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

Peng Zhao* **Yu-Xiang Wang†** **Zhi-Hua Zhou***
* National Key Laboratory for Novel Software Technology, Nanjing University, China
† Department of Computer Science, UC Santa Barbara, USA
{zhaop, zhouzh}@lamda.nju.edu.cn; yuxiangw@cs.ucsb.edu

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. 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 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.

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, optimization, and time series analysis, 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 the *policy regret* [Dekel et al., 2012], defined as

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

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

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

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 environments [Sugiyama and Kawanabe, 2012, Gama et al., 2014, Zhao et al., 2020a, 2021b]. 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 the 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

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

The upper bound of $\text{D-Reg}_T(\mathbf{v}_{1:T})$ 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 level of non-stationarity. Our approach builds upon recent advance of non-stationary online learning [Daniely et al., 2015, Zhang et al., 2018a, Zhao et al., 2020b] to hedge the uncertainty via meta-base aggregation, 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 structure. Intuitively, for dynamic online algorithms, it is necessary to keep some probability of aggressive movement to catch up with potential changes of the non-stationary environments, which results in tensions between dynamic regret and switching cost. We elegantly address the difficulty by proposing 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 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.

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 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. We anticipate that our techniques for OCO with memory will have broader applications in online decision-making problems.

We summarize the main contributions 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* $\mathcal{O}(\sqrt{T(1+P_T)})$ dynamic policy regret. 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. We finally conclude the paper in Section 6. 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 Non-stochastic Control. 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., 2021], 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, 2021] and robust control [Goel and Hassibi, 2020].

Dynamic Regret. Zinkevich [2003] pioneers the dynamic regret of OCO problems 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 $\mathcal{O}(\sqrt{T(1+P_T)})$ bound. Recent works achieve problem-dependent guarantee by exploiting smoothness [Zhao et al., 2020b] and improved rate by exploiting exp-concavity [Baby and Wang, 2021]. Dynamic regret of bandit convex optimization is studies in [Zhao et al., 2021a]. 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, and we refer the reader to [Zhang et al., 2018a] for more discussions.

More Discussions. Online non-stochastic control in non-stationary environments is also recently studied via the measure of *adaptive regret* [Hazan and Seshadhri, 2009] — the regret compared to the best policy on any interval in 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., 2021], 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 studies 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. To the best of our knowledge, dynamic regret and adaptive regret reflect different perspectives of environments and their relationship is still unclear even for the standard OCO setting [Zhang, 2020, Section 5. Open Problems]. **(2)** Techniques: optimizing either dynamic regret or adaptive regret requires the meta-base 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] follow 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 bound [Zhang et al., 2021] 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 proposed controller.

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.

1: **for** $t = m+1, \dots, T$ **do**
2: the player chooses a decision $\mathbf{w}_t \in \mathcal{W}$;
3: the adversary reveals the loss $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$ that applies to last $m+1$ decisions;
4: the player suffers a loss of $f_t(\mathbf{w}_{t-m}, \dots, \mathbf{w}_t)$;
5: **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 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.

Assumptions. Next, we introduce several standard assumptions used in the analysis [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., $|f_t(\mathbf{x}_0, \dots, \mathbf{x}_m) - f_t(\mathbf{y}_0, \dots, \mathbf{y}_m)| \le L \sum_{i=0}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|_2$.

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 $t \in [T]$, $\|\nabla \widetilde{f}_t(\mathbf{w})\|_2 \le G$.

Assumption 3 (bounded domain)**.** The domain $\mathcal{W}$ is convex, closed, and satisfies $\|\mathbf{w} - \mathbf{w}'\|_2 \le D$ for $\forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$. 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,

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

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 $\sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \le \mathcal{O}(\eta T)$, 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:

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

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 functions $\widetilde{f}_1, \ldots, \widetilde{f}_T$ ensures*

$$\text{D-Reg}_T(\mathbf{v}_{1:T}) \leq \mathcal{O}\Big(\eta T + \frac{1+P_T}{\eta} + P_T\Big) \qquad (4)$$

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 parameter-free online algorithms to achieve optimal bound, especially due to the switching cost arising in the OCO with memory setting. In Section 4.3, we will present our solution by introducing several novel algorithmic ingredients.

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., 2020b], inspired by which we employ the meta-base framework to design a two-layer approach for optimizing 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 framework. The framework admits a two-layer structure and is essentially an online ensemble method. 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 (3) we have

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

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

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

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 $\mathcal{H}=\{\eta_i=\mathcal{O}(2^i\cdot T^{-\frac{1}{2}}), i\in[N]\}$ such that $\eta_1=\mathcal{O}(T^{-\frac{1}{2}})$ 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 online methods, we propose the following new meta-base regret decomposition to avoid directly controlling switching cost of final predictions or controlling switching cost of every base-learner:

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

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

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

Intuitively, the base-learner's switching cost is now taken into account when evaluating the base 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 the switching cost $\sum_{t=2}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t-1}\|_1$.

We present settings of step size pool, base-algorithm, and meta-algorithm to realize 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,

Algorithm 1 Scream

Input: step size pool $\mathcal{H} = \{\eta_1, \ldots, \eta_N\}$, learning rate of meta-algorithm ε

1: 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 $p_{m,i} \propto 1/(i^2+i)$, $\forall i \in [N]$
2: **for** $t = m+1$ **to** T **do**
3: Receive $\mathbf{w}_{t,i}$ from base-learner $\mathcal{B}_i$ for $i \in [N]$
4: Submit the decision $\mathbf{w}_t = \sum_{i=1}^N p_{t,i}\mathbf{w}_{t,i}$
5: Suffer a loss of $f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t)$
6: Observe the online function $f_t : \mathcal{W}^{m+1} \mapsto \mathbb{R}$ that applies to last $m+1$ decisions
7: Construct linearized loss $g_t(\mathbf{w}) = \langle \nabla \widetilde{f}_t(\mathbf{w}_t), \mathbf{w} \rangle$
8: Construct switching-cost-regularized loss $\boldsymbol{\ell}_t \in \mathbb{R}^N$ with $\ell_{t,i} = g_t(\mathbf{w}_{t,i}) + \lambda \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$
9: Update the weight $\boldsymbol{p}_{t+1} \in \Delta_N$ according to $p_{t+1,i} \propto p_{t,i} \exp(-\varepsilon \ell_{t,i})$
10: Base-learner $\mathcal{B}_i$ updates the local decision by $\mathbf{w}_{t+1,i} = \Pi_{\mathcal{W}}[\mathbf{w}_{t,i} - \eta_i \nabla \widetilde{f}_t(\mathbf{w}_t)]$, $\forall i \in [N]$
11: **end for**

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 $\sum_{t=1}^T \big(\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i}\big) + \lambda D \sum_{t=2}^T \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1$, 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) [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 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. Our algorithm enjoys an *optimal* dynamic policy regret, striking a good balance between regret and switching cost.

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*

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

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

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, the attained dynamic policy regret 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. The dynamic policy regret in Theorem 2 exhibits a quadratic dependence on the memory length m, while the best static policy regret for OCO with memory only exhibits a linear dependence [Anava et al., 2015] (see discussions in Appendix B.5 for details). Recall the upper bound decomposition of dynamic policy regret in (3), though we cannot reduce the memory dependence in front of switching of comparators, it remains unclear whether it is possible to achieve a linear memory dependence for dynamic regret of unary functions and switching cost of final decisions. We leave this as future work.

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., 2021].

5.1 Problem Statement

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

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{9}$$

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,

$$\mathrm{Reg}_T = \sum_{t=1}^{T} c_t(x_t, u_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} c_t(x_t^\pi, u_t^\pi). \tag{10}$$

The comparator could be chosen with complete foreknowledge of the disturbance and loss functions. A variety of control algorithms have been proposed to optimize the measure under different settings. However, we argue that competing with a fixed controller may be not appropriate, especially because the unknown disturbance 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 facilitate the online controller with capability of competing with *time-varying* controllers to adapt to those changes. To this end, we generalize the standard measure (10) to the *dynamic policy regret* to benchmark the algorithm with a sequence of *time-varying* controllers $\pi_1, \ldots, \pi_T \in \Pi$,

$$\text{D-Reg}_T(\pi_{1:T}) = \sum_{t=1}^{T} c_t(x_t, u_t) - \sum_{t=1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t}), \tag{11}$$

The measure clearly subsumes standard policy regret (10) when choosing 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 (cf. 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 policy class called *Disturbance-Action Controller* (DAC) class, which parametrizes the executed action as a linear function of 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.

Definition 1 (Disturbance-Action Controller, DAC)**.** A disturbance-action controller $\pi(K, M)$ with a 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 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 can be implemented because the disturbance can be perfectly recovered by $w_t = x_{t+1} - Ax_t - Bu_t$ as system dynamics A and B are supposed to be known.

The following proposition due to Agarwal et al. [2019] presents an important property of DAC policies.

Proposition 3. *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 $x_t^K(M_{0:t-1}) = \sum_{i=0}^{H+t-1} \Psi_{t-1,i}^{K,t-1}(M_{0:t-1}) w_{t-1-i}$ and $u_t^K(M_{0:t}) = -Kx_t^K(M_{0:t-1}) + \sum_{i=1}^{H} M_t^{[i]} w_{t-i}$, where $\widetilde{A}_K = A - BK$ and $\Psi_{t,i}^{K,h}(M_{t-h:t}) = \widetilde{A}_K^i \mathbf{1}_{i \le h} + \sum_{j=0}^{h} \widetilde{A}_K^j B M_{t-j}^{[i-j]} \mathbf{1}_{1 \le i-j \le H}$.*

Evidently, both state x_t and control signal u_t are linear functions of DAC parameters $M_0, \ldots, M_t$, 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 the OCO with memory setting. To this end, Agarwal et al. [2019] further propose a truncated method that truncates the state with a fixed memory length H and defines the 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 $y_{t+1}^K = \sum_{i=0}^{2H} \Psi_{t,i}^{K,H}(M_{t-H:t}) w_{t-i}$ and $v_{t+1}^K = -Ky_{t+1}^K(M_{t-H:t}) + \sum_{i=1}^{H} M_{t+1}^{[i]} w_{t+1-i}$.

It can be proved that the error introduced by the truncation (the gap between f_t and c_t) can be precisely controlled. Therefore, it remains to feed 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.

Algorithm 2 Scream.Control

Input: step size pool $\mathcal{H} = \{\eta_1, \dots, \eta_N\}$; learning rate of meta-algorithm ε; memory length H; linear controller K; feasible set $\mathcal{M}$

Initialization: $u_1, \dots, u_H$, any feasible output control signals for the first H rounds;
Initialization: base decisions of the H-th round $M_{H,1}, M_{H,2}, \dots M_{H,N} \in \mathcal{M}$; 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 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 truncated loss $f_t : \mathcal{M}^{H+2} \mapsto \mathbb{R}$ by Definition 2 and $g_t(M) = \langle \nabla \widetilde{f}_t(M_t), M \rangle$
  Compute the switching-cost-regularized loss $\boldsymbol{\ell}_t \in \mathbb{R}^N$ with $\ell_{t,i} = \lambda \|M_{t,i} - M_{t-1,i}\|_{\mathrm{F}} + g_t(M_{t,i})$
  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 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

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 $u_t = \pi(K, M_t)$ to parametrize the space and define the unary loss of the truncated loss, i.e., $\widetilde{f}_t : \mathcal{M} \mapsto \mathbb{R}$ with $\widetilde{f}_t(M) = f_t(M, \dots, M)$, defined in Definition 2.
(2) meta-base 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 $M_{t,1}, \dots, M_{t,N}$ from all base-learners to produce the final parameter M_t.

Algorithm 2 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.

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

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

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

(1) K is (κ, γ)-strongly stable, whose precise definition is in Definition 3 of Appendix A.2;
(2) $M \in \mathcal{M}$ such that $\mathcal{M} = \{M = (M^{[1]}, \dots, M^{[H]}) \mid \|M^{[i]}\|_{\mathrm{op}} \le \kappa_B \kappa^3 (1-\gamma)^i\}$.

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

$$\mathcal{H} = \left\{ \eta_i \;\middle|\; \eta_i = 2^{i-1} \cdot \sqrt{\frac{D_f^2}{(\lambda G_f + G_f^2) T}}, i \in [N] \right\}, \quad (12)$$

where $N = \lceil \frac{1}{2} \log_2(1+T) \rceil + 1 = \mathcal{O}(\log T)$ 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 20 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

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

where the comparators can be any feasible policies in $\Pi = \{\pi(K, M) \mid M \in \mathcal{M}\}$ 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.

6 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 aggregation 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.

Acknowledgment

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

References

N. Agarwal, B. Bullins, E. Hazan, S. M. Kakade, and K. Singh. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 111–119, 2019.

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

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

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

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

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

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

L. Chen, Q. Yu, H. Lawrence, and A. 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.

N. Chen, G. Goel, and A. 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.

A. Cohen, A. Hasidim, T. Koren, N. Lazic, Y. Mansour, and K. Talwar. Online linear quadratic control. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1029–1038, 2018.

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

A. Daniely and Y. 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.

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

O. Dekel, A. Tewari, and R. Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

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

Y. Freund and R. 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.

J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014.

S. Geulen, B. Vöcking, and M. 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.

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

G. Goel and B. Hassibi. Competitive control. *ArXiv preprint*, arXiv:2107.13657, 2021.

G. Goel and A. 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.

G. Goel, Y. Lin, H. Sun, and A. 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.

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

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

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

E. 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.

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

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

R. E. Kalman. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mexicana*, 5(2):102–119, 1960.

N. Merhav, E. Ordentlich, G. Seroussi, and M. J. Weinberger. On sequential strategies for loss functions with memory. *IEEE Transactions on Information Theory*, 48(7):1947–1958, 2002.

A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. 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.

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

G. Shi, Y. Lin, S. Chung, Y. Yue, and A. Wierman. Online optimization with memory and competitive control. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

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

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

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

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

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

L. Zhang, T. Yang, J. Yi, R. Jin, and Z.-H. Zhou. Improved dynamic regret for non-degenerate functions. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 732–741, 2017.

L. Zhang, S. Lu, and Z.-H. Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1330–1340, 2018a.

L. Zhang, T. Yang, R. Jin, and Z.-H. Zhou. Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5877–5886, 2018b.

Y.-J. Zhang, P. Zhao, and Z.-H. 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.

Z. Zhang, A. Cutkosky, and I. C. Paschalidis. Adversarial tracking control via strongly adaptive online learning with memory. *ArXiv preprint*, arXiv:2102.01623, 2021.

P. Zhao and L. 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.

P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 746–755, 2020a.

P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 12510–12520, 2020b.

P. Zhao, G. Wang, L. Zhang, and Z.-H. Zhou. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1–45, 2021a.

P. Zhao, X. Wang, S. Xie, L. Guo, and Z.-H. Zhou. Distribution-free one-pass learning. *IEEE Transaction on Knowledge and Data Engineering*, 33:951–963, 2021b.

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

A 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., 2020b].

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

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

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,

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t)]. \tag{13}$$

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 5. *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 (13) enjoys the following dynamic regret,*

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

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

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

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 (41),

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

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,

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

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

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

Combining all above inequalities, we have

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

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 3. 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:

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

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

A.3 Technical Lemmas

The following lemma plays an important role in analyzing algorithms based on the mirror descent.

Lemma 6 (Lemma 3.2 of Chen and Teboulle [1993])**.** *Let $\mathcal{X}$ be a convex set in a Banach space $\mathcal{B}$. Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a closed proper convex function on $\mathcal{X}$. Given a convex regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$, we denote its induced Bregman divergence by $\mathcal{D}_\psi(\cdot, \cdot)$. Then, 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 7. *If the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is λ-strongly convex with respect to a norm $\|\cdot\|$, then we have the following lower bound for the induced Bregman divergence: $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) \geq \frac{\lambda}{2}\|\mathbf{x} - \mathbf{y}\|^2$.*

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. ■

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. We finally discuss the memory dependence in Appendix B.5.

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

$$\sum_{t=m}^{T} f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t) - \sum_{t=m}^{T} \widetilde{f}_t(\mathbf{w}_t) \leq m^2 L \sum_{t=m}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2, \tag{14}$$

and the dynamic policy regret can be thus upper bounded by

$$\begin{aligned} \text{D-Reg}_T(\mathbf{v}_1, \ldots, \mathbf{v}_T) &= \sum_{t=1}^{T} f_t(\mathbf{w}_{t-m}, \ldots, \mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{v}_{t-m}, \ldots, \mathbf{v}_t) \\ &\overset{(14)}{\leq} \underbrace{\sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t)}_{\texttt{dynamic regret over unary loss}} + \underbrace{\lambda \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2}_{\texttt{switching cost of decisions}} + \underbrace{\lambda \sum_{t=1}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2}_{\texttt{switching cost of comparators}}, \end{aligned} \tag{15}$$

where we define $\lambda := m^2 L$ 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 5, we know that

$$\sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) \leq \frac{\eta}{2} G^2 T + \frac{1}{2\eta}(D^2 + 2DP_T), \tag{16}$$

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 (15) 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:

$$\sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 = \sum_{t=1}^{T} \|\Pi_{\mathcal{W}}[\mathbf{w}_{t-1} - \eta \mathbf{g}_{t-1}] - \mathbf{w}_{t-1}\|_2 \leq \eta \sum_{t=1}^{T} \|\mathbf{g}_{t-1}\|_2 \leq \eta G T. \tag{17}$$

Combining above two inequalities (17) and (16) yields

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

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) of the main paper.

Lemma 8. *The switching cost of meta-base outputs can be upper bounded in the following way:*

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

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

$$\begin{aligned}
\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 &= \left\|\sum_{i=1}^{N} p_{t,i}\mathbf{w}_{t,i} - \sum_{i=1}^{N} p_{t-1,i}\mathbf{w}_{t-1,i}\right\|_2 \\
&\le \left\|\sum_{i=1}^{N} p_{t,i}\mathbf{w}_{t,i} - \sum_{i=1}^{N} p_{t,i}\mathbf{w}_{t-1,i}\right\|_2 + \left\|\sum_{i=1}^{N} p_{t,i}\mathbf{w}_{t-1,i} - \sum_{i=1}^{N} p_{t-1,i}\mathbf{w}_{t-1,i}\right\|_2 \\
&\le \sum_{i=1}^{N} p_{t,i}\|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 + D\sum_{i=1}^{N}|p_{t,i} - p_{t-1,i}| \\
&= \sum_{i=1}^{N} p_{t,i}\|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 + D\|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1,
\end{aligned} \tag{18}$$

where the first inequality holds due to the triangle inequality and the second inequality is true owing 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 $\{h_t\}_{t=1,\dots,T}$ 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:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \eta\langle\nabla h_t(\mathbf{w}_t), \mathbf{w}\rangle + \mathcal{D}_\psi(\mathbf{w}, \mathbf{w}_t), \tag{19}$$

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 9. *Online Mirror Descent* (19) *satisfies that*

$$\sum_{t=1}^{T} h_t(\mathbf{w}_t) - \sum_{t=1}^{T} h_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\| \leq \frac{1}{\eta}\left(R^2 + \gamma P_T\right) + \eta(\lambda G + G^2)T, \tag{20}$$

provided that $\mathcal{D}_\psi(\mathbf{x}, \mathbf{z}) - \mathcal{D}_\psi(\mathbf{y}, \mathbf{z}) \leq \gamma \|\mathbf{x} - \mathbf{y}\|$ *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 9 holds against *any* comparator sequence in the domain, in particular, we can set comparators as the best fixed decision in hindsight and thus obtain static regret with switching cost, $\sum_{t=1}^{T} h_t(\mathbf{w}_t) - \sum_{t=1}^{T} h_t(\mathbf{w}^*) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\| \leq R^2/\eta + \eta(\lambda G + G^2)T$, 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 9 exhibits a general analysis for the dynamic regret and switching cost of OMD algorithm. By flexibly choosing the regularizer ψ and comparator sequence $\mathbf{v}_1, \ldots, \mathbf{v}_T$, we can obtain 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 9, we first analyze the switching cost of the online mirror descent, as demonstrated in the following stability lemma.

Lemma 10. *For Online Mirror Descent* (19), *the instantaneous switching cost is at most*

$$\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq \eta \|\nabla h_t(\mathbf{w}_t)\|_*. \tag{21}$$

Proof From the update procedure of OMD (19) and Lemma 6, we know that

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

which implies

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

Since the regularizer ψ is chosen as a 1-strongly convex function with respect to the norm $\|\cdot\|$, by Lemma 7 we have

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

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

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

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

Based on the above stability lemma, we can now prove Theorem 9 regarding dynamic regret with switching cost for OMD.

Proof [of Theorem 9] Notice that the dynamic regret can be decomposed in the following way:

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

From Lemma 10 and Hölder's inequality, we have

$$\texttt{term (a)} \leq \sum_{t=1}^{T} \|\nabla h_t(\mathbf{w}_t)\|_* \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq \eta \sum_{t=1}^{T} \|\nabla h_t(\mathbf{w}_t)\|_*^2. \tag{22}$$

Next, we investigate the term (b):

$$\begin{aligned}\texttt{term (b)} &\leq \frac{1}{\eta}\sum_{t=1}^{T}\left(\mathcal{D}_\psi(\mathbf{v}_t,\mathbf{w}_t)-\mathcal{D}_\psi(\mathbf{v}_t,\mathbf{w}_{t+1})-\mathcal{D}_\psi(\mathbf{w}_{t+1},\mathbf{w}_t)\right)\\ &\leq \frac{1}{\eta}\sum_{t=2}^{T}\left(\mathcal{D}_\psi(\mathbf{v}_t,\mathbf{w}_t)-\mathcal{D}_\psi(\mathbf{v}_{t-1},\mathbf{w}_t)\right)+\mathcal{D}_\psi(\mathbf{v}_1,\mathbf{w}_1)\\ &\leq \frac{\gamma}{\eta}\sum_{t=2}^{T}\|\mathbf{v}_t-\mathbf{v}_{t-1}\|+\frac{1}{\eta}R^2, \end{aligned} \tag{23}$$

where the first inequality holds due to Lemma 6, 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 $\mathcal{D}_\psi(\mathbf{x},\mathbf{z})-\mathcal{D}_\psi(\mathbf{y},\mathbf{z})\leq\gamma\|\mathbf{x}-\mathbf{y}\|$ holds for any $\mathbf{x},\mathbf{y},\mathbf{z}\in\mathcal{W}$.

Furthermore, the switching cost can be bounded by Lemma 10,

$$\sum_{t=2}^{T}\|\mathbf{w}_t-\mathbf{w}_{t-1}\|\leq\eta\sum_{t=2}^{T}\|\nabla h_{t-1}(\mathbf{w}_{t-1})\|_*. \tag{24}$$

Combining (22), (23), and (24), 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 9, 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 11. *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*

$$\lambda\sum_{t=2}^{T}\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2+\sum_{t=1}^{T}\widetilde{f}_t(\mathbf{w}_t)-\sum_{t=1}^{T}\widetilde{f}_t(\mathbf{v}_t)\leq(G^2+\lambda G)\eta T+\frac{1}{2\eta}(D^2+2DP_T), \tag{25}$$

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 12. *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*

$$\lambda\sum_{t=2}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t-1}\|_1+\sum_{t=1}^{T}\langle\boldsymbol{p}_t,\boldsymbol{\ell}_t\rangle-\sum_{t=1}^{T}\ell_{t,i}\leq\frac{\ln(1/p_{1,i})}{\varepsilon}+\varepsilon(\lambda G+G^2)T. \tag{26}$$

Proof From the proof of Theorem 9, 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 (15), 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. Since the meta-algorithm actually performs Hedge over the switching-cost-regularized loss $\boldsymbol{\ell}_t \in \mathbb{R}^N$, Corollary 12 implies that for any $i \in [N]$,

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

It can be verified that $G_{\text{meta}} = \max_{t\in[T]} \|\boldsymbol{\ell}_t\|_\infty \leq (\lambda + G)D$. Moreover, 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 $\varepsilon = \varepsilon^* = \sqrt{\frac{2}{(2\lambda+G)(\lambda+G)D^2T}}$, we can obtain the following upper bound for the meta-regret,

$$
\sum_{t=1}^T \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \sum_{t=1}^T \ell_{t,i} + \lambda' \sum_{t=2}^T \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 \leq D\sqrt{2(2\lambda + G)(\lambda + G)T}(1 + \ln(i+1)). \tag{27}
$$

Note that the dependence of learning rate tuning on T can be removed by either a time-varying tuning or doubling trick.

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 9 implies that the base-regret satisfies

$$\sum_{t=1}^{T} g_t(\mathbf{w}_{t,i}) - \sum_{t=1}^{T} g_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2 \leq (G^2 + \lambda G)\eta_i T + \frac{1}{2\eta_i}(D^2 + 2DP_T), \tag{28}$$

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 $\eta_{i^*} \leq \eta_* \leq \eta_{i^*+1} = 2\eta_{i^*}$. Therefore, we have

$$i^* \leq \left\lceil \frac{1}{2} \log_2 \left(1 + \frac{2P_T}{D}\right) \right\rceil + 1. \tag{29}$$

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 (27) and base-regret (28).

$$\begin{aligned}
&\sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \\
&\leq \underbrace{\sum_{t=1}^{T} \left(\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i^*}\right) + \lambda D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1}_{\texttt{meta-regret}} + \underbrace{\sum_{t=1}^{T} \left(g_t(\mathbf{w}_{t,i^*}) - g_t(\mathbf{v}_t)\right) + \lambda \sum_{t=2}^{T} \|\mathbf{w}_{t,i^*} - \mathbf{w}_{t-1,i}\|_2}_{\texttt{base-regret}} \\
&\leq D\sqrt{2(2\lambda + G)(\lambda + G)T}(1 + \ln(i^* + 1)) + (G^2 + \lambda G)\eta_{i^*} T + \frac{1}{2\eta_{i^*}}(D^2 + 2DP_T) \\
&\leq D\sqrt{2(2\lambda + G)(\lambda + G)T}(1 + \ln(i^* + 1)) + (G^2 + \lambda G)\eta_* T + \frac{1}{\eta_*}(D^2 + 2DP_T) \\
&\leq \underbrace{2D(\lambda + G)\sqrt{T}\left(1 + \ln\left(\left\lceil \log_2(1 + 2P_T/D)\right\rceil + 2\right)\right)}_{\leq \mathcal{O}(\sqrt{T}(1 + \log\log P_T))} + \underbrace{2\sqrt{2}\sqrt{(G^2 + \lambda G)(D^2 + 2DP_T)T}}_{\leq \mathcal{O}(\sqrt{T(1+P_T)})} \\
&\leq \mathcal{O}(\sqrt{T(1 + P_T)}).
\end{aligned}$$

Combining the upper bound of the dynamic policy regret exhibited in (3), we can achieve that

$$\begin{aligned}
\text{D-Reg}_T(\mathbf{v}_{1:T}) &\leq \lambda \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 + \lambda \sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2 + \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) \\
&\leq \mathcal{O}(\sqrt{T(1 + P_T)}) + \mathcal{O}(P_T) \\
&\leq \mathcal{O}(\sqrt{T(1 + P_T) + P_T^2}) && (\sqrt{a} + \sqrt{b} \leq \sqrt{2(a + b)}) \\
&= \mathcal{O}(\sqrt{T + (T + P_T)P_T}) \\
&\leq \mathcal{O}(\sqrt{T(1 + P_T)}),
\end{aligned}$$

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 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 13. *Under Assumptions 1–3, running OGD over the unary loss achieves* $\sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \min_{\mathbf{v}\in\mathcal{W}} \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}) \leq (G^2 + m^2LG)\eta T + \frac{2D^2}{\eta}$. *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 $\mathcal{O}(m^{\frac{3}{4}}\sqrt{T})$ 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 7 (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 7, 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 7 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 7. We adapt their results to our notations to ease the understanding.

Theorem 14 (Theorem 3.1 of Anava et al. [2015])**.** *Under Assumptions 2, 3, and the assumption that the online functions are* $\bar{L}$*-Lipschitz (Assumption 7), running OGD over the unary loss achieves*

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \min_{\mathbf{v}\in\mathcal{W}} \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}) \leq 2\eta G^2 T + \frac{2D^2}{\eta} + 2\bar{L}m^{\frac{3}{2}}\eta GT. \tag{30}$$

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 exhibits a linear dependence on the memory length.

Finally, we discuss the memory dependence issue in the dynamic policy regret. In Theorem 2, we show an $\mathcal{O}(m\sqrt{T(1+P_T)} + m^2\sqrt{T} + m^2P_T)$ dynamic policy regret. Therefore, the overall memory dependence is quadratic. In particular, the result implies an $\mathcal{O}(m^2\sqrt{T})$ static policy regret when comparing with a fixed comparator (now $P_T = 0$). Notably, this is worse than the $\mathcal{O}(m\sqrt{T})$ bound obtained specifically by the (one-layer) method for static regret minimization [Anava et al., 2015], due to the two-layer structure of our approach. Recall the upper bound decomposition of dynamic policy regret:

$$\text{D-Reg}_T(\mathbf{v}_{1:T}) \leq \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widetilde{f}_t(\mathbf{v}_t) + m^2L\sum_{t=2}^{T}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 + m^2L\sum_{t=2}^{T}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2. \tag{31}$$

The last term $m^2L\sum_{t=2}^{T}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2 = \mathcal{O}(m^2P_T)$ is irrelevant to the algorithm, thus we cannot optimize the memory dependence further. However, for the remaining terms (dynamic regret over unary loss and switching cost of decisions), it remains unclear whether it is possible to further reduce and achieve a linear memory dependence. On the other hand, it would be also interesting to see whether it is necessary to go through the expansion (31) for achieving dynamic policy regret.

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

In this section, we present omitted details for Section 5 online non-stochastic control, including the proofs of Proposition 3 and Theorem 4.

C.1 Proof of Proposition 3

We will prove the following statement that gives the state recurrence for any $h \leq t$, which is essentially a strengthened result of Proposition 3.

Proposition 15. *Suppose one chooses the DAC controller $\pi(M_t, K)$ at iteration t, the reaching state is*

$$x_{t+1} = \widetilde{A}_K^{h+1} x_{t-h} + \sum_{i=0}^{H+h} \Psi_{t,i}^{K,h}(M_{t-h:t}) w_{t-i}, \tag{32}$$

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

$$\Psi_{t,i}^{K,h}(M_{t-h:t}) = \widetilde{A}_K^i \mathbf{1}_{i\le h} + \sum_{j=0}^{h} \widetilde{A}_K^j B M_{t-j}^{[i-j]} \mathbf{1}_{1\le i-j\le H}. \tag{33}$$

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

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

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

Exchanging the summation index yields,

$$\begin{aligned} \sum_{j=0}^{h} \sum_{i=1}^{H} \widetilde{A}_K^j B M_{t-j}^{[i]} w_{t-j-i} &= \sum_{i=1}^{H} \sum_{k=i}^{i+h} \widetilde{A}_K^{k-i} B M_{t-k+i}^{[i]} w_{t-k} && (34) \\ &= \sum_{k=1}^{H+h} \sum_{i=k-h}^{k} \widetilde{A}_K^{k-i} B M_{t-k+i}^{[i]} w_{t-k} \mathbf{1}_{1\le i\le H} && (35) \\ &= \sum_{k=1}^{H+h} \sum_{l=0}^{h} \widetilde{A}_K^{h-l} B M_{t+l-h}^{[l+k-h]} w_{t-k} \mathbf{1}_{1\le l+(k-h)\le H} && (36) \\ &= \sum_{k=1}^{H+h} \sum_{m=0}^{h} \widetilde{A}_K^{m} B M_{t-m}^{[k-m]} w_{t-k} \mathbf{1}_{1\le k-m\le H} && (37) \\ &= \sum_{i=1}^{H+h} \sum_{j=0}^{h} \widetilde{A}_K^{j} B M_{t-j}^{[i-j]} w_{t-i} \mathbf{1}_{1\le i-j\le H}, && (38) \end{aligned}$$

where (34) holds by defining a third variable $k = j + i$, and (35) is obtained by exchanging the summation index i and k and the new range of i is from inequality $i \le k \le i + h$. Moreover, (36) is obtained by another change of variable $l = i - k + h$, (37) is obtained by replacing l by $h - m$, and (38) is true by setting $i = k, j = m$. Therefore, we can 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 4

To prove the dynamic policy regret of online non-stochastic control (Theorem 4), 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 4 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 16 (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 $\|M_t^{[i]}\|_{\mathrm{op}} \leq \tau(1-\gamma)^i$, the approximation error between original loss and truncated loss is at most*

$$\left| \sum_{t=1}^{T} c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - \sum_{t=1}^{T} f_t(M_{t-1-H:t}) \right| \leq 2TG_c D^2 \kappa^3 (1-\gamma)^{H+1}, \tag{39}$$

where

$$D := \frac{W\kappa^3(1+H\kappa_B\tau)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{W\tau}{\gamma}. \tag{40}$$

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

$$\begin{aligned}
& c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - f_t(M_{t-H-1:t}) \\
&= c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - c_t(y_t^K(M_{t-H-1:t-1}), v_t^K(M_{t-H-1:t})) \\
&\leq G_c D \left(\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-H-1:t-1})\| + \|u_t^K(M_{0:t}) - v_t^K(M_{t-H-1:t})\| \right) \\
&\leq G_c D(\kappa^2(1-\gamma)^{H+1} D + \kappa^3(1-\gamma)^{H+1} D) \\
&\leq 2G_c D^2 \kappa^3 (1-\gamma)^{H+1},
\end{aligned}$$

where the last two inequalities use the Lipschitzness and the boundedness presented in Lemma 19. 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 OGD 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,

$$M_{t+1} = \Pi_{\mathcal{M}}[M_t - \eta \nabla_M \widetilde{f}_t(M_t)] \tag{41}$$

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 17. *Suppose the function $\widetilde{f} : \mathcal{M} \mapsto \mathbb{R}$ is convex; the gradient norm $\|\nabla_M \widetilde{f}_t(M)\|_{\mathrm{F}} \leq G_f$ holds for any $M \in \mathcal{M}$ and $t \in [T]$; 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 (41) satisfies that*

$$\lambda \sum_{t=2}^{T} \|M_{t-1} - M_t\|_{\mathrm{F}} + \sum_{t=1}^{T} \widetilde{f}_t(M_t) - \sum_{t=1}^{T} \widetilde{f}_t(M_t^*) \leq \frac{\eta}{2}(G_f^2 + 2\lambda G_f)T + \frac{1}{2\eta}(D_f^2 + 2D_f P_T), \tag{42}$$

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,

$$\begin{aligned} & \sum_{t=1}^{T} \left(\|M_t - M_t^*\|_{\mathrm{F}}^2 - \|M_{t+1} - M_t^*\|_{\mathrm{F}}^2 \right) \\ \leq & \sum_{t=1}^{T} \|M_t - M_t^*\|_{\mathrm{F}}^2 - \sum_{t=2}^{T} \|M_t - M_{t-1}^*\|_{\mathrm{F}}^2 \\ \leq & \|M_1 - M_1^*\|_{\mathrm{F}}^2 + \sum_{t=2}^{T} \left(\|M_t - M_t^*\|_{\mathrm{F}}^2 - \|M_t - M_{t-1}^*\|_{\mathrm{F}}^2 \right) \\ = & \|M_1 - M_1^*\|_{\mathrm{F}}^2 + \sum_{t=2}^{T} \langle M_{t-1}^* - M_t^*, 2M_t - M_{t-1}^* - M_t^* \rangle \\ \leq & D_f^2 + 2D_f \sum_{t=2}^{T} \|M_{t-1}^* - M_t^*\|_{\mathrm{F}}. \end{aligned}$$

Hence, combining all above inequalities, we have

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

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

$$\begin{aligned} \sum_{t=2}^{T} \|M_t - M_{t-1}\|_{\mathrm{F}} =& \ \|\Pi_{\mathcal{M}}[M_{t-1} - \eta G_{t-1}] - M_{t-1}\|_{\mathrm{F}}^2 \\ &\leq \|M_{t-1} - \eta G_{t-1} - M_{t-1}\|_{\mathrm{F}} \leq \eta G_f T, \end{aligned}$$

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

C.2.3 Proof of Theorem 4

Proof We begin with the following dynamic policy regret decomposition,

$$\sum_{t=1}^{T} c_t(x_t, u_t) - \sum_{t=1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t})$$

$$
\begin{aligned}
&= \sum_{t=1}^{T} c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - \sum_{t=1}^{T} c_t(x_t^K(M^*_{0:t-1}), u_t^K(M^*_{0:t})) \\
&= \underbrace{\sum_{t=1}^{T} c_t(x_t^K(M_{0:t-1}), u_t^K(M_{0:t})) - \sum_{t=1}^{T} f_t(M_{t-1-H:t})}_{:=A_T} \\
&\quad + \underbrace{\sum_{t=1}^{T} f_t(M_{t-1-H:t}) - \sum_{t=1}^{T} f_t(M^*_{t-1-H:t})}_{:=B_T} + \underbrace{\sum_{t=1}^{T} f_t(M^*_{t-1-H:t}) - \sum_{t=1}^{T} c_t(x_t^K(M^*_{0:t-1}), u_t^K(M^*_{0:t}))}_{:=C_T}. \qquad (43)
\end{aligned}
$$

Notice that both A_T and C_T essentially represent the approximation error introduced by the truncated loss, so we can apply Theorem 16 and obtain that

$$
A_T + C_T \le 4TG_cD^2\kappa^3(1-\gamma)^{H+1}. \qquad (44)
$$

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

$$
\begin{aligned}
B_T &= \sum_{t=1}^{T} f_t(M_{t-1-H:t}) - \sum_{t=1}^{T} f_t(M^*_{t-1-H:t}) \\
&\le \sum_{t=1}^{T} \widetilde{f}_t(M_t) - \sum_{t=1}^{T} \widetilde{f}_t(M_t^*) + \lambda \sum_{t=2}^{T} \|M_{t-1} - M_t\|_{\mathrm{F}} + \lambda \sum_{t=2}^{T} \|M^*_{t-1} - M_t^*\|_{\mathrm{F}} \\
&\le \sum_{t=1}^{T} \langle \nabla_M \widetilde{f}_t(M_t), M_t - M_t^* \rangle + \lambda \sum_{t=2}^{T} \|M_{t-1} - M_t\|_{\mathrm{F}} + \lambda \sum_{t=2}^{T} \|M^*_{t-1} - M_t^*\|_{\mathrm{F}} \\
&= \sum_{t=1}^{T} g_t(M_t) - \sum_{t=1}^{T} g_t(M_t^*) + \lambda \sum_{t=2}^{T} \|M_{t-1} - M_t\|_{\mathrm{F}} + \lambda \sum_{t=2}^{T} \|M^*_{t-1} - M_t^*\|_{\mathrm{F}}, \qquad (45)
\end{aligned}
$$

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 $\{g_t\}_{t=1,\dots,T}$ 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 17 ensures the base-regret satisfies that

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

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 8 of Section B.2, we can show that the meta-regret satisfies that

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

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 make use the static regret with switching cost of online mirror descent for the prediction with expert advice setting (c.f. Corollary 12 in Appendix B.3) and obtain that

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

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 $\eta_{i^*} \le \eta_* \le 2\eta_{i^*}$. As a result, we have

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

Combining this result with the regret decomposition (43) and the upper bounds (44), (45), we have

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

The specific values of D, L_f, G_f, D_f can be found in Lemma 20. By setting $H = \mathcal{O}(\log T)$, we can ensure the final dynamic policy regret is at most $\widetilde{\mathcal{O}}(\sqrt{T(1+P_T)})$ and hence complete the proof. ■

C.3 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 work of Agarwal et al. [2019], and we adapt them to our notations and provide the proofs to achieve self-containedness. Specifically,

- Lemma 18 establishes the norm relations between the ℓ_1, op norm and Frobenius norm used in the $\mathcal{M}$-space.
- Lemma 19 checks the boundedness of several variables of interest.
- Lemma 20 shows several properties of the truncated functions $\{f_t\}$ and the feasible set $\mathcal{M}$.
- Lemma 21 provides an upper bound for the norm of transfer matrix.
- Lemma 22 connects the DAC class and the strongly linear controller class.

Lemma 18 (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*

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

We then have the following inequalities on their relations:

$$\|M\|_{\ell_1,\mathrm{op}} \le \sqrt{H}\|M\|_{\mathrm{F}}, \text{ and } \|M\|_{\mathrm{F}} \le \sqrt{d}\|M\|_{\ell_1,\mathrm{op}}, \tag{46}$$

where $d = \min\{d_u, d_x\}$.

Proof [of Lemma 18] Recall the matrix norm relations, 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

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

We thus complete the proof. ■

Lemma 19. *Suppose K and $K^\star$ are two (κ,γ)-strongly stable linear controllers (cf. Definition 3). Define*

$$D := \frac{W(\kappa^3 + H\kappa_B\kappa^3\tau)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} + \frac{W\tau}{\gamma}. \tag{47}$$

Suppose there exists a $\tau > 0$ such that for every $i \in \{0,\ldots,H-1\}$ and every $t \in [T]$, $\|M_t^{[i]}\|_{\mathrm{F}} \le \tau(1-\gamma)^i$. 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* $\|v_t^K(M_{t-H-1:t})\| \le D$.
- $\|x_t^K(M_{0:t-1}) - y_t^K(M_{t-1-H:t-1})\| \le \kappa^2(1-\gamma)^{H+1}D$.
- $\|u_t^K(M_{0:t}) - v_t^K(M_{t-1-H:t})\| \le \kappa^3(1-\gamma)^{H+1}D$.

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 3, 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 22.

Proof [of Lemma 19] We first study the state.

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

$$\leq \kappa^2(1-\gamma)^{H+1}\|x_{t-H}^K(M_{0:t-H-1})\| + W(\kappa^2 + H\kappa_B\kappa^2\tau)/\gamma$$
$$\leq \frac{W(\kappa^2 + H\kappa_B\kappa^2\tau)}{\gamma(1-\kappa^2(1-\gamma)^{H+1})} \leq D, \tag{48}$$

where inequality (48) 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).

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

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

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

■

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 $\{f_t\}$ and the feasible set $\mathcal{M}$ such that we can further apply the results of OCO with memory.

Lemma 20. *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 (40), and we restate it below*

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

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

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

Besides,

$$L_f \le 3\sqrt{H}G_cDW\kappa_B\kappa^3.$$

(ii) The gradient norm of surrogate loss $\widetilde{f}_t : \mathcal{M} \mapsto \mathbb{R}$ is bounded by G_f, namely, $\|\nabla_M \widetilde{f}_t(M)\|_{\mathrm{F}} \le G_f$ holds for any $M \in \mathcal{M}$ and any $t \in [T]$. Besides,

$$G_f \le 3Hd^2G_cW\kappa_B\kappa^3\gamma^{-1}.$$

(iii) The diameter of the feasible set is at most D_f, namely, $\|M-M'\|_{\mathrm{F}} \le D_f$ holds for any $M, M' \in \mathcal{M}$. Besides,

$$D_f \le 2\sqrt{d}\kappa_B\kappa^3\gamma^{-1}.$$

Proof [of Lemma 20] We first prove the claim (i), i.e., the L_f-coordinate-wise Lipschitz continuity. For simplicity, we will make use of the following definitions in the following arguments.

$$\begin{aligned}
M_{t-H-1:t} &:= \{M_{t-H-1}\dots M_{t-k}\dots M_t\}\\
M_{t-H-1:t-1} &:= \{M_{t-H-1}\dots M_{t-k}\dots M_{t-1}\}\\
\widetilde{M}_{t-H-1:t} &:= \{M_{t-H-1}\dots \widetilde{M}_{t-k}\dots M_t\}\\
\widetilde{M}_{t-H-1:t-1} &:= \{M_{t-H-1}\dots \widetilde{M}_{t-k}\dots M_{t-1}\}
\end{aligned}$$

By representing f_t using c_t, we have

$$\begin{aligned}
&f_t(M_{t-H-1:t}) - f_t(\widetilde{M}_{t-H-1:t})\\
&= c_t\left(y_t^K(M_{t-H-1:t-1}), v_t^K(M_{t-H-1:t})\right) - c_t\left(y_t^K(\widetilde{M}_{t-H-1:t-1}), v_t^K(\widetilde{M}_{t-H-1:t})\right)\\
&\le G_cD\|y_t^K - \widetilde{y}_t^K\| + G_cD\|v_t^K - \widetilde{v}_t^K\|,
\end{aligned} \tag{49}$$

where for convenience we use the notations $y_t^K := y_t^K(\widetilde{M}_{t-H-1:t-1}), \widetilde{y}_t^K := y_t^K(\widetilde{M}_{t-H-1:t-1})$ and $v_t^K := v_t^K(M_{t-H-1:t}, \widetilde{v}_t^K := \widetilde{v}_t^K(M_{t-H-1:t}$. 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 19.

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

$$\begin{aligned}
\|y_t^K - \widetilde{y}_t^K\| &= \left\|\sum_{i=0}^{2H}\left(\Psi_{t-1,i}^{K,H}(M_{t-H-1:t-1}) - \Psi_{t-1,i}^{K,H}(\widetilde{M}_{t-H-1:t-1})\right)w_{t-1-i}\right\|\\
&= \left\|\widetilde{A}_K^k B\sum_{i=0}^{2H}\left(M_{t-k}^{[i-k]} - \widetilde{M}_{t-k}^{[i-k]}\right)\mathbf{1}_{i-k\in[H]}w_{t-1-i}\right\|\\
&\le \kappa_B\kappa^2(1-\gamma)^kW\sum_{i=1}^{H}\|M_{t-k}^{[i]} - \widetilde{M}_{t-k}^{[i]}\|\\
&\le \kappa_B\kappa^2W\|M_{t-k} - \widetilde{M}_{t-k}\|,
\end{aligned} \tag{50}$$

and we have

$$\begin{aligned}
\|v_t^K - \widetilde{v}_t^K\| &= \left\|-K(y_t^K - \widetilde{y}_t^K) + \mathbf{1}_{k=0}\sum_{i=1}^{H}\left(M_{t-k}^{[i]} - \widetilde{M}_{t-k}^{[i]}\right)\right\|\\
&\le (\kappa_B\kappa^3W+1)\|M_{t-k} - \widetilde{M}_{t-k}\|\\
&\le 2\kappa_B\kappa^3W\|M_{t-k} - \widetilde{M}_{t-k}\|.
\end{aligned} \tag{51}$$

Combining (49), (50), and (51), we obtain

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

So we have $L_f \le 3G_c DW\kappa_B\kappa^3$.

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 $r \in \{0, \dots, H-1\}$,

$$\left|\nabla_{M_{p,q}^{[r]}}\widetilde{f}_t(M)\right| \le G_c \left\|\frac{\partial y_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} + G_c \left\|\frac{\partial v_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}}. \tag{52}$$

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

$$\begin{aligned} \left\|\frac{\partial y_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} &\le \left\|\sum_{i=0}^{2H}\sum_{j=0}^{H}\left[\frac{\partial \widetilde{A}_K^j B M^{[i-j]}}{\partial M_{p,q}^{[r]}}\right] w_{t-1-i}\mathbf{1}_{i-j\in[H]}\right\|_{\mathrm{F}} \\ &\le \sum_{i=r+1}^{r+H+1}\left\|\frac{\partial \widetilde{A}_K^{i-r-1} B M^{[r]}}{\partial M_{p,q}^{[r]}} w_{t-1-i}\right\|_{\mathrm{F}} \\ &\le W\kappa_B\kappa^2\left\|\frac{\partial M^{[r]}}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} \sum_{i=r+1}^{r+H+1}(1-\gamma)^{i-r-1} \\ &\le \frac{W\kappa_B\kappa^2}{\gamma}\left\|\frac{\partial M^{[r]}}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} \\ &\le \frac{W\kappa_B\kappa^2}{\gamma} \end{aligned} \tag{53}$$

$$\begin{aligned} \left\|\frac{\partial v_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} &\le \kappa\left\|\frac{\partial y_t^K(M)}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} + \sum_{i=1}^{H}\left\|\frac{\partial M^{[i]}}{\partial M_{p,q}^{[r]}} w_{t-i}\right\|_{\mathrm{F}} \\ &\le \frac{W\kappa_B\kappa^3}{\gamma} + W\left\|\frac{\partial M^{[r]}}{\partial M_{p,q}^{[r]}}\right\|_{\mathrm{F}} \\ &\le W\left(\frac{\kappa_B\kappa^3}{\gamma} + 1\right) \end{aligned} \tag{54}$$

Combining (52), (53), and (54), we obtain

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

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

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 $\forall i, 0 \le i \le H-1$, $\|M\|_{\mathrm{op}}^{[i]} \le \kappa_B\kappa^3(1-\gamma)^i$. Therefore, we have

$$\begin{aligned} \max_{M_1,M_2\in\mathcal{M}}\|M_1 - M_2\|_{\mathrm{F}} &\overset{(46)}{\le} \sqrt{d}\max_{M_1,M_2\in\mathcal{M}}\|M_1 - M_2\|_{\ell_1,\mathrm{op}} \\ &\le \sqrt{d}\max_{M_1,M_2\in\mathcal{M}}(\|M_1\|_{\ell_1,\mathrm{op}} + \|M_2\|_{\ell_1,\mathrm{op}}) \end{aligned}$$

$$
\begin{aligned}
&= \sqrt{d} \max_{M_1, M_2 \in \mathcal{M}} \left(\sum_{i=0}^{H-1} \|M_1^{[i]}\|_{\mathrm{op}} + \|M_2^{[i]}\|_{\mathrm{op}} \right) \\
&\le \sqrt{d} \max_{M_1, M_2 \in \mathcal{M}} \left(2 \sum_{i=0}^{H-1} \kappa_B \kappa^3 (1-\gamma)^i \right) \\
&= 2\sqrt{d} \kappa_B \kappa^3 \sum_{i=0}^{H-1} (1-\gamma)^i \\
&\le 2\sqrt{d} \kappa_B \kappa^3 \gamma^{-1}.
\end{aligned}
$$

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

The following lemma provides an upper bound for the norm of transfer matrix.

Lemma 21. *Suppose K is (κ, γ)-strongly stable as defined in Definition 3. Suppose there exists a $\tau > 0$ such that for every $i \in \{0, \dots, H-1\}$ and every $t \in [T]$, $\|M_t^{[i]}\|_{\mathrm{F}} \le \tau (1-\gamma)^i$. Then, we have*

$$
\|\Psi_{t,i}^{K,h}\| \le \kappa^2 (1-\gamma)^i \mathbf{1}_{i \le h} + H \kappa_B \kappa^2 \tau (1-\gamma)^{i-1}. \tag{55}
$$

Proof [of Lemma 21] We first expand $\Psi_{t,i}^{K,h}$ by its definition (cf. Proposition 3 for its formal definition):

$$
\begin{aligned}
\|\Psi_{t,i}^{K,h}\| &= \left\| \widetilde{A}_K^i \mathbf{1}_{i \le h} + \sum_{j=0}^{h} \widetilde{A}_K^j B M_{t-j}^{[i-j]} \mathbf{1}_{1 \le i-j \le H} \right\| \\
&\le \|\widetilde{A}_K^i\| \mathbf{1}_{i \le h} + \sum_{j=1}^{H} \|\widetilde{A}_K^j B M_{t-j}^{[i-j]}\| \qquad (56) \\
&\le \kappa^2 (1-\gamma)^i + \sum_{j=1}^{H} \kappa^2 (1-\gamma)^j \kappa_B \tau (1-\gamma)^{i-j-1} \\
&\le \kappa^2 (1-\gamma)^i + \kappa^2 \kappa_B \tau \sum_{j=1}^{H} (1-\gamma)^{i-1} \\
&= \kappa^2 (1-\gamma)^i + H \kappa^2 \kappa_B \tau (1-\gamma)^{i-1},
\end{aligned}
$$

where inequality (56) has to be emphasized here that no matter what the index i is, once i is fixed, to satisfy the condition $1 \le i - j \le H$, there is at most H different values which j can take. And that is why we can take j in range $[H]$ as an upper bound. ■

In the following lemma, 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 [Agarwal et al., 2019, Lemma 5.2].

Lemma 22. *With $K, K^\star$ chosen as the (κ, γ)-strongly stable linear controllers as defined in Definition 3 and under Assumption 5, there exists a DAC policy $\pi(M_\Delta, K)$ with $M_\Delta = (M_\Delta^{[1]}, \dots, M_\Delta^{[H]})$ defined by*

$$
M_\Delta^{[i]} = (K - K^\star)(A - BK^\star)^i \tag{57}
$$

such that

$$
\sum_{t=1}^{T} c_t(x_t^K(M_\Delta), u_t^K(M_\Delta)) - \sum_{t=1}^{T} c_t(x_t^{K^\star}, u_t^{K^\star}) \le T \cdot 4 G_c D W H \kappa_B^2 \kappa^6 (1-\gamma)^{H-1} \gamma^{-1}, \tag{58}
$$

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 22] The coordinate-wise Lipschitzness of the cost functions implies that

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

By the linear dynamical equation (9), we have

$$x_{t+1}^{K^\star} = \sum_{i=0}^{t}(A - BK^\star)^i w_{t-i} = \sum_{i=0}^{t} \widetilde{A}_{K^\star}^i w_{t-i} \tag{59}$$

By the property of the DAC policy (Proposition 3), 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

$$\begin{aligned}
\Psi_{t,i}^{K,h}(M_\Delta) &= \widetilde{A}_K^i \mathbf{1}_{i \leq h} + \sum_{j=0}^{h} \widetilde{A}_K^j B M_\Delta^{[i-j]} \mathbf{1}_{i-j \in [H]} \\
&= \widetilde{A}_K^i + \sum_{k=1}^{i} \widetilde{A}_K^{i-k} B M_\Delta^{[k]} && (60) \\
&= \widetilde{A}_K^i + \sum_{k=1}^{i} \widetilde{A}_K^{i-k} B (K - K^\star) \widetilde{A}_{K^\star}^{k-1} && (61) \\
&= \widetilde{A}_K^i + \sum_{k=1}^{i} \widetilde{A}_K^{i-k} (\widetilde{A}_{K^\star} - \widetilde{A}_K) \widetilde{A}_{K^\star}^{k-1} \\
&= \widetilde{A}_K^i + \sum_{k=1}^{i} \widetilde{A}_K^{i-k} \widetilde{A}_{K^\star}^{k} - \widetilde{A}_K^{i-k+1} \widetilde{A}_{K^\star}^{k-1} \\
&= \widetilde{A}_K^i + \widetilde{A}_{K^\star}^i - \widetilde{A}_K^i \\
&= \widetilde{A}_{K^\star}^i,
\end{aligned}$$

where (60) holds by introducing a new index $k = i - j$ and (61) can be obtained by plugging the construction of $M_\Delta^{[i]}$ (57). So we achieve the conclusion that

$$x_{t+1}^{K}(M_\Delta) = \sum_{i=0}^{H} \widetilde{A}_{K^\star}^i w_{t-i} + \sum_{i=H+1}^{t} \Psi_{t,i}^{K,t}(M_\Delta) w_{t-i}. \tag{62}$$

Combining (59) and (62) yields

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

$$
\begin{aligned}
&\leq \kappa^2 W(1-\gamma)^H\gamma^{-1}\left(2(1-\gamma)+H\kappa_B^2\kappa^3\right)\\
&\leq H\kappa_B^2\kappa^5 W(1-\gamma)^H\gamma^{-1}(2(1-\gamma)+1)\\
&\leq 2WH\kappa_B^2\kappa^5(1-\gamma)^H\gamma^{-1},
\end{aligned}
$$

where the second inequality makes use of Lemma 19. Next, we investigate the difference between the control signals,

$$
\begin{aligned}
\|u_{t+1}^{K^\star} - u_{t+1}^{K}(M_\Delta)\| &= \left\| -K^\star x_{t+1}^{K^\star} - \left(-K x_{t+1}^{K}(M_\Delta) + \sum_{i=1}^{H} M_\Delta^{[i]} w_{t+1-i} \right) \right\| \\
&= \left\| -K^\star x_{t+1}^{K^\star} + K x_{t+1}^{K}(M_\Delta) - \sum_{i=1}^{H} (K-K^\star)\widetilde{A}_{K^\star}^{i-1} w_{t+1-i} \right\| \\
&= \left\| -K^\star \left(x_{t+1}^{K^\star} - \sum_{i=0}^{H-1} \widetilde{A}_{K^\star}^{i} w_{t-i} \right) + K \left(x_{t+1}^{K}(M_\Delta) - \sum_{i=0}^{H-1} \widetilde{A}_{K^\star}^{i} w_{t-i} \right) \right\| \\
&= \left\| -K^\star \sum_{i=H}^{t} \widetilde{A}_{K^\star}^{i} w_{t-i} + K \sum_{i=H}^{t} \Psi_{t,i}^{K,h}(M_\Delta) w_{t-i} \right\| \\
&\leq 2WH\kappa_B^2\kappa^6(1-\gamma)^{H-1}\gamma^{-1}.
\end{aligned}
$$

Using above inequalities and Lipschitz assumption as well as the boundedness result (Lemma 19), we complete the proof. ■