setting of $\varepsilon = \mathcal{O}(1/\sqrt{T})$ and our analysis shows that $G_{\mathrm{meta}} \leq \mathcal{O}(\sqrt{\lambda})$.

From (10), we can see that the switching cost of meta-algorithm exhibits an undesirable memory dependence of order $\mathcal{O}(\lambda^{3/4}) = \mathcal{O}(m^{3/2})$, whereas our desired one is linear in m. Therefore, it is natural to ask for an improved meta-algorithm that can enjoy a better memory dependence. However, we present the following theorem to negatively show that when the loss of meta-algorithm lies in the range of $[-C, C]$ for some $C > 0$, *any* algorithm must incur a regret of $\Omega(\sqrt{\lambda C T})$. The proof is deferred to Appendix B.5.

[1] Note that the $\log\log P_T$ term can be dominated by $\sqrt{P_T}$ and is thus absorbed within the $\mathcal{O}(\cdot)$-notation.

Algorithm 2 **Lazy Scream**

```
Input: Scream $\mathcal{A}$ (Algorithm 1), epoch number $B$, epoch length $\Delta$
 1: Initialization: $\mathbf{w}_1$ from Scream $\mathcal{A}$
 2: for $k = 1$ to $K$ do
 3:    Initialize $\nabla_k = \mathbf{0}$
 4:    for $t = (k-1)\Delta + 1$ to $k\Delta$ do
 5:       Submit the decision $\mathbf{w}_t = \mathring{\mathbf{w}}_k$
 6:       Suffer a loss of $f_t(\mathbf{w}_{t-m:t})$
 7:       $\nabla_k = \nabla_k + \nabla \widetilde{f}_t(\mathbf{w}_t) = \nabla_k + \nabla \widetilde{f}_t(\mathring{\mathbf{w}}_k)$
 8:    end for
 9:    Send $\nabla_k$ to Scream $\mathcal{A}$ for update and receive $\mathring{\mathbf{w}}_{k+1}$
10: end for
```

Theorem 3. *Consider a T-round prediction with expert advice problem with λ-switching cost. Given $\lambda > 0$ and $C > 0$, there exists a sequence of loss functions $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$ satisfying $\boldsymbol{\ell}_t \in [-C, C]^N$ for all $t \in [T]$ such that any feasible expert algorithm (whose output is $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T \in \Delta_N$) incurs the following regret*

$$\sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{p}_t \rangle - \min_{i \in [N]} \sum_{t=1}^{T} \ell_{t,i} + \lambda \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 \geq \Omega(\sqrt{\lambda C T}).$$

In our case, we have $|\ell_{t,i}| \leq GD + \sqrt{\lambda}$ (see the argument in (33) for details). Therefore, by applying Theorem 3, we know that the meta-algorithm will incur at least $\Omega(\lambda^{3/4}\sqrt{T})$ regret, which prohibits Scream from achieving the desired $\mathcal{O}(\sqrt{\lambda})$ memory dependence.

An Improved Algorithm. To address this memory dependence issue, we propose an improved algorithm called **Lazy Scream**, presented in Algorithm 2, which is a simple variant of the vanilla Scream algorithm (see Algorithm 1). Specifically, Lazy Scream builds upon Scream with episodic updates, and proceeds in K epochs (Line 2). The k-th epoch is denoted by $\mathcal{I}_k$ such that $|\mathcal{I}_k| = \Delta$, for all $k \in [K]$. Specifically, the algorithm updates at the epoch-level, for each epoch $\mathcal{I}_k$, the learner submits the same decision (Line 5) and computes the cumulative loss gradient (Line 7), and at the end of each epoch, the learner sends the cumulative gradient to the original Scream algorithm (Algorithm 1) for update (Line 9). The next theorem shows that Lazy Scream attains an improved dynamic policy regret in terms of memory length, whose proof can be found in Appendix B.6.

Theorem 4. *Under the same assumptions as Theorem 2, by setting the learning rate of meta-algorithm optimally and the step size pool $\mathcal{H}$ as* (8)*, our proposed Lazy Scream (Algorithm 2) with epoch length $\Delta = \sqrt{\lambda}$ ensures that*

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) \leq \mathcal{O}\big(\sqrt{\lambda T(1 + P_T)} + \lambda P_T\big),$$

for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$, where $\lambda = m^2 L$ and $P_T = \sum_{t=2}^{T} \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2$.

Theorem 4 implies an $\mathcal{O}(\sqrt{\lambda T(1 + P_T)})$ dynamic regret for OCO with switching cost. Below we further prove that our result is *minimax optimal* in switching-cost coefficient λ, time horizon T, and path length P_T.

Theorem 5. *Given a real value $\tau \in [0, DT]$ and a parameter $\lambda > 0$, there exist (1) a sequence of convex loss functions $h_1, \ldots, h_T$ with $h_t : \mathcal{W} \mapsto \mathbb{R}$ for $t \in [T]$, which satisfy Assumption 2 and some feasible domain $\mathcal{W} \subseteq \mathbb{R}^d$ with Assumption 3; and (2) a sequence of comparators $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathbb{R}^d$ whose path length $P_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) = \sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2 \le \tau$, such that any online algorithm returning $\mathbf{w}_1, \ldots, \mathbf{w}_T \in \mathcal{W}$ satisfies*

$$\sum_{t=1}^{T} h_t(\mathbf{w}_t) - \sum_{t=1}^{T} h_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \ge \Omega(\sqrt{\lambda \tau T}). \tag{11}$$

Theorem 5 demonstrates the minimax optimality of the obtained $\mathcal{O}(\sqrt{\lambda T(1 + P_T)})$ dynamic regret bound for OCO with switching cost, which is optimal in terms of switching-cost coefficient λ, time horizon T, and path length P_T. The corresponding proof can be found in Appendix B.7.

5. Online Non-stochastic Control

In this section, we apply the results of OCO with memory to an important online decision-making problem, online non-stochastic control (Agarwal et al., 2019), which draws much attention from researchers in online learning and control theory communities (Agarwal et al., 2019; Simchowitz et al., 2020; Hazan et al., 2020; Simchowit, 2020; Gradu et al., 2020a; Cassel and Koren, 2020; Gradu et al., 2020b; Zhang et al., 2022b).

5.1 Problem Statement

Problem Setting. We study the online control of the linear dynamical system (LDS) governed by

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{12}$$

where at iteration t, the controller provides the control u_t upon the observed dynamical state x_t and suffers a cost $c_t(x_t, u_t)$ with convex function $c_t : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \mapsto \mathbb{R}$. Following the notational convention of previous works, throughout the section we will use unbold fonts to denote vectors (including control signal, state, disturbance, etc.). We focus on online *non-stochastic* control (Agarwal et al., 2019), that is, the disturbance can be generated arbitrarily and no statistical assumption is imposed on its distribution; additionally, cost functions can be chosen adversarially. The adversarial nature of the disturbance and online cost functions hinders an a priori computation of the optimal policy as in settings of classical control theory (Kalman, 1960) and therefore requires techniques from modern online learning to tackle adversarial environments.

Policy Regret. The standard measure for online non-stochastic control is the *policy regret* (Agarwal et al., 2019), defined as the difference between cumulative loss of the designed controller $\mathcal{A}$ and that of the compared controller $\pi \in \Pi$, namely,

$$\text{Regret}_T = \sum_{t=1}^{T} c_t(x_t, u_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} c_t(x_t^\pi, u_t^\pi). \tag{13}$$

The comparator could be chosen with complete foreknowledge of the disturbance and loss functions. Recently, a variety of control algorithms have been proposed to optimize this measure under different settings (Agarwal et al., 2019; Hazan et al., 2020; Simchowitz et al., 2020; Cassel and Koren, 2020; Gradu et al., 2020a; Foster and Simchowitz, 2020). However, we argue that competing with a fixed controller may be not appropriate, especially because the unknown disturbances and cost functions can change arbitrarily in the non-stochastic control setting so that the optimal controller of each round would also change accordingly. Therefore, it is necessary to enable the online controller to compete with *time-varying* controllers to adapt to those changes. To this end, we generalize the standard measure (13) to the *dynamic policy regret* to benchmark the algorithm with a sequence of *time-varying* controllers $\pi_1, \ldots, \pi_T \in \Pi$, formally,

$$\text{D-Regret}_T(\pi_{1:T}) = \sum_{t=1}^{T} c_t(x_t, u_t) - \sum_{t=1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t}). \tag{14}$$

The measure clearly subsumes the standard policy regret (13) when choosing the compared controllers as a fixed one, i.e., $\pi_* \in \arg\min_{\pi \in \Pi} \sum_{t=1}^{T} c_t(x_t^{\pi}, u_t^{\pi})$. In this work, the benchmark set Π is chosen as the class of disturbance-action controllers (see Definition 1), which encompasses many controllers of interest.

5.2 Reduction to OCO with Memory

Following the pioneering work (Agarwal et al., 2019), we will work on the *Disturbance-Action Controller* (DAC) policy class, which parametrizes the executed action as a linear function of the past disturbances. By doing so, we can reduce online non-stochastic control to OCO with memory so that the results of Section 4 can be leveraged to design robust controllers with provable dynamic policy regret guarantee.

Definition 1 (Disturbance-Action Controller, DAC)**.** A disturbance-action controller, denoted by $\pi(K, M)$, with memory length H is specified by a fixed matrix K and parameters $M = (M^{[1]}, \ldots, M^{[H]})$. At each iteration t, the controller $\pi(K, M)$ chooses the action as a linear map of the past disturbances with an offset linear controller, formally, $u_t = -Kx_t + \sum_{i=1}^{H} M^{[i]} w_{t-i}$.

For convenience, we define $w_i = 0$ for $i < 0$. The DAC policy is implementable because the disturbance can be recovered by $w_t = x_{t+1} - Ax_t - Bu_t$ as system dynamics A and B are supposed to be known. Our method can also extend to the scenario of online non-stochastic control with unknown systems, which is presented at the end of this section. The following proposition by Agarwal et al. (2019) presents an important property of DAC policy.

Proposition 6 (Lemma 4.3 of Agarwal et al. (2019))**.** *Suppose the initial state is $x_0 = 0$ and one chooses the DAC controller $\pi(K, M_t)$ at iteration t, the reaching state and the corresponding DAC control are*

$$\begin{aligned} x_t^K(M_{0:t-1}) = & \sum_{i=0}^{H+t-1} \Psi_{t-1,i}^{K,t-1}(M_{0:t-1}) w_{t-1-i}, \\ u_t^K(M_{0:t}) = & -Kx_t^K(M_{0:t-1}) + \sum_{i=1}^{H} M_t^{[i]} w_{t-i}, \end{aligned}$$

where $\widetilde{A}_K = A - BK$ *and*

$$\Psi_{t,i}^{K,h}(M_{t-h:t}) = \widetilde{A}_K^i \mathbf{1}_{i \leq h} + \sum_{j=0}^{h} \widetilde{A}_K^j B M_{t-j}^{[i-j]} \mathbf{1}_{1 \leq i-j \leq H}.$$

Evidently, both state x_t and control u_t are linear functions of DAC parameters $M_{0:t}$, so the cost $c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t}))$ is a function of historical parameters $M_{0:t}$. Thereby, the remaining challenge is to handle this *memory* issue due to the state transition of online control, which can be addressed by OCO with memory studied in Section 4. Note that there is one big caveat in applying the technique—the current memory length is not fixed but growing with time, which is not feasible in OCO with memory. To this end, Agarwal et al. (2019) further propose a truncation operation that truncates the state with a fixed memory length H and defines the following truncated loss.

Definition 2 (Truncated Loss)**.** For the cost function $c_t : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \mapsto \mathbb{R}$ and DAC policies $\{\pi(K, M_t)\}_{t=1,\ldots,T}$, given memory length H, the induced truncated loss $f_t : \mathcal{M}^{H+2} \mapsto \mathbb{R}$ is defined as

$$f_t(M_{t-1-H:t}) = c_t(y_t^K(M_{t-1-H:t-1}), v_t^K(M_{t-1-H:t})),$$

where the truncated state and truncated DAC control are

$$y_{t+1}^K = \sum_{i=0}^{2H} \Psi_{t,i}^{K,H}(M_{t-H:t}) w_{t-i}, \quad \text{and } v_{t+1}^K = -K y_{t+1}^K(M_{t-H:t}) + \sum_{i=1}^{H} M_{t+1}^{[i]} w_{t+1-i}.$$

It can be proved that the error introduced by the truncation operation (the gap between f_t and c_t) can be precisely controlled. Therefore, by feeding the truncated loss f_t to the OCO with memory framework with a memory length of $H+2$, we finish the reduction from online non-stochastic control to OCO with memory.

5.3 Dynamic Policy Regret of Online Non-stochastic Control

The above reduction enables us to leverage results of OCO with memory (Section 4) to design online controllers competitive with time-varying compared policies. We propose the **Scream.Control** algorithm, consisting of the following two components:

(1) DAC parameterization for reduction: using DAC control $u_t = \pi(K, M_t)$ for parameterization and define the unary loss of the truncated loss, i.e., $\widetilde{f}_t : \mathcal{M} \mapsto \mathbb{R}$ with $\widetilde{f}_t(M) = f_t(M, \ldots, M)$ (see Definition 2).

(2) meta-base online ensemble structure for OCO with memory: performing Scream algorithm of Section 4 over unary loss $\widetilde{f}_t$, and using meta-algorithm to combine intermediate parameters $M_{t,1}, \ldots, M_{t,N}$ from all base-learners to produce the final M_t.

Algorithm 3 describes our proposed algorithm for optimizing dynamic policy regret of online non-stochastic control. We further provide its theoretical guarantee. We begin with several standard assumptions used in the literature (Agarwal et al., 2019; Hazan et al., 2020; Gradu et al., 2020a) and next present the main result.

Algorithm 3 Scream.Control

```
Input: step size pool $\mathcal{H} = \{\eta_1, \ldots, \eta_N\}$; learning rate of meta-learner $\varepsilon$; memory length $H$; linear controller $K$; feasible domain $\mathcal{M}$
Initialization: $u_1, \ldots, u_H$, any feasible output control signals for the first $H$ rounds;
Initialization: base decisions of the $H$-th round $M_{H,1}, M_{H,2}, \ldots M_{H,N} \in \mathcal{M}$; non-uniform weight $\boldsymbol{p}_{H+1} \in \Delta_N$ with $p_{H+1,i} \propto 1/(i^2 + i)$, $\forall i \in [N]$
for $t = H + 1$ to $T$ do
    Receive $M_{t,i}$ from base-learner $\mathcal{B}_i$ for $i \in [N]$
    Obtain the policy parameter $M_t = \sum_{i=1}^{N} p_{t,i} M_{t,i}$
    Output $u_t = -Kx_t + \sum_{i=1}^{H} M_t^{[i]} w_{t-i}$
    Suffer a loss of $c_t(x_t, u_t)$ and observe the cost function $c_t : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \mapsto \mathbb{R}$
    Construct the truncated loss $f_t : \mathcal{M}^{H+2} \mapsto \mathbb{R}$ by Definition 2 and the linearized loss by $g_t(M) = \langle \nabla \widetilde{f}_t(M_t), M \rangle$
    Compute the switching-cost-regularized loss $\boldsymbol{\ell}_t \in \mathbb{R}^N$ with $\ell_{t,i} = \lambda \|M_{t,i} - M_{t-1,i}\|_{\mathrm{F}} + g_t(M_{t,i})$ for $i \in [N]$
    Update the weight to $\boldsymbol{p}_{t+1} \in \Delta_N$ via $p_{t+1,i} \propto p_{t,i} \exp(-\varepsilon \ell_{t,i})$
    Base-learner $\mathcal{B}_i$ updates the local parameter by $M_{t+1,i} = \Pi_{\mathcal{M}}[M_{t,i} - \eta_i \nabla \widetilde{f}_t(M_t)]$
    Observe the new state $x_{t+1}$ and calculate the disturbance $w_t = x_{t+1} - Ax_t - Bu_t$
end for
```

Assumption 4. The system matrices are bounded, i.e., $\|A\|_{\mathrm{op}} \leq \kappa_A$ and $\|B\|_{\mathrm{op}} \leq \kappa_B$. Besides, the disturbance $\|w_t\| \leq W$ holds for any $t \in [T]$.

Assumption 5. The cost function $c_t(x, u)$ is convex. Further, when $\|x\|, \|u\| \leq D$, it holds that $|c_t(x, u)| \leq \beta D^2$ and $\|\nabla_x c_t(x, u)\|, \|\nabla_u c_t(x, u)\| \leq G_c D$.

Assumption 6. DAC controller $\pi(K, M)$ satisfies:

(1) K is (κ, γ)-strongly stable, whose precise definition is in Definition 4 of Appendix A.2;

(2) $M \in \mathcal{M}$ where $\mathcal{M} = \{M = (M^{[1]}, \ldots, M^{[H]}) \mid \|M^{[i]}\|_{\mathrm{op}} \leq \kappa_B \kappa^3 (1 - \gamma)^i\}$.

Theorem 7. *Under Assumptions 4–6, we set learning rate optimally and the step size pool $\mathcal{H}$ as*

$$\mathcal{H} = \left\{ \eta_i \;\middle|\; \eta_i = 2^{i-1} \cdot \sqrt{\frac{D_f^2}{(\lambda G_f + G_f^2)T}}, i \in [N] \right\}, \tag{15}$$

where $N = \left\lceil \frac{1}{2} \log_2(1 + T) \right\rceil + 1 = \mathcal{O}(\log T)$ is the number of base-learners, and $\lambda = (H + 2)^2 L_f$. The parameters L_f, G_f, D_f are defined in Lemma 29 and only depend on natural parameters of the linear dynamical system and truncated memory length H. By choosing $H = \Theta(\log T)$, our Scream.Control algorithm enjoys

$$\sum_{t=1}^{T} c_t(x_t, u_t) - \sum_{t=1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t}) \leq \widetilde{\mathcal{O}}\big(\sqrt{T(1 + P_T)}\big),$$

where $\pi_1, \ldots, \pi_T \in \Pi$ can be any comparator sequence in the compared DAC policy class $\Pi = \{\pi(K, M) \mid M \in \mathcal{M}\}$ with $\pi_t = \pi(K, M_t^*)$ for $t \in [T]$. The path length $P_T = \sum_{t=2}^{T} \|M_{t-1}^* - M_t^*\|_{\mathrm{F}}$ measures the cumulative variation of comparators.

Algorithm 4 System Identification via Random Inputs (Hazan et al., 2020)

```
Input: rounds of exploration $T_0$.
1: for $t = 1, \dots, T_0$ do
2:     Execute the control $u_t = -Kx_t + \widetilde{u}_t$ with $\widetilde{u}_t \sim_{i.i.d.} \{\pm 1\}^{d_u}$
3:     Record the observed state $x_{t+1}$
4: end for
5: Declare $N_j = \frac{1}{T_0 - k} \sum_{t=0}^{T_0-k-1} x_{t+j+1} \widetilde{u}_t^\top$, for all $j \in [k]$
6: Define $\widehat{C}_0 = [N_0, \dots, N_{k-1}], \widehat{C}_1 = [N_1, \dots, N_k]$ and return estimation $\widehat{A}, \widehat{B}$ as
       $$\widehat{B} = N_0, \quad \widehat{A}_K \triangleq \widehat{C}_1 \widehat{C}_0^\top \left( \widehat{C}_0 \widehat{C}_0^\top \right)^{-1}, \quad \widehat{A} = \widehat{A}_K + \widehat{B}K.$$
```

Till now, we assume the knowledge of the underlying system A and B. By further adopting the system identification via random inputs developed by Hazan et al. (2020), our result can be extended to online non-stochastic control with unknown systems. Indeed, when the system is unknown, i.e., A and B are not known in advance, we follow the explore-then-commit method of Hazan et al. (2020) to identify the underlying dynamics and then deploy the control algorithm based on the estimated system dynamics. The algorithmic descriptions are summarized in Algorithm 4. In the exploration phase, the identification algorithm (Hazan et al., 2020, Algorithm 2) uses some random inputs to approximately recover the system dynamics. Specifically, given an estimation budget $T_0 < T$, in the first T_0 rounds, we input the control signal $u_t = -Kx_t + \widetilde{u}_t$ with the random inputs $\widetilde{u}_t \sim \{\pm 1\}^{d_u}$ and then observe the corresponding state x_{t+1}. Then, by the estimation method presented in Line 6 of Algorithm 4, we can show that the estimation regret overhead is $\widetilde{\mathcal{O}}(T^{2/3})$ when choosing $T_0 = \Theta(T^{2/3})$.

To give the formal regret analysis and ensure finite-sample convergence rate, we focus on the system with strong controllability following the work of Hazan et al. (2020).

Definition 3 (Strong Controllability)**.** For a linear dynamical system and a strongly stable linear controller K, for $k \geq 1$, define a matrix $C_k \in \mathbb{R}^{d_x \times k d_u}$ as

$$C_k = \left[B, \widetilde{A}_K B, \dots, \widetilde{A}_K^{k-1} B \right], \tag{16}$$

where $\widetilde{A}_K = A - BK$. A linear dynamical system is controllable with controllability index k if C_k has full row-rank. In addition, such a system is also (k, κ_c)-strongly controllable if $\|\left(C_k C_k^\top\right)^{-1}\| \leq \kappa_c$.

Assumption 7 (Strong Controllability)**.** The dynamical system $x_{t+1} = Ax_t + Bu_t + w_t$ is (k, κ_c)-strongly controllable.

Theorem 8. *Under the same assumptions of Theorem 7 except that system matrices A and B are now unknown, and suppose the systems are strongly controllable (see Assumption 7) and the time horizon T is sufficiently large, Scream.Control with system identification (Algorithm 4) ensures that with high probability,*

$$\sum_{t=1}^{T} c_t(x_t, u_t) - \sum_{t=1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t}) \leq \widetilde{\mathcal{O}}(\sqrt{T(1 + P_T)} + T^{2/3}),$$

where $\pi_1, \ldots, \pi_T \in \Pi$ *can be any comparator sequence in the compared DAC policy class* $\Pi = \{\pi(K, M) \mid M \in \mathcal{M}\}$ *with* $\pi_t = \pi(K, M_t^*)$ *for* $t \in [T]$. *The path length* $P_T = \sum_{t=2}^{T} \|M_{t-1}^* - M_t^*\|_{\mathrm{F}}$ *measures the cumulative variation of comparators.*

Finally, we note that our obtained dynamic policy regret bound in Theorem 7 can recover the $\widetilde{\mathcal{O}}(\sqrt{T})$ static policy regret for non-stochastic control with known systems (Agarwal et al., 2019), and the result in Theorem 8 implies an $\widetilde{\mathcal{O}}(T^{2/3})$ high-probability static policy regret for non-stochastic control with unknown systems (Hazan et al., 2020).

Corollary 9. *For known systems, under the same assumptions of Theorem 7, it holds that Scream.Control enjoys a static policy regret at most*

$$\sum_{t=1}^{T} c_t(x_t, u_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} c_t(x_t^\pi, u_t^\pi) \leq \widetilde{\mathcal{O}}(\sqrt{T}).$$

For unknown systems, under the same assumptions of Theorem 8, Scream.Control with system identification ensures that with high probability,

$$\sum_{t=1}^{T} c_t(x_t, u_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} c_t(x_t^\pi, u_t^\pi) \leq \widetilde{\mathcal{O}}(T^{2/3}).$$

In above, the comparator set Π *can be chosen as either the set of DAC policies or the set of strongly linear controllers.*

6. Experiment

Although our paper mainly focuses on the theoretical investigation, in this section, we further present empirical studies to support our theoretical findings. We report the results of OCO with memory in Section 6.1 and online non-stochastic control in Section 6.2.

6.1 OCO with Memory

Since OCO with memory is essentially tackled by optimizing the upper bound of the policy regret, which consists of the vanilla regret over the unary functions and the switching cost, as explained in (9) for dynamic policy regret. Thus, in the empirical studies, we directly investigate the performance of different algorithms in optimizing this upper bound, i.e., the unary regret with switching cost. More specifically, we consider the following OCO with switching cost problem: at each round, the player predicts $\mathbf{w}_t \in \mathcal{W}$ and the environments choose the loss function $f_t : \mathcal{W} \mapsto \mathbb{R}$. The player will then suffer a loss of $f_t(\mathbf{w}_t)$ as well as a switching cost of $\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2$, and thus the overall loss is $f_t(\mathbf{w}_t) + \lambda\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2$ with some $\lambda > 0$ as the trade-off parameter.

Settings. We simulate the online learning scenario by the following setting: the player sequentially receives the feature of data item and then predicts its label. The data item of each round is denoted by $(\mathbf{x}_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a d-dimensional ball with diameter Γ and $\mathcal{Y} \in \mathbb{R}$ is the space of real values. The time horizon is set as $T = 50000$ and the dimension is set as $d = 10$. To simulate the distribution changes, we generate the output according to $\mathbf{y}_t = \mathbf{x}_t^\top \mathbf{w}_t^* + \varepsilon_t$, where $\mathbf{w}_t^* \in \mathbb{R}^d$ is the underlying model and $\varepsilon_t \in [0, 0.1]$ is a random noise. The underlying model $\mathbf{w}_t^*$ will change every 1000 rounds, randomly sampled from a d-dimensional ball with diameter $D/2$, so there are in total $S = 50$ changes. We the squared loss as loss functions, defined as $f_t(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^\top \mathbf{x}_t - y_t)^2$ and thus the gradient is $\nabla f_t(\mathbf{w}) = (\mathbf{w}^\top \mathbf{x}_t - y_t) \cdot \mathbf{x}_t$. The feasible set $\mathcal{W}$ is also set as d-dimensional ball with diameter $D/2$, and thus from all above settings, we know that $\|\mathbf{x}_t\|_2 \leq \Gamma$, $\|\mathbf{w}\|_2 \leq D/2$, and $\|\nabla f_t(\mathbf{w})\|_2 \leq D\Gamma^2$. We set $\Gamma = 1$ and $D = 2$, so the gradient norm is upper bounded by $G = D\Gamma^2 = 2$.

Contenders and Measure. We benchmark our proposed Scream algorithm with the following two algorithms: (1) OGD (Zinkevich, 2003), is the online gradient descent algorithm. The work of Anava et al. (2015) proves that this simple *static* regret minimization algorithm also enjoys a low switching cost when choosing the step size as $\eta = \mathcal{O}(1/\sqrt{T})$. (2) Ader (Zhang et al., 2018a), is the online algorithm designed in non-stationary online convex optimization. Ader is also in a meta-base structure to optimize the dynamic regret, but the algorithm does *not* consider the switching cost. Thus its switching cost might be huge (as analyzed in Section 4.2).

We examine the performance of all compared algorithms via the following three measures: (1) the overall cost $\sum_{t=1}^T f_t(\mathbf{w}_t) + \lambda \sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2$, (2) the cumulative loss $\sum_{t=1}^T f_t(\mathbf{w}_t)$, and (3) the switching cost $\lambda \sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2$. Here, we set the regularizer coefficient $\lambda = \alpha G$, where G is the gradient norm upper bound, with the purpose of matching the magnitude of cumulative loss and the switching cost. We consider three cases with different regularizer coefficients that impose different levels of penalty on the switching cost:

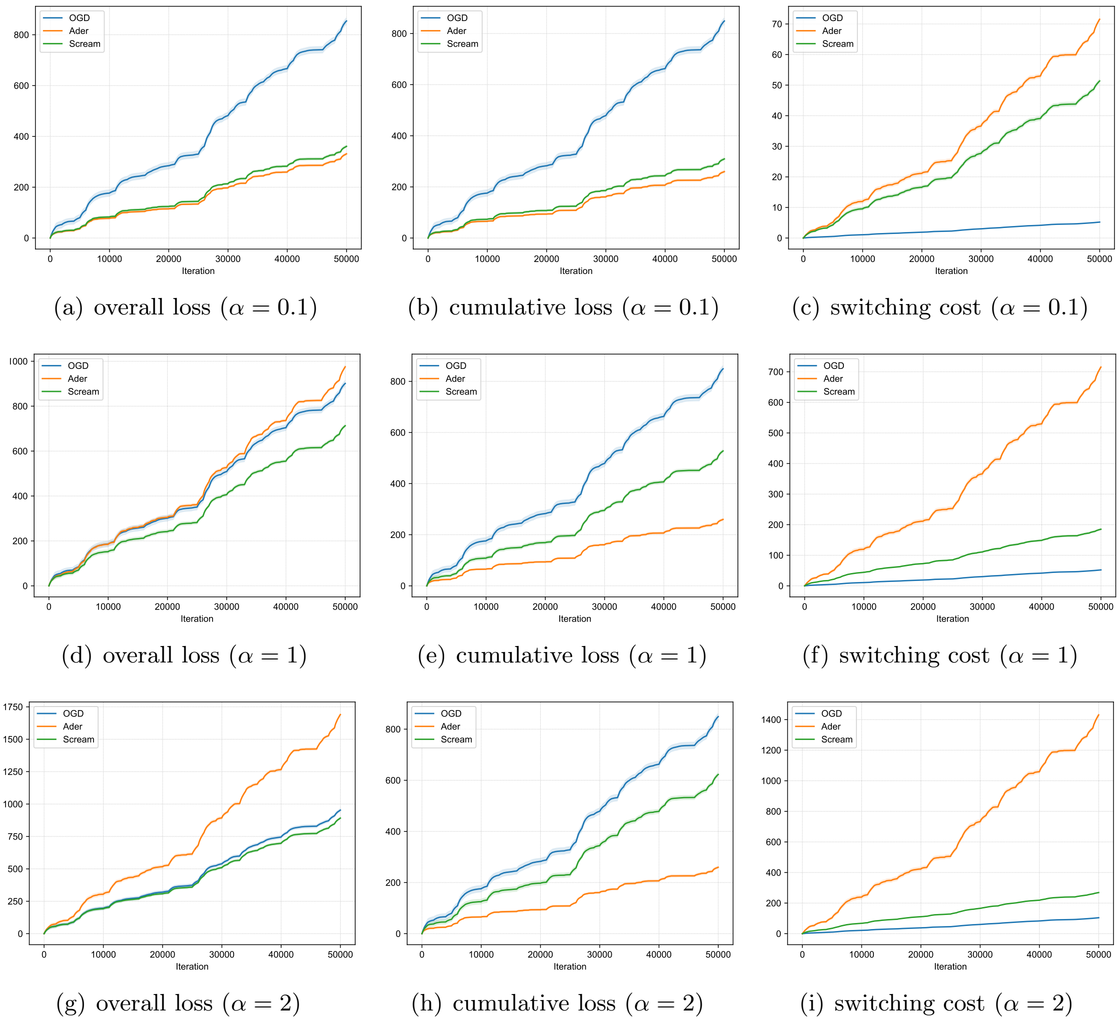
(i) small regularizer ($\alpha = 0.1$): in this case the switching cost is small so that optimizing the dynamic regret would dominate the performance;

(ii) medium regularizer ($\alpha = 1$): in this case the algorithm needs to have a good balance of dynamic regret and switching cost in order to behave well;

(iii) large regularizer ($\alpha = 2$): in this case dynamic regret is small so that optimizing the switching cost would dominate the performance.

We repeat the experiments five times and report the mean and standard deviation of different algorithms with respect to three performance measures (overall loss, cumulative loss, and switching cost).

Results. Figure 1 plots performance comparisons of three algorithms (OGD, Ader, Scream) under different regularizer coefficients. There are in total nine sub-figures, where each row presents the performance under a particular regularizer coefficient ($\alpha = 0.1, 1, 2$), and each column reports the performance in terms of a specific measure (overall loss, cumulative loss, and switching cost). For instance, Figure 1(d) plots the overall loss under the setting of $\lambda = \alpha G$ with $\alpha = 0.1$. We first focus on the measure of overall loss. From the results of overall loss (Figures 1(a), 1(d), 1(g)), we can see that under the case of small regularizer ($\alpha = 0.1$), Ader achieves the best, and Scream is comparable, while the performance of OGD is not good; with the medium regularizer ($\alpha = 1$), Scream evidently ranks the first, whereas Ader and OGD are not well-behaved; under the case of large regularizer ($\alpha = 2$), OGD performs surprisingly well, and Scream is comparable, whereas the performance of Ader is not desired. The results accord to our theory well, especially after a further examination of corresponding cumulative loss (Figures 1(b), 1(e), 1(h)) and switching cost (Figures 1(c),



(a) overall loss ($\alpha = 0.1$)

(b) cumulative loss ($\alpha = 0.1$)

(c) switching cost ($\alpha = 0.1$)

(d) overall loss ($\alpha = 1$)

(e) cumulative loss ($\alpha = 1$)

(f) switching cost ($\alpha = 1$)

(g) overall loss ($\alpha = 2$)

(h) cumulative loss ($\alpha = 2$)

(i) switching cost ($\alpha = 2$)

Figure 1: Performance comparisons of OGD, Ader, Scream, under different regularizer coefficients ($\lambda = \alpha G$, G is the gradient norm upper bound). The performance is evaluated by three measures: overall loss, cumulative loss, and switching cost.

1(f), 1(i)). Indeed, we can observe that Ader focuses on optimizing the dynamic regret (i.e., cumulative loss) but fails to control the switching cost; and OGD indeed yields a sequence of slow-moving decisions, but it fails to optimize the dynamic regret. Consequently, when the regularizer is small, one can optimize the overall loss by simply forgetting about the switching cost, and this is why Ader could behave well in this setting. Moreover, the switching cost plays a more important role in the overall loss with a large regularizer. Therefore, the algorithm can optimize the overall loss by simply producing a sequence of slow-moving decisions regardless of regret minimization. This is why OGD could achieve a surprisingly good performance in this setting. However, under the non-degenerate settings (for example, with medium regularizer), the two compared methods behave badly and Scream achieves the best. It is because our proposed Scream algorithm strikes a good balance between minimizing the dynamic regret and controlling the switching cost, owing to the novel online ensemble structure via the introduced switching-cost-regularized loss. Therefore, the above empirical studies demonstrate the effectiveness of our proposed algorithm and its algorithmic components.

6.2 Online Non-stochastic Control

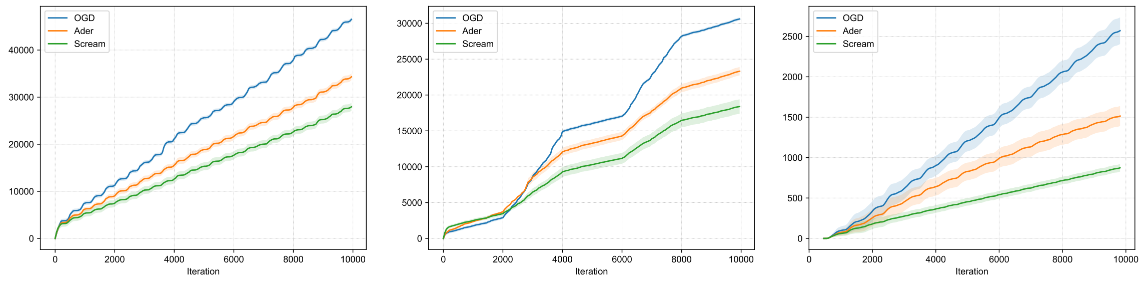
This part further examines the performance of our proposed algorithm in online non-stochastic control.

Settings. We conduct the experiments in synthetic linear dynamical system (LDS) environments and a real inverted pendulum environment. For the synthetic environment, we consider a time-varying LDS governed by $x_{t+1} = A_t x_t + B_t u_t + w_t$, where w_t is the Gaussian noise, A_t and B_t are the time-varying system matrices to be specified later. It is generally challenging to control time-varying systems, and we here consider a special case that can be handled by the online non-stochastic control framework. Specifically, we design the system matrices as $A_t = A + \Delta_{t,A}$ and $B_t = B + \Delta_{t,B}$, where A and B are fixed, and $\Delta_{t,A}, \Delta_{t,B}$ are time-varying zero-mean Gaussian random matrices. Notably, when applying online non-stochastic control methods, we only need to access A and B, and the changes of system matrices can be treated as a part of disturbance. Indeed, we have $x_{t+1} = Ax_t + Bu_t + (w_t + \Delta_{t,A}x_t + \Delta_{t,B}u_t) = Ax_t + Bu_t + \widetilde{w}_t$, where $\widetilde{w}_t$ is the effective disturbance of this time-varying system. Moreover, we choose the quadratic loss as the online cost function, defined as $c_t(x_t, u_t) = x_t^\top Q_t x_t + u_t^\top R_t u_t$, where $Q_t = a_t I$ and $R_t = b_t I$ change over time. By setting different a_t and b_t, we simulate the following two environments. (1) gradual change: in which $a_t = \sin(t/(10\pi))$ and $b_t = \sin(t/(20\pi))$; (2) abrupt change: the whole time horizon is divides into five stages, and the cost functions only change between different stages. In addition, we examine the performance in the real inverted pendulum environment, which is a commonly used benchmark consisting of a nonlinear and unstable system. The goal of this task is to balance the inverted pendulum by applying torque that will stabilize it in a vertically upright position. The state is a 2-dimensional vector denoted by $x_t = [\theta_t, \dot{\theta}_t]^\top$, where the first entry θ_t is the deviation angle normalized between $[-\pi, \pi]$ and the second entry $\dot{\theta}_t$ is the rotational velocity. The action is a 1-dimensional $u_t = \ddot{\theta}_t$ representing the torque applied on the system. The inverted pendulum environment is a non-linear dynamical system with transitions

$$x_{t+1} = \begin{bmatrix} \theta_{t+1} \\ \dot{\theta}_{t+1} \end{bmatrix} = \begin{bmatrix} \theta_t + c\dot{\theta}_t \\ \dot{\theta}_t + a\sin(\theta_t + \pi) + b\ddot{\theta}_t \end{bmatrix}.$$

and the online cost function is set as $c_t(x_t, u_t) = a_t\theta_t^2 + b_t\dot{\theta}_t^2 + c_t\ddot{\theta}_t^2$, where $a_t = \sin(t/(10\pi))$, $b_t = \sin(t/(20\pi))$, and $c_t = \sin(t/(20\pi))$ are slowly evolving parameters.

Contenders and Measure. We benchmark our proposed Scream.Control algorithm with the following two algorithms: (1) OGD.Control, which uses the OGD algorithm for the online non-stochastic control (Agarwal et al., 2019); (2) Ader.Control, Ader is an OCO algorithm (Zhang et al., 2018a) that admits a two-layer structure and enjoys dynamic regret guarantee. Although it cannot deal with the OCO with memory problem (see discussions



(a) LDS, gradual

(b) LDS, abrupt

(c) Inverted Pendulum

Figure 2: Performance comparisons of different algorithms. The performance is measured by the cumulative loss, the smaller the better. From left to right: (a) synthetic time-varying LDS with gradual changes; (b) synthetic time-varying LDS with abrupt changes; (c) real pendulum environments.

in Section 4.2), we apply it for online-non-stochastic control, serving to validate the effectiveness of our proposed switching-cost-regularized surrogate loss. We denote the three control algorithms simply as “OGD”, “Ader”, and “Scream” when there is no confusion. We record the cumulative loss as the performance measure, namely, $\sum_{t=1}^{T} c_t(x_t, u_t)$. We repeat the experiments five times and report the mean and standard deviation.

Results. Figure 2 plots the performance comparison of three algorithms (OGD, Ader, Scream) in terms of the cumulative cost. The result shows that our proposed algorithm outperforms the other two contenders, which validates that the meta-base structure (compared with OGD) and the switching-cost-regularizer (compared with Ader) are necessary for online non-stochastic control problems in non-stationary environments.

7. Conclusion

This paper investigates the dynamic policy regret of online convex optimization with memory and online non-stochastic control. For OCO with memory, we propose the Scream algorithm and prove an optimal $\mathcal{O}(\sqrt{T(1+P_T)})$ dynamic policy regret, where P_T is the path length of comparators that reflects the environmental non-stationarity. Our approach admits the meta-base online ensemble structure to handle uncertain environments and introduces a novel meta-base decomposition via switching-cost regularized loss to algorithmically address the tension between dynamic regret and switching cost. The approach is further used to design robust controllers for online non-stochastic control, where the underlying disturbance and cost functions could be chosen adversarially. We adopt the DAC parameterization and design the Scream.Control algorithm that provably achieves an $\widetilde{\mathcal{O}}(\sqrt{T(1+P_T)})$ dynamic policy regret, where P_T is the path length of compared controllers. Minimizing dynamic policy regret facilitates our controller with more robustness, since it can compete with any sequence of time-varying controllers instead of a fixed one.

In the future, we will explore the possibility of extension to *bandit* feedback, where the only feedback to the controller is the loss value (Cassel and Koren, 2020; Gradu et al., 2020a).

Moreover, it would be also intriguing to investigate whether dynamic policy regret can be improved when the cost functions are *strongly convex* or *exponentially concave* (Foster and Simchowitz, 2020; Baby and Wang, 2021, 2022).

Acknowledgments

Peng Zhao, Yu-Hu Yan, and Zhi-Hua Zhou were supported by the National Science Foundation of China (61921006, 62206125) and JiangsuSF (BK20220776) and the Collaborative Innovation Center of Novel Software Technology and Industrialization. Peng Zhao was also supported by National Postdoctoral Program for Innovative Talent. Yu-Xiang Wang was supported by a startup grant from UCSB CS Department. Part of this work was conducted while Peng Zhao remotely visited UCSB in summer 2020. The authors thank Ming Yin and Dheeraj Baby for helpful discussions. We are also grateful for the anonymous reviewers for their insightful comments.

Appendix A. Preliminaries

In this section, we present the preliminaries, including the dynamic regret results of memoryless online convex optimization, additional notions, and some technical lemmas.

A.1 Dynamic Regret of Memoryless OCO

In this part we present the dynamic regret analysis of the online gradient descent (OGD) algorithm for memoryless online convex optimization (Zinkevich, 2003; Zhang et al., 2018a; Zhao et al., 2020).

We first specify the problem settings and notations of memoryless online convex optimization. Specifically, the player iteratively selects a decision $\mathbf{w} \in \mathcal{W}$ from a convex set $\mathcal{W} \subseteq \mathbb{R}^d$ and then suffers a loss of $f_t(\mathbf{w}_t)$, in which the loss function $f_t : \mathcal{W} \mapsto \mathbb{R}$ is assumed to be convex and chosen adversarially by the environments. The performance measure we are concerned with is the *dynamic regret*, defined as

$$\text{D-Regret}_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{v}_t),$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$ is the comparator sequence arbitrarily chosen in the domain by the environments. The critical advantage of the above measure is that it supports to compete with a sequence of *time-varying* comparators, instead of a fixed one as specified in the standard (static) regret.

In the development of dynamic regret of memoryless OCO, one of the most crucial building blocks is the well-known Online Gradient Descent (OGD) algorithm (Zinkevich, 2003), which starts from any $\mathbf{w}_1 \in \mathcal{W}$ and performs the following update,

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t)]. \tag{17}$$

Here, $\eta > 0$ is the step size and $\Pi_{\mathcal{W}}[\cdot]$ denotes the Euclidean projection onto the nearest point in the feasible domain $\mathcal{W}$. The standard textbooks of online convex optimization (Shalev-Shwartz, 2012; Hazan, 2016) show that OGD can achieves an optimal $\mathcal{O}(\sqrt{T})$ static regret for convex functions, providing with appropriate step size settings. Furthermore, such a simple algorithm actually also enjoys the following dynamic regret guarantee (Zinkevich, 2003, Theorem 2), and we supply the proof for self-containedness.

Theorem 10. *Let $\mathcal{W} \in \mathbb{R}^d$ be a bounded convex and compact set in Euclidean space, and we denote by D an upper bound of the diameter of the domain, i.e., $\|\mathbf{w} - \mathbf{w}'\|_2 \leq D$ holds for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$. Suppose the gradient norm of f_t over $\mathcal{W}$ is bounded by G, i.e., $\|\nabla f_t(\mathbf{w})\|_2 \leq G$ holds for any $\mathbf{w} \in \mathcal{W}$ and $t \in [T]$. Then, OGD (17) enjoys the following dynamic regret,*

$$\text{D-Regret}_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) \leq \frac{\eta}{2} G^2 T + \frac{1}{2\eta}(D^2 + 2DP_T),$$

which holds for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$, and $P_T = \sum_{t=2}^T \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2$ is the path length that measures the cumulative movements of the comparator sequence.

Proof Since the online functions are convex, we have

$$\text{D-Regret}_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) = \sum_{t=1}^T f_t(\mathbf{w}_t) - \sum_{t=1}^T f_t(\mathbf{v}_t) \leq \sum_{t=1}^T \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}_t \rangle.$$

Thus, it suffices to bound the sum of $\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}_t \rangle$ over iterations. Note that from the update rule in (52),

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{v}_t\|_2^2 &= \|\Pi_{\mathcal{X}}[\mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t)] - \mathbf{v}_t\|_2^2 \\ &\leq \|\mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 \\ &= \eta^2 \|\nabla f_t(\mathbf{w}_t)\|_2^2 - 2\eta \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}_t \rangle + \|\mathbf{w}_t - \mathbf{v}_t\|_2^2 \end{aligned}$$

The inequality holds due to Pythagorean theorem (Hazan, 2016, Theorem 2.1). After rearranging, we obtain

$$\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}_t \rangle \leq \frac{\eta}{2} \|\nabla f_t(\mathbf{w}_t)\|_2^2 + \frac{1}{2\eta} \left(\|\mathbf{w}_t - \mathbf{v}_t\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{v}_t\|_2^2 \right).$$

Summing the above inequality from $t = 1$ to T yields,

$$\text{D-Regret}_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) \leq \frac{\eta}{2} \sum_{t=1}^T \|\nabla f_t(\mathbf{w}_t)\|_2^2 + \frac{1}{2\eta} \sum_{t=1}^T \left(\|\mathbf{w}_t - \mathbf{v}_t\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{v}_t\|_2^2 \right).$$

We further provide an upper bound for the second term on the right-hand side. Indeed,

$$\begin{aligned} & \sum_{t=1}^T \left(\|\mathbf{w}_t - \mathbf{v}_t\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{v}_t\|_2^2 \right) \\ \leq\ & \sum_{t=1}^T \|\mathbf{w}_t - \mathbf{v}_t\|_2^2 - \sum_{t=2}^T \|\mathbf{w}_t - \mathbf{v}_{t-1}\|_2^2 \\ \leq\ & \|\mathbf{w}_1 - \mathbf{v}_1\|_2^2 + \sum_{t=2}^T \left(\|\mathbf{w}_t - \mathbf{v}_t\|_2^2 - \|\mathbf{w}_t - \mathbf{v}_{t-1}\|_2^2 \right) \end{aligned}$$

$$= \|\mathbf{w}_1 - \mathbf{v}_1\|_2^2 + \sum_{t=2}^{T} \langle \mathbf{v}_{t-1} - \mathbf{v}_t, 2\mathbf{w}_t - \mathbf{v}_{t-1} - \mathbf{v}_t \rangle \leq D^2 + 2D \sum_{t=2}^{T} \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2.$$

Combining all above inequalities, we have

$$\begin{aligned} \text{D-Regret}_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) &\leq \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{w}_t)\|_2^2 + \frac{1}{2\eta} \left(D^2 + 2D \sum_{t=2}^{T} \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2 \right) \\ &\leq \frac{\eta}{2} G^2 T + \frac{1}{2\eta}(D^2 + 2DP_T). \end{aligned}$$

Hence, we complete the proof. ■

A.2 Additional Notions

We introduce the formal definition of strongly stable linear controllers (Cohen et al., 2018; Agarwal et al., 2019). Indeed, the stable condition can guarantee the convergence, but nothing can be ensured about the rate of convergence. While working on the class of strongly stable controllers, we can establish the non-asymptotic convergence rate.

Definition 4. A linear controller K is (κ, γ)-strongly stable if there exist matrices L, H satisfying $A - BK = HLH^{-1}$, such that the following two conditions are satisfied:

(i) The spectral norm of L satisfies $\|L\| \leq 1 - \gamma$.

(ii) The controller and transforming matrices are bounded, i.e., $\|K\|, \|H\|, \|H^{-1}\| \leq \kappa$.

A.3 Technical Lemmas

The following lemmas are important in analyzing algorithms based on the mirror descent.

Lemma 11 (Lemma 3.2 of Chen and Teboulle (1993))**.** *Let $\mathcal{X}$ be a convex set in a Banach space $\mathcal{B}$ and $f : \mathcal{X} \mapsto \mathbb{R}$ be a closed proper convex function on $\mathcal{X}$. Given a convex regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ and its induced Bregman divergence $\mathcal{D}_\psi(\cdot, \cdot)$, any update of the form*

$$\mathbf{x}_k = \arg\min_{\mathbf{x} \in \mathcal{X}} \{ f(\mathbf{x}) + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_{k-1}) \}$$

satisfies the following inequality for any $\mathbf{u} \in \mathcal{X}$,

$$f(\mathbf{x}_k) - f(\mathbf{u}) \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{k-1}) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_k) - \mathcal{D}_\psi(\mathbf{x}_k, \mathbf{x}_{k-1}).$$

Lemma 12. *If the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is λ-strongly convex with respect to a norm $\|\cdot\|$, then the induced Bregman divergence is lower-bounded as $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) \geq \frac{\lambda}{2}\|\mathbf{x} - \mathbf{y}\|^2$.*

Proof By the definition of strong convexity, we know that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\psi(\mathbf{x}) \geq \psi(\mathbf{y}) + \nabla\psi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{y}\|^2$. Reformulating the inequality and combining the definition of Bregman divergence, we know that $D_\psi(\mathbf{x}, \mathbf{y}) \triangleq \psi(\mathbf{x}) - \psi(\mathbf{y}) + \nabla\psi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) \geq \frac{\lambda}{2}\|\mathbf{x} - \mathbf{y}\|^2$, which ends the proof. ■

The following concentration inequality is used in analyzing dynamic policy regret for non-stochastic control with unknown systems.

Lemma 13 (Azuma-Hoeffding's Inequality for Vectors (Hayes, 2005, Theorem 1.8))**.** *Suppose that $S_m = \sum_{t=1}^m X_t$ is a martingale where $X_1, \ldots, X_m$ take values in $\mathbb{R}^n$ and are such that $\mathbb{E}[X_t] = \mathbf{0}$ and $\|X_t\|_2 \leq D$ for all t, for $t > 0$. Then for every $\varepsilon > 0$,*

$$\Pr[\|S_m\|_2 \geq \varepsilon] \leq 2e^2 e^{-\frac{\varepsilon^2}{2mD^2}}.$$

Appendix B. Omitted Details for Section 4 (OCO with Memory)

In this section, we present omitted details for Section 4 OCO with memory, including proofs of Theorem 1 (in Appendix B.1) and Theorem 2 (in Appendix B.4). Moreover, we provide the proof of the switching cost decomposition (5) in Appendix B.2 and supply more details for the online mirror descent in Appendix B.3. The proofs of Theorem 2, Theorem 3, Theorem 4, Theorem 5 are listed in the following sections. We finally discuss the memory dependence in Appendix B.8.

B.1 Proof of Theorem 1

Proof The coordinate-Lipschitz continuity of f_t (Assumption 1) implies that

$$|f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t) - \widetilde{f}_t(\mathbf{w}_t)| \leq L \cdot \sum_{i=1}^m \|\mathbf{w}_t - \mathbf{w}_{t-i}\|_2 \leq mL \sum_{i=1}^m \|\mathbf{w}_{t-i+1} - \mathbf{w}_{t-i}\|_2.$$

Therefore, we have

$$\sum_{t=m}^T f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t) - \sum_{t=m}^T \widetilde{f}_t(\mathbf{w}_t) \leq m^2 L \sum_{t=m}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2, \tag{18}$$

and the dynamic policy regret can be thus upper bounded by

$$\begin{aligned} \text{D-Regret}_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) &= \sum_{t=1}^T f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t) - \sum_{t=1}^T f_t(\mathbf{v}_{t-m}, \ldots, \mathbf{v}_t) \\ &\overset{(18)}{\leq} \underbrace{\sum_{t=1}^T \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^T \widetilde{f}_t(\mathbf{v}_t)}_{\texttt{dynamic regret over unary loss}} + \underbrace{\lambda \sum_{t=1}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2}_{\texttt{switching cost of decisions}} + \underbrace{\lambda \sum_{t=1}^T \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2}_{\texttt{switching cost of comparators}}, \end{aligned} \tag{19}$$

where we define $\lambda \triangleq m^2 L$ for notational convenience. Note that the first term is the dynamic regret over the unary loss, which is optimized by OGD over the unary loss. Since the sequence of unary loss $\{\widetilde{f}_t\}_{t=1}^T$ is convex and *memoryless*, from the standard dynamic regret analysis (Zinkevich, 2003; Zhang et al., 2018a), as shown in Theorem 10, we get

$$\sum_{t=1}^T \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^T \widetilde{f}_t(\mathbf{v}_t) \leq \frac{\eta}{2} G^2 T + \frac{1}{2\eta}(D^2 + 2DP_T), \tag{20}$$

where $P_T = \sum_{t=2}^T \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2$ is the path length measuring the fluctuation of the comparator sequence $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T$. Next, the last term of (19) is the switching cost of the comparators, which is exactly the path length λP_T.

So we only need to further examine the switching cost of the decisions, i.e., $\sum_{t=2}^{T} \|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2$, as well as the dynamic regret over the unary loss, i.e., $\sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t)$. By the non-expansive property of the projection operator, we can derive an upper bound for the switching cost:

$$\sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 = \sum_{t=1}^{T} \|\Pi_{\mathcal{W}}[\mathbf{w}_{t-1} - \eta \nabla \widetilde{f}_t(\mathbf{w}_{t-1})] - \mathbf{w}_{t-1}\|_2 \leq \eta \sum_{t=1}^{T} \|\nabla \widetilde{f}_t(\mathbf{w}_{t-1})\|_2 \leq \eta G T. \tag{21}$$

Combining above two inequalities (21) and (20) yields

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{v}_{t-m}, \ldots, \mathbf{v}_t) \leq \frac{\eta}{2}(G^2 + 2\lambda G)T + \frac{1}{2\eta}(D^2 + 2DP_T) + \lambda P_T,$$

with $\lambda = m^2 L$. We thus compete the proof. ■

B.2 Proof of Switching Cost Decomposition

The following lemma restates the switching cost decomposition presented in (5).

Lemma 14. *The switching cost of meta-base outputs can be upper bounded as*

$$\sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \leq D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \sum_{t=2}^{T} \sum_{i=1}^{N} p_{t,i} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2.$$

Proof By the meta-base structure, the final decision of each round is $\mathbf{w}_t = \sum_{i=1}^{N} p_{t,i} \mathbf{w}_{t,i}$. Therefore, we can expand the switching cost of the final prediction sequence as

$$\begin{aligned}
\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 &= \left\| \sum_{i=1}^{N} p_{t,i} \mathbf{w}_{t,i} - \sum_{i=1}^{N} p_{t-1,i} \mathbf{w}_{t-1,i} \right\|_2 \\
&\leq \left\| \sum_{i=1}^{N} p_{t,i} \mathbf{w}_{t,i} - \sum_{i=1}^{N} p_{t,i} \mathbf{w}_{t-1,i} \right\|_2 + \left\| \sum_{i=1}^{N} p_{t,i} \mathbf{w}_{t-1,i} - \sum_{i=1}^{N} p_{t-1,i} \mathbf{w}_{t-1,i} \right\|_2 \\
&\leq \sum_{i=1}^{N} p_{t,i} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 + D \sum_{i=1}^{N} |p_{t,i} - p_{t-1,i}| \\
&= \sum_{i=1}^{N} p_{t,i} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 + D \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1,
\end{aligned} \tag{22}$$

where the second step holds due to the triangle inequality and the third step is true due to the boundedness of the feasible domain (Assumption 3). Hence, we complete the proof. ■

B.3 Additional Results for Online Mirror Descent

In this section, we present additional results and descriptions for Online Mirror Descent (OMD), which enables a unified view for algorithm design of both meta-algorithm and base-algorithm.

Consider the standard online convex optimization setting, and the sequence of online convex functions are $\{h_t\}_{t=1,\ldots,T}$ with $h_t : \mathcal{W} \mapsto \mathbb{R}$. Online mirror descent starts from any $\mathbf{w}_1 \in \mathcal{W}$, and at iteration t, the algorithm performs the following update:

$$\mathbf{w}_{t+1} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}} \eta\langle \nabla h_t(\mathbf{w}_t), \mathbf{w}\rangle + \mathcal{D}_\psi(\mathbf{w}, \mathbf{w}_t), \tag{23}$$

where $\eta > 0$ is the step size. The regularizer $\psi : \mathcal{W} \mapsto \mathbb{R}$ is a differentiable convex function defined on $\mathcal{W}$ and is assumed (without loss of generality) to be 1-strongly convex w.r.t. some norm $\|\cdot\|$ over $\mathcal{W}$. The induced Bregman divergence $\mathcal{D}_\psi$ is defined by $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$.

The following generic result gives an upper bound of dynamic regret with switching cost of OMD, which can be regarded as a generalization of Theorem 1 from gradient descent (for Euclidean norm) to mirror descent (for general primal-dual norm).

Theorem 15. *Online Mirror Descent* (23) *satisfies that*

$$\sum_{t=1}^T h_t(\mathbf{w}_t) - \sum_{t=1}^T h_t(\mathbf{v}_t) + \lambda \sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\| \le \frac{1}{\eta}\left(R^2 + \gamma P_T\right) + \eta(\lambda G + G^2)T, \tag{24}$$

provided that $\mathcal{D}_\psi(\mathbf{x}, \mathbf{z}) - \mathcal{D}_\psi(\mathbf{y}, \mathbf{z}) \le \gamma\|\mathbf{x} - \mathbf{y}\|$ holds for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{W}$. In above, $R^2 = \sup_{\mathbf{x},\mathbf{y}\in\mathcal{W}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y})$, and $G = \sup_{\mathbf{w}\in\mathcal{W},t\in[T]} \|\nabla h_t(\mathbf{w})\|_*$. Note that the above result holds for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$.

Remark 3. The dynamic regret of Theorem 15 holds against *any* comparator sequence in the domain. In particular, we can set them as the best fixed decision in hindsight and thus obtain static regret with switching cost, $\sum_{t=1}^T h_t(\mathbf{w}_t) - \sum_{t=1}^T h_t(\mathbf{w}^*) + \lambda\sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\| \le R^2/\eta + \eta(\lambda G + G^2)T$, that holds for any $\mathbf{w}^* \in \mathcal{W}$. A technical caveat is that when deriving the static regret, the Bregman divergence is not required to satisfy the Lipschitz condition.

Theorem 15 exhibits a general analysis for the dynamic regret and switching cost of OMD. By flexibly choosing the regularizer ψ and comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T$, we have the following two implications, which correspond to base-regret (dynamic regret with switching cost of OGD) and meta-regret (static regret with switching cost of Hedge) respectively.

Before presenting the proof of Theorem 15, we first analyze the switching cost of the online mirror descent, as demonstrated in the following stability lemma.

Lemma 16. *For Online Mirror Descent* (23)*, the instantaneous switching cost is at most*

$$\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \le \eta\|\nabla h_t(\mathbf{w}_t)\|_*. \tag{25}$$

Proof From the update procedure of OMD (23) and Lemma 11, we know that

$$\langle \mathbf{w}_{t+1} - \mathbf{w}_t, \eta\nabla h_t(\mathbf{w}_t)\rangle \le \mathcal{D}_\psi(\mathbf{w}_t, \mathbf{w}_t) - \mathcal{D}_\psi(\mathbf{w}_t, \mathbf{w}_{t+1}) - \mathcal{D}_\psi(\mathbf{w}_{t+1}, \mathbf{w}_t),$$

which implies

$$\mathcal{D}_\psi(\mathbf{w}_t, \mathbf{w}_{t+1}) + \mathcal{D}_\psi(\mathbf{w}_{t+1}, \mathbf{w}_t) \le \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \eta \nabla h_t(\mathbf{w}_t) \rangle.$$

Since the regularizer ψ is chosen as a 1-strongly convex function with respect to the norm $\|\cdot\|$, by Lemma 12 we have

$$\mathcal{D}_\psi(\mathbf{w}_t, \mathbf{w}_{t+1}) + \mathcal{D}_\psi(\mathbf{w}_{t+1}, \mathbf{w}_t) \ge \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2.$$

Combining above two inequalities and further applying the Hölder's inequality, we obtain

$$\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \le \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \eta \nabla h_t(\mathbf{w}_t) \rangle \le \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \|\eta \nabla h_t(\mathbf{w}_t)\|_*.$$

Therefore, we conclude that $\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \le \eta \|\nabla h_t(\mathbf{w}_t)\|_*$ and finish the proof. ■

Based on the above stability lemma, we can now prove Theorem 15 regarding dynamic regret with switching cost for OMD.

Proof [of Theorem 15] Notice that the dynamic regret can be decomposed as follows:

$$\begin{aligned}\sum_{t=1}^T h_t(\mathbf{w}_t) - \sum_{t=1}^T h_t(\mathbf{v}_t) \le {} & \sum_{t=1}^T \langle \nabla h_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}_t \rangle \\ = {} & \underbrace{\sum_{t=1}^T \langle \nabla h_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_{t+1} \rangle}_{\texttt{term (a)}} + \underbrace{\sum_{t=1}^T \langle \nabla h_t(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{v}_t \rangle}_{\texttt{term (b)}}.\end{aligned}$$

From Lemma 16 and Hölder's inequality, we have

$$\texttt{term (a)} \le \sum_{t=1}^T \|\nabla h_t(\mathbf{w}_t)\|_* \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \le \eta \sum_{t=1}^T \|\nabla h_t(\mathbf{w}_t)\|_*^2. \tag{26}$$

Next, we investigate the term (b):

$$\begin{aligned}\texttt{term (b)} & \le \frac{1}{\eta} \sum_{t=1}^T \left(\mathcal{D}_\psi(\mathbf{v}_t, \mathbf{w}_t) - \mathcal{D}_\psi(\mathbf{v}_t, \mathbf{w}_{t+1}) - \mathcal{D}_\psi(\mathbf{w}_{t+1}, \mathbf{w}_t)\right) \\ & \le \frac{1}{\eta} \sum_{t=2}^T \left(\mathcal{D}_\psi(\mathbf{v}_t, \mathbf{w}_t) - \mathcal{D}_\psi(\mathbf{v}_{t-1}, \mathbf{w}_t)\right) + \mathcal{D}_\psi(\mathbf{v}_1, \mathbf{w}_1) \\ & \le \frac{\gamma}{\eta} \sum_{t=2}^T \|\mathbf{v}_t - \mathbf{v}_{t-1}\| + \frac{1}{\eta} R^2,\end{aligned} \tag{27}$$

where the first inequality holds due to Lemma 11, and the second inequality makes uses of the non-negativity of the Bregman divergence. The last inequality holds due to the assumption of Lipschitz property that $\mathcal{D}_\psi(\mathbf{x}, \mathbf{z}) - \mathcal{D}_\psi(\mathbf{y}, \mathbf{z}) \le \gamma \|\mathbf{x} - \mathbf{y}\|$ holds for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{W}$. Furthermore, the switching cost can be bounded by Lemma 16,

$$\sum_{t=2}^T \|\mathbf{w}_t - \mathbf{w}_{t-1}\| \le \eta \sum_{t=2}^T \|\nabla h_{t-1}(\mathbf{w}_{t-1})\|_*. \tag{28}$$

Combining (26), (27), and (28), we can attain that

$$\begin{aligned}
& \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\| + \sum_{t=1}^{T} h_t(\mathbf{w}_t) - \sum_{t=1}^{T} h_t(\mathbf{v}_t) \\
\leq {} & \frac{1}{\eta}(R^2 + \gamma P_T) + \eta \sum_{t=1}^{T} (\lambda \|\nabla h_t(\mathbf{w}_t)\|_* + \|\nabla h_{t-1}(\mathbf{w}_{t-1})\|_*^2) \\
\leq {} & \frac{1}{\eta}(R^2 + \gamma P_T) + \eta(\lambda G + G^2)T,
\end{aligned}$$

which finishes the proof. ■

As we mentioned earlier, Theorem 1 can be regarded as a corollary of Theorem 15, by specifying the Euclidean norm and $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$. We give a formal statement in the following corollary.

Corollary 17. *Setting the ℓ_2 regularizer $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ and step size $\eta > 0$ for OMD, suppose $\|\nabla \widetilde{f}_t(\mathbf{w})\|_2 \leq G$ and $\|\mathbf{w} - \mathbf{w}'\|_2 \leq D$ hold for all $\mathbf{w} \in \mathcal{W}$ and $t \in [T]$, then we have*

$$\lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 + \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) \leq (G^2 + \lambda G)\eta T + \frac{1}{2\eta}(D^2 + 2DP_T), \tag{29}$$

which holds for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$, and $P_T = \sum_{t=2}^{T} \|\mathbf{v}_{t-1} - \mathbf{v}_t\|_2$ is the path length that measures the cumulative movements of the comparator sequence.

Further, we present a corollary regarding the static regret with switching cost for the meta-algorithm, which is essentially a specialization of OMD algorithm by setting the negative-entropy regularizer.

Corollary 18. *Setting the negative-entropy regularizer $\psi(\boldsymbol{p}) = \sum_{i=1}^{N} p_i \log p_i$ and learning rate $\varepsilon > 0$ for OMD, suppose $\|\boldsymbol{\ell}_t\|_\infty \leq G$ holds for any $t \in [T]$ and the algorithm starts from the initial weight $p_1 \in \Delta_N$, then we have*

$$\lambda \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \sum_{t=1}^{T} \ell_{t,i} \leq \frac{\ln(1/p_{1,i})}{\varepsilon} + \varepsilon(\lambda G + G^2)T. \tag{30}$$

Proof From the proof of Theorem 15, we can easily obtain that

$$\lambda \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \sum_{t=1}^{T} \ell_{t,i} \leq \frac{\mathcal{D}_\psi(\mathbf{e}_i, \boldsymbol{p}_1)}{\varepsilon} + \varepsilon(\lambda G + G^2)T.$$

When choosing the negative-entropy regularizer, the induced Bregman divergence becomes Kullback-Leibler divergence, i.e., $\mathcal{D}_\psi(\boldsymbol{q}, \boldsymbol{p}) = \mathrm{KL}(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^{N} q_i \ln(q_i/p_i)$. Therefore, $\mathcal{D}_\psi(\boldsymbol{e}_i, \boldsymbol{p}_1) = \ln(1/p_{1,i})$, which implies the desired result. ■

B.4 Proof of Theorem 2

Proof As indicated in (19), the dynamic policy regret can be upper bounded by three terms, including dynamic regret over the unary regret, switching cost of decisions, and switching cost of comparators. The third term is essentially the path length of the comparators, and we focus on the first two terms.

$$\begin{aligned}
&\sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \\
\overset{(5)}{\leq}\; & \sum_{t=1}^{T} \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}_t \rangle + \lambda D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \lambda \sum_{t=2}^{T} \sum_{i=1}^{N} p_{t,i} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 \\
=\; & \sum_{t=1}^{T} \sum_{i=1}^{N} p_{t,i} \Big(\langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w}_{t,i} \rangle + \lambda \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 \Big) - \sum_{t=1}^{T} \Big(\langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w}_{t,i} \rangle + \lambda \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 \Big) \\
& \qquad + \lambda D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \sum_{t=1}^{T} \Big(\langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w}_{t,i} \rangle - \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{v}_t \rangle \Big) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 \\
=\; & \underbrace{\sum_{t=1}^{T} \big(\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i} \big) + \lambda D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1}_{\texttt{meta-regret}} + \underbrace{\sum_{t=1}^{T} \big(g_t(\mathbf{w}_{t,i}) - g_t(\mathbf{v}_t) \big) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2}_{\texttt{base-regret}},
\end{aligned}$$

where the last step uses the convexity of $\widetilde{f}_t$ and the definition of linearized loss $g_t(\mathbf{w}) = \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w} \rangle$. We will formally prove that our proposed algorithm optimizes the right-hand side of above inequality.

Bounding Meta-regret. Denote by $\mathbf{e}_i$ the i-th standard basis of $\mathbb{R}^N$-space and by $\lambda' = \lambda D$ for simplicity. Denote by $G_{\text{meta}} = \max_{t \in [T]} \|\boldsymbol{\ell}_t\|_\infty$ the maximum scale of the loss of meta-algorithm. Since the meta-algorithm actually performs Hedge over the switching-cost-regularized loss $\boldsymbol{\ell}_t \in \mathbb{R}^N$, Corollary 18 implies that for any $i \in [N]$,

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \sum_{t=1}^{T} \ell_{t,i} + \lambda' \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 &\leq \varepsilon(\lambda' G_{\text{meta}} + G_{\text{meta}}^2) T + \frac{\mathcal{D}_\psi(\mathbf{e}_i, \boldsymbol{p}_1)}{\varepsilon} \\
&= \varepsilon(\lambda D + G_{\text{meta}}) G_{\text{meta}} T + \frac{\ln(1/p_{1,i})}{\varepsilon} \qquad (31) \\
&\leq \varepsilon(\lambda D + G_{\text{meta}}) G_{\text{meta}} T + \frac{2\ln(i+1)}{\varepsilon},
\end{aligned}$$

where the last step holds because we adopt a non-uniform weight initialization with the initial weight $\boldsymbol{p}_1 \in \Delta_N$ set as $p_{1,i} = \frac{1}{i(i+1)} \cdot \frac{N+1}{N}$ for any $i \in [N]$. By choosing the learning rate as $\varepsilon = \varepsilon^* = \sqrt{\frac{2}{G_{\text{meta}}(\lambda D + G_{\text{meta}})T}}$, we can obtain the following upper bound for the meta-regret,

$$\sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \sum_{t=1}^{T} \ell_{t,i} + \lambda' \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 \leq \sqrt{2G_{\text{meta}}(\lambda D + G_{\text{meta}})T}\,(1 + \ln(i+1)). \quad (32)$$

Note that the dependence of learning rate tuning on T can be removed by either a time-varying tuning or doubling trick. We now present an upper bound for G_{meta}, indeed,

$$\begin{aligned}\ell_{t,i} &= \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w}_{t,i}\rangle + \lambda\|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 \le \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w}_{t,i}\rangle + \lambda\eta_i\|\nabla\widetilde{f}_t(\mathbf{w}_t)\|_2 \\ &\le GD + \lambda\eta_i G \le GD+\lambda\eta_N G \le GD\left(1+2\lambda\sqrt{\frac{1}{\lambda G+G^2}}\right) = \mathcal{O}(\sqrt{\lambda}). \end{aligned}\tag{33}$$

Bounding Base-regret. As specified by our algorithm, there are multiple base-learners, each performing OGD over the linearized loss with a particular step size $\eta_i \in \mathcal{H}$ for base-learner $\mathcal{B}_i$:

$$\mathbf{w}_{t+1,i} = \Pi_{\mathcal{W}}[\mathbf{w}_{t,i} - \eta_i \nabla g_t(\mathbf{w}_{t,i})] = \Pi_{\mathcal{W}}[\mathbf{w}_{t,i} - \eta_i\nabla\widetilde{f}_t(\mathbf{w}_t)].$$

As a result, Theorem 15 implies that the base-regret satisfies

$$\sum_{t=1}^T g_t(\mathbf{w}_{t,i}) - \sum_{t=1}^T g_t(\mathbf{v}_t) + \lambda\sum_{t=2}^T \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 \le (G^2+\lambda G)\eta_i T + \frac{1}{2\eta_i}(D^2+2DP_T), \tag{34}$$

which holds for any comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$ as well as any base-learner $i \in [N]$.

Bounding Overall Dynamic Regret. Due to the boundedness of the path length, we know that the optimal step size η_* provably lies in the range of $[\eta_1, \eta_N]$. Furthermore, by the construction of the pool of candidate step sizes, we can confirm that there exists an index $i^* \in [N]$ ensuring $\eta_{i^*} \le \eta_* \le \eta_{i^*+1} = 2\eta_{i^*}$. Therefore, we have

$$i^* \le \left\lceil \frac{1}{2}\log_2\left(1+\frac{2P_T}{D}\right)\right\rceil + 1. \tag{35}$$

Notice that the meta-base decomposition at the beginning of the proof holds for any index of base-learners $i \in [N]$. Thus, in particular, we can choose the index i^* and achieve the following result by using the upper bounds of meta-regret (32) and base-regret (34).

$$\begin{aligned}&\sum_{t=1}^T \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^T \widetilde{f}_t(\mathbf{v}_t) + \lambda\sum_{t=2}^T\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \\ &\le \underbrace{\sum_{t=1}^T \big(\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle - \ell_{t,i^*}\big) + \lambda D\sum_{t=2}^T\|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1}_{\texttt{meta-regret}} + \underbrace{\sum_{t=1}^T \big(g_t(\mathbf{w}_{t,i^*}) - g_t(\mathbf{v}_t)\big) + \lambda\sum_{t=2}^T\|\mathbf{w}_{t,i^*} - \mathbf{w}_{t-1,i^*}\|_2}_{\texttt{base-regret}} \\ &\le \sqrt{2G_{\text{meta}}(\lambda D + G_{\text{meta}})T}\,(1+\ln(i^*+1)) + (G^2+\lambda G)\eta_{i^*}T + \frac{1}{2\eta_{i^*}}(D^2+2DP_T) \\ &\le \sqrt{2G_{\text{meta}}(\lambda D + G_{\text{meta}})T}\,(1+\ln(i^*+1)) + (G^2+\lambda G)\eta_* T + \frac{1}{\eta_*}(D^2+2DP_T) \\ &\lesssim \sqrt{2(GD+\sqrt{\lambda})(\lambda D + GD + \sqrt{\lambda})T}\,(1+\ln(i^*+1)) + \sqrt{(G^2+\lambda G)(D^2+2DP_T)T} \qquad (36) \\ &\le \mathcal{O}\left(\lambda^{\frac{3}{4}}\sqrt{T}(1+\log\log P_T)\right) + \mathcal{O}\left(\sqrt{\lambda T(1+P_T)}\right),\end{aligned}$$

where in (36), we use $a \lesssim b$ to represent $a = \mathcal{O}(b)$. Therefore, we have

$$\begin{aligned}\text{D-Regret}_T(\mathbf{v}_{1:T}) \leq{}& \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 + \lambda \sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2 \\ \leq{}& \mathcal{O}\left(\lambda^{\frac{3}{4}}\sqrt{T}(1 + \log\log P_T) + \sqrt{\lambda T(1 + P_T)} + \lambda P_T\right) \leq \mathcal{O}(\sqrt{T(1 + P_T)}).\end{aligned}$$

The last step omits the dependence on λ. Moreover, the inequality holds due to the following observation:

$$\begin{aligned}\text{D-Regret}_T(\mathbf{v}_{1:T}) &\leq \mathcal{O}(\sqrt{T(1 + P_T)}) + \mathcal{O}(P_T) \\ &\leq \mathcal{O}(\sqrt{T(1 + P_T) + P_T^2}) \qquad\qquad (\sqrt{a} + \sqrt{b} \leq \sqrt{2(a + b)}) \\ &= \mathcal{O}(\sqrt{T + (T + P_T)P_T}) \\ &\leq \mathcal{O}(\sqrt{T(1 + P_T)}),\end{aligned}$$

where the last step holds as $P_T = \sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2 \leq DT$ due to the boundedness of the domain. We hence complete the proof of Theorem 2. ■

B.5 Proof of Theorem 3

Proof First, we show that for any $\lambda > 0$ and $C > 0$, the original online learning problem can be reduced to optimize the following one through a shifting operation,

$$\sum_{t=1}^{T} \langle \boldsymbol{\ell}'_t, \boldsymbol{p}'_t \rangle - \sum_{t=1}^{T} \ell'_{t,i^*} + \lambda \sum_{t=2}^{T} \|\boldsymbol{p}'_t - \boldsymbol{p}'_{t-1}\|_1, \tag{37}$$

where $\ell'_{t,i} \triangleq \ell_{t,i} + C, \boldsymbol{p}'_t \triangleq \boldsymbol{p}_t$ for all $t \in [T], i \in [N]$, and evidently $\ell'_{t,i} \in [0, 2C]$. The above equivalence can be simply proven by plugging the definition of $\ell'_{t,i}$ and $\boldsymbol{p}'_t$ into (37). Formally,

$$\begin{aligned}&\sum_{t=1}^{T} \langle \boldsymbol{\ell}'_t, \boldsymbol{p}'_t \rangle - \sum_{t=1}^{T} \ell'_{t,i^*} + \lambda \sum_{t=2}^{T} \|\boldsymbol{p}'_t - \boldsymbol{p}'_{t-1}\|_1 \\ ={}& \sum_{t=1}^{T} \langle \boldsymbol{\ell}_t + [C, \ldots, C]^\top, \boldsymbol{p}_t \rangle - \sum_{t=1}^{T} (\ell_{t,i^*} + C) + \lambda \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 \\ ={}& \sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{p}_t \rangle - \sum_{t=1}^{T} \ell_{t,i^*} + \lambda \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1.\end{aligned}$$

Next, we prove that there exists a sequence of loss functions $\boldsymbol{\ell}'_1, \ldots, \boldsymbol{\ell}'_T$ satisfying $\boldsymbol{\ell}'_t \in [0, 2C]^N$ for all $t \in [T]$ such that any feasible expert algorithm (whose output is $\boldsymbol{p}'_1, \ldots, \boldsymbol{p}'_T \in \Delta_N$) incurs the following regret

$$\sum_{t=1}^{T} \langle \boldsymbol{\ell}'_t, \boldsymbol{p}'_t \rangle - \sum_{t=1}^{T} \ell'_{t,i^*} + \lambda \sum_{t=2}^{T} \|\boldsymbol{p}'_t - \boldsymbol{p}'_{t-1}\|_1 \geq \Omega(\sqrt{\lambda C T}).$$

The remaining proof borrows the intuition from Theorem 13 of Altschuler and Talwar (2018). First we give a hard constraint on the switching cost, e.g., $\sum_{t=2}^{T} \|\boldsymbol{p}'_t - \boldsymbol{p}'_{t-1}\|_1 = S$. Then we divide the time horizon T into $B = 4S^2/(a^2 \log N)$ blocks, each of uniform length T/B, where a is some constant to be specified later. For each block $b \in [B]$, assign to each expert $i \in [N]$ a loss sampled from $2C \cdot \mathrm{Ber}(1/2)$, i.e., $2C$ with probability $1/2$ and otherwise 0, for each iteration in that block. Clearly this adversary is oblivious.

Note that the cumulative loss of the i-th expert, namely, $\sum_{t=1}^{T} \ell'_t(i)$, is equal in distribution to T/B times a $2C \cdot \mathrm{Bin}(B, 1/2)$ random variable. In the following, we first consider the expected cumulative loss of the best expert. Suppose there are N variables drawn i.i.d. from $\mathrm{Bin}(B, 1/2)$, then the minimum one has the following upper bound.

Lemma 19. *There exists a universal constant $c > 0$ such that for all $B, N \in \mathbb{N}_+$,*

$$\mathbb{E}\left[\min_{i \in [N]} Z_i\right] \leq \frac{B}{2} - c\sqrt{B \log N}$$

where $\{Z_i\}_{i \in [N]}$ are i.i.d. from $\mathrm{Bin}(B, 1/2)$.

The adversary chooses a to be the constant that makes Lemma 19 holds. Thus the loss of the best expert satisfies that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \ell'_t(i^*)\right] &\leq 2C \cdot \frac{T}{B}\left(\frac{B}{2} - a\sqrt{B \log N}\right) \\ = 2C \cdot \left(\frac{T}{2} - aT\sqrt{\frac{\log N}{B}}\right) &= 2C \cdot \left(\frac{T}{2} - \frac{a^2 T \log N}{2S}\right).\end{aligned} \tag{38}$$

Now let us compute the expected loss of any algorithm $\mathcal{A}$ whose switching cost is at most S. It is simple to see that the following strategy is optimal: in the first round of each block, randomly assign the weights since there is no information about the losses of the experts; then convert the weight on the bad experts (with loss $2C$) to the good experts (with loss 0) if the current switching cost is still less than S. Let the random variable W denote the total weights that the algorithm assigns to the bad experts in the blocks' first iteration. Clearly $\mathbb{E}[W] = B/2$. Then the random variable $\min\{W, S/2\}$ is equal to the weights that algorithm $\mathcal{A}$ can convert from bad experts to good expert ($S/2$ dues to that converting weight of $S/2$ will suffers S switching cost). Thus, we have

$$\begin{aligned}&\mathbb{E}[\text{cumulative loss of } \mathcal{A}] = 2C \cdot \mathbb{E}[\mathcal{A}\text{'s weights on bad experts}] \\ &= 2C \cdot \mathbb{E}\left[\min\left\{W, \frac{S}{2}\right\} + \frac{T}{B} \cdot \left(W - \min\left\{W, \frac{S}{2}\right\}\right)\right] \\ &\geq 2C \cdot \frac{T}{B} \cdot \mathbb{E}\left[W - \frac{S}{2}\right] \geq 2C \cdot \frac{T}{B}\left(\frac{B}{2} - 2S\right) \\ &= 2C \cdot \left(\frac{T}{2} - \frac{2ST}{B}\right) = 2C \cdot \left(\frac{T}{2} - \frac{a^2 T \log N}{2S}\right).\end{aligned} \tag{39}$$

Combining (38) and (39), we conclude that any algorithm for λ-switching cost and S-switching cost budget suffers an expected regret at least $a^2 CT \log N / S = \Omega(CT/S)$. As a result, the regret of (37) is at least $\Omega(CT/S + \lambda S) = \Omega(\sqrt{\lambda CT})$, which finishes the proof. ■

B.6 Proof of Theorem 4

Proof We begin the proof by decomposing the dynamic regret of OCO with switching cost, and will then prove the theorem by exploiting the property of Scream algorithm.

Regret Decomposition. We divide the time horizon T into K epochs of equal length Δ, where the k-th epoch is denoted by $\mathcal{I}_k \triangleq \{t_{k,1}, \ldots, t_{k,\Delta}\}$ (Δ, K to be specified later). Without loss of generality, we assume $T = K \cdot \Delta$. Since in Algorithm 2, the meta-learner and base-learners do *not* update within each epoch, we denote by $\mathring{\mathbf{w}}_1, \ldots, \mathring{\mathbf{w}}_K$ the decisions of K epochs. Thus the dynamic regret of OCO with switching cost can be decomposed as

$$
\begin{aligned}
&\sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \le \sum_{t=1}^{T} \langle \nabla_t, \mathbf{w}_t - \mathbf{v}_t \rangle + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \\
&= \sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} \langle \nabla_t, \mathring{\mathbf{w}}_k - \mathbf{v}_t \rangle + \lambda \sum_{k=2}^{K} \|\mathring{\mathbf{w}}_k - \mathring{\mathbf{w}}_{k-1}\|_2 \\
&= \underbrace{\sum_{k=1}^{K} \left\langle \sum_{t \in \mathcal{I}_k} \nabla_t, \mathring{\mathbf{w}}_k - \mathring{\mathbf{v}}_k \right\rangle + \lambda \sum_{k=2}^{K} \|\mathring{\mathbf{w}}_k - \mathring{\mathbf{w}}_{k-1}\|_2}_{\texttt{term (A)}} + \underbrace{\sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} \langle \nabla_t, \mathring{\mathbf{v}}_k - \mathbf{v}_t \rangle}_{\texttt{term (B)}},
\end{aligned}
$$

where $\nabla_t \triangleq \nabla \widetilde{f}_t(\mathbf{w}_t)$ and the k-th comparator $\mathring{\mathbf{v}}_k \triangleq \mathbf{v}_{t_{k,1}}$ is chosen as the first one in the k-th epoch. Define $\mathbf{g}_k \triangleq \sum_{t \in \mathcal{I}_k} \nabla_t$ the loss of the k-epoch. Intuitively, term (A) is the dynamic regret of OCO with switching cost in K rounds with the loss sequence $\mathbf{g}_{1:K}$ and comparator sequence $\mathring{\mathbf{v}}_{1:K}$. Since the new comparator sequence is artificially constructed, we need to measure its difference from the original sequence $\mathbf{v}_{1:T}$, i.e., term (B). Term (B) can be simply bounded using the sub-additivity property of vector norms, formally,

$$
\texttt{term (B)} \le G \sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} \|\mathring{\mathbf{v}}_k - \mathbf{v}_t\|_2 \le G \sum_{k=1}^{K} |\mathcal{I}_k| \sum_{t \in \mathcal{I}_k} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2 \le G \Delta P_T.
$$

Black-box Use of Scream. Term (A) is actually the dynamic regret of OCO with switching cost in K rounds. Plugging in the regret bound of Scream (36), it holds that

$$
\begin{aligned}
&\texttt{term (A)} = \sum_{k=1}^{K} \langle \mathbf{g}_k, \mathring{\mathbf{w}}_k - \mathring{\mathbf{v}}_k \rangle + \lambda \sum_{k=2}^{K} \|\mathring{\mathbf{w}}_k - \mathring{\mathbf{w}}_{k-1}\|_2 \\
&\lesssim \sqrt{2(G'D + \sqrt{\lambda})(\lambda D + G'D + \sqrt{\lambda})K\,(1 + \ln(i_K^* + 1))} + \sqrt{\left(G'^2 + \lambda G'\right)(D^2 + 2DP_K)K} \\
&\lesssim \sqrt{2(\Delta G D + \sqrt{\lambda})(\lambda D + \Delta G D)\frac{T}{\Delta}\,(1 + \ln(i_K^* + 1))} + \sqrt{(\Delta^2 G^2 + \lambda \Delta G)(D^2 + 2DP_K)\frac{T}{\Delta}} \\
&= \sqrt{2D(\Delta G D + \sqrt{\lambda})\left(\frac{\lambda}{\Delta} + G\right)T\,(1 + \ln(i_K^* + 1))} + \sqrt{(\Delta G^2 + \lambda G)(D^2 + 2DP_K)T} \\
&\le \mathcal{O}(\sqrt{\lambda T(1 + P_T)}),
\end{aligned}
$$

where the path length in K epochs $P_K \triangleq \sum_{k=2}^{K}\|\mathring{\mathbf{v}}_k - \mathring{\mathbf{v}}_{k-1}\|_2 \leq P_T$, the gradient upper bound $G' = \max_{k\in[K]}\|\mathbf{g}_k\|_2 \leq \Delta G$ and $a \lesssim b$ means $a = \mathcal{O}(b)$. The last step is due to the property of the best base learner, that is,

$$i_K^* \overset{(35)}{\leq} \left\lceil \frac{1}{2}\log_2\big(1 + \frac{2P_K}{D}\big)\right\rceil + 1 \leq \left\lceil \frac{1}{2}\log_2\big(1 + \frac{2P_T}{D}\big)\right\rceil + 1,$$

and by choosing $\Delta = \sqrt{\lambda}$. Combining the above inequality with the upper bound of $\texttt{term (B)} \leq G\sqrt{\lambda}P_T = \mathcal{O}(\sqrt{\lambda T(1+P_T)})$ finishes the proof. ■

B.7 Proof of Theorem 5

Proof Overall the proof consists of two parts. First, we propose a lower bound for static regret of OCO with switching cost. Second, building upon the static regret lower bound, we give a lower bound for dynamic regret of OCO with switching cost to complete the proof.

Static Regret Lower Bound. To give a static regret lower bound, we first consider a T-round prediction with expert advice problem with λ-switching cost. Theorem 3 shows that given $\lambda > 0$ and $C > 0$, there exists a sequence of loss functions $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$ satisfying $\boldsymbol{\ell}_t \in [-C, C]^N$ for all $t \in [T]$ such that any feasible expert algorithm (whose output is $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T \in \Delta_N$) incurs the following regret

$$\sum_{t=1}^{T}\langle \boldsymbol{\ell}_t, \boldsymbol{p}_t\rangle - \min_{i\in[N]}\sum_{t=1}^{T}\ell_{t,i} + \lambda\sum_{t=2}^{T}\|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 \geq \Omega(\sqrt{\lambda CT}). \tag{40}$$

Consequently, given a parameter $\lambda > 0$, we choose the feasible domain as $\mathcal{W} = C_1\Delta_N$, where $C_1 = \min\{1, D/\sqrt{2}\}$. It is easy to observe that $\mathcal{W}$ satisfies Assumption 3, because for any $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \Delta_N$, $\|C_1\boldsymbol{p}_1 - C_1\boldsymbol{p}_2\|_2 \leq C_1 \cdot \sqrt{2} \leq D$ holds. Choose $C_2 = G/\sqrt{N}$ and loss functions as $h_t(\mathbf{w}) = \langle \boldsymbol{\ell}_t, \mathbf{w}\rangle$, where $\boldsymbol{\ell}_{1:T}$ is the loss sequence that makes (40) holds given C_2 and λ. Since for any $\mathbf{w} \in \mathcal{W}$, $\|\nabla h_t(\mathbf{w})\|_2 = \|\boldsymbol{\ell}_t\|_2 \leq C_2\sqrt{N} \leq G$, the loss functions $h_1, \ldots, h_T$ satisfy Assumption 2. Thus any online algorithm returning $\mathbf{w}'_1 \triangleq C_1\mathbf{w}_1, \ldots, \mathbf{w}'_T \triangleq C_1\mathbf{w}_T \in \mathcal{W}$ satisfies

$$\begin{aligned}
&\sum_{t=1}^{T} h_t(\mathbf{w}'_t) - \min_{\mathbf{v}\in\mathcal{W}}\sum_{t=1}^{T} h_t(\mathbf{v}) + \lambda\sum_{t=2}^{T}\|\mathbf{w}'_t - \mathbf{w}'_{t-1}\|_2 \\
={}& C_1\left(\sum_{t=1}^{T}\langle \boldsymbol{\ell}_t, \mathbf{w}_t\rangle - \min_{\mathbf{v}\in\Delta_N}\sum_{t=1}^{T}\langle \boldsymbol{\ell}_t, \mathbf{v}\rangle + \lambda\sum_{t=2}^{T}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2\right) \\
\geq{}& C_1\left(\sum_{t=1}^{T}\langle \boldsymbol{\ell}_t, \mathbf{w}_t\rangle - \min_{i\in[N]}\sum_{t=1}^{T}\ell_{t,i} + \frac{\lambda}{\sqrt{d}}\sum_{t=2}^{T}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_1\right) \\
\overset{(40)}{\geq}{}& C_1\Omega(\sqrt{\lambda C_2 T}) = \Omega(\sqrt{\lambda T}),
\end{aligned} \tag{41}$$

where the first step is by plugging in the definition of $h_1, \ldots, h_T$ and $\mathbf{w}'_1, \ldots, \mathbf{w}'_T$, the second step is because the optimizer in a simplex is on one of its vertices and the relationship between ℓ_1-norm and ℓ_2-norm, formally, $\|\mathbf{x} - \mathbf{y}\|_1 \leq \sqrt{d}\cdot\|\mathbf{x} - \mathbf{y}\|_2$, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, where d denotes the dimension.

Dynamic Regret Lower Bound. We consider two cases according to the value of τ. When $\tau \leq D$, we can always find a comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathcal{W}$ such that

$$\begin{aligned}
&\sum_{t=1}^{T} h_t(\mathbf{w}_t) - \sum_{t=1}^{T} h_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \\
\geq &\sum_{t=1}^{T} h_t(\mathbf{w}_t) - \min_{\mathbf{v} \in \mathcal{W}} \sum_{t=1}^{T} h_t(\mathbf{v}) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \overset{(41)}{\geq} \Omega(\sqrt{\lambda T}) = \Omega(\sqrt{\lambda \tau T}),
\end{aligned}$$

where the last step holds since $\tau \leq D$ can be seen as a constant and thus will not affect the order. Next, we consider the case $\tau \in (D, DT]$. Without loss of generality, we assume $\lceil \tau \rceil$ divides T and let $K = T/\lceil \tau \rceil$. To proceed, we construct the following piecewise-stationary comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T$: for any $i \in [\lceil \tau \rceil]$, denote by $\mathcal{I}_i = [(i-1)K + 1, iK]$ the i-th interval, the comparators within the interval are set as

$$\mathbf{v}_{(i-1)K+1} = \mathbf{v}_{(i-1)K+2} = \cdots = \mathbf{v}_{iK} \in \arg\min_{\mathbf{v} \in \mathcal{W}} \sum_{t \in \mathcal{I}_i} h_t(\mathbf{v}).$$

Note that the path length of this comparator sequence does not exceeds τD. Thus, the dynamic regret competing with the comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T$ can be evaluated as,

$$\begin{aligned}
&\sum_{t=1}^{T} h_t(\mathbf{w}_t) - \sum_{t=1}^{T} h_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \\
\geq &\sum_{i=1}^{\lceil \tau \rceil} \left(\sum_{t \in \mathcal{I}_i} h_t(\mathbf{w}_t) - \min_{\mathbf{v} \in \mathcal{W}} \sum_{t \in \mathcal{I}_i} h_t(\mathbf{v}) + \lambda \sum_{t=(i-1)K+2}^{iK} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \right) \\
\overset{(41)}{\gtrsim} &\sum_{i=1}^{\lceil \tau \rceil} \sqrt{\lambda |\mathcal{I}_i|} = \lceil \tau \rceil \sqrt{\lambda \cdot \frac{T}{\lceil \tau \rceil}} \geq \sqrt{\lambda \tau T},
\end{aligned}$$

where the first inequality is true by ignoring the switching cost between two consecutive pieces. In addition, $a \gtrsim b$ means $a = \Omega(b)$. Hence, we complete the proof. ■

B.8 Discussion on Memory Dependence

In this part, we examine a subtle issue: the memory dependence of our static policy regret bound (an implication of the dynamic policy regret bound in Theorem 2) and that of existing work (Anava et al., 2015).

First, we state our attained static policy regret for OCO with memory via performing OGD over the unary loss with an optimal step size tuning (which is feasible as there is no dependence on the path length P_T).

Theorem 20. *Under Assumptions 1–3, running OGD over the unary loss achieves*

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \min_{\mathbf{v} \in \mathcal{W}} \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}) \leq (G^2 + m^2 LG)\eta T + \frac{2D^2}{\eta}.$$

Setting the step size optimally as $\eta = \eta^* = \sqrt{\frac{2D^2}{(G^2+m^2LG)T}}$, we attain an $\mathcal{O}(m\sqrt{T})$ static policy regret.

Anava et al. (2015) present an $\mathcal{O}(m^{3/4}\sqrt{T})$ static policy regret for OCO with memory, which seems better than ours at the first glance. However, we point it out that this is due to the different assumptions imposing over the Lipschitz continuity. Their assumption is presented as follows.

Assumption 8 (Lipschitzness of Anava et al. (2015))**.** The function $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$ is $\bar{L}$-Lipschitz, i.e.,

$$|f_t(\mathbf{x}_0, \ldots, \mathbf{x}_m) - f_t(\mathbf{y}_0, \ldots, \mathbf{y}_m)| \leq \bar{L}\|(\mathbf{x}_0, \ldots, \mathbf{x}_m) - (\mathbf{y}_0, \ldots, \mathbf{y}_m)\|_2 = \bar{L}\sqrt{\sum_{i=0}^{m}\|\mathbf{x}_i - \mathbf{y}_i\|_2^2}.$$

We compare this definition of Lipschitzness with the version used in our paper, namely, the coordinate-wise Lipschitzness defined in Assumption 1. Indeed, their definition imposes a *stronger* requirement on the function than ours. Clearly, when the online function f_t satisfies $\bar{L}$-Lipschitz assumption as specified in Assumption 8, it is also $\bar{L}$-coordinate-wise Lipschitz due to the simple fact that $\sqrt{\sum_{i=0}^{m}\|\mathbf{x}_i - \mathbf{y}_i\|_2^2} \leq \sum_{i=0}^{m}\|\mathbf{x}_i - \mathbf{y}_i\|_2$. On the other hand, when the online function f_t is L-coordinate-wise Lipschitz as required by Assumption 1, we thus conclude that it is Lipschitz in the sense of Assumption 8 with the Lipschitz coefficient $\bar{L} = \sqrt{m}L$, due to the following inequality (by Cauchy-Schwarz inequality) $L\sum_{i=0}^{m}\|\mathbf{x}_i - \mathbf{y}_i\|_2 \leq L\sqrt{m}\sqrt{\sum_{i=0}^{m}\|\mathbf{x}_i - \mathbf{y}_i\|_2}$.

In the following, we restate the static regret bound of Anava et al. (2015) under Assumption 8. We adapt their results to our notations to ease the understanding.

Theorem 21 (Theorem 3.1 of Anava et al. (2015))**.** *Under Assumptions 2, 3, and the assumption that the online functions are $\bar{L}$-Lipschitz (Assumption 8), running OGD over the unary loss achieves*

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \min_{\mathbf{v}\in\mathcal{W}}\sum_{t=1}^{T}\widetilde{f}_t(\mathbf{v}) \leq 2\eta G^2 T + \frac{2D^2}{\eta} + 2\bar{L}m^{\frac{3}{2}}\eta GT. \tag{42}$$

Setting the step size optimally yields an $\mathcal{O}(\bar{L}^{1/2}m^{3/4}\sqrt{T})$ static policy regret.

Therefore, when the online functions are only L-coordinate-wise Lipschitz as considered in this paper, applying above theorem immediately obtains an $\mathcal{O}(\bar{L}^{1/2}m^{3/4}\sqrt{T}) = \mathcal{O}((\sqrt{m}L)^{1/2}m^{3/4}\sqrt{T}) = \mathcal{O}(L^{1/2}m\sqrt{T})$, which exhibiting a linear memory dependence.

Appendix C. Omitted Details for Section 5 (Non-stochastic Control)

In this section, we present omitted details for Section 5 online non-stochastic control, including the proofs of Proposition 6, Theorem 7, Theorem 8, and Corollary 9.

C.1 Proof of Proposition 6

We will prove the following statement that gives the state recurrence for any $h \leq t$, which is essentially a strengthened result of Proposition 6.

Proposition 22. *Suppose one chooses the DAC controller $\pi(M_t, K)$ at iteration t, the reaching state is*

$$x_{t+1} = \widetilde{A}_K^{h+1} x_{t-h} + \sum_{i=0}^{H+h} \Psi_{t,i}^{K,h}(M_{t-h:t}) w_{t-i}, \tag{43}$$

where $\widetilde{A}_K = A - BK$, and $\Psi_{t,i}^{K,h}(M_{t-h:t})$ is the transfer matrix defined as

$$\Psi_{t,i}^{K,h}(M_{t-h:t}) = \widetilde{A}_K^{i} \mathbf{1}_{i \leq h} + \sum_{j=0}^{h} \widetilde{A}_K^{j} B M_{t-j}^{[i-j]} \mathbf{1}_{1 \leq i-j \leq H}. \tag{44}$$

The evolving equation holds for any $h \in \{0, \ldots, t\}$.

Proof First, by substituting the DAC policy into the dynamics equation, we have

$$\begin{aligned} x_{t+1} = Ax_t + Bu_t + w_t &= (A - BK)x_t + \sum_{i=1}^{H} BM_t^{[i]} w_{t-i} + w_t \\ &= \widetilde{A}_K^{h+1} x_{t-h} + \sum_{j=0}^{h} \widetilde{A}_K^{j} \left(\sum_{i=1}^{H} BM_{t-j}^{[i]} w_{t-j-i} + w_{t-j} \right) \\ &= \widetilde{A}_K^{h+1} x_{t-h} + \sum_{j=0}^{h} \sum_{i=1}^{H} \widetilde{A}_K^{j} BM_{t-j}^{[i]} w_{t-j-i} + \sum_{j=0}^{h} \widetilde{A}_K^{j} w_{t-j}. \end{aligned}$$

Exchanging the summation index yields,

$$\begin{aligned} \sum_{j=0}^{h} \sum_{i=1}^{H} \widetilde{A}_K^{j} BM_{t-j}^{[i]} w_{t-j-i} &= \sum_{i=1}^{H} \sum_{k=i}^{i+h} \widetilde{A}_K^{k-i} BM_{t-k+i}^{[i]} w_{t-k} && (45) \\ &= \sum_{k=1}^{H+h} \sum_{i=k-h}^{k} \widetilde{A}_K^{k-i} BM_{t-k+i}^{[i]} w_{t-k} \mathbf{1}_{1 \leq i \leq H} && (46) \\ &= \sum_{k=1}^{H+h} \sum_{l=0}^{h} \widetilde{A}_K^{h-l} BM_{t+l-h}^{[l+k-h]} w_{t-k} \mathbf{1}_{1 \leq l+(k-h) \leq H} && (47) \\ &= \sum_{k=1}^{H+h} \sum_{m=0}^{h} \widetilde{A}_K^{m} BM_{t-m}^{[k-m]} w_{t-k} \mathbf{1}_{1 \leq k-m \leq H} && (48) \\ &= \sum_{i=1}^{H+h} \sum_{j=0}^{h} \widetilde{A}_K^{j} BM_{t-j}^{[i-j]} w_{t-i} \mathbf{1}_{1 \leq i-j \leq H}, && (49) \end{aligned}$$

where (45) holds by defining a third variable $k = j + i$, and (46) is obtained by exchanging the summation index i and k and the new range of i is from inequality $i \leq k \leq i + h$.

Moreover, (47) is obtained by another change of variable $l = i - k + h$, (48) is obtained by replacing l by $h - m$, and (49) is true by setting $i = k, j = m$. Therefore, we obtain that

$$\begin{aligned} x_{t+1} &= \widetilde{A}_K^{h+1} x_{t-h} + \sum_{j=0}^{h} \sum_{i=1}^{H} \widetilde{A}_K^j B M_{t-j}^{[i]} w_{t-j-i} + \sum_{j=0}^{h} \widetilde{A}_K^j w_{t-j} \\ &= \widetilde{A}_K^{h+1} x_{t-h} + \sum_{i=0}^{H+h} \sum_{j=0}^{h} \widetilde{A}_K^j B M_{t-j}^{[i-j]} w_{t-i} \mathbf{1}_{1 \le i-j \le H} + \sum_{i=0}^{h} \widetilde{A}_K^i w_{t-i} \\ &= \widetilde{A}_K^{h+1} x_{t-h} + \sum_{i=0}^{H+h} \left(\widetilde{A}_K^i \mathbf{1}_{i \le h} + \sum_{j=0}^{h} \widetilde{A}_K^j B M_{t-j}^{[i-j]} \mathbf{1}_{1 \le i-j \le H} \right) w_{t-i} \end{aligned}$$

and hence complete the proof. ■

C.2 Proof of Theorem 7

To prove the dynamic policy regret of online non-stochastic control (Theorem 7), we will first present theoretical analysis of the reduction to OCO with memory in Appendix C.2.1, then give the dynamic regret analysis over the $\mathcal{M}$-space in Appendix C.2.2, and finally present the overall proof of Theorem 7 in Appendix C.2.3.

C.2.1 Approximation Error

In Section 5.2 of the main paper, we have presented how to reduce from online non-stochastic control to OCO with memory, by employing the DAC parameterization and introducing the truncated loss functions. In this part, we introduce the following theorem that discloses that the truncation loss f_t approximates the original cost function c_t well.

Theorem 23 (Theorem 5.3 of Agarwal et al. (2019)). *Suppose the disturbance are bounded by W. For any (κ, γ)-strongly stable linear controller K, and any $\tau > 0$ such that the sequence of $M_1, \ldots, M_T$ satisfies $\|M_t^{[i]}\|_{\text{op}} \le \tau(1-\gamma)^i, \forall i \in [H]$, the approximation error between original loss and truncated loss is at most*

$$\left| \sum_{t=1}^{T} c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - \sum_{t=1}^{T} f_t(M_{t-1-H:t}) \right| \le 2TG_c D^2 \kappa^3 (1-\gamma)^{H+1}, \tag{50}$$

where

$$D \triangleq \frac{W\kappa^3(1 + H\kappa_B \tau)}{\gamma(1 - \kappa^2(1-\gamma)^{H+1})} + \frac{W\tau}{\gamma}. \tag{51}$$

Proof By Lipschitzness and definition of the truncated loss, we get that

$$\begin{aligned} & c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - f_t(M_{t-H-1:t}) \\ =\ & c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - c_t(y_t^K(M_{t-H-1:t-1}), v_t^K(M_{t-H-1:t})) \\ \le\ & G_c D \left(\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1})\| + \|u_t^K(M_{0:t}) - v_t^K(M_{t-H-1:t})\| \right) \end{aligned}$$

$$\leq G_c D(\kappa^2(1-\gamma)^{H+1}D + \kappa^3(1-\gamma)^{H+1}D) \leq 2G_cD^2\kappa^3(1-\gamma)^{H+1},$$

where the last two inequalities use the Lipschitzness and the boundedness presented in Lemma 28. We complete the proof by summing over the iterations from $t = 1, \ldots, T$. ■

C.2.2 Dynamic Regret Analysis over $\mathcal{M}$-space

In previous sections, we have analyzed the dynamic regret of our method over the $\mathbb{R}^d$-space. However, after reducing online non-stochastic control to OCO with memory, we need to apply their results to the $\mathcal{M}$-space and thus require to generalize the arguments of previous sections from Euclidean norm for $\mathbb{R}^d$-space to Frobenius norm for $\mathcal{M}$-space. For completeness, we present the proof here.

At the first place, we analyze the dynamic regret of the online gradient descent (OGD) algorithm over the $\mathbb{R}^d$-space. OGD begins with any $M_1 \in \mathcal{M}$ and performs the following update procedure,

$$M_{t+1} = \Pi_{\mathcal{M}}[M_t - \eta \nabla_M \widetilde{f}_t(M_t)] \tag{52}$$

where $\eta > 0$ is the step size and $\Pi_{\mathcal{M}}[\cdot]$ denotes the projection onto the nearest point in the feasible set $\mathcal{M}$. We have the following dynamic regret regarding its dynamic regret.

Theorem 24. *Suppose the function $\widetilde{f} : \mathcal{M} \mapsto \mathbb{R}$ is convex, the gradient norm satisfies $\max_{M\in\mathcal{M}} \max_{t\in[T]} \|\nabla_M \widetilde{f}_t(M)\|_{\mathrm{F}} \leq G_f$ and the Euclidean diameter of $\mathcal{M}$ is at most D_f, i.e., $\sup_{M,M'\in\mathcal{M}} \|M - M'\|_{\mathrm{F}} \leq D_f$. Then, OGD with a step size $\eta > 0$ as shown in* (52) *satisfies that*

$$\lambda \sum_{t=2}^{T} \|M_{t-1} - M_t\|_{\mathrm{F}} + \sum_{t=1}^{T} \widetilde{f}_t(M_t) - \sum_{t=1}^{T} \widetilde{f}_t(M_t^*) \leq \frac{\eta}{2}(G_f^2 + 2\lambda G_f)T + \frac{1}{2\eta}(D_f^2 + 2D_f P_T), \tag{53}$$

which holds for any comparator sequence $M_1^*, \ldots, M_T^* \in \mathcal{M}$. Besides, the path length $P_T = \sum_{t=2}^{T} \|M_{t-1}^* - M_t^*\|_{\mathrm{F}}$ measures the non-stationarity of the comparator sequence.

Proof Denote the gradient by $G_t = \nabla_M \widetilde{f}_t(M_t)$. The convexity of online surrogate loss functions implies that

$$\sum_{t=1}^{T} \widetilde{f}_t(M_t) - \sum_{t=1}^{T} \widetilde{f}_t(M_t^*) \leq \sum_{t=1}^{T} \langle G_t, M_t - M_t^* \rangle.$$

Thus, it suffices to bound the sum of $\langle G_t, M_t - M_t^* \rangle$. From the OGD update rule and the non-expensive property, we have

$$\begin{aligned}\|M_{t+1} - M_t^*\|_{\mathrm{F}}^2 &= \|\Pi_{\mathcal{M}}[M_t - \eta G_t] - M_t^*\|_{\mathrm{F}}^2 \leq \|M_t - \eta G_t - M_t^*\|_{\mathrm{F}}^2 \\ &= \eta^2 \|G_t\|_{\mathrm{F}}^2 - 2\eta \langle G_t, M_t - M_t^* \rangle + \|M_t - M_t^*\|_{\mathrm{F}}^2\end{aligned}$$

After rearranging, we obtain

$$\langle G_t, M_t - M_t^* \rangle \leq \frac{\eta}{2} \|G_t\|_{\mathrm{F}}^2 + \frac{1}{2\eta}\left(\|M_t - M_t^*\|_{\mathrm{F}}^2 - \|M_{t+1} - M_t^*\|_{\mathrm{F}}^2\right).$$

Next, we turn to analyze the second term on the right-hand side. Indeed,

$$\sum_{t=1}^{T}\left(\|M_t - M_t^*\|_{\mathrm{F}}^2 - \|M_{t+1} - M_t^*\|_{\mathrm{F}}^2\right) \le \sum_{t=1}^{T}\|M_t - M_t^*\|_{\mathrm{F}}^2 - \sum_{t=2}^{T}\|M_t - M_{t-1}^*\|_{\mathrm{F}}^2$$

$$\le \|M_1 - M_1^*\|_{\mathrm{F}}^2 + \sum_{t=2}^{T}\left(\|M_t - M_t^*\|_{\mathrm{F}}^2 - \|M_t - M_{t-1}^*\|_{\mathrm{F}}^2\right)$$

$$= \|M_1 - M_1^*\|_{\mathrm{F}}^2 + \sum_{t=2}^{T}\langle M_{t-1}^* - M_t^*, 2M_t - M_{t-1}^* - M_t^*\rangle \le D_f^2 + 2D_f\sum_{t=2}^{T}\|M_{t-1}^* - M_t^*\|_{\mathrm{F}}.$$

Hence, combining all above inequalities, we have

$$\begin{aligned}\sum_{t=1}^{T}\widetilde{f}_t(M_t) - \sum_{t=1}^{T}\widetilde{f}_t(M_t^*) &\le \frac{\eta}{2}\sum_{t=1}^{T}\|G_t\|_{\mathrm{F}}^2 + \frac{1}{2\eta}\left(D_f^2 + 2D_f\sum_{t=2}^{T}\|M_{t-1}^* - M_t^*\|_{\mathrm{F}}\right)\\ &\le \frac{\eta}{2}G_f^2T + \frac{1}{2\eta}(D_f^2 + 2D_fP_T).\end{aligned}$$

On the other hand, the switching cost can be bounded by

$$\|M_t - M_{t-1}\|_{\mathrm{F}} = \|\Pi_{\mathcal{M}}[M_{t-1} - \eta G_{t-1}] - M_{t-1}\|_{\mathrm{F}}^2 \le \|M_{t-1} - \eta G_{t-1} - M_{t-1}\|_{\mathrm{F}} \le \eta G_f,$$

which together with the previous dynamic regret bound yields the desired result. ■

C.2.3 Proof of Theorem 7

Proof We begin with the following dynamic policy regret decomposition,

$$\begin{aligned}&\sum_{t=1}^{T}c_t(x_t,u_t) - \sum_{t=1}^{T}c_t(x_t^{\pi_t},u_t^{\pi_t})\\ =&\sum_{t=1}^{T}c_t(x_t^K(M_{0:t-1}),u_t^K(M_{0:t})) - \sum_{t=1}^{T}c_t(x_t^K(M_{0:t-1}^*),u_t^K(M_{0:t}^*))\\ =&\underbrace{\sum_{t=1}^{T}c_t(x_t^K(M_{0:t-1}),u_t^K(M_{0:t})) - \sum_{t=1}^{T}f_t(M_{t-1-H:t})}_{\triangleq A_T} + \underbrace{\sum_{t=1}^{T}f_t(M_{t-1-H:t}) - \sum_{t=1}^{T}f_t(M_{t-1-H:t}^*)}_{\triangleq B_T}\\ &\quad + \underbrace{\sum_{t=1}^{T}f_t(M_{t-1-H:t}^*) - \sum_{t=1}^{T}c_t(x_t^K(M_{0:t-1}^*),u_t^K(M_{0:t}^*))}_{\triangleq C_T}. \end{aligned} \tag{54}$$

Notice that both A_T and C_T essentially represent the approximation error introduced by the truncated loss, so we can apply Theorem 23 and obtain

$$A_T + C_T \le 4TG_cD^2\kappa^3(1-\gamma)^{H+1}. \tag{55}$$

We now focus on the quantity B_T, which is the dynamic policy regret over the truncated loss functions $\{f_t\}_{t=1,\ldots,T}$. Indeed,

$$\begin{aligned} B_T &= \sum_{t=1}^T f_t(M_{t-1-H:t}) - \sum_{t=1}^T f_t(M^*_{t-1-H:t}) \\ &\leq \sum_{t=1}^T \widetilde{f}_t(M_t) - \sum_{t=1}^T \widetilde{f}_t(M^*_t) + \lambda \sum_{t=2}^T \|M_{t-1} - M_t\|_{\mathrm{F}} + \lambda \sum_{t=2}^T \|M^*_{t-1} - M^*_t\|_{\mathrm{F}} \\ &\leq \sum_{t=1}^T \langle \nabla_M \widetilde{f}_t(M_t), M_t - M^*_t\rangle + \lambda \sum_{t=2}^T \|M_{t-1} - M_t\|_{\mathrm{F}} + \lambda \sum_{t=2}^T \|M^*_{t-1} - M^*_t\|_{\mathrm{F}} \\ &= \sum_{t=1}^T g_t(M_t) - \sum_{t=1}^T g_t(M^*_t) + \lambda \sum_{t=2}^T \|M_{t-1} - M_t\|_{\mathrm{F}} + \lambda \sum_{t=2}^T \|M^*_{t-1} - M^*_t\|_{\mathrm{F}}, \end{aligned} \tag{56}$$

where $\lambda = (H+2)^2 L_f$ and $g_t(M) = \langle \nabla_M \widetilde{f}_t(M_t), M\rangle$ is the surrogate linearized loss. As a consequence, we are reduced to proving an dynamic regret over the sequence of functions $\{g_t\}_{t=1,\ldots,T}$ with switching cost, namely, the first three terms on the right-hand side. We thus make use of the techniques developed in Appendix B.4 (dynamic policy regret minimization for OCO with memory) to decompose the terms into meta-regret and base-regret:

$$\begin{aligned} &\sum_{t=1}^T g_t(M_t) - \sum_{t=1}^T g_t(M^*_t) + \lambda \sum_{t=2}^T \|M_{t-1} - M_t\|_{\mathrm{F}} \\ = &\underbrace{\left(\lambda \sum_{t=2}^T \|M_{t-1} - M_t\|_{\mathrm{F}} + \sum_{t=1}^T g_t(M_t)\right) - \left(\lambda \sum_{t=2}^T \|M_{t-1,i} - M_{t,i}\|_{\mathrm{F}} + \sum_{t=1}^T g_t(M_{t,i})\right)}_{\texttt{meta-regret}} \\ &\qquad + \underbrace{\left(\lambda \sum_{t=2}^T \|M_{t-1,i} - M_{t,i}\|_{\mathrm{F}} + \sum_{t=1}^T g_t(M_{t,i}) - \sum_{t=1}^T g_t(M^*_t)\right)}_{\texttt{base-regret}}. \end{aligned}$$

We remark that the regret decomposition holds for any base-learner index $i \in [N]$. We now provide the upper bounds for the meta-regret and base-regret, respectively. First, Theorem 24 ensures the base-regret satisfies that

$$\texttt{base-regret} \leq \frac{\eta_i}{2}(G_f^2 + 2\lambda G_f)T + \frac{1}{2\eta_i}(D_f^2 + 2D_f P_T),$$

where $P_T = \sum_{t=2}^T \|M^*_{t-1} - M^*_t\|_{\mathrm{F}}$ is the path length of the comparator sequence. On the other hand, similar to Lemma 14 of Section B.2, we can show that the meta-regret satisfies

$$\texttt{meta-regret} \leq \lambda' \sum_{t=2}^T \|\boldsymbol{p}_{t-1} - \boldsymbol{p}_t\|_1 + \sum_{t=1}^T \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle - \sum_{t=1}^T \ell_{t,i},$$

where the surrogate loss vector $\boldsymbol{\ell}_t \in \Delta_N$ of the meta-algorithm is defined as

$$\ell_{t,i} = \lambda \|M_{t-1,i} - M_{t,i}\|_{\mathrm{F}} + g_t(M_{t,i}), \text{ for } i \in [N].$$

Then, we can use the static regret with switching cost of online mirror descent for the prediction with expert advice setting (c.f. Corollary 18 in Appendix B.3) and obtain that

$$\begin{aligned}\texttt{meta-regret} &\leq \varepsilon(2\lambda + G_f)(\lambda_f + G_f)D_f^2 T + \frac{\ln(1/p_{1,i})}{\varepsilon}\\ &= D_f\sqrt{2(2\lambda + G_f)(\lambda + G_f)T}\big(1 + \ln(1+i)\big),\end{aligned}$$

where the equation can be obtained by an appropriate setting of the learning rate ε.

Since the above decomposition and the upper bounds of meta-regret and base-regret all hold for any base-learner index $i \in [N]$, we will choose the best index denoted by i^* to make the regret bound tightest possible. Specifically, from the construction of the step size pool, we can ensure that there exists a step size η_{i^*} such that the optimal step size provably satisfies $\eta_{i^*} \leq \eta_* \leq 2\eta_{i^*}$. As a result, we have

$$\begin{aligned}&\sum_{t=1}^T g_t(M_t) - \sum_{t=1}^T g_t(M_t^*) + \lambda\sum_{t=2}^T \|M_{t-1} - M_t\|_{\mathrm{F}}\\ &\leq \frac{\eta_{i^*}}{2}(G_f^2 + 2\lambda G_f)T + \frac{1}{2\eta_{i^*}}(D_f^2 + 2D_f P_T) + D_f\sqrt{2(2\lambda + G_f)(\lambda + G_f)T}\big(1 + \ln(1+i)\big)\\ &\leq \frac{\eta_*}{2}(G_f^2 + 2\lambda G_f)T + \frac{1}{\eta_*}(D_f^2 + 2D_f P_T) + D_f\sqrt{2(2\lambda + G_f)(\lambda + G_f)T}\big(1 + \ln(1+i)\big)\\ &\leq \frac{3}{2}\sqrt{(G_f^2 + 2\lambda G_f)(D_f^2 + 2D_f P_T)T}\\ &\qquad + D_f\sqrt{2(2\lambda + G_f)(\lambda + G_f)T}\,(1 + \ln(\lceil\log_2(1 + 2P_T/D)\rceil + 2))\,.\end{aligned}$$

Combining this result with the regret decomposition (54) and the upper bounds (55), (56), we have

$$\begin{aligned}&\sum_{t=1}^T c_t(x_t, u_t) - \sum_{t=1}^T c_t(x_t^{\pi_t}, u_t^{\pi_t})\\ &\leq 4TG_c D^2\kappa^3(1-\gamma)^{H+1} + \frac{3}{2}\sqrt{(G_f^2 + 2\lambda G_f)(D_f^2 + 2D_f P_T)T}\\ &\qquad + D_f\sqrt{2(2\lambda + G_f)(\lambda + G_f)T}\,(1 + \ln(\lceil\log_2(1 + 2P_T/D)\rceil + 2)) + \lambda P_T.\end{aligned}$$

The specific values of D, L_f, G_f, D_f can be found in Lemma 29. By setting $H = \mathcal{O}(\log T)$, we obtain an $\widetilde{\mathcal{O}}(\sqrt{T(1 + P_T)})$ dynamic policy regret and hence complete the proof. ■

C.3 Proof of Theorem 8

In this part, we present the proof of Theorem 8. Specifically, we provide the main proof of Theorem 8 in Appendix C.3.1 and the proofs of some key lemmas in Appendix C.3.2.

Notations. We define some notations for convenience. Define ε_w an upper bound for the gap between the true disturbance w_t and the estimated one $\widehat{w}_t$, i.e., $\|w_t - \widehat{w}_t\|_2 \leq \varepsilon_w$, and define a universal upper bound W_0 for ε_w and disturbance bound W (cf. Assumption 4) as $W, \varepsilon_w \leq W_0$. We also define $d_{\min} = \min\{d_x, d_u\}, \widetilde{A}_K = A - BK, \widehat{A}_K = \widehat{A} - \widehat{B}K$ for notational convenience.

C.3.1 Proof of Theorem 8

Proof The overall dynamic regret is at most

$$\sum_{t=1}^{T} c_t(x_t, u_t) - \sum_{t=1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t}) \leq \underbrace{\sum_{t=1}^{T_0} c_t(x_t, u_t)}_{\texttt{term (A)}} + \underbrace{\sum_{t=T_0+1}^{T} c_t(x_t, u_t) - \sum_{t=T_0+1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t})}_{\texttt{term (B)}},$$

where term (A) is the cumulative cost during the system identification procedure and term (B) is the dynamic regret caused by Scream.Control algorithm over the rest rounds. Note that term (A) enjoys a trivial upper bound of $\mathcal{O}(T_0)$, and term (B) can be decomposed into two parts:

$$\begin{aligned}\texttt{term (B)} = & \underbrace{\sum_{t=T_0+1}^{T} c_t(x_t, u_t) - \sum_{t=T_0+1}^{T} c_t(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S}))}_{\texttt{term (b-1)}} \\ & + \underbrace{\sum_{t=T_0+1}^{T} c_t(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S})) - \sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(S), u_t^{\pi_t}(S)\right)}_{\texttt{term (b-2)}}.\end{aligned}$$

Here, $(x_t^{\pi_t}(S), u_t^{\pi_t}(S))$ is the state-action pair produced by the policy π_t on the true system $S = (A, B, \{w\})$, whereas $(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S}))$ is the state-action pair produced by the policy π_t on the estimated system $\widehat{S} = (\widehat{A}, \widehat{B}, \{\widehat{w}\})$. Summarizing, term (b-1) is the dynamic regret on the estimated system and term (b-2) is the gap between the cumulative cost of the true system and that of the estimated system. From Theorem 7, it holds that $\texttt{term (b-1)} \leq \widetilde{\mathcal{O}}(\sqrt{T(1+P_T)})$. From Lemma 26, we can bound term (b-2) as $\texttt{term (b-2)} \leq \mathcal{O}(\varepsilon_{A,B}T)$. Overall, with probability at least $1-\delta$, the total dynamic regret is at most

$$\begin{aligned}\text{D-Regret}_T & \leq \mathcal{O}(T_0) + \widetilde{\mathcal{O}}(\sqrt{T(1+P_T)}) + \mathcal{O}(\varepsilon_{A,B}T) \\ & = \mathcal{O}(\varepsilon_{A,B}^{-2} + \varepsilon_{A,B}T) + \widetilde{\mathcal{O}}(\sqrt{T(1+P_T)}) \\ & \leq \mathcal{O}(T^{2/3}) + \widetilde{\mathcal{O}}(\sqrt{T(1+P_T)}).\end{aligned}$$

The second step makes use of the relationship between the system identification rounds T_0 and the estimation error $\|\widehat{A} - A\|_{\text{op}}, \|\widehat{B} - B\|_{\text{op}} \leq \varepsilon_{A,B}$, as demonstrated in Lemma 25. The last step holds by setting the rounds of exploration to ensure $\varepsilon_{A,B} = \min\{10^{-3}\kappa^{-10}\gamma^2, T^{-1/3}\}$, which is realized when total time horizon is large enough, i.e., $T \geq 10^9\kappa^{30}\gamma^{-6}$. ■

C.3.2 Key Lemmas in Unknown Systems

The proof of Theorem 8 relies on the two key lemmas (Lemma 25 and Lemma 26). In the following, we provide the formal statements and corresponding proofs.

Lemma 25 establishes the relationship between the estimation accuracy $\varepsilon_{A,B}$ and the number of estimation rounds T_0. This lemma is firstly due to Hazan et al. (2020) and is restated here for self-containedness.

Lemma 25 (Theorem 19 of Hazan et al. (2020))**.** *Under Assumptions 4, 6, 7, when Algorithm 4 runs for T_0 rounds, if the output pair $(\widehat{A}, \widehat{B})$ satisfies, with probability at least $1-\delta$, that $\|\widehat{A} - A\|_{\mathrm{op}}, \|\widehat{B} - B\|_{\mathrm{op}} \le \varepsilon_{A,B}$, then it holds that $T_0 = \mathcal{O}(\varepsilon_{A,B}^{-2})$.*

Proof [of Lemma 25] Based on the observation, we have the following two equations:

$$\widetilde{A}_K C_k = (\widetilde{A}_K C_k), \quad \widehat{A}_K \widehat{C}_0 = \widehat{C}_1.$$

Using Lemma 36, it holds that

$$\|\widetilde{A}_K - \widehat{A}_K\|_{\mathrm{op}} \le \frac{\|\widetilde{A}_K C_k - \widehat{C}_1\|_{\mathrm{op}} + \|C_k - \widehat{C}_0\|_{\mathrm{op}} \|\widetilde{A}_K\|_{\mathrm{op}}}{\sigma_{\min}(C_k) - \|C_k - \widehat{C}_0\|_{\mathrm{op}}}. \tag{57}$$

Lemma 31 tells that with probability at least $1-\delta$, $\|N_j - \widetilde{A}_K^j B\|_{\mathrm{F}} \le \varepsilon$, where

$$\varepsilon \triangleq 3\kappa_B \kappa^2 d_u W \gamma^{-1} \sqrt{\frac{2 d_{\min} \log(2e^2 k \delta^{-1})}{T_0 - k}}. \tag{58}$$

Owing to the benign high-probability guarantee, we only need to focus on the successful event, that is, under the case when $\|N_j - \widetilde{A}_K^j B\|_{\mathrm{F}} \le \varepsilon$ is true. We then try to bound $\|C_k - C_0\|_{\mathrm{op}}, \|\widetilde{A}_K C_k - C_1\|_{\mathrm{op}}$,

$$\begin{aligned} \|C_k - \widehat{C}_0\|_{\mathrm{op}} \le \|C_k - \widehat{C}_0\|_{\mathrm{F}} &= \left\| \left[N_0 - B, \ldots, N_{k-1} - \widetilde{A}_K^{k-1} B \right] \right\|_{\mathrm{F}} \\ &= \sqrt{\sum_{i=0}^{k-1} \|N_i - \widetilde{A}_K^i B\|_{\mathrm{F}}^2} \le \sqrt{k\varepsilon^2} = \varepsilon\sqrt{k}, \end{aligned} \tag{59}$$

$$\begin{aligned} \|\widetilde{A}_K C_k - \widehat{C}_1\|_{\mathrm{op}} \le \|\widetilde{A}_K C_k - \widehat{C}_1\|_{\mathrm{F}} &= \left\| \left[N_1 - \widetilde{A}_K B, \ldots, N_k - \widetilde{A}_K^k B \right] \right\|_{\mathrm{F}} \\ &= \sqrt{\sum_{i=1}^{k} \|N_i - \widetilde{A}_K^i B\|_{\mathrm{F}}^2} \le \sqrt{k\varepsilon^2} = \varepsilon\sqrt{k}. \end{aligned} \tag{60}$$

Using Lemma 34 to upper-bound $\sigma_{\min}(C_k)$, and plugging (59) and (60) into (57), we have

$$\|\widetilde{A}_K - \widehat{A}_K\|_{\mathrm{op}} \le \frac{\varepsilon\sqrt{k} + \varepsilon\sqrt{k} \cdot \kappa^2(1-\gamma)}{1/\sqrt{\kappa_c} - \varepsilon\sqrt{k}}.$$

The gap between A and $\widehat{A}$ can be bounded as

$$\begin{aligned} \|A - \widehat{A}\|_{\mathrm{op}} &= \|\widetilde{A}_K + BK - \widehat{A}_K - \widehat{B}K\|_{\mathrm{op}} \\ &\le \|\widetilde{A}_K - \widehat{A}_K\|_{\mathrm{op}} + \|K\|_{\mathrm{op}} \|B - \widehat{B}\|_{\mathrm{op}} \\ &\le \frac{\varepsilon\sqrt{k} + \varepsilon\sqrt{k} \cdot \kappa^2(1-\gamma)}{1/\sqrt{\kappa_c} - \varepsilon\sqrt{k}} + \kappa\varepsilon \le \frac{3\varepsilon\kappa^{5/2}}{\sqrt{1/\kappa_c} - \varepsilon\sqrt{\kappa}}. \end{aligned}$$

If we want $\|\widehat{A}-A\|_{\mathrm{F}}, \|\widehat{B}-B\|_{\mathrm{F}} \leq \varepsilon_{A,B}$, the following equations should hold:

$$\begin{aligned} &\|\widehat{A}-A\|_{\mathrm{F}} \leq \sqrt{d_x}\|\widehat{A}-A\|_{\mathrm{op}} \leq \sqrt{d_x}\left(\frac{3\varepsilon\kappa^{5/2}}{\sqrt{1/\kappa_c}-\varepsilon\sqrt{\kappa}}\right) \triangleq \varepsilon_A \leq \varepsilon_{A,B}, \\ &\|\widehat{B}-B\|_{\mathrm{F}} \leq \sqrt{d_{\min}}\|\widehat{B}-B\|_{\mathrm{op}} \leq \sqrt{d_{\min}}\varepsilon \triangleq \varepsilon_B \leq \varepsilon_{A,B}. \end{aligned} \tag{61}$$

Besides, it is easy to see that $\varepsilon_B = \sqrt{d_{\min}}\varepsilon \leq \sqrt{d_x}\varepsilon \leq \varepsilon_A$, thus conditions in (61) can be simplified as $\varepsilon_A \leq \varepsilon_{A,B}$. Finally, combining the above inequality with the value of ε (c.f. (58)), we can obtain that $T_0 = \mathcal{O}(\varepsilon_{A,B}^{-2})$. ■

Lemma 26 measures the difference of the cumulative costs of a policy between the true system and the estimated one. This result holds for both strongly stable linear controllers and non-stationary DAC policy and here we only give a proof of the latter, for the former result, we refer readers to Hazan et al. (2020, Lemma 16).

Lemma 26 (Identification Accuracy). *Under Assumptions 4-6, suppose $\|\widehat{A}-A\|_{\mathrm{op}}, \|\widehat{B}-B\|_{\mathrm{op}} \leq \varepsilon_{A,B} \leq 0.25\kappa^{-3}\gamma$ and let K be any (κ,γ)-strongly stable linear controller with respect to (A,B). Then for any non-stationary DAC policy $\pi_{1:T}$ parameterized via $M_{1:T}$,*

$$\left|\sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S})\right) - \sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(S), u_t^{\pi_t}(S)\right)\right| \leq \mathcal{O}\left(\varepsilon_{A,B}T + \varepsilon_{A,B}^2 T\right),$$

where $(x_t^{\pi_t}(S), u_t^{\pi_t}(S))$ is the state-action pair produced by policy π_t on the true system $S = (A, B, \{w\})$ and $(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S}))$ is produced on the estimated system $\widehat{S} = (\widehat{A}, \widehat{B}, \{\widehat{w}\})$.

Proof [of Lemma 26] If the policy is a non-stationary DAC policy parameterized via $M_{1:T}$, in system $(A, B, \{w\})$, it holds that

$$\begin{aligned} \|x_{t+1}^{\pi_t}(S)\|_2 &\leq W\sum_{i=0}^{H+t}\|\Psi_{t,i}^{K,t}(M_{0:t})\|_{\mathrm{op}} \\ &= W\sum_{i=0}^{H+t}\|\widetilde{A}_K^i\mathbf{1}_{i\leq t} + \sum_{j=0}^{t}\widetilde{A}_K^j B M_{t-j}^{[i-j]}\mathbf{1}_{1\leq i-j\leq H}\|_{\mathrm{op}} \\ &\leq W\left(\kappa^2\sum_{i=0}^{H+t}(1-\gamma)^i + \kappa_B^2\kappa^3\sum_{i=0}^{H+t}\sum_{j=0}^{t}\|\widetilde{A}_K^j\mathbf{1}_{1\leq i-j\leq H}\|_{\mathrm{op}}\right) \\ &\leq W\left(\kappa^2\gamma^{-1} + \kappa_B^2\kappa^3\sum_{i=0}^{H+t}\sum_{j=i-H}^{i-1}\|\widetilde{A}_K^j\|_{\mathrm{op}}\mathbf{1}_{0\leq j\leq t}\right) \\ &\leq W\left(\kappa^2\gamma^{-1} + \kappa_B^2\kappa^5\sum_{i=0}^{H+t}\sum_{j=i-H}^{i-1}(1-\gamma)^j\mathbf{1}_{0\leq j\leq t}\right) \\ &\leq W\left(\kappa^2\gamma^{-1} + \kappa_B^2\kappa^5 H\sum_{i=0}^{t}(1-\gamma)^i\right) \end{aligned}$$

$$
\begin{aligned}
&\leq W\left(\kappa^2\gamma^{-1} + \kappa_B^2\kappa^5 H\gamma^{-1}\right)\\
&\leq 2W\kappa_B^2\kappa^5\gamma^{-1}H.
\end{aligned}
$$

By Lemma 32, a linear controller K is $\left(\kappa, \gamma - 2\kappa^3\varepsilon_{A,B}\right)$-strongly stable with respect to the estimated system $\widehat{S} = (\widehat{A}, \widehat{B}, \{\widehat{w}\})$ if it is (κ,γ)-strongly stable for the true system $S = (A, B, \{w\})$. Thus it can be easily verified that

$$1 - \gamma + 2\kappa^3\varepsilon_{A,B} \leq 1 - \gamma + 2\kappa^3 \cdot 0.25\kappa^{-3}\gamma = 1 - \gamma/2.$$

For simplicity, we can say that linear controller K is $(\kappa, \gamma/2)$-strongly stable for the estimated system $\widehat{S}$. Further, let $\|\widehat{B}\|_{\text{op}} \leq \kappa_{\widehat{B}}$, it holds that

$$\kappa_{\widehat{B}} = \|\widehat{B}\|_{\text{op}} = \|(\widehat{B} - B) + B\|_{\text{op}} \leq \varepsilon_{A,B} + \kappa_B \leq 2\kappa_B.$$

As a result, we can bound $\|x_{t+1}^{\pi_t}(\widehat{S})\|_2$ as

$$\|x_{t+1}^{\pi_t}(\widehat{S})\|_2 \leq 2(\varepsilon_w + W)(2\kappa_B)^2\kappa^5(\gamma/2)^{-1}H = 32W_0\kappa_B^2\kappa^5\gamma^{-1}H.$$

As for the action $u_t^{\pi_t}(\widehat{S})$, we can bound it as

$$
\begin{aligned}
\|u_t^{\pi_t}(\widehat{S})\|_2 \leq \|-Kx_t^{\pi_t}(\widehat{S})\|_2 + \left\|\sum_{i=1}^{H} M_t^{[i]}\widehat{w}_{t-i}\right\|_2 &\leq 32W_0\kappa_B^2\kappa^6\gamma^{-1}H + 2W_0\kappa_B\kappa^3\gamma^{-1}\\
&\leq 34W_0\kappa_B^2\kappa^6\gamma^{-1}H.
\end{aligned}
$$

Thus, the diameter of the state-action domain in the estimated system, denoted as $\widehat{D}$, is at most $\widehat{D} \triangleq \max_{t\in[T]}\max\{\|x_t(\widehat{S})\|_2, \|u_t(\widehat{S})\|_2\} = 34W_0\kappa_B^2\kappa^6\gamma^{-1}H$. The gap of the cumulative costs between the true system and the estimated system can be bounded as

$$
\begin{aligned}
&\left|\sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S})\right) - \sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(S), u_t^{\pi_t}(S)\right)\right|\\
&\leq G_c\widehat{D}\sum_{t=1}^{T}\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 + G_c\widehat{D}\sum_{t=1}^{T}\|u_t^{\pi_t}(\widehat{S}) - u_t^{\pi_t}(S)\|_2.
\end{aligned}
\tag{62}
$$

We start by analyzing $\|u_t^{\pi_t}(\widehat{S}) - u_t^{\pi_t}(S)\|_2$:

$$
\begin{aligned}
\|u_t^{\pi_t}(\widehat{S}) - u_t^{\pi_t}(S)\|_2 &= \left\|\left(-Kx_t^{\pi_t}(\widehat{S}) + \sum_{i=1}^{H} M_t^{[i]}\widehat{w}_{t-i}\right) - \left(-Kx_t^{\pi_t}(S) + \sum_{i=1}^{H} M_t^{[i]}w_{t-i}\right)\right\|_2\\
&\leq \kappa\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 + \sum_{i=1}^{H}\|M_t^{[i]}(\widehat{w}_{t-i} - w_{t-i})\|\\
&\leq \kappa\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 + \varepsilon_w\kappa_B\kappa^3\sum_{i=1}^{H}(1-\gamma)^i\\
&\leq \kappa\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 + \varepsilon_w\kappa_B\kappa^3\gamma^{-1}.
\end{aligned}
\tag{63}
$$

Plugging (63) into (62), it holds that

$$\begin{aligned} &\left| \sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S})\right) - \sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(S), u_t^{\pi_t}(S)\right) \right| \\ &\le 2\kappa G_c \widehat{D} \sum_{t=1}^{T} \|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 + G_c \widehat{D} \varepsilon_w \kappa_B \kappa^3 \gamma^{-1} T. \end{aligned} \tag{64}$$

This motivates the need to analyze $\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2$. To begin with, we define $\widehat{\Psi}^{K,h}_{t,i}(M_{t-h:t}) = \widehat{A}_K^i \mathbf{1}_{i\le h} + \sum_{j=0}^{h} \widehat{A}_K^j \widehat{B} M_{t-j}^{[i-j]} \mathbf{1}_{1\le i-j\le H}$, where $\widehat{A}_K \triangleq \widehat{A} - \widehat{B}K$. Expanding $x_t^{\pi_t}(\widehat{S})$ and $x_t^{\pi_t}(S)$ using Proposition 6, it holds that

$$\begin{aligned} &\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 = \left\| \sum_{i=0}^{H+t} \Psi^{K,t}_{t,i}(M_{1:t}) w_{t-i} - \sum_{i=0}^{H+t} \widehat{\Psi}^{K,t}_{t,i}(M_{1:t}) \widehat{w}_{t-i} \right\|_2 \\ &\le \underbrace{\left\| \sum_{i=0}^{H+t} \Psi^{K,t}_{t,i}(M_{1:t}) w_{t-i} - \sum_{i=0}^{H+t} \Psi^{K,t}_{t,i}(M_{1:t}) \widehat{w}_{t-i} \right\|_2}_{\texttt{term (i)}} + \underbrace{\left\| \sum_{i=0}^{H+t} \Psi^{K,t}_{t,i}(M_{1:t}) \widehat{w}_{t-i} - \sum_{i=0}^{H+t} \widehat{\Psi}^{K,t}_{t,i}(M_{1:t}) \widehat{w}_{t-i} \right\|_2}_{\texttt{term (ii)}}. \end{aligned} \tag{65}$$

First, we analyze term (i):

$$\texttt{term (i)} \le \varepsilon_w \sum_{i=0}^{H+t} \|\Psi^{K,t}_{t,i}(M_{1:t})\|_{\mathrm{op}} \le 2\varepsilon_w \kappa_B^2 \kappa^5 \gamma^{-1} H. \tag{66}$$

Second, we investigate term (ii):

$$\begin{aligned} \texttt{term (ii)} &\le (W + \varepsilon_w) \sum_{i=0}^{H+t} \left\| \Psi^{K,t}_{t,i}(M_{1:t}) - \widehat{\Psi}^{K,t}_{t,i}(M_{1:t}) \right\|_{\mathrm{op}} \\ &\le 2W_0 \sum_{i=0}^{H+t} \left(\left\| \left(\widetilde{A}_K^i - \widehat{A}_K^i \right) \mathbf{1}_{i\le t} \right\|_{\mathrm{op}} + \kappa_B \kappa^3 \sum_{j=0}^{t} \|\widetilde{A}_K^j B - \widehat{A}_K^j \widehat{B}\|_{\mathrm{op}} \mathbf{1}_{1\le i-j\le H} \right) \\ &\le 2W_0\kappa^2 \underbrace{\sum_{i=0}^{t} \|L^i - \widehat{L}^i\|_{\mathrm{op}}}_{\texttt{term (a)}} + 2W_0 \kappa_B \kappa^3 \underbrace{\sum_{i=0}^{H+t} \sum_{j=0}^{t} \|\widetilde{A}_K^j B - \widehat{A}_K^j \widehat{B}\|_{\mathrm{op}} \mathbf{1}_{1\le i-j\le H}}_{\texttt{term (b)}}. \end{aligned} \tag{67}$$

For term (a), using Lemma 35, it holds that

$$\sum_{i=0}^{t} \|L^i - \widehat{L}^i\|_{\mathrm{op}} \le 3\gamma^{-2} \|L - \widehat{L}\|_{\mathrm{op}} \le 3\gamma^{-2} \cdot 2\kappa^3 \varepsilon_{A,B} = 6\kappa^3 \gamma^{-2} \varepsilon_{A,B}.$$

For term (b), by inserting an intermediate term, we have

$$\sum_{i=0}^{H+t} \sum_{j=0}^{t} \|\widetilde{A}_K^j B - \widehat{A}_K^j \widehat{B}\|_{\mathrm{op}} \mathbf{1}_{1\le i-j\le H}$$

$$
\begin{aligned}
&\le \sum_{i=0}^{H+t}\sum_{j=0}^{t}\|\widetilde{A}_K^j B - \widetilde{A}_K^j \widehat{B}\|_{\mathrm{op}}\mathbf{1}_{1\le i-j\le H} + \sum_{i=0}^{H+t}\sum_{j=0}^{t}\|\widetilde{A}_K^j \widehat{B} - \widehat{A}_K^j \widehat{B}\|_{\mathrm{op}}\mathbf{1}_{1\le i-j\le H}\\
&\le \varepsilon_{A,B}\sum_{i=0}^{H+t}\sum_{j=0}^{t}\|\widetilde{A}_K^j\|_{\mathrm{op}}\mathbf{1}_{1\le i-j\le H} + \kappa_{\widehat{B}}\sum_{i=0}^{H+t}\sum_{j=0}^{t}\|\widetilde{A}_K^j - \widehat{A}_K^j\|_{\mathrm{op}}\mathbf{1}_{1\le i-j\le H}\\
&\le \varepsilon_{A,B}H\gamma^{-1} + 2\kappa_B\kappa^2 H\sum_{i=0}^{t}\|L^i - \widehat{L}^i\|_{\mathrm{op}}\\
&\le \varepsilon_{A,B}H\gamma^{-1} + 2\kappa_B\kappa^2 H\cdot 6\kappa^3\gamma^{-2}\varepsilon_{A,B}.
\end{aligned}
$$

Plugging term (a) and term (b) into (67), we have

$$
\begin{aligned}
\texttt{term (ii)} &\le 2W_0\kappa^2\cdot 6\kappa^3\gamma^{-2}\varepsilon_{A,B} + 2W_0\kappa_B\kappa^3\cdot(\varepsilon_{A,B}H\gamma^{-1} + 2\kappa_B\kappa^2 H\cdot 6\kappa^3\gamma^{-2}\varepsilon_{A,B})\\
&\le 38W_0\kappa_B^2\kappa^8\gamma^{-2}H\varepsilon_{A,B}.
\end{aligned}
$$

Plugging the bounds of (66) and (67) into (65), we have

$$
\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 \le 2\varepsilon_w\kappa_B^2\kappa^5\gamma^{-1}H + 38W_0\kappa_B^2\kappa^8\gamma^{-2}H\varepsilon_{A,B}. \tag{68}
$$

Furthermore, by Lemma 33, we have

$$
W_0 \le 2\sqrt{d_u}\kappa^3\gamma^{-1}W, \quad \varepsilon_w \le 42\sqrt{d_u}\kappa^{12}\gamma^{-3}W\varepsilon_{A,B}
$$

Plugging W_0 and ε_w into (68), it holds that

$$
\|x_t^{\pi_t}(\widehat{S}) - x_t^{\pi_t}(S)\|_2 \le \mathcal{O}(\varepsilon_{A,B} + \varepsilon_{A,B}^2).
$$

Plugging the above bound into (64), we have

$$
\left|\sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(\widehat{S}), u_t^{\pi_t}(\widehat{S})\right) - \sum_{t=T_0+1}^{T} c_t\left(x_t^{\pi_t}(S), u_t^{\pi_t}(S)\right)\right| \le \mathcal{O}\left(\varepsilon_{A,B}T + \varepsilon_{A,B}^2 T\right),
$$

which finishes the proof. ■

C.4 Proof of Corollary 9

We now present the proof of Corollary 9, i.e., the static policy regret of the controller. Corollary 9 states that when the system dynamics are known, Scream.Control enjoys the following static policy regret,

$$
\sum_{t=1}^{T} c_t(x_t, u_t) - \min_{\pi\in\Pi}\sum_{t=1}^{T} c_t(x_t^{\pi}, u_t^{\pi}) \le \widetilde{\mathcal{O}}(\sqrt{T}), \tag{69}
$$

where the comparator set Π can be chosen as either the set of DAC policies or the set of strongly linear controllers. Let us denote the two comparator sets as Π_{DAC} and Π_{SLC}, respectively. Moreover, when the system dynamics are unknown, using the identification algorithm of Hazan et al. (2020), we can achieve an $\widetilde{\mathcal{O}}(T^{2/3})$ static regret, which also holds for either the set of DAC policies or the set of strongly linear controllers. Therefore, in the following we will prove the statement for two comparator sets separately.

Proof [of Corollary 9] When the comparator set Π is chosen as the set of DAC policies, i.e., $\pi \in \Pi_{\mathrm{DAC}} = \{\pi(K, M) \mid M \in \mathcal{M}\}$, the result of (69) can be easily obtained from Theorem 7 by setting $\pi_1 = \ldots = \pi_T = \pi_* \in \arg\min_{\pi \in \Pi} \sum_{t=1}^T c_t(x_t^\pi, u_t^\pi)$. Under such a case, the path length $P_T = \sum_{t=2}^T \|M_{t-1} - M_t\|_{\mathrm{F}} = 0$, and thus

$$\sum_{t=1}^T c_t(x_t, u_t) - \min_{\pi \in \Pi_{\mathrm{DAC}}} \sum_{t=1}^T c_t(x_t^\pi, u_t^\pi) \leq \widetilde{\mathcal{O}}(\sqrt{T}).$$

On the other hand, when choosing the comparator set Π as Π_{SL}, i.e., $\pi = K \in \Pi_{\mathrm{SL}} = \{K \mid K \text{ is } (\kappa, \gamma)\text{-strongly stable}\}$, we will need some efforts to prove the statement.

We show that the statement can be obtained by further incorporating Lemma 30, which demonstrates that minimizing static policy regret over the DAC class is sufficient to deliver a policy regret competing with the strongly linear controller class (Agarwal et al., 2019, Lemma 5.2). In fact, denote by $\pi^* = K^\star = \arg\min_{K \in \Pi_{\mathrm{SL}}} \sum_{t=1}^T c_t(x_t^K, u_t^K)$, and we have

$$\begin{aligned}
&\sum_{t=1}^T c_t(x_t, u_t) - \min_{\pi \in \Pi_{\mathrm{SLC}}} \sum_{t=1}^T c_t(x_t^\pi, u_t^\pi) \\
&= \sum_{t=1}^T c_t(x_t, u_t) - \min_{\pi \in \Pi_{\mathrm{DAC}}} \sum_{t=1}^T c_t(x_t^\pi, u_t^\pi) + \min_{\pi \in \Pi_{\mathrm{DAC}}} \sum_{t=1}^T c_t(x_t^\pi, u_t^\pi) - \sum_{t=1}^T c_t(x_t^{K^*}, u_t^{K^*}) \\
&\leq \widetilde{\mathcal{O}}(\sqrt{T}) + \sum_{t=1}^T c_t(x_t^{\pi(M_\Delta, K)}, u_t^{\pi(M_\Delta, K)}) - \sum_{t=1}^T c_t(x_t^{K^*}, u_t^{K^*}) \\
&\leq \widetilde{\mathcal{O}}(\sqrt{T}) + T \cdot 4 G_c D W H \kappa_B^2 \kappa^6 (1-\gamma)^{H-1} \gamma^{-1} \leq \widetilde{\mathcal{O}}(\sqrt{T}),
\end{aligned}$$

where the first inequality uses the optimality of $\arg\min_{\pi \in \Pi_{\mathrm{DAC}}} \sum_{t=1}^T c_t(x_t^\pi, u_t^\pi)$ and $\pi(M_\Delta, K)$ is a DAC policy with $M_\Delta = (M_\Delta^{[1]}, \ldots, M_\Delta^{[H]})$ defined by $M_\Delta^{[i]} = (K - K^\star)(A - BK^\star)^i$. The second inequality holds by Lemma 30, and the final inequality sets $H = \mathcal{O}(\log T)$.

The above arguments hold for the known system setting. On the other hand, when the system dynamics are unknown, using the system identification yields an additional estimation overhead of order $\widetilde{\mathcal{O}}(T^{2/3})$ no matter which comparator set is chosen. Therefore, the overall regret remains $\widetilde{\mathcal{O}}(T^{2/3})$ for unknown systems. Hence, we complete the proof. ■

C.5 Supporting Lemmas

In this part, we provide several supporting lemmas used frequently in the analysis of online non-stochastic control. Most of them are due to the pioneering works (Agarwal et al., 2019; Hazan et al., 2020), and we adapt them to our notations and provide the proofs to achieve self-containedness. Specifically,

- Lemma 27 establishes the norm relations between the ℓ_1, op norm and Frobenius norm used in the $\mathcal{M}$-space.
- Lemma 28 checks the boundedness of several variables of interest.
- Lemma 29 shows several properties of the truncated functions $\{f_t\}_{t=1}^T$ and the feasible set $\mathcal{M}$.
- Lemma 30 connects the DAC class and the strongly linear controller class.
- Lemma 31 – Lemma 36 are useful for analysis in unknown systems.

Lemma 27 (Norm Relations)**.** *For any $M = (M^{[1]}, \ldots, M^{[H]}) \in \mathcal{M} \subseteq (\mathbb{R}^{d_u \times d_x})^H$, its ℓ_1, op norm and Frobenius norm are defined by*

$$\|M\|_{\ell_1,\mathrm{op}} \triangleq \sum_{i=1}^{H} \|M^{[i]}\|_{\mathrm{op}}, \text{ and } \|M\|_{\mathrm{F}} \triangleq \sqrt{\sum_{i=1}^{H} \|M^{[i]}\|_{\mathrm{F}}^2}.$$

Denoting by $d = \min\{d_u, d_x\}$, we then have the following inequalities on their relations:

$$\|M\|_{\ell_1,\mathrm{op}} \le \sqrt{H}\|M\|_{\mathrm{F}}, \text{ and } \|M\|_{\mathrm{F}} \le \sqrt{d}\|M\|_{\ell_1,\mathrm{op}}.$$

Proof [of Lemma 27] We know that for any matrix $X \in \mathbb{R}^{m\times n}$, $\|X\|_{\mathrm{op}} \le \|X\|_{\mathrm{F}} \le \sqrt{d}\|X\|_{\mathrm{op}}$. Therefore, by definition and Cauchy-Schwarz inequality, we obtain

$$\|M\|_{\ell_1,\mathrm{op}} = \sum_{i=1}^{H} \|M^{[i]}\|_{\mathrm{op}} \le \sum_{i=1}^{H} \|M^{[i]}\|_{\mathrm{F}} \le \sqrt{H}\|M\|_{\mathrm{F}}.$$

On the other hand, we have

$$\|M\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{H} \|M^{[i]}\|_{\mathrm{F}}^2} \le \sum_{i=1}^{H} \|M^{[i]}\|_{\mathrm{F}} \le \sum_{i=1}^{H} \sqrt{d}\|M^{[i]}\|_{\mathrm{op}} = \sqrt{d}\|M\|_{\ell_1,\mathrm{op}},$$

which completes the proof. ■

Lemma 28 (Lemma 5.5 of Agarwal et al. (2019))**.** *Suppose K and $K^\star$ are two (κ, γ)-strongly stable linear controllers (cf. Definition 4). Define*

$$D \triangleq \frac{W(\kappa^3 + H\kappa_B\kappa^3\tau)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{W\tau}{\gamma}. \tag{70}$$

Suppose there exists a $\tau > 0$ such that for all $i \in [H]$ and $t \in [T]$, $\|M_t^{[i]}\|_{\mathrm{F}} \le \tau(1-\gamma)^i$. Then, we have

- *$\|x_t^K(M_{0:t-1})\| \le D$, $\|y_t^K(M_{t-H-1:t-1})\| \le D$, and $\|x_t^{K^\star}\| \le D$.*
- *$\|u_t^K(M_{0:t})\| \le D$, and $\|v_t^K(M_{t-H-1:t})\| \le D$.*

- $\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-1-H:t-1})\| \leq \kappa^2(1-\gamma)^{H+1}D.$
- $\|u_t^K(M_{0:t}) - v_t^K(M_{t-1-H:t})\| \leq \kappa^3(1-\gamma)^{H+1}D.$

In above, the definitions of state $x_t^K(M_{0:t-1})$ and corresponding DAC control $u_t^K(M_{0:t})$ can be found in Proposition 6, and the definitions of truncated state $x_t^K(M_{0:t-1})$ and corresponding DAC control $v_t^K(M_{0:t})$ can be found in Definition 2. The definitions of state $x_t^{K^\star}$ can be found (and will be used) in Lemma 30.

Proof [of Lemma 28] We first study the state.

$$\begin{aligned}
\|x_t^K(M_{0:t-1})\| &= \left\| \widetilde{A}_K^{H+1} x_{t-H-1}^K(M_{0:t-H-2}) + \sum_{i=0}^{2H} \Psi_{t-1,i}^{K,H}(M_{t-H-1:t-1}) w_{t-1-i} \right\| \\
&\leq \kappa^2(1-\gamma)^{H+1}\|x_{t-H-1}^K(M_{0:t-H-2})\| + W\sum_{i=0}^{2H}\|\Psi_{t-1,i}^{K,H}(M_{t-H-1:t-1})\| \\
&\leq \kappa^2(1-\gamma)^{H+1}\|x_{t-H-1}^K(M_{0:t-H-2})\| + W\sum_{i=0}^{2H}\left(\kappa^2(1-\gamma)^i + H\kappa_B\kappa^2\tau(1-\gamma)^{i-1}\right) \\
&\leq \kappa^2(1-\gamma)^{H+1}\|x_{t-H}^K(M_{0:t-H-1})\| + W(\kappa^2 + H\kappa_B\kappa^2\tau)/\gamma \\
&\leq \frac{W(\kappa^2 + H\kappa_B\kappa^2\tau)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} \leq D, \qquad (71)
\end{aligned}$$

where inequality (71) is a summation of geometric series and the ratio of this series is $\kappa^2(1-\gamma)^{H+1}$. Similarly,

$$\begin{aligned}
\|y_t^K(M_{t-1-H:t-1})\| &= \left\|\sum_{i=0}^{2H}\Psi_{t-1,i}^{K,H}(M_{t-1-H:t-1})w_{t-1-i}\right\| \\
&\leq W\sum_{i=0}^{2H}\|\Psi_{t-1,i}^{K,H}(M_{t-1-H:t-1})\| \\
&\leq W\sum_{i=0}^{2H}\left(\kappa^2(1-\gamma)^i + H\kappa_B\kappa^2\tau(1-\gamma)^{i-1}\right) \\
&\leq W\left(\frac{\kappa^2 + H\kappa_B\kappa^2\tau}{\gamma}\right) \leq D.
\end{aligned}$$

Besides,

$$\|x_t^{K^\star}\| = \left\|\sum_{i=0}^{t-1}\widetilde{A}_{K^\star}^i w_{t-1-i}\right\| \leq W\sum_{i=0}^{t-1}\kappa^2(1-\gamma)^i \leq \frac{W\kappa^2}{\gamma} \leq D.$$

So the difference can be evaluated as follows:

$$\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1})\| = \|\widetilde{A}_K^{H+1}x_{t-H-1}^K(M_{0:t-H-1})\| \leq \kappa^2(1-\gamma)^{H+1}D.$$

We now consider the action (or control signal).

$$\|u_t^K(M_{0:t})\| = \left\|-Kx_t^K(M_{0:t-1}) + \sum_{i=1}^{H} M_t^{[i]}w_{t-i}\right\|$$

$$
\begin{aligned}
&\leq \kappa \|x_t^K(M_{0:t-1})\| + \sum_{i=1}^{H} W\tau(1-\gamma)^{i-1} \\
&\leq \frac{W(\kappa^3 + H\kappa_B\kappa^3\tau)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{W\tau}{\gamma} \leq D.
\end{aligned}
$$

Similarly,

$$\|v_t^K(M_{t-H-1:t})\| \leq \kappa \|y_t^K(M_{t-H-1:t-1})\| + \sum_{i=1}^{H} W\tau(1-\gamma)^{i-1} \leq D.$$

The difference of the actions is

$$\|u_t^K(M_{0:t-1}) - v_t^K(M_{t-H-1:t-1})\| = \|-K(x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1}))\| \leq \kappa^3(1-\gamma)^{H+1}D,$$

which finishes the proof. ■

To reduce the online non-stochastic control to OCO with memory, in Definition 2 we define the truncated loss $f_t : \mathcal{M}^{H+2} \mapsto \mathbb{R}$ as

$$f_t(M_{t-1-H:t}) = c_t(y_t^K(M_{t-1-H:t-1}), v_t^K(M_{t-1-H:t})),$$

where $y_{t+1}^K(M_{t-H:t}) = \sum_{i=0}^{2H} \Psi_{t,i}^{K,H}(M_{t-H:t})w_{t-i}$ and $v_{t+1}^K(M_{t-H:t+1}) = -Ky_{t+1}(M_{t-H:t}) + \sum_{i=1}^{H} M_{t+1}^{[i]} w_{t+1-i}$. In the following lemma, we show several properties of the truncated functions $\{f_t\}_{t=1}^T$ and the feasible set $\mathcal{M}$ such that we can further apply the results of OCO with memory.

Lemma 29. *The truncated loss $f_t : \mathcal{M}^{H+2} \mapsto \mathbb{R}$ and the feasible set $\mathcal{M}$ satisfy the following properties. For notational convenience, we first let D be defined the same as* (51)*, and we restate it below*

$$D \triangleq \frac{W\kappa^3(1 + H\kappa_B\tau)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{W\tau}{\gamma}.$$

(i) The function is L_f-coordinate-wise Lipschitz with respect to the Euclidean (i.e., Frobenius) norm, namely,

$$|f_t(M_{t-H-1}, \ldots, M_{t-k}, \ldots, M_t)| - |f_t(M_{t-H-1}, \ldots, \widetilde{M}_{t-k}, \ldots, M_t)| \leq L_f \|M_{t-k} - \widetilde{M}_{t-k}\|_{\mathrm{F}},$$

where $L_f \leq 3\sqrt{H} G_c D W \kappa_B \kappa^3$.

(ii) The gradient norm of surrogate loss $\widetilde{f}_t : \mathcal{M} \mapsto \mathbb{R}$ is bounded by G_f, i.e., $\|\nabla_M \widetilde{f}_t(M)\|_{\mathrm{F}} \leq G_f$ holds for any $M \in \mathcal{M}$ and any $t \in [T]$, where $G_f \leq 3Hd^2 G_c W \kappa_B \kappa^3 \gamma^{-1}$.

(iii) The diameter of the feasible set is at most D_f, namely, $\|M - M'\|_{\mathrm{F}} \leq D_f$ holds for any $M, M' \in \mathcal{M}$, where $D_f \leq 2\sqrt{d}\kappa_B\kappa^3\gamma^{-1}$.

Proof [of Lemma 29] We first prove the claim (i), i.e., the L_f-coordinate-wise Lipschitz continuity. For simplicity, we use the following definitions in the following arguments.

$$M_{t-H-1:t} \triangleq \{M_{t-H-1} \dots M_{t-k} \dots M_t\}, \quad M_{t-H-1:t-1} \triangleq \{M_{t-H-1} \dots M_{t-k} \dots M_{t-1}\},$$
$$\widetilde{M}_{t-H-1:t} \triangleq \{M_{t-H-1} \dots \widetilde{M}_{t-k} \dots M_t\}, \quad \widetilde{M}_{t-H-1:t-1} \triangleq \{M_{t-H-1} \dots \widetilde{M}_{t-k} \dots M_{t-1}\}.$$

By representing f_t using c_t, we have

$$\begin{aligned} & f_t(M_{t-H-1:t}) - f_t(\widetilde{M}_{t-H-1:t}) \\ ={}& c_t\left(y_t^K(M_{t-H-1:t-1}), v_t^K(M_{t-H-1:t})\right) - c_t\left(y_t^K(\widetilde{M}_{t-H-1:t-1}), v_t^K(\widetilde{M}_{t-H-1:t})\right) \\ \leq{}& G_c D\|y_t^K - \widetilde{y}_t^K\| + G_c D\|v_t^K - \widetilde{v}_t^K\|, \end{aligned} \tag{72}$$

where for convenience we use the notations $y_t^K \triangleq y_t^K(\widetilde{M}_{t-H-1:t-1}), \widetilde{y}_t^K \triangleq y_t^K(\widetilde{M}_{t-H-1:t-1})$ and $v_t^K \triangleq v_t^K(M_{t-H-1:t}), \widetilde{v}_t^K \triangleq \widetilde{v}_t^K(M_{t-H-1:t})$. Besides, the last inequality holds because the norm of $\|y_t^K\|$, $\|\widetilde{y}_t^K\|$, $\|v_t^K\|$, $\|\widetilde{v}_t^K\|$ are all bounded by D, as shown in Lemma 28.

Then we try to bound $\|y_t^K - \widetilde{y}_t^K\|$ and $\|v_t^K - \widetilde{v}_t^K\|$.

$$\begin{aligned} \|y_t^K - \widetilde{y}_t^K\| &= \left\|\sum_{i=0}^{2H}\left(\Psi_{t-1,i}^{K,H}(M_{t-H-1:t-1}) - \Psi_{t-1,i}^{K,H}(\widetilde{M}_{t-H-1:t-1})\right) w_{t-1-i}\right\| \\ &= \left\|\widetilde{A}_K^k B \sum_{i=0}^{2H}\left(M_{t-k}^{[i-k]} - \widetilde{M}_{t-k}^{[i-k]}\right)\mathbf{1}_{i-k\in[H]} w_{t-1-i}\right\| \\ &\leq \kappa_B\kappa^2(1-\gamma)^k W\sum_{i=1}^{H}\|M_{t-k}^{[i]} - \widetilde{M}_{t-k}^{[i]}\| \\ &\leq \kappa_B\kappa^2 W\|M_{t-k} - \widetilde{M}_{t-k}\|, \end{aligned} \tag{73}$$

and we have

$$\begin{aligned} \|v_t^K - \widetilde{v}_t^K\| &= \left\|-K(y_t^K - \widetilde{y}_t^K) + \mathbf{1}_{k=0}\sum_{i=1}^{H}\left(M_{t-k}^{[i]} - \widetilde{M}_{t-k}^{[i]}\right)\right\| \\ &\leq (\kappa_B\kappa^3 W + 1)\|M_{t-k} - \widetilde{M}_{t-k}\| \\ &\leq 2\kappa_B\kappa^3 W\|M_{t-k} - \widetilde{M}_{t-k}\|. \end{aligned} \tag{74}$$

Combining (72), (73), and (74), we obtain

$$\begin{aligned} f_t(M_{t-H-1:t}) - f_t(\widetilde{M}_{t-H-1:t}) &\leq G_c D\|y_t^K - \widetilde{y}_t^K\| + G_c D\|v_t^K - \widetilde{v}_t^K\| \\ &\leq G_c D\kappa_B\kappa^2 W\|M_{t-k} - \widetilde{M}_{t-k}\| + G_c D 2\kappa_B\kappa^3 W\|M_{t-k} - \widetilde{M}_{t-k}\| \\ &\leq 3G_c D\kappa_B\kappa^3 W\|M_{t-k} - \widetilde{M}_{t-k}\|. \end{aligned}$$

So we have $L_f \leq 3G_c D W\kappa_B\kappa^3$.

Next, we prove the claim (ii), i.e., the boundedness of the gradient norm. Indeed, we will try to bound $\nabla_{M_{p,q}^{[r]}}\widetilde{f}_t(M)$ for every $p \in [d_u], q \in [d_x]$ and $r \in \{0, \dots, H-1\}$,

$$\left|\nabla_{M_{p,q}^{[r]}}\widetilde{f}_t(M)\right| \leq G_c\left\|\frac{\partial y_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} + G_c\left\|\frac{\partial v_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}}. \tag{75}$$

So we will bound the two terms of the right-hand side respectively.

$$\begin{aligned}\left\|\frac{\partial y_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} &\le \left\|\sum_{i=0}^{2H}\sum_{j=0}^{H}\left[\frac{\partial \widetilde{A}_K^j B M^{[i-j]}}{\partial M_{p,q}^{[r]}}\right] w_{t-1-i}\mathbf{1}_{i-j\in[H]}\right\|_{\mathrm{F}}\\ &\le \sum_{i=r+1}^{r+H+1}\left\|\frac{\partial \widetilde{A}_K^{i-r-1}BM^{[r]}}{\partial M_{p,q}^{[r]}} w_{t-1-i}\right\|_{\mathrm{F}}\\ &\le W\kappa_B\kappa^2\left\|\frac{\partial M^{[r]}}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}}\sum_{i=r+1}^{r+H+1}(1-\gamma)^{i-r-1}\\ &\le \frac{W\kappa_B\kappa^2}{\gamma}\left\|\frac{\partial M^{[r]}}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} \le \frac{W\kappa_B\kappa^2}{\gamma} \end{aligned} \tag{76}$$

$$\begin{aligned}\left\|\frac{\partial v_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} &\le \kappa\left\|\frac{\partial y_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} + \sum_{i=1}^{H}\left\|\frac{\partial M^{[i]}}{\partial M_{p,q}^{[r]}} w_{t-i}\right\|_{\mathrm{F}}\\ &\le \frac{W\kappa_B\kappa^3}{\gamma} + W\left\|\frac{\partial M^{[r]}}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} \le W\left(\frac{\kappa_B\kappa^3}{\gamma}+1\right)\end{aligned} \tag{77}$$

Combining (75), (76), and (77), we obtain

$$\left|\nabla_{M_{p,q}^{[r]}}\widetilde{f}_t(M)\right| \le G_c\frac{W\kappa_B\kappa^2}{\gamma} + G_cW\left(\frac{\kappa_B\kappa^3}{\gamma}+1\right) \le 3G_cW\kappa_B\kappa^3\gamma^{-1}.$$

Thus, $\|\nabla_M\widetilde{f}_t(M)\|_{\mathrm{F}}$ is at most $3Hd^2G_cW\kappa_B\kappa^3\gamma^{-1}$.

Finally, we prove the claim (iii), i.e., the upper bound of diameter of the feasible set. Actually, the construction of feasible set $\mathcal{M}$ ensures that $\forall i \in [H]$, $\|M\|_{\mathrm{op}}^{[i]} \le \kappa_B\kappa^3(1-\gamma)^i$. Therefore, we have

$$\begin{aligned}&\max_{M_1,M_2\in\mathcal{M}}\|M_1-M_2\|_{\mathrm{F}} \overset{\text{(Lemma 27)}}{\le} \sqrt{d}\max_{M_1,M_2\in\mathcal{M}}\|M_1-M_2\|_{\ell_1,\mathrm{op}}\\ &\le \sqrt{d}\max_{M_1,M_2\in\mathcal{M}}(\|M_1\|_{\ell_1,\mathrm{op}}+\|M_2\|_{\ell_1,\mathrm{op}}) = \sqrt{d}\max_{M_1,M_2\in\mathcal{M}}\left(\sum_{i=1}^{H}\|M_1^{[i]}\|_{\mathrm{op}}+\|M_2^{[i]}\|_{\mathrm{op}}\right)\\ &\le \sqrt{d}\max_{M_1,M_2\in\mathcal{M}}\left(2\sum_{i=1}^{H}\kappa_B\kappa^3(1-\gamma)^i\right) = 2\sqrt{d}\kappa_B\kappa^3\sum_{i=1}^{H}(1-\gamma)^i \le 2\sqrt{d}\kappa_B\kappa^3\gamma^{-1}.\end{aligned}$$

Hence, we finish the proof of all three claims in the statement. ■

In the following, we show that minimizing the static policy regret over the DAC class is sufficient to deliver a policy regret competing with the strongly linear controller class.

Lemma 30 (Lemma 5.2 of Agarwal et al. (2019))**.** *With $K, K^\star$ chosen as the (κ,γ)-strongly stable linear controllers as defined in Definition 4 and under Assumption 5, there exists a DAC policy $\pi(M_\Delta, K)$ with $M_\Delta = (M_\Delta^{[0]}, \ldots, M_\Delta^{[H-1]})$ defined by*

$$M_\Delta^{[i]} = (K - K^\star)(A - BK^\star)^i$$

such that

$$\sum_{t=1}^T c_t(x_t^K(M_\Delta), u_t^K(M_\Delta)) - \sum_{t=1}^T c_t(x_t^{K^\star}, u_t^{K^\star}) \leq T \cdot 4G_c DWH\kappa_B^2 \kappa^6 (1-\gamma)^{H-1}\gamma^{-1},$$

where $x_t^{K^\star}$ is the state attained by executing a linear controller $K^\star$ which chooses the action $u_t^{K^\star} = -K^\star x_t^{K^\star}$.

Proof [of Lemma 30] The coordinate-wise Lipschitzness of the cost functions implies that

$$c_t\left(x_t^K(M_\Delta), u_t^K(M_\Delta)\right) - c_t\left(x_t^{K^\star}, u_t^{K^\star}\right) \leq G_c D \left\| x_t^K(M_\Delta) - x_t^{K^\star} \right\| + G_c D \left\| u_t^K(M_\Delta) - u_t^{K^\star} \right\|.$$

By the linear dynamical equation (12), we have

$$x_{t+1}^{K^\star} = \sum_{i=0}^t (A - BK^\star)^i w_{t-i} = \sum_{i=0}^t \widetilde{A}_{K^\star}^i w_{t-i} \tag{78}$$

By the property of the DAC policy (Proposition 6), we have

$$x_{t+1}^K(M_\Delta) = \widetilde{A}_K^{h+1} x_{t-h}^K(M_\Delta) + \sum_{i=0}^{H+h} \Psi_{t,i}^{K,h}(M_\Delta) w_{t-i}.$$

Setting $h = t$ and combining the assumption that the starting state $x_0 = \mathbf{0}$, we achieve the following equation,

$$x_{t+1}^K(M_\Delta) = \sum_{i=0}^H \Psi_{t,i}^{K,t}(M_\Delta) w_{t-i} + \sum_{i=H+1}^t \Psi_{t,i}^{K,t}(M_\Delta) w_{t-i}.$$

Now we turn to calculate the transfer matrix $\Psi_{t,i}^{K,h}(M_\Delta)$ explicitly. Actually, for any $i \in \{0, \ldots, H\}$, $h \geq H$, i.e., $0 \leq i \leq H \leq h$, by definition we have

$$\begin{aligned}
\Psi_{t,i}^{K,h}(M_\Delta) &= \widetilde{A}_K^i \mathbf{1}_{i \leq h} + \sum_{j=0}^h \widetilde{A}_K^j B M_\Delta^{[i-j]} \mathbf{1}_{i-j \in [H]} \\
&= \widetilde{A}_K^i + \sum_{k=1}^i \widetilde{A}_K^{i-k} B M_\Delta^{[k]} \qquad (79) \\
&= \widetilde{A}_K^i + \sum_{k=1}^i \widetilde{A}_K^{i-k} B(K - K^\star) \widetilde{A}_{K^\star}^{k-1} \qquad (80)
\end{aligned}$$

$$\begin{aligned}
&= \widetilde{A}_K^i + \sum_{k=1}^{i} \widetilde{A}_K^{i-k}(\widetilde{A}_{K^\star} - \widetilde{A}_K)\widetilde{A}_{K^\star}^{k-1} \\
&= \widetilde{A}_K^i + \sum_{k=1}^{i} \widetilde{A}_K^{i-k}\widetilde{A}_{K^\star}^{k} - \widetilde{A}_K^{i-k+1}\widetilde{A}_{K^\star}^{k-1} \\
&= \widetilde{A}_K^i + \widetilde{A}_{K^\star}^i - \widetilde{A}_K^i \\
&= \widetilde{A}_{K^\star}^i,
\end{aligned}$$

where (79) holds by introducing a new index $k = i - j$ and (80) can be obtained by plugging the construction of $M_\Delta^{[i]}$ (30). So we achieve the conclusion that

$$x_{t+1}^K(M_\Delta) = \sum_{i=0}^{H} \widetilde{A}_{K^\star}^i w_{t-i} + \sum_{i=H+1}^{t} \Psi_{t,i}^{K,t}(M_\Delta) w_{t-i}. \tag{81}$$

Combining (78) and (81) yields

$$\begin{aligned}
\left\| x_{t+1}^{K^\star} - x_{t+1}^K(M_\Delta) \right\| &= \left\| \sum_{i=H+1}^{t} \left(\Psi_{t,i}^{K,t}(M_\Delta) - \widetilde{A}_{K^\star}^i \right) w_{t-i} \right\| \\
&\le W \left(\sum_{i=H+1}^{t} \|\Psi_{t,i}^{K,t}(M_\Delta)\| + \sum_{i=H+1}^{t} \|\widetilde{A}_{K^\star}^i\| \right) \\
&\le W \left(\sum_{i=H+1}^{t} \left(2\kappa^2(1-\gamma)^i + H\kappa_B^2\kappa^5(1-\gamma)^{i-1} \right) \right) \\
&\le W \left(2\kappa^2(1-\gamma)^{H+1}\gamma^{-1} + H\kappa_B^2\kappa^5(1-\gamma)^H\gamma^{-1} \right) \\
&\le \kappa^2 W(1-\gamma)^H\gamma^{-1} \left(2(1-\gamma) + H\kappa_B^2\kappa^3 \right) \\
&\le H\kappa_B^2\kappa^5 W(1-\gamma)^H\gamma^{-1}(2(1-\gamma)+1) \\
&\le 2WH\kappa_B^2\kappa^5(1-\gamma)^H\gamma^{-1},
\end{aligned}$$

where the second inequality makes use of Lemma 28. Next, we investigate the difference between the control signals,

$$\begin{aligned}
\|u_{t+1}^{K^\star} - u_{t+1}^K(M_\Delta)\| &= \left\| -K^\star x_{t+1}^{K^\star} - \left(-K x_{t+1}^K(M_\Delta) + \sum_{i=1}^{H} M_\Delta^{[i]} w_{t+1-i} \right) \right\| \\
&= \left\| -K^\star x_{t+1}^{K^\star} + K x_{t+1}^K(M_\Delta) - \sum_{i=1}^{H} (K - K^\star)\widetilde{A}_{K^\star}^{i-1} w_{t+1-i} \right\| \\
&= \left\| -K^\star \left(x_{t+1}^{K^\star} - \sum_{i=0}^{H-1} \widetilde{A}_{K^\star}^i w_{t-i} \right) + K \left(x_{t+1}^K(M_\Delta) - \sum_{i=0}^{H-1} \widetilde{A}_{K^\star}^i w_{t-i} \right) \right\| \\
&= \left\| -K^\star \sum_{i=H}^{t} \widetilde{A}_{K^\star}^i w_{t-i} + K \sum_{i=H}^{t} \Psi_{t,i}^{K,h}(M_\Delta) w_{t-i} \right\|
\end{aligned}$$

$$\leq 2WH\kappa_B^2\kappa^6(1-\gamma)^{H-1}\gamma^{-1}.$$

Using above inequalities and Lipschitz assumption as well as the boundedness result (Lemma 28), we complete the proof. ■

The remaining part of this section lists useful supporting lemmas for studying non-stochastic control in unknown systems. Lemma 31 gives a high-probability bound about the estimation accuracy in unknown systems.

Lemma 31 (Moment Recovery (Hazan et al., 2020, Lemma 21))**.** *Under Assumption 6, Algorithm 4 satisfies for all $j \in [k]$, with probability at least $1-\delta$, it holds that*

$$\|N_j - \widetilde{A}_K^j B\|_{\mathrm{F}} \leq 3\kappa_B\kappa^2 d_u W\gamma^{-1}\sqrt{\frac{2d_{\min}\log\left(2e^2k\delta^{-1}\right)}{T_0-k}}. \tag{82}$$

Proof [of Lemma 31] When the control inputs are chosen as $u_t = -Kx_t + \widetilde{u}_t$, using the transition equation of linear dynamical systems, it holds that

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t + w_t = Ax_t + B\left(-Kx_t + \widetilde{u}_t\right) + w_t = \widetilde{A}_K x_t + B\widetilde{u}_t + w_t\\
&= \widetilde{A}_K\left(Ax_{t-1} + Bu_{t-1} + w_{t-1}\right) = \widetilde{A}_K\left(\widetilde{A}_K x_{t-1} + B\widetilde{u}_{t-1} + w_{t-1}\right) + B\widetilde{u}_t + w_t\\
&= \widetilde{A}_K^2 x_{t-1} + \widetilde{A}_K\left(B\widetilde{u}_{t-1} + w_{t-1}\right) + \left(B\widetilde{u}_t + w_t\right) = \dots\\
&= \sum_{i=0}^{t}\widetilde{A}_K^{t-i}\left(B\widetilde{u}_i + w_i\right).
\end{aligned}$$

Let $N_{j,t} = x_{t+j+1}\widetilde{u}_t^\top$, we can prove that

$$\begin{aligned}
\mathbb{E}\left[N_{j,t}\right] = \mathbb{E}\left[x_{t+j+1}\widetilde{u}_t^\top\right] &= \mathbb{E}\left[\sum_{i=0}^{t+j}\widetilde{A}_K^{t+j-i}\left(B\widetilde{u}_i + w_i\right)\widetilde{u}_t^\top\right]\\
&= \sum_{i=0}^{t+j}\widetilde{A}_K^{t+j-i}\cdot\mathbb{E}\left[\left(B\widetilde{u}_i + w_i\right)\widetilde{u}_t^\top\right] = \widetilde{A}_K^j\cdot\mathbb{E}\left[\left(B\widetilde{u}_t + w_t\right)\widetilde{u}_t^\top\right]\\
&= \widetilde{A}_K^j B\cdot\mathbb{E}\left[\widetilde{u}_t\widetilde{u}_t^\top\right] + \widetilde{A}_K^j w_t\cdot\mathbb{E}\left[\widetilde{u}_t^\top\right] = \widetilde{A}_K^j B,
\end{aligned}$$

where the second last equation is due to the fact that $\widetilde{u}_i$ and $\widetilde{u}_j$ are independent when $i \neq j$, and the last step is true because $\mathbb{E}_{\widetilde{u}_t}\left[\widetilde{u}_t\widetilde{u}_t^\top\right] = I, \mathbb{E}_{\widetilde{u}_t}\left[\widetilde{u}_t\right] = \mathbf{0}$. Consequently, we can prove that $\mathbb{E}[N_j] = \frac{1}{T_0-k}\sum_{t=0}^{T_0-k-1}\mathbb{E}\left[N_{j,t}\right] = \widetilde{A}_K^j B$. Note that for $0 \leq t_1, t_2 \leq T_0-k-1$ and $t_1 \neq t_2$, N_{j,t_1} and N_{j,t_2} are not independent because they contains the same random variables η, so we cannot use Hoeffding's inequality here.

For each index $j \in [k]$, we can define a sequence of variables $\widetilde{N}_{j,t} \triangleq N_{j,t} - \widetilde{A}_K^j B$, we can prove that $\{\widetilde{N}_{j,t}\}_{t=0}^{T_0-k-1}$ is a *martingale difference sequence* w.r.t. the sequence $\{\widetilde{u}_t\}_{t=0}^{T_0-k-1}$:

$$\mathbb{E}\left[\widetilde{N}_{j,t}\,\middle|\,\widetilde{u}_{0:t-1}\right] = \mathbb{E}\left[N_{j,t}\mid\widetilde{u}_{0:t-1}\right] - \widetilde{A}_K^j B$$

$$\begin{aligned}
&= \mathbb{E}\left[\sum_{i=0}^{t+j} \widetilde{A}_K^{t+j-i}\left(B\widetilde{u}_i + w_i\right)\widetilde{u}_t^\top \,\middle|\, \widetilde{u}_{0:t-1}\right] - \widetilde{A}_K^j B \\
&= \mathbb{E}\left[\sum_{i=0}^{t-1} \widetilde{A}_K^{t+j-i}\left(B\widetilde{u}_i + w_i\right)\widetilde{u}_t^\top \,\middle|\, \widetilde{u}_{0:t-1}\right] + \mathbb{E}\left[\sum_{i=t}^{t+j} \widetilde{A}_K^{t+j-i}\left(B\widetilde{u}_i + w_i\right)\widetilde{u}_t^\top\right] - \widetilde{A}_K^j B \\
&= \mathbb{E}\left[\widetilde{A}_K^j\left(B\widetilde{u}_t + w_t\right)\widetilde{u}_t^\top\right] - \widetilde{A}_K^j B = \mathbf{0}.
\end{aligned}$$

For all $j \in [k], t = 0, \ldots, T_0 - k - 1$, the operator norm of $N_{j,t}$ can be bounded by

$$\|N_{j,t}\|_{\text{op}} \le \|x_{t+j+1}\|_{\text{op}}\|\widetilde{u}_t\|_{\text{op}} \le \|x_{t+j+1}\|_2\|\widetilde{u}_t\|_2 \le 2\kappa_B\kappa^2\sqrt{d_u}W\gamma^{-1}\cdot\sqrt{d_u} = 2\kappa_B\kappa^2 d_u W\gamma^{-1}.$$

Also, for $\widetilde{N}_{j,t}$, we can prove that

$$\begin{aligned}
\|\widetilde{N}_{j,t}\|_{\text{op}} \le \|N_{j,t}\|_{\text{op}} + \|\widetilde{A}_K^j B\|_{\text{op}} \le 2\kappa_B\kappa^2 d_u W\gamma^{-1} + \kappa_B\kappa^2(1-\gamma)^j \le 3\kappa_B\kappa^2 d_u W\gamma^{-1}, \\
\|\widetilde{N}_{j,t}\|_{\text{F}} \le \sqrt{d_{\min}}\|\widetilde{N}_{j,t}\|_{\text{op}} \le 3\sqrt{d_{\min}}\kappa_B\kappa^2 d_u W\gamma^{-1} \triangleq D_N.
\end{aligned}$$

Using Lemma 13, we have $\Pr\left[\|\sum_{t=0}^{T_0-k}\widetilde{N}_{j,t}\|_{\text{F}} \ge x\right] \le 2e^2\exp\left(\frac{-x^2}{2(T_0-k)D_N^2}\right)$. By substituting $\widetilde{N}_{j,t}$ by $N_{j,t} - \widetilde{A}_K^j B$, it holds that $\Pr\left[\|N_j - \widetilde{A}_K^j B\|_{\text{F}} \ge \frac{x}{T_0-k}\right] \le 2e^2\exp\left(\frac{-x^2}{2(T_0-k)D_N^2}\right)$. Finally, let $\varepsilon = \frac{x}{T_0-k}$, we have

$$\Pr\left[\|N_j - \widetilde{A}_K^j B\|_{\text{F}} \ge \varepsilon\right] \le 2e^2\exp\left(\frac{-(T_0-k)\varepsilon^2}{2D_N^2}\right)$$

We set $2e^2\exp\left(\frac{-(T_0-k)\varepsilon^2}{2D_N^2}\right) = \frac{\delta}{k}$ to make above concentration inequality holds for each $j \in [k]$ with probability at least $1-\delta$, which implies that

$$\varepsilon = 3\kappa_B\kappa^2 d_u W\gamma^{-1}\sqrt{\frac{2d_{\min}\log\left(2e^2k\delta^{-1}\right)}{T_0-k}}.$$

Hence, we complete the proof. ■

Lemma 32 (Preservation of Stability). *Under Assumption 6, if K is (κ,γ)-strongly stable for a linear dynamical system $S = (A, B, \{w\})$, i.e., $A - BK = QLQ^{-1}$, and $\|A - \widehat{A}\|_{\text{F}}, \|A - \widehat{A}\|_{\text{F}} \le \varepsilon_{A,B}$, then the same linear controller K is $\left(\kappa, \gamma - 2\kappa^3\varepsilon_{A,B}\right)$-strongly stable for the estimated system $\widehat{S} = (\widehat{A}, \widehat{B}, \{\widehat{w}\})$, i.e., $\widehat{A} - \widehat{B}K = Q\widehat{L}Q^{-1}$, where $\|\widehat{L}\| \le 1 - \gamma + 2\kappa^3\varepsilon_{A,B}$.*

Proof [of Lemma 32] First, we try to express the strong stability of K with respect to $(\widehat{A}, \widehat{B})$ as

$$\begin{aligned}
\widehat{A} - \widehat{B}K &= A - BK + (\widehat{A} - A) - (\widehat{B} - B)K \\
&= QLQ^{-1} + (\widehat{A} - A) - (\widehat{B} - B)K \\
&= Q\left(L + Q^{-1}\left((\widehat{A} - A) - (\widehat{B} - B)K\right)Q\right)Q^{-1} \triangleq \widehat{Q}\widehat{L}\widehat{Q}^{-1},
\end{aligned}$$

where the last equality is by defining $\widehat{L} = L + Q^{-1}((\widehat{A} - A) - (\widehat{B} - B)K)Q$. Further, the operator norm of $\widehat{L}$ can be bounded as

$$\begin{aligned}\|\widehat{L}\|_{\mathrm{op}} &= \|L + Q^{-1}\left((\widehat{A} - A) - (\widehat{B} - B)K\right)Q\|_{\mathrm{op}} \\ &\le \|L\|_{\mathrm{op}} + \|Q^{-1}\|_{\mathrm{op}}\left(\|\widehat{A} - A\|_{\mathrm{op}} + \|K\|_{\mathrm{op}}\|\widehat{B} - B\|_{\mathrm{op}}\right)\|Q\|_{\mathrm{op}} \\ &\le (1-\gamma) + \kappa \cdot (\varepsilon_{A,B} + \kappa \cdot \varepsilon_{A,B}) \cdot \kappa \le 1 - \gamma + 2\kappa^3 \varepsilon_{A,B}.\end{aligned}$$

By definition of strong stability, it holds that K is $\left(\kappa, \gamma - 2\kappa^3\varepsilon_{A,B}\right)$-strongly stable for the estimated system $\widehat{S} = (\widehat{A}, \widehat{B}, \{\widehat{w}\})$. ■

Lemma 33 below provides boundedness results in the fictitious system.

Lemma 33 (Lemma 18 of Hazan et al. (2020)). *Under Assumption 4 and Assumption 6, if it holds that $\varepsilon_{A,B} \le 10^{-3}\kappa^{-10}\gamma^2$, then for any $t \ge T_0 + 1$, we have*

$$\|x_t\|_2 \le 20\sqrt{d_u}\kappa^{11}\gamma^{-3}W, \quad \|w_t - \widehat{w}_t\|_2 \le 42\sqrt{d_u}\kappa^{12}\gamma^{-3}W\varepsilon_{A,B}, \quad \|\widehat{w}_{t-1}\|_2 \le 2\sqrt{d_u}\kappa^3\gamma^{-1}W.$$

Lemma 34. *Under Assumption 7, $\sigma_{\min}(C_k) \ge 1/\sqrt{\kappa_c}$, where C_k is defined in* (16).

Proof [of Lemma 34] Under Assumption 7, it holds that $\|(C_k C_k^\top)^{-1}\|_{\mathrm{op}} \le \kappa_c$, i.e.,

$$\sigma_{\max}((C_k C_k^\top)^{-1}) \le \kappa_c.$$

It is apparent that $\left((C_k C_k^\top)^{-1}\right)^\top = \left((C_k C_k^\top)^\top\right)^{-1} = (C_k C_k^\top)^{-1}$, i.e., $(C_k C_k^\top)^{-1}$ is a symmetric matrix. Then we have

$$\begin{aligned}\sigma_{\max}((C_k C_k^\top)^{-1}) &= \lambda_{\max}\left((C_k C_k^\top)^{-1}\left((C_k C_k^\top)^{-1}\right)^\top\right) = \lambda_{\max}\left((C_k C_k^\top)^{-1}(C_k C_k^\top)^{-1}\right) \\ &= \lambda_{\max}^2\left((C_k C_k^\top)^{-1}\right) \le \kappa_c.\end{aligned}$$

Finally we have $\sigma_{\min}(C_k) = \lambda_{\min}(C_k C_k^\top) \ge 1/\sqrt{\kappa_c}$, which finishes the proof. ■

Lemma 35 (Lemma 17 of Hazan et al. (2020)). *For any matrix pair $L, \widehat{L}$, such that $\|L\|_{\mathrm{op}}, \|\widehat{L}\|_{\mathrm{op}} \le 1 - \gamma, \gamma \in (0,1)$, we have $\sum_{t=0}^{\infty}\|L^t - \widehat{L}^t\|_{\mathrm{op}} \le 3\gamma^{-2}\|L - \widehat{L}\|_{\mathrm{op}}$.*

Lemma 36 (Perturbation Analysis (Hazan et al., 2020, Lemma 22)). *Let $x^\star$ be the solution to linear system $Ax = b$, and $\widehat{x}$ be the solution to $(A + \Delta A)x = b + \Delta b$, then if it holds that $\|\Delta A\| \le \sigma_{\min}(A)$, it is true that*

$$\|x^\star - \widehat{x}\| \le \frac{\|\Delta b\| + \|\Delta A\|\|x^\star\|}{\sigma_{\min}(A) - \|\Delta A\|_{\mathrm{op}}}.$$

References

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

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

Jason Altschuler and Kunal Talwar. Online learning over a finite action set with limited switching. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 1569–1573, 2018.

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

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Bandits with feedback graphs and switching costs. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 10397–10407, 2019.

Dheeraj Baby and Yu-Xiang Wang. Online forecasting of total-variation-bounded sequences. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 11071–11081, 2019.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pages 359–409, 2021.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in proper online learning with strongly convex losses and beyond. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1805–1845, 2022.

Dheeraj Baby, Saurabh Garg, Tzu-Ching Yen, Sivaraman Balakrishnan, Zachary Chase Lipton, and Yu-Xiang Wang. Online label shift: Optimal dynamic regret meets practical algorithms. *ArXiv preprint*, arXiv:2305.19570, 2023.

Yong Bai, Yu-Jie Zhang, Peng Zhao, Masashi Sugiyama, and Zhi-Hua Zhou. Adapting to online label shift with provable guarantees. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 29960–29974, 2022.

Omar Besbes, Yonatan Gur, and Assaf J. Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Avrim Blum and Adam Kalai. Universal portfolios with and without transaction costs. *Machine Learning*, 35(3):193–205, 1999.

Olivier Bousquet and Manfred K. Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, 2002.

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

Asaf B. Cassel, Alon Cohen, and Tomer Koren. Efficient online linear control with stochastic convex costs and unknown dynamics. In *Proceedings of 35th Conference on Learning Theory (COLT)*, volume 178, pages 3589–3604, 2022a.

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

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

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 989–997, 2012.

Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1160–1168, 2013.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Lin Chen, Qian Yu, Hannah Lawrence, and Amin Karbasi. Minimax regret of switching-constrained online convex optimization: No phase transition. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 3477–3486, 2020.

Niangjun Chen, Gautam Goel, and Adam Wierman. Smoothed online convex optimization in high dimensions via online balanced descent. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 1574–1594, 2018.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1029–1038, 2018.

Ashok Cutkosky. Parameter-free, dynamic, and strongly-adaptive online learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 2250–2259, 2020.

Amit Daniely and Yishay Mansour. Competitive ratio vs regret minimization: Achieving the best of both worlds. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory (ALT)*, pages 333–368, 2019.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1405–1411, 2015.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

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

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 459–467, 2014.

Claude-Nicolas Fiechter. PAC adaptive control of linear systems. In *Proceedings of the 10th Annual Conference on Computational Learning Theory (COLT)*, pages 72–80, 1997.

Dylan J. Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3211–3221, 2020.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Sascha Geulen, Berthold Vöcking, and Melanie Winkler. Regret minimization for online buffering problems using the weighted majority algorithm. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 132–143, 2010.

Gautam Goel and Babak Hassibi. Regret-optimal control in dynamic environments. *ArXiv preprint*, arXiv:2010.10473, 2020.

Gautam Goel and Babak Hassibi. Online estimation and control with optimal pathlength regret. In *Proceedings of the 4th Learning for Dynamics and Control Conference (L4DC)*, pages 404–414, 2022a.

Gautam Goel and Babak Hassibi. Competitive control. *IEEE Transactions on Automatic Control*, in press, 2022b.

Gautam Goel and Adam Wierman. An online algorithm for smoothed regression and LQR control. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2504–2513, 2019.

Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 1873–1883, 2019.

Eyal Gofer. Higher-order regret bounds with switching costs. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, pages 210–243, 2014.

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

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

Lei Guo and Lennart Ljung. Performance analysis of general tracking algorithms. *IEEE Transactions on Automatic Control*, 40(8):1388–1402, 1995.

András György and Gergely Neu. Near-optimal rates for limited-delay universal lossy source coding. *IEEE Transactions on Information Theory*, 60(5):2823–2834, 2014.

András György and Csaba Szepesvári. Shifting regret, mirror descent, and matrices. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2943–2951, 2016.

Thomas P. Hayes. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2005.

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

Elad Hazan and C. Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 393–400, 2009.

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

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32 (2):151–178, 1998.

Mark Herbster and Manfred K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 398–406, 2015.

Rudolf Emil Kalman. Contributions to the theory of optimal control. *Boletín de la Sociedad Matemática Mexicana*, 5(2):102–119, 1960.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of the 28th Annual Conference Computational Learning Theory (COLT)*, pages 1286–1304, 2015.

Haipeng Luo, Mengxiao Zhang, Peng Zhao, and Zhi-Hua Zhou. Corralling a larger band of bandits: A case study on switching regret for linear bandits. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, pages 3635–3684, 2022.

Neri Merhav, Erik Ordentlich, Gadiel Seroussi, and Marcelo J. Weinberger. On sequential strategies for loss functions with memory. *IEEE Transactions on Information Theory*, 48 (7):1947–1958, 2002.

Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, pages 7195–7201, 2016.

Arkadij S. Nemirovsky and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Uri Sherman and Tomer Koren. Lazy OCO: Online convex optimization on a switching budget. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pages 3972–3988, 2021.

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

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

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

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2645–2653, 2011.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation.* The MIT Press, 2012.

Guanghui Wang, Yuanyu Wan, Tianbao Yang, and Lijun Zhang. Online convex optimization with continuous switching constraint. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 28636–28647, 2021.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3972–3980, 2016.

Yu-Hu Yan, Peng Zhao, and Zhi-Hua Zhou. Fast rates in time-varying strongly monotone games. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 39138–39164, 2023.

Lijun Zhang. Online learning in changing environments. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5178–5182, 2020. Early Career.

Lijun Zhang, Tianbao Yang, Jinfeng Yi, Rong Jin, and Zhi-Hua Zhou. Improved dynamic regret for non-degenerate functions. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 732–741, 2017.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1330–1340, 2018a.

Lijun Zhang, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5877–5886, 2018b.

Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26772–26808, 2022a.

Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. A simple online algorithm for competing with dynamic comparators. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 390–399, 2020.

Zhiyu Zhang, Ashok Cutkosky, and Ioannis Ch. Paschalidis. Adversarial tracking control via strongly adaptive online learning with memory. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 8458–8492, 2022b.

Zhiyu Zhang, Ashok Cutkosky, and Yannis Paschalidis. Optimal comparator adaptive online learning with switching cost. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022c.

Peng Zhao. *Online Ensemble Theories and Methods for Robust Online Learning.* PhD thesis, Nanjing University, Nanjing, China, 2021. Advisor: Zhi-Hua Zhou.

Peng Zhao and Lijun Zhang. Improved analysis for dynamic regret of strongly convex and smooth functions. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control (L4DC)*, pages 48–59, 2021.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 12510–12520, 2020.

Peng Zhao, Guanghui Wang, Lijun Zhang, and Zhi-Hua Zhou. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1–45, 2021a.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *ArXiv preprint*, arXiv:2112.14368, 2021b.

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

Peng Zhao, Yu-Xiang Wang, and Zhi-Hua Zhou. Non-stationary online learning with memory and non-stochastic control. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2101–2133, 2022b.

Kai Zheng, Haipeng Luo, Ilias Diakonikolas, and Liwei Wang. Equipping experts/bandits with long-term memory. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 5927–5937, 2019.

Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms.* Chapman & Hall/CRC Press, 2012.

Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8): nwac123, 07 2022.

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