Rate-Optimal Online Convex Optimization in Adaptive Linear Control

Asaf Cassel[*] and **Alon Cohen**[†] and **Tomer Koren**[‡]

Abstract

We consider the problem of controlling an unknown linear dynamical system under adversarially changing convex costs and full feedback of both the state and cost function. We present the first computationally-efficient algorithm that attains an optimal $\sqrt{T}$-regret rate compared to the best stabilizing linear controller in hindsight, while avoiding stringent assumptions on the costs such as strong convexity. Our approach is based on a careful design of non-convex lower confidence bounds for the online costs, and uses a novel technique for computationally-efficient regret minimization of these bounds that leverages their particular non-convex structure.

1. Introduction

We study a general setting of online adaptive linear control, where a learner attempts to stabilize an initially unknown discrete-time linear dynamical system while minimizing its cumulative cost with respect to an arbitrary sequence of convex loss functions. The system dynamics evolve according to

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t,$$

where $x_t \in \mathbb{R}^{d_x}$, $u_t \in \mathbb{R}^{d_u}$ are the (fully observable) system's state and learner's control at time step t, and $w_t \in \mathbb{R}^{d_x}$ is the system noise added at step t which is a zero-mean i.i.d. Gaussian random variable. The matrices $A_\star \in \mathbb{R}^{d_x \times d_x}$ and $B_\star \in \mathbb{R}^{d_x \times d_u}$ are the system parameters, which are assumed to be unknown ahead of time and need to be learned adaptively. The goal is to minimize regret with respect to a sequence of convex loss functions $c_1, \ldots, c_T$ over T time steps, namely, the difference between the learner's cumulative control cost $\sum_{t=1}^{T} c_t(x_t, u_t)$ and the best cumulative cost achieved by a control policy from a given set of benchmark policies.

This general framework encapsulates numerous variations of learning in linear control that have been studied extensively in the literature. When the system parameters are known ahead of time and the costs are fixed and known (convex) quadratics, this amounts to the classical "planning" formulation of linear-quadratic (LQ) stochastic control; see (Bertsekas, 1995). The special case where the costs are fixed and known quadratics but the system parameters are unknown has been addressed much more recently (Abbasi-Yadkori and Szepesvári, 2011; Cohen et al., 2019; Mania et al., 2019). This was recently extended to allow for a fixed and known convex cost (Plevrakis and Hazan, 2020) and later for stochastic i.i.d. costs (Cassel et al., 2022). On the other hand, the case where the system parameters are known but the quadratic costs are allowed to vary arbitrarily between rounds was first addressed in (Cohen et al., 2018), and has been later extended in various ways to allow for arbitrarily-varying convex costs (Agarwal et al., 2019a,b; Simchowitz et al., 2020; Cassel and Koren, 2020). In all of these special cases, we now know of efficient algorithms with rate-optimal $\sqrt{T}$ regret guarantees.

For the online adaptive linear control problem in its full generality, however, no regret-optimal algorithms are presently known. The state-of-the-art is due to (Simchowitz et al., 2020) that achieved $T^{2/3}$-regret using a simple explore-then-exploit strategy: in the exploration phase, their algorithm estimates the dynamics parameters by exciting the system with noise; then, in the exploitation phase it runs an online procedure for known dynamics using the estimated transitions. This simple strategy has also been shown to achieve the optimal $\sqrt{T}$-regret when the online costs are additionally *strongly convex*, demonstrating that the stringent strong convexity assumption allows one to circumvent the challenge of balancing exploration and exploitation in online adaptive linear control.

*. Blavatnik School of Computer Science, Tel Aviv University; acassel@mail.tau.ac.il.
†. School of Electrical Engineering, Tel Aviv University, and Google Research; alonco@tauex.tau.ac.il.
‡. Blavatnik School of Computer Science, Tel Aviv University, and Google Research; tkoren@tauex.tau.ac.il.

In this paper, we resolve this gap and give the first rate-optimal algorithm for the general online adaptive linear control problem, accommodating arbitrarily changing general convex (and Lipschitz) costs and unknown system parameters. Our algorithm is computationally efficient and attains a $\sqrt{T}$ regret guarantee with polynomial dependence on the natural parameters of the problem.

Techniques. Our approach builds upon a combination of recent techniques in online linear control. First, we rely on the Disturbance Action Policies (DAPs) of Agarwal et al. (2019a): our algorithm generates DAPs that choose the control at each time step as a linear transformation of past noise terms; the DAPs themselves are maintained by online convex optimization algorithms that generate slowly-changing decisions for guaranteeing the stability of the system throughout the learning process. Moreover, since the dynamics are unknown, our algorithm estimates the noise terms on-the-fly, and uses these estimates in place of the true noise vectors (this is akin to a technique in (Plevrakis and Hazan, 2020)).

Second, following the recent developments of Cassel et al. (2022) for the case of stochastic costs, we perform regret minimization with respect to optimistic lower confidence bounds of the online costs. However, these confidence bounds turn out to be inherently nonconvex. To maintain computational efficiency, we adapt a trick of Dani et al. (2008) (in the context of stochastic linear bandits) for relaxing the nonconvex objectives so as to assume the form of a minimum of a small number of convex objectives; then, we hedge over multiple copies of online gradient descent as "experts" in a meta-algorithm, where each copy minimizes regret with respect to one of these convex objectives.

Even so, the decisions of the hedging meta-algorithm are random and can thus change abruptly, interfering with the slowly-moving nature of the DAPs that is crucial for the stability of the system. We address this issue by using a lazy version of Follow the Perturbed Leader in place of the meta-algorithm (due to (Altschuler and Talwar, 2018)) that employs only a small number of switches between experts. Overall, this results in a computationally efficient scheme that maintains the $\sqrt{T}$ regret rate of the individual gradient-based experts.

Related work. The problem of adaptive linear-quadratic control has a long history (e.g., Bertsekas, 1995). Recent years have seen a renewed interest in this problem through the modern view of regret minimization—building on classic asymptotic results to obtain finite-time guarantees (Abbasi-Yadkori and Szepesvári, 2011; Abeille and Lazaric, 2018; Arora et al., 2018; Dean et al., 2018; Faradonbeh et al., 2017; Ibrahimi et al., 2012; Ouyang et al., 2017). More recently Cohen et al. (2019); Mania et al. (2019) provided polynomial-time algorithms obtaining an optimal $\sqrt{T}$ regret rate. The optimality of the $\sqrt{T}$ rate was proved concurrently by Cassel et al. (2020); Simchowitz and Foster (2020).

More recently, Plevrakis and Hazan (2020) gave an efficient algorithm with $\sqrt{T}$ regret for learning the dynamics under a fixed known convex cost. Plevrakis and Hazan (2020) also observed that the problem of learning both dynamics and stochastic convex costs under bandit feedback is reducible to an instance of stochastic bandit convex optimization for which complex, yet polynomial-time, generic algorithms exist (Agarwal et al., 2011). Cassel et al. (2022) later study the problem of learning the dynamics and stochastic convex costs under full-information feedback. Unlike the approach of (Plevrakis and Hazan, 2020), their algorithm is based on an "optimism in the face of uncertainty" principle and is thus conceptually simpler and more efficient to implement.

Our approach relies on the standard assumption that the controller is provided with some initial stabilizing policy. First proposed in Dean et al. (2018), such an assumption yields regret that is polynomial in the problem dimensions, and was later shown to be necessary by Chen and Hazan (2021).

Past work has also considered adaptive LQG control, namely linear-quadratic control under partial observability of the state (for example, Simchowitz et al., 2020). However, it turned out that in the stochastic setting, learning the optimal partial-observation linear controller is in a sense easier than learning the full-observation controller. It is in fact possible to obtain $\mathrm{poly}(\log T)$ regret for adaptive LQG (Lale et al., 2020). This result is facilitated by simplifying assumptions on both the noise distribution as well as the benchmark policy, assumptions which we do not make in this work.

Most works on regret minimization in adaptive control are model-based; meaning, the algorithm attempts to estimate the model parameters. Previous literature also considered the alternative approach of model-free control (e.g., Abbasi-Yadkori et al., 2019; Cassel and Koren, 2021; Fazel et al., 2018; Malik et al., 2020; Tu and Recht, 2019). These works, however, rely heavily on the assumption of quadratic strongly-convex costs and do not apply to general convex costs.

Lastly, Cassel and Koren (2020); Gradu et al. (2020); Plevrakis and Hazan (2020) consider control under bandit feedback. These results are unfortunately impeded by the state-of-the-art in Bandit Convex Optimization, that is either not efficient in practice (namely, high-degree polynomial runtime) or requires further assumptions on the curvature of the cost functions. For this reason we focus here on full-information feedback, with the hope that our techniques can be adapted to bandit feedback in subsequent work, contingent on future advancements in BCO.

2. Preliminaries

2.1 Linear control background

A discrete-time linear control system is one whose dynamics are governed by the following rule:

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t,$$

where $A_\star \in \mathbb{R}^{d_x \times d_x}$, $B_\star \in \mathbb{R}^{d_x \times d_u}$, and where $w_t \in \mathbb{R}^{d_x}$ is zero-mean i.i.d. In the planning version of the problem the controller knows $A_\star, B_\star$ and, at each time t, can choose u_t as a function of $x_1, \ldots, x_t$. After choosing u_t, the controller incurs a known cost $c(x_t, u_t)$. Classic results pertain to quadratic costs, and state that the control rule that minimizes the steady state cost $J(\pi) = \lim_{T\to\infty} \mathbb{E}_\pi[\frac{1}{T}\sum_{t=1}^{T} c(x_t, u_t)]$, chooses $u_t = Kx_t$ for some matrix $K \in \mathbb{R}^{d_u \times d_x}$. Moreover, the optimal rule $\pi_\star$ stabilizes the system, implying that $J(\pi_\star)$ is finite and well-defined for any quadratic cost function.

We require the following notion of strong stability (Cohen et al., 2018), which is standard in the literature and whose purpose is to quantify the classic notion of (asymptotic) stability.

Definition 1 (Strong stability). A controller K for the system $(A_\star, B_\star)$ is $(\kappa, \gamma)-$strongly stable ($\kappa \ge 1, 0 < \gamma \le 1$) if there exist matrices Q, L such that $A_\star + B_\star K = QLQ^{-1}$, $\|L\| \le 1 - \gamma$, and $\|K\|, \|Q\|\|Q^{-1}\| \le \kappa$.

2.2 Problem setup

We address the problem of controlling an unknown linear dynamical system subject to general adversarial convex costs with full state and cost observation. In particular, the system parameters $A_\star$, $B_\star$ are initially unknown and the learner repeatedly interacts with the system as follows:

(1) The player observes state x_t;
(2) The player chooses control u_t;
(3) The player observes the cost function $c_t : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}$, and incurs cost $c_t(x_t, u_t)$.

Note that $(w_t)_{t=1}^{\infty}$ are unobserved, and the cost c_t is revealed only after selecting u_t. Our goal is to minimize regret with respect to any policy π in a benchmark policy class Π. To that end, denote by x_t^π, u_t^π the state and action sequence resulting when following a policy π; then the regret compared to π is defined as

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

and we seek to bound this quantity with high probability for a fixed $\pi \in \Pi$. We focus on the benchmark policy class of strongly stable linear policies that choose $u_t = Kx_t$. i.e., $\Pi_{\text{lin}} = \{K \in \mathbb{R}^{d_u \times d_x} \ : \ K \text{ is } (\kappa, \gamma)\text{-strongly stable}\}$.

We make the following assumptions on our learning problem:

- **Non-stochastic convex and Lipschitz costs.** The costs c_t are arbitrarily determined by an oblivious adversary[1] such that each $c_t(x, u)$ is convex in the pair (x, u) and for any $(x, u), (x', u')$ we have $|c_t(x, u) - c_t(x', u')| \le \|(x - x', u - u')\|$;[2]
- **i.i.d. Gaussian noise.** $(w_t)_{t=1}^{T}$ is a sequence of i.i.d. random variables such that $w_t \sim \mathcal{N}(0, \sigma^2 I)$;
- **Stabilizable system.** $A_\star$ is $(\kappa, \gamma)-$strongly stable, and $\|B_\star\| \le R_B$.

Note the assumption that $A_\star$ is strongly stable is without loss of generality. Otherwise, given access to a stabilizing controller K_0, we show in Appendix A a generic black-box reduction that takes any learning algorithm that assumes strongly-stable $A_\star$, augments its observations and adds $K_0 x_t$ to its predicted actions. This essentially replaces $A_\star$ with $A_\star + B_\star K_0$, which is $(\kappa, \gamma)-$strongly stable as desired, and only incurs a 2κ multiplicative factor in the regret.

2.3 Disturbance Action Policies

We use the, now standard, class of Disturbance Action Policies (DAPs) first proposed by Agarwal et al. (2019a). This class is parameterized by a sequence of matrices $\{M^{[h]} \in \mathbb{R}^{d_u \times d_x}\}_{h=1}^{H}$. For brevity of notation, these are concatenated

1. An oblivious adversary does not use past random choices of the learner to select its loss functions.
2. In Appendix B we also explain how to accommodate quadratic losses via an appropriate choice of a normalizing constant.

into a single matrix $M \in \mathbb{R}^{d_u \times H d_x}$ defined as $M = \left(M^{[1]} \cdots M^{[H]}\right)$. A DAP π_M chooses actions

$$u_t = \sum_{h=1}^{H} M^{[h]} w_{t-h},$$

where recall that the w_t are system disturbances. Consider the benchmark policy class[3]

$$\Pi_{\mathrm{DAP}} = \{\pi_M \ : \ \|M\|_F \le R_{\mathcal{M}}\}.$$

There are two main reasons for considering the DAP parameterization. First, the loss functions are convex in M, a fact which is generally untrue for K in a linear policy $u_t = K x_t$. This paves the way for tools from the online convex optimization literature. Second, as shown in (Agarwal et al., 2019a, Lemma 5.2), if $H \in \Omega(\gamma^{-1} \log T)$ and $R_{\mathcal{M}} \in \Omega(\kappa^2 \sqrt{d_u/\gamma})$ then Π_{DAP} is a good approximation for Π_{lin} in the sense that a regret guarantee with respect to Π_{DAP} gives the same guarantee with respect to Π_{lin} up to a constant additive factor. In light of the above, our regret guarantee will be given with respect to Π_{DAP}.

Bounded memory representation. As observed in recent literature, the linear dynamics have an infinitely long memory, i.e., all past actions have some effect on the current state, and as such on the losses. However, due to the stability of $A_\star$, the effective memory of the system, H, is essentially a constant. To see this, unroll the transition model to get that

$$x_t = A_\star^H x_{t-H} + \sum_{i=1}^{H} \left(A_\star^{i-1} B_\star u_{t-i} + A_\star^{i-1} w_{t-i}\right) = A_\star^H x_{t-H} + \Psi_\star \tilde{\rho}_{t-1} + w_{t-1}, \tag{1}$$

where $\Psi_\star = [A_\star^{H-1} B_\star, \ldots, A_\star B_\star, B_\star, A_\star^{H-1}, \ldots, A_\star] \in \mathbb{R}^{d_x \times d_\Psi}$ and $\tilde{\rho}_t = [u_{t-H}^\mathsf{T}, \ldots, u_t^\mathsf{T}, w_{t-H}^\mathsf{T}, \ldots, w_{t-1}^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{d_\Psi}$, where $d_\Psi := H d_u + (H-1) d_x$. Now, since $A_\star$ is strongly stable, the term $A_\star^H x_{t-H}$ quickly becomes negligible. Following the notation set by Cassel et al. (2022), this observation is combined with the DAP policy parameterization to define the following bounded memory representations. For an arbitrary sequence of disturbances $w = \{w_t\}_{t \ge 1}$ define

$$\begin{aligned}
u_t(M; w) &= \textstyle\sum_{h=1}^{H} M^{[h]} w_{t-h}; \\
P(M) &= \begin{pmatrix}
M^{[H]} & M^{[H-1]} & \cdots & M^{[1]} & & & & \\
& M^{[H]} & M^{[H-1]} & \cdots & M^{[1]} & & & \\
& & \ddots & \ddots & & \ddots & & \\
& & & M^{[H]} & M^{[H-1]} & \cdots & M^{[1]} & \\
& & & & I & & & \\
& & & & & \ddots & & \\
& & & & & & I &
\end{pmatrix} \qquad (2) \\
\rho_t(M; w) &= \left(u_{t+1-H}(M; w)^\mathsf{T}, \ldots u_t(M; w)^\mathsf{T}, w_{t+1-H}, \ldots, w_{t-1}\right)^\mathsf{T} = P(M) w_{t+1-2H:t-1}; \\
x_t(M; \Psi, w) &= \Psi \rho_{t-1}(M; w) + w_{t-1}. \qquad (3)
\end{aligned}$$

Notice that u_t, ρ_t, x_t do not depend on the entire sequence w, but only $w_{t-H:t-1}, w_{t+1-2H:t-1}$, and $w_{t-2H:t-1}$ respectively. Importantly, this means that we can compute these functions with knowledge of only the last (at most) $2H$ disturbances. While this notation does not reveal this fact explicitly, it helps with both brevity and clarity.

3. Algorithm and Main Result

In this section we present our algorithm for regret minimization in linear systems with unknown dynamics and adversarial convex costs; see Algorithm 1. We denote by $\Pi_{\mathcal{M}}$ the projection onto $\mathcal{M}$.

3. We note that a more common definition uses $\sum_{h=1}^{H} \|M^{[h]}\|$ to measure the size of the class. We chose the Frobenius norm for simplicity of the analysis, but replacing it would not change the results significantly.

Algorithm 1 OCO in Adaptive Linear Control

1: **input**: confidence parameter δ, memory length H, optimism parameter α, regularization parameters λ_Ψ, λ_w, learning rate η_G, noise bound W.

2: **set** $i = 1, \tau = 1, V_1 = \lambda_\Psi I, M_1 = 0$ and $\hat{w}_t = 0, \tilde{w}_t, u_t = 0$ for all $t < 1$.

3: **define** loss scaling function:

$$C_M(\Psi) := \sqrt{8} W R_{\mathcal{M}} H \|\Psi\|_F + \alpha\sqrt{2/H}(2 + R_{\mathcal{M}}^{-1}\sqrt{d_x}).$$

4: **for** $t = 1, 2, \ldots, T$ **do**

5: **play** $u_t = \sum_{i=1}^{H} M_t^{[h]} \hat{w}_{t-h}$.

6: **set** $V_{t+1} = V_t + \rho_t \rho_t^\mathsf{T}$ for $\rho_t = (u_{t+1-H}^\mathsf{T}, \ldots, u_t^\mathsf{T}, \hat{w}_{t+1-H}^\mathsf{T}, \ldots, \hat{w}_{t-1}^\mathsf{T})^\mathsf{T}$.

7: **observe** x_{t+1} and cost function c_t.

8: **calculate**

$$(A_t \ B_t) = \mathop{\arg\min}_{(A\ B) \in \mathbb{R}^{d_x \times (d_x + d_u)}} \sum_{s=1}^{t} \|(A\ B)z_s - x_{s+1}\|^2 + \lambda_w \|(A\ B)\|_F^2, \quad \text{where } z_s = \begin{pmatrix} x_s \\ u_s \end{pmatrix}.$$

9: **estimate noise** $\hat{w}_t = \Pi_{\{\|w\| \le W\}}[x_{t+1} - A_t x_t - B_t u_t]$.

10: **sample** $\tilde{w}_t \sim \mathcal{N}(0, \sigma^2 I_{d_x})$.

11: **if** $\det(V_{t+1}) > 2\det(V_{\tau_i})$ **then**

12: **start new epoch**: $i = i + 1, \tau_i = t + 1$.

13: **estimate system parameters**

$$\Psi_{\tau_i} = \mathop{\arg\min}_{\Psi \in \mathbb{R}^{d_x \times d_\Psi}} \left\{ \sum_{s=1}^{t} \|\Psi \rho_s - x_{s+1}\|^2 + \lambda_\Psi \|\Psi\|_F^2 \right\}.$$

14: **initialize** $\mathcal{A}$ = new instance of $\text{BFPL}^\star_{\delta/6}$, and set $M_{\tau_i} = \ldots = M_{\tau_i + 2H} = 0$.

15: **else if** $t \ge \tau_i + 2H$ **then**

16: **define** expert loss functions: $\forall k \in [d_\Psi] \times [(2H-1)d_x], \chi \in \{\pm 1\}$

$$\tilde{f}_t(M; k, \chi) = c_t(x_t(M; \Psi_{\tau_i}, \tilde{w}), u_t(M; \tilde{w})) - \alpha\sigma\chi \cdot \left(V_{\tau_i}^{-1/2} P(M)\right)_k.$$

17: **define** loss vector $\tilde{\ell}_t \in \mathbb{R}^{2(2H-1)d_x d_\Psi^2}$ s.t. $(\tilde{\ell}_t)_{k,\chi} = \tilde{f}_t(M_t(k,\chi); k, \chi)/C_M(\Psi_{\tau_i})$.

18: **update** experts: $\forall k \in [d_\Psi] \times [(2H-1)d_x], \chi \in \{\pm 1\}$

$$M_{t+1}(k, \chi) = \Pi_{\mathcal{M}}\left[M_t(k,\chi) - \eta_G \nabla_M \tilde{f}_t(M_t(k,\chi); k, \chi)\right].$$

19: **update** prediction $(k_{t+1}, \chi_{t+1}) = \mathcal{A}(\tilde{\ell}_t)$ and **set** $M_{t+1} = M_{t+1}(k_{t+1}, \chi_{t+1})$

The algorithm mediates between least squares estimation of the system dynamics (Lines 8 and 13), and optimizing the policy w.r.t. adversarially-changing cost functions. For OCO, the algorithm uses a combination of Online Gradient Descent Zinkevich (2003) (Line 18) and $\text{BFPL}^\star_\delta$ Altschuler and Talwar (2018) (Line 14)—an experts algorithm that also guarantees an overall small number of switches with probability at least $1-\delta$. Our algorithm uses DAP parameterization (Line 5; see Section 2.3 and notations therein), and feeds the aforementioned online optimization algorithms with lower confidence bounds of the online costs. See below for further details on the algorithm's operation.

We have the following guarantee for Algorithm 1. The proof is deferred to Appendix C.

Theorem 2 (Simplified version of Theorem 7 in Appendix C). *Let $\delta \in (0,1)$ and suppose that we run Algorithm 1 with parameters $R_{\mathcal{M}}, R_B \ge 1$ and for proper choices of $W, H, \lambda_w, \lambda_\Psi, \eta_G, \alpha$. If $T \ge 8$ then for any $\pi \in \Pi_{\text{DAP}}$, with probability at least $1-\delta$, $\text{regret}_T(\pi) \le \text{poly}(\kappa, \gamma^{-1}, \sigma, R_B, R_{\mathcal{M}}, d_x, d_u, \log(T/\delta))\sqrt{T}$.*

Our algorithm is comprised of multiple components working in tandem. We now give a brief overview of each of the components and how they play together.

3.1 Prerequisites: system estimation and DAP parameterization

Parameter estimation: The algorithm proceeds in epochs. At the beginning of each epoch, it estimates the unrolled model via least squares using all past observations (Line 13), and the estimate Ψ_{τ_i} is then kept fixed throughout the epoch. The epoch ends when the determinant of V_t is doubled (Line 11); intuitively, when the confidence of the unrolled model increases substantially.[4] Throughout the epoch, the algorithm maintains estimates of the transition noise $(\hat{w}_t)_{t=1}^T$ (Lines 8 and 9). We observe that these noise estimates are essentially produced for "free" and no explicit exploration is needed.

DAP implementation: While the benefits of Π_{DAP} are clear, notice that it cannot be implemented as is since we do not have access to the system disturbances w_t nor can we accurately recover them (due to the uncertainty in the transition model). Similarly to previous works, our algorithm thus uses estimated disturbances $\hat{w}_t$ to compute its actions. At each time step t, the algorithm chooses u_t as a linear function of the past H noise estimates, and parameterized by M_t (Line 5). M_t itself is updated using OCO on surrogate cost functions that are formed as a composition between $c_t(x, u)$ and the bounded memory representations $u_t(M; \hat{w}), x_t(M; \Psi_{\tau_i}, \hat{w})$, implicitly assuming that M_t was kept fixed for the last H time steps. It is therefore crucial that these representations closely reflect the state and action that are actually observed, hence the OCO procedure has to make sure that the sequence $(M_t)_{t=1}^T$ changes slowly (more on this below).

Construction of lower confidence bounds: The algorithm uses the estimated unrolled model to minimize regret with respect to lower confidence bounds of the form:

$$c_t(x_t(M; \Psi_{\tau_i}, \hat{w}), u_t(M; \hat{w})) - \alpha' \cdot \|V_{\tau_i}^{-1/2} \rho_{t-1}(M; \hat{w})\|. \tag{4}$$

This lower confidence bound follows immediately by combining the Lipschitzness of c_t and standard self-normalizing concentration bounds Abbasi-Yadkori and Szepesvári (2011). In our analysis, we show that it indeed lower bounds $c_t(x_t, u_t)$. Such lower confidence bounds are used extensively in multi-armed bandit and reinforcement learning literature to efficiently combine exploration and exploitation Auer et al. (2002, 2008). Intuitively, their minimization steers the resulting policy towards state-action pairs that either yield low cost, or are insufficiently explored.

3.2 Key idea: making the algorithm efficient

The functions in Eq. (4) are, unfortunately, nonconvex (being a difference of two convex functions), and thus cannot be used in OCO algorithms in their current form. However, we overcome this by relaxing the functions in Eq. (4); we do so in two steps. First, we move to an expected, amortized notion of optimism. We can do this since since $\hat{w} \approx w$, which are i.i.d, and thus standard concentration arguments imply that the realized bonus term is close to its conditional expectation, which takes the form:

$$\sqrt{\mathbb{E}\left\|V_{\tau_i}^{-1/2} \rho_{t-1}(M; w)\right\|^2} = \sigma \left\|V_{\tau_i}^{-1/2} P(M)\right\|_F.$$

Second, building on a trick from Dani et al. (2008) in the context of linear bandit optimization, we further bound $\left\|V_{\tau_i}^{-1/2} P(M)\right\|_F \le d_\Psi \left\|V_{\tau_i}^{-1/2} P(M)\right\|_\infty$ (where $\|\cdot\|_\infty$ is the entry-wise matrix infinity norm). Due to an adaptivity issue (more on this below), we also replace the estimated noises $\hat{w}$ in the cost term with random simulated noises $\tilde{w} \sim \mathcal{N}(0, \sigma^2 I)$. After this relaxation, the resulting $\tilde{f}_t$ can be written as a minimum of convex function of the form

$$\tilde{f}_t(M; k, \chi) = c_t(x_t(M; \Psi_{\tau_i}, \hat{w}), u_t(M; \hat{w})) - \alpha\sigma\chi \cdot \left(V_{\tau_i}^{-1/2} P(M)\right)_k, \tag{5}$$

where $\alpha = d_\Psi \alpha'$. Crucial to this trick is the fact that, unlike Eq. (4), the linearized non-convex term is independent of the time index t. This observation yields computationally-efficient regret minimization via a two-tier approach described as follows. We run a different copy of Online Gradient Descent Zinkevich (2003) for each value of k, χ, maintaining a different set of DAP parameters $M_t(k, \chi)$, and fed with $\tilde{f}_t(\cdot; k, \chi)$ (Line 18). On top of the OGD algorithms, we run an experts meta-algorithm to minimize $\tilde{f}_t(M_t(k, \chi); k, \chi)$ over k, χ (Line 19), treating the output of each OGD algorithm as an expert.

4. Concretely, the volume of the confidence ellipsoid around the unrolled model decreases by a constant factor.

Observe that having initially taken expectation over the noises yields an exploration bonus term that, for fixed M, is fixed throughout each epoch. This makes sure that our OCO algorithms, that are restarted at every epoch, can compare against $M_\star$ (the best in hindsight) with k and χ being fixed at the start of the epoch.

3.3 Additional challenges

Stabilizing the meta-algorithm: Our hedging approach nevertheless comes at a price. The choices of the meta-algorithm are inherently random, thus M_t might change abruptly between consecutive rounds (recall that DAP require slowly-changing M_t). We therefore use a version of Follow the Lazy Leader (BFPL*; Altschuler and Talwar (2018)) that guarantees, with high probability, both no-regret and a small number of switches. The small number of switches in conjunction with the fact that each of the expert algorithms generate slowly-changing decisions, guarantee that M_t itself is slowly-changing overall.

Mitigating adaptivity in costs: Even so, the guarantees of BFPL* hold only against *oblivious* adversaries (and this limitation is inherent, as Altschuler and Talwar (2018) discuss extensively), yet the loss sequence constituting of the functions in Eq. (5) is unfortunately *not* oblivious. This is because the noise estimate $\hat{w}$ were generated using policies derived from previous choices of BFPL*. We overcome this hindrance relying on the fact that the noise vectors are drawn from a known (Gaussian) distribution. This allows to sample i.i.d. copies of the noise vectors $\tilde{w}$ (Line 10) that we use in $\tilde{w}$ instead of $\hat{w}$, arriving at the functions defined in Line 16, and ensuring that BFPL* receives obliviously-generated losses.

4. Analysis

In this section we give a (nearly) complete proof of Theorem 2 in a simplified setup, inspired by Plevrakis and Hazan (2020), where $A_\star = 0$. The analysis in the general case is significantly more technical and thus deferred from this extended abstract (see Appendix C for full details).

Suppose that $A_\star = 0$ and thus $x_{t+1} = B_\star u_t + w_t$, assume that $c_t(x,u) = c_t(x)$, i.e., the costs do not depend on u, and aim to minimize the pseudo regret, $\max_{u:\|u\|\le R_u} \sum_{t=1}^{T} [J_t(B_\star u_t) - J_t(B_\star u)]$, where $J_t(x) = \mathbb{E}_w c_t(x+w)$, is the expected instantaneous cost, which can be computed from $c_t(x)$ for a known noise distribution. The resulting problem is an instance of the following variant of online convex optimization, which we now define with clean notation as to avoid confusion with our general setting.

4.1 Simplified setting: OCO with a Hidden Linear Transform

Consider the following setting of online convex optimization. Let $\mathcal{S} \subseteq \mathbb{R}^{d_a}$ be a convex decision set. (We denote by $\Pi_{\mathcal{S}}$ the projection onto $\mathcal{S}$.) At round t the learner:

(i) predicts $a_t \in \mathcal{S}$;
(ii) observes cost function $\ell_t : \mathbb{R}^{d_y} \to \mathbb{R}$ and state $y_{t+1} = Q_\star a_t + \epsilon_t$;
(iii) incurs cost $\ell_t(Q_\star a_t)$.

We have that $\epsilon_t \in \mathbb{R}^{d_y}$ are i.i.d. noise terms, $Q_\star \in \mathbb{R}^{d_y \times d_a}$ is an unknown linear transform, and $y_t \in \mathbb{R}^{d_y}$ are noisy observations. The cost functions are chosen by an oblivious adversary, and we consider minimizing the regret, defined as

$$\text{regret}_T = \max_{a\in\mathcal{S}} \sum_{t=1}^{T} [\ell_t(Q_\star a_t) - \ell_t(Q_\star a)].$$

Assumptions. We make the following assumptions:

- $\ell_t(\cdot)$ are convex and $1-$Lipschitz;
- There exist known $W, R_Q \ge 0$ such that $\|\epsilon_t\| \le W$, and $\|Q_\star\| \le R_Q$.
- For all $a \in \mathcal{S}$ we have $\|a\| \le R_a/2$.

Algorithm 2 OCO with a hidden linear transform

input: optimism parameter α, regularizer λ, learning rates η_G, η_M
set: $V_1 = \lambda I, \widehat{Q}_1 = 0, i = 1, \tau_1 = 1$, and $a_1(k, \chi) \in \mathcal{S}, p_t(k, \chi) = 1/2d_a \quad \forall k \in [d_a], \chi \in \{\pm 1\}$.
for $t = 1, 2, \ldots, T$ **do**
  draw $(k_t, \chi_t) \sim p_t$, and **play** $a_t = a_t(k_t, \chi_t)$.
  observe $y_{t+1} = Q_\star a_t + w_t$ and cost function ℓ_t, and **set** $V_{t+1} = V_t + a_t a_t^\mathsf{T}$.
  if $\det(V_{t+1}) > 2\det(V_{\tau_i})$ **then**
    start new episode $i = i + 1, \tau_i = t + 1$, and **set** $p_{t+1}(k, \chi) = 1/2d_a, a_{t+1}(k, \chi) = a_t(k, \chi)$.
    estimate parameters: $\widehat{Q}_{\tau_i} = \arg\min_{Q \in \mathbb{R}^{d_y \times d_a}} \sum_{s=1}^{t} \{\|Qa_s - y_{s+1}\|^2 + \lambda \|Q\|_F^2\}$.
  else
    define expert loss functions: $\bar{\ell}_t(a; k, \chi) = \ell_t(\widehat{Q}_{\tau_i} a) - \alpha \chi \cdot (V_{\tau_i}^{-1/2} a)_k$.
    update experts: $a_{t+1}(k, \chi) = \Pi_{\mathcal{S}}\big[a_t(k, \chi) - \eta_G \nabla_a \bar{\ell}_t(a_t(k, \chi); k, \chi)\big]$. ▷ OGD
    update prediction: $p_{t+1}(k, \chi) \propto p_t(k, \chi) \exp(-\eta_M \bar{\ell}_t(a_t(k, \chi); k, \chi))$. ▷ MW

Algorithm. Our algorithm for this simplified setup is detailed in Algorithm 2. Unlike the full control setting, the adversarial costs here have no memory, thus enable the following simplifications compared to Algorithm 1. First, we can forgo the DAP parameterization and directly optimize the prediction a_t. This both removes the need to estimate the disturbances, and simplifies the construction of the lower confidence bound. Moreover, the lack of memory obviates the need to make our predictions change slowly over time, and we replace the BFPL* sub-routine with Multiplicative Weights (MW) (see Arora et al., 2012).

4.2 Analysis

The main result of this section bounds the regret of Algorithm 2 with high probability.

Theorem 3. *Let $\delta \in (0,1)$ and suppose that we run Algorithm 2 with parameters*

$$\eta_G = \frac{R_a}{(2\alpha R_a^{-1} + R_Q)\sqrt{T}}, \eta_M = \frac{\sqrt{\log(2d_a)}}{2(2\alpha + R_a R_Q)\sqrt{T}}, \lambda = R_a^2, \alpha = \sqrt{d_a}\Big(W d_y \sqrt{8 \log \tfrac{2T}{\delta}} + \sqrt{2} R_a R_Q\Big).$$

If $T \geq 8$ then with probability at least $1 - \delta$,

$$\text{regret}_T \leq 77 d_a^{3/2} \left(W d_y \sqrt{8 \log \tfrac{2T}{\delta}} + R_a R_Q\right) \sqrt{T \log^2 \frac{4 d_a T^2}{\delta}}.$$

The proof of Theorem 3 is composed of two main lemmas. Similarly to the control setting, we first define an optimistic loss $\bar{\ell}_t(a) = \ell_t(\widehat{Q}_{\tau_{i(t)}} a) - \alpha \|V_{\tau_{i(t)}}^{-1/2} a\|_\infty$, where $i(t) = \max\{i : \tau_i \leq t\}$. The following lemma shows that the optimistic loss lower bounds the true loss, and bounds the error between the two (See proof in Appendix D).

Lemma 4 (optimism). *Suppose that $\sqrt{d_a}\|\widehat{Q}_{\tau_{i(t)}} - Q_\star\|_{V_{\tau_{i(t)}}} \leq \alpha$. Then for any $a \in \mathbb{R}^{d_a}$,*

$$\bar{\ell}_t(a) \leq \ell_t(Q_\star a) \leq \bar{\ell}_t(a) + 2\alpha \sqrt{a^\mathsf{T} V_{\tau_{i(t)}}^{-1} a}.$$

Next, the following result bounds the regret with respect to the optimistic cost functions.

Lemma 5. *Define $G_i = \|\widehat{Q}_{\tau_i}\| + \alpha \lambda^{-1/2}$ and $\bar{G} = 2\alpha\lambda^{-1/2} + R_Q$. With probability at least $1 - \delta$, for all epochs $i \geq 1$ simultaneously:*

$$\sum_{t=\tau_i}^{\tau_{i+1}-1} \big(\bar{\ell}_t(a_t) - \bar{\ell}_t(a)\big) \leq 3R_a\big(\bar{G} + \bar{G}^{-1} G_i^2\big) \sqrt{T \log \frac{2 d_a T^2}{\delta}}.$$

Proof. First, fix an epoch i and notice that $a_t(k, \chi)$ are the result of running Online Gradient Descent (OGD) on the functions $\bar{\ell}_t(\cdot; k, \chi)$, which are G_i Lipschitz. A classic regret bound for OGD (see Lemma 25 in Appendix E.1) then gives us that for all $a \in \mathcal{S}$ and $\tau_i \le s \le T$

$$\sum_{t=\tau_i}^{s} \bar{\ell}_t(a_t(k,\chi);k,\chi) - \bar{\ell}_t(a;k,\chi) \le \frac{1}{2} R_a(\bar{G} + G_i^2 \bar{G}^{-1})\sqrt{T}.$$

Next, note that MW is invariant to a constant shift in the loss vectors. Letting $a_0 \in \mathcal{S}$ be arbitrary, we have that p_t is updated according to the MW rule with the loss of each expert being $\bar{\ell}_t(a_t(k,\chi);k,\chi) - \ell_t(\widehat{Q}_{\tau_i} a_0)$. Using the Lipschitz property of ℓ_t, these are bounded as

$$|\bar{\ell}_t(a_t(k,\chi);k,\chi) - \ell_t(\widehat{Q}_{\tau_i} a_0)| \le \|\widehat{Q}_{\tau_i}\| \|a_t(k,\chi) - a_0\| + \alpha\lambda^{-1/2}\|a_t(k,\chi)\| \le G_i R_a.$$

A standard regret guarantee of MW (Lemma 26 in Appendix E.1) thus gives us that with probability at least $1-\delta$,

$$\sum_{t=\tau_i}^{s} \bar{\ell}_t(a_t(k_t,\chi_t);k_t,\chi_t) - \bar{\ell}_t(a_t(k,\chi);k,\chi) \le R_a\big(\bar{G} + \bar{G}^{-1}G_i^2\big)\sqrt{6T\log\frac{2d_aT}{\delta}},$$

for all $k \in [d_a], \chi \in \{-1,1\}$, and $\tau_i \le s \le T$.

Now, let $k_t^*(a), \chi_t^*(a)$ be such that $\bar{\ell}_t(a) = \bar{\ell}_t(a;k_t^*(a),\chi_t^*(a))$ for all $\tau_i \le t < \tau_{i+1}$. Importantly, notice that $k_t^*(a), \chi_t^*(a)$ are independent of the time index t. This is because the minimum in $\bar{\ell}_t$ is taken over the optimism term, which is independent of t inside a given epoch. For ease of notation, the following will omit the dependence of k^*, χ^* on a, which will be kept as a fixed (arbitrary) comparator. Combining the above, with probability $\ge 1-\delta$ we have that for all $a \in \mathcal{S}$:

$$\begin{aligned}
&\sum_{t=\tau_i}^{\tau_{i+1}-1} \bar{\ell}_t(a_t) - \bar{\ell}_t(a) \le \sum_{t=\tau_i}^{\tau_{i+1}-1} \bar{\ell}_t(a_t;k_t,\chi_t) - \bar{\ell}_t(a;k^*,\chi^*) \qquad (\bar{\ell}_t(\cdot) \le \bar{\ell}_t(\cdot;k,\chi)) \\
&= \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(\bar{\ell}_t(a_t(k_t,\chi_t);k_t,\chi_t) - \bar{\ell}_t(a_t(k^*,\chi^*);k^*,\chi^*)\big) + \sum_{t=\tau_i}^{\tau_{i+1}-1} (\bar{\ell}_t(a_t(k^*,\chi^*);k^*,\chi^*) - \bar{\ell}_t(a;k^*,\chi^*)) \\
&\le 3R_a\big(\bar{G} + \bar{G}^{-1}G_i^2\big)\sqrt{T\log\frac{2d_aT}{\delta}}.
\end{aligned}$$

Repeating the above with δ/T and taking a union bound over the epochs concludes the proof. ■

We are now ready to prove Theorem 3. We focus here on the main ideas, deferring some details to Appendix D.

Proof of Theorem 3. We decompose the regret as

$$\text{regret}_T(a) \le \underbrace{\sum_{t=1}^{T} \ell_t(Q_\star a_t) - \bar{\ell}_t(a_t)}_{R_1} + \underbrace{\sum_{t=1}^{T} \bar{\ell}_t(a_t) - \bar{\ell}_t(a)}_{R_2} + \underbrace{\sum_{t=1}^{T} \bar{\ell}_t(a) - \ell_t(Q_\star a)}_{R_3},$$

and conclude the proof by bounding each term on the following good event. Suppose Lemma 5 holds for all epochs with $\delta/2T$, and that $\sqrt{d_a}\|\widehat{Q}_t - Q_\star\|_{V_t} \le \alpha$ for all $t \le T$, which follows from a standard least squares estimation bound (see Lemma 22). Taking a union bound, this event holds with probability at least $1-\delta$. We conclude that Lemma 4 holds and thus $R_3 \le 0$. Moreover, we get that

$$R_1 \le \sum_{i=1}^{N} \sum_{t=\tau_i}^{\tau_{i+1}-1} 2\alpha\sqrt{a_t^\mathsf{T} V_{\tau_i}^{-1} a_t} \le 2\alpha \sum_{t=1}^{T} \sqrt{2a_t^\mathsf{T} V_t^{-1} a_t} \le 2\alpha\sqrt{2T\sum_{t=1}^{T} a_t^\mathsf{T} V_t^{-1} a_t} \le 2\alpha\sqrt{10Td_a\log T},$$

where the second inequality uses Lemma 27 of Cohen et al. (2019), which states that for $V_1 \succeq V_2 \succeq 0$ we have $V_1 \preceq V_2(\det(V_1)/\det(V_2))$, the third is due to Jensen's inequality, and the fourth is a standard algebraic argument (see Lemma 23).

Now, an immediate corollary (see Eq. (10)) of the least square error bound is that $\|\widehat{Q}_t\| \le \alpha\lambda^{-1/2} + R_Q$. We thus have that $G_i \le \bar{G}$ for all $i \le N$. Next, notice that the number of epochs satisfies $N \le 2d_a \log T$ (Lemma 24). We thus use Lemma 5 to conclude that

$$R_2 = \sum_{i=1}^{N} \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(\bar{\ell}_t(a_t) - \bar{\ell}_t(a)\big) \le \sum_{i=1}^{N} 3R_a\big(\bar{G} + \bar{G}^{-1}G_i^2\big)\sqrt{T\log\frac{4d_aT^2}{\delta}} \le 12d_a(2\alpha + R_aR_Q)\sqrt{T\log^2\frac{4d_aT^2}{\delta}} \blacksquare$$

Acknowledgments

This work was partially supported by the Deutsch Foundation, by the Israeli Science Foundation (ISF) grant 2549/19, by the Len Blavatnik and the Blavatnik Family foundation, by the Yandex Initiative in Machine Learning, and by the Israeli VATAT data science scholarship.

References

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

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3108–3117. PMLR, 2019.

Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9. PMLR, 2018.

Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 24, 2011.

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

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pages 10175–10184, 2019b.

Jason Altschuler and Kunal Talwar. Online learning over a finite action set with limited switching. In *Conference On Learning Theory*, pages 1569–1573. PMLR, 2018.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

Sanjeev Arora, Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Towards provable control for unknown linear dynamical systems. 2018.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

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

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning*, pages 1328–1337. PMLR, 2020.

Asaf Cassel, Alon Cohen, and Tomer Koren. Efficient online linear control with stochastic convex costs and unknown dynamics. *arXiv preprint arXiv:2203.01170*, 2022.

Asaf B Cassel and Tomer Koren. Online policy gradient for model free learning of linear quadratic regulators with $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1304–1313. PMLR, 2021.

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

Alexey Chernov and Fedor Zhdanov. Prediction with expert advice under discounted loss. In *International Conference on Algorithmic Learning Theory*, pages 255–269. Springer, 2010.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pages 1029–1038, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309, 2019.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

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

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time analysis of optimal adaptive policies for linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2018.

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

Morteza Ibrahimi, Adel Javanmard, and Benjamin Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. *Advances in Neural Information Processing Systems*, 25, 2012.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L. Bartlett, and Martin J. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *Journal of Machine Learning Research*, 21(21):1–51, 2020.

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

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE, 2017.

Orestis Plevrakis and Elad Hazan. Geometric exploration for online control. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7637–7647. Curran Associates, Inc., 2020.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pages 8210–8219. PMLR, 2020.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

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

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083. PMLR, 2019.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Appendix A. Black-box reduction from unstable system

In this section we show that any algorithm that works under the assumption that $A_\star$ is stable can be turned into one that instead receives as input a controller K_0 that stabilizes the system $(A_\star, B_\star)$, and incurs the same regret up to a factor of 2κ. Importantly, our result is not tailored to our specific algorithm and holds for any algorithm and a wide variety of benchmark policy classes. This will show that our simplifying assumption that $A_\star$ is stable is indeed without loss of generality. This has previously been considered to be true, but we could not find a formal proof in the literature; for completeness, we provide one here.

Formal setup. Formally, let $\mathcal{A}$ be an online control algorithm such that

$$u_t = \mathcal{A}(x_{1:t}, u_{1:t-1}, c_{1:t-1}, \zeta).$$

We model $\mathcal{A}$ as a deterministic function, and assume ζ to be its input random bits. Next, let Π be any benchmark policy class that any $\pi \in \Pi$ satisfies that

$$u_t = \pi(w_{1:t-1}, x_0, \xi, t),$$

i.e., a potentially stochastic (in ξ) time dependent policy that makes decisions solely based on past disturbances. Notice that this is almost without loss of generality since, given knowledge of the system, past state and actions can be recovered from the disturbances and used to compute the next action. The limitation of such classes is that they are non-adaptive in the sense that their policies choose the same actions regardless of the underlying system parameters and cost functions. By definition, Π_{DAP} satisfies this assumption.

Instead of assuming that $A_\star$ is $(\kappa, \gamma)-$strongly stable, here we assume that we are given a controller $K_0 \in \mathbb{R}^{d_x \times d_u}$ such that $A_\star + B_\star K_0$ is $(\kappa, \gamma)-$strongly stable.

The reduction. Given $\mathcal{A}$ and K_0, we define a meta-algorithm that at each time t:

(i) **calculates** $\tilde{u}_t = \mathcal{A}(x_{1:t}, \tilde{u}_{1:t-1}, \tilde{c}_{1:t-1}, \zeta)$ where $\tilde{c}_t(x, u) = c_t(x, u + K_0 x)/2\kappa$;

(ii) **plays** $u_t = K_0 x_t + \tilde{u}_t$ and **observes** x_{t+1}, c_t.

The following is our main result for the reduction to stable $A_\star$.

Proposition 6. *Suppose that $\mathcal{A}$ has a regret upper bound of $C_T(\Pi)$ for $A_\star$ stable and benchmark policy class Π. Then given a stabilizing controller K_0, our meta algorithm has regret guarantee of $2\kappa C_T(\Pi)$ against the benchmark class $\Pi_{K_0} = \{\pi_{K_0} : \pi \in \Pi\}$, where*

$$\pi_{K_0}(x_t, w_{1:t-1}, x_0, \xi, t) = K_0 x_t + \pi(w_{1:t-1}, x_0, \xi, t).$$

Proof. Notice that

$$x_t = A_\star x_{t-1} + B_\star u_{t-1} + w_{t-1} = (A_\star + B_\star K_0) x_{t-1} + B_\star \tilde{u}_{t-1} + w_{t-1}.$$

This implies that from the perspective of $\mathcal{A}$ the underlying system is $(\tilde{A}_\star, B_\star)$ where $\tilde{A}_\star = A_\star + B_\star K_0$ is (κ, γ)-strongly stable, and the disturbances w_t are unchanged. Next, note that $\tilde{c}_t$ is convex as it is formed as a composition of c_t with an affine function. Moreover, the construction of $\tilde{c}_t$ is purely deterministic, thus from the point of view of the algorithm, the loss sequence $\tilde{c}_t$ is oblivious. Finally, we have that

$$\begin{aligned}
|\tilde{c}_t(x, u) - \tilde{c}_t(x', u')|^2 &= (2\kappa)^{-2} |c_t(x, u + K_0 x) - c_t(x', u' + K_0 x')|^2 \\
&\le (2\kappa)^{-2} (\|x - x'\|^2 + \|(u - u') + K_0(x - x')\|^2) && (c_t \text{ Lipschitz}) \\
&\le (2\kappa)^{-2} (2\|u - u'\|^2 + (1 + 2\|K_0\|^2)\|x - x'\|^2) \\
&\le \|x - x'\|^2 + \|u - u'\|^2, && (\|K_0\| \le \kappa)
\end{aligned}$$

i.e., $\tilde{c}_t$ are 1-Lipschitz. Thus, given these conditions, we can use the regret guarantee for $\mathcal{A}$ to get that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \Big[\tilde{c}_t(x_t, \tilde{u}_t) - \tilde{c}_t(x_t^\pi(\tilde{A}_\star, B_\star), u_t^\pi) \Big] \le C_T(\Pi), \quad \forall \pi \in \Pi,$$

where $x_t^\pi(A,B)$ is the state sequence that arises when following policy $\pi \in \Pi$ on the system (A,B). We note that u_t^π is the previously defined action sequence that results from following π, which does not depend on (A,B) due to our assumption on the class Π.

Now, for any $\pi_{K_0} \in \Pi_{K_0}$ let $u_t^{\pi_{K_0}}$ be the action sequence when following π_{K_0} on the system $(A_\star, B_\star)$. For its underlying policy $\pi \in \Pi$ we thus have

$$\begin{aligned} u_t^{\pi_{K_0}} &= \pi_{K_0}(x_t^{\pi_{K_0}}(A_\star,B_\star), w_{1:t-1}, x_0, \xi, t) \\ &= K_0 x_t^{\pi_{K_0}}(A_\star,B_\star) + \pi(w_{1:t-1}, x_0, \xi, t) = K_0 x_t^{\pi_{K_0}}(A_\star,B_\star) + u_t^\pi. \end{aligned}$$

Next, we prove by induction that $x_t^\pi(\tilde{A}_\star, B_\star) = x_t^{\pi_{K_0}}(A_\star, B_\star)$ for all $t \geq 1$. This holds trivially for the initial state $t = 1$. Now, assume this is true up to $t-1$, then we have that

$$\begin{aligned} x_t^\pi(\tilde{A}_\star, B_\star) &= \tilde{A}_\star x_{t-1}^\pi(\tilde{A}_\star, B_\star) + B_\star u_{t-1}^\pi + w_{t-1} \\ &= (A_\star + B_\star K_0) x_{t-1}^{\pi_{K_0}}(A_\star, B_\star) + B_\star u_{t-1}^\pi + w_{t-1} && \text{(induction hypothesis)} \\ &= A_\star x_{t-1}^{\pi_{K_0}}(A_\star, B_\star) + B_\star(u_{t-1}^\pi + K_0 x_{t-1}^{\pi_{K_0}}(A_\star, B_\star)) + w_{t-1} \\ &= A_\star x_{t-1}^{\pi_{K_0}}(A_\star, B_\star) + B_\star u_{t-1}^{\pi_{K_0}} + w_{t-1} \\ &= x_t^{\pi_{K_0}}(A_\star, B_\star), \end{aligned}$$

thus proving the inductive claim. We conclude that under the above event, for any $\pi_{K_0} \in \Pi_{K_0}$ we have that

$$\begin{aligned} &\sum_{t=1}^T \big[c_t(x_t, u_t) - c_t(x_t^{\pi_{K_0}}(A_\star, B_\star), u_t^{\pi_{K_0}})\big] \\ &\quad = \sum_{t=1}^T \Big[c_t(x_t, \tilde{u}_t + K_0 x_t) - c_t(x_t^\pi(\tilde{A}_\star, B_\star), u_t^\pi + K_0 x_t^\pi(\tilde{A}_\star, B_\star))\Big] \\ &\quad = 2\kappa \sum_{t=1}^T \Big[\tilde{c}_t(x_t, \tilde{u}_t) - \tilde{c}_t(x_t^\pi(\tilde{A}_\star, B_\star), u_t^\pi)\Big] \\ &\quad \leq 2\kappa C_T(\Pi), \end{aligned}$$

thus concluding the proof. ■

Appendix B. Extensions

In this section we elaborate on how to extend our results to Gaussian noise, and to handle quadratic costs.

Gaussian noise. In Theorem 7 (Appendix C) we will analyze a slight modification of Algorithm 1. Instead of $w_t \sim \mathcal{N}(0, \sigma^2 I)$ we will make a simplifying assumption that w_t is zero-mean, has a known distribution with covariance Σ, and is bounded as $\|w_t\| \leq W$. This assumption will also modify the generated noise in Line 10 of Algorithm 1.

Now, we claim that we can run Algorithm 1 as is and obtain nearly the same guarantees. To that end, we follow the reduction proposed in Cassel and Koren (2021). First, notice that Algorithm 1 does not require an accurate estimate of Σ. In fact, we can replace Σ in the lower confidence bound Line 16 with any $\hat{\Sigma}$ satisfying

$$\Sigma \preceq \hat{\Sigma} \preceq 2\Sigma,$$

and the regret guarantee would change by at most a factor of 2. To see this, one needs to examine the proof of Lemma 15 and in particular that of Lemma 17, and replace $\hat{\Sigma}$ with either its lower or upper bounds appropriately.

Now, suppose that we run Algorithm 1 with the noises $w_t, \tilde{w}_t$ replaced with

$$\begin{aligned} \bar{w}_t &= w_t \mathbb{1}_{\left\{\|\Sigma^{-1/2} w_t\|^2 \leq 5 d_x \log(2T/\delta)\right\}} \\ \bar{\tilde{w}}_t &= \tilde{w}_t \mathbb{1}_{\left\{\|\Sigma^{-1/2} \tilde{w}_t\|^2 \leq 5 d_x \log(2T/\delta)\right\}}, \end{aligned}$$

and denote the covariance of this truncated noise $\bar{\Sigma} = \mathbb{E}\bar{w}\bar{w}^{\mathsf{T}}$. As shown in Cassel and Koren (2021), we have that for $T \geq 12$

$$\begin{aligned} \mathbb{E}\bar{w}_t = \mathbb{E}\bar{\tilde{w}}_t &= 0 \\ \max \|\bar{w}_t\|, \|\bar{\tilde{w}}_t\| &\leq W \\ \bar{\Sigma} \preceq \Sigma &\preceq 2\bar{\Sigma}, \end{aligned}$$

where $W = \sqrt{5d_x\|\Sigma\| \log(2T/\delta)}$. They also use standard tail inequalities to show that with probability at least $1-\delta$, both $\bar{w}_t = w_t$ and $\tilde{w}_t = \bar{\tilde{w}}_t$. On this event we have that the regret of Algorithm 1 with or without the truncation is the same. Since $\bar{w}_t$ satisfy the assumptions of Theorem 7, we can use it to bound the regret with respect to $\bar{w}_t$ with probability at least $1-\delta$. Using a union bound, we conclude that the regret of the original algorithm is bounded with probability at least $1-2\delta$ where the parameter W is set to $\sqrt{5d_x\|\Sigma\| \log(2T/\delta)}$.

Quadratic costs. We now consider the case in which the cost functions are of the form:

$$c_t(x, u) = x^{\mathsf{T}} Q_t x + u^{\mathsf{T}} R_t u, \quad \text{where } \|Q_t\|, \|R_t\| \leq 1.$$

This could also be replaced with the notion of sub-quadratic Lipschitz costs (see, e.g., Simchowitz et al., 2020)). Now, let $R_{\max}$ be an upper bound on $\|(x_t, u_t)\|$ (see Lemma 10). Define the loss functions $\tilde{c}_t$ that coincide with c_t on $\|(x_t, u_t)\| \leq R_{\max}$ and outside of the region they are extrapolated such that they are globally convex and $2R_{\max}$ Lipschitz. By design, there is no difference between running Algorithm 1 on either $c_t/2R_{\max}$ and $\tilde{c}_t/2R_{\max}$. Since $\tilde{c}_t/2R_{\max}$ satisfy the assumptions for Theorem 7, we get that its regret bound holds for quadratic (up to a $2R_{\max}$ multiplicative factor).

Appendix C. Proof of Theorem 2

In this section we prove a regret bound for a slight modification of Algorithm 1. Concretely, we replace the assumption that $w_t \sim \mathcal{N}(0, \sigma^2 I)$ with the assumption that w_t is zero-mean, has a known distribution $\mathcal{W}$ with covariance Σ, and is bounded as $\|w_t\| \leq W$. This results in the following modifications to Algorithm 1:

1. The lower confidence bound $\alpha\sigma\chi \cdot \left(V_{\tau_i}^{-1/2} P(M)\right)_k$ in Line 16 is replaced with

$$\alpha\chi \cdot \left(V_{\tau_i}^{-1/2} P(M)\Sigma_{2H-1}\right)_k,$$

   where $\Sigma_{2H-1} = I_{2H-1} \otimes \Sigma$, and $\otimes$ is the Kronecker product;

2. The generated noises $\tilde{w}_t$ are sampled from $\mathcal{W}$.

In Appendix B we explained how this modification is applicable to Algorithm 1 without altering the Gaussian noise assumption. The following is our main result, which bounds the regret of Algorithm 1 under the above modifications and the bounded noise assumption.

Theorem 7 (restatement of Theorem 2). *Let $\delta \in (0,1)$ and suppose that we run Algorithm 1 with parameters $W, R_{\mathcal{M}}, R_B \geq 1$ and*

$$H = \gamma^{-1}\log T, \quad \lambda_w = 5\kappa^2 W^2 R_{\mathcal{M}}^2 R_B^2 H \gamma^{-1}, \quad \lambda_\Psi = 2W^2 R_{\mathcal{M}}^2 H^2, \quad \eta_G = R_{\mathcal{M}}^2 \alpha^{-1}\sqrt{2H/T},$$
$$\alpha = 21 W R_{\mathcal{M}} R_B \kappa^2 (d_x + d_u)\sqrt{H^3\gamma^{-3}(d_x^2\kappa^2 + d_u R_B^2)\log\frac{24T^2}{\delta}}.$$

If $T \geq 8$ then for any $\pi \in \Pi_{\mathrm{DAP}}$, with probability at least $1-\delta$

$$\mathrm{regret}_T(\pi) \leq 43261 W R_{\mathcal{M}}^2 R_B^2 \kappa^3 \gamma^{-8}(d_x^2\kappa^2 + d_u R_B^2)\log^6\left(\frac{48T^2}{\delta}\right)\sqrt{T d_x (d_x+d_u)^3 \log(6d_\Psi^2)}.$$

Structure. We begin with a preliminaries section (Appendix C.1) that states several results that will be used throughout, and are either technical or adaptated from existing results. Next, in Appendix C.2 we provide the body of the proof, decomposing the regret into logical terms, and stating the bound for each one. Finally, in Appendices C.3 to C.5 we prove the bounds for each term.

C.1 Preliminaries

Disturbance estimation. The success of our algorithm relies on the estimation of the system disturbances, which was first bounded in Plevrakis and Hazan (2020). Here we use the following statement, due to Cassel et al. (2022).

Lemma 8. *Suppose that* $\lambda_w = 5\kappa^2 W^2 R_{\mathcal{M}}^2 R_B^2 H \gamma^{-1}$, $H > \log T$, *and* $T \geq d_x$. *With probability at least* $1-\delta$

$$\sqrt{\sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2} \leq C_w, \quad \text{where} \quad C_w = 10 W \kappa R_{\mathcal{M}} R_B \gamma^{-1} \sqrt{H(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{T}{\delta}}.$$

As noted by Plevrakis and Hazan (2020), the quality of the disturbance estimation does not depend on the choices of the algorithm, i.e., we can recover the noise without any need for exploration. This is in stark contrast to the estimation of the system matrices $A_\star, B_\star$, which requires exploration.

Estimating the unrolled model. Here we bound the difference between Ψ_t, the least squares estimate, and the real model $\Psi_\star$. Notice that some standard manipulations on Eq. (1) yield that

$$x_t = \Psi_\star \rho_{t-1} + w_{t-1} + e_{t-1},$$

where $\rho_{t-1} = [u_{t-H}^\mathsf{T}, \ldots, u_{t-1}^\mathsf{T}, \hat{w}_{t-H}^\mathsf{T}, \ldots, \hat{w}_{t-2}^\mathsf{T}]$ are the observations defined in Line 6 of Algorithm 1, and $e_{t-1} = A_\star^H x_{t-H} + \sum_{h=1}^{H} A_\star^{h-1}(w_{t-h} - \hat{w}_{t-h})$ is a bias term, which, while small, is not negligible. The following result bounds the least squares error for observations of this form. It takes the least squares estimation error bound of Abbasi-Yadkori and Szepesvári (2011), and augments it with a sensitivity analysis with respect to the biased observations.

Lemma 9 (Cassel et al. (2022)). *Let* $\Delta_t = \Psi_\star - \Psi_t$, *and suppose that* $\|\rho_t\|^2 \leq \lambda_\Psi, T \geq d_x$. *With probability at least* $1-\delta$, *we have for all* $1 \leq t \leq T$

$$\|\Delta_t\|_{V_t}^2 \leq \mathrm{Tr}\big(\Delta_t^\mathsf{T} V_t \Delta_t\big) \leq 16 W^2 d_x^2 \log\left(\frac{T}{\delta}\right) + 4\lambda_\Psi \|\Psi_\star\|_F^2 + 2\sum_{s=1}^{t-1} \|e_s\|^2.$$

If we also have that $\lambda_\Psi = 2W^2 R_{\mathcal{M}}^2 H^2$, *and that* $\sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2 \leq C_w^2$ *(see Lemma 8) then*

$$\|\Delta_t\|_{V_t} \leq \sqrt{\mathrm{Tr}\big(\Delta_t^\mathsf{T} V_t \Delta_t\big)} \leq 21 W R_{\mathcal{M}} R_B \kappa^2 H \sqrt{\gamma^{-3}(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{T}{\delta}},$$

and $\|(\Psi_t \; I)\|_F \leq 17 R_B \kappa^2 \sqrt{\gamma^{-3}(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{T}{\delta}}$.

DAP bounds and properties. We need several properties that relate to the DAP parameterization and will be useful throughout. The following lemma is due to (Cassel et al., 2022, Lemma 11).

Lemma 10. *We have that for all* w *such that* $\|w_t\| \leq W$, $M \in \mathcal{M}$, *and* $t \leq T$

1. $\|(\Psi_\star \; I)\|_F \leq R_B \kappa \sqrt{2d_x/\gamma}$;
2. $\|u_t(M; w)\| \leq W R_{\mathcal{M}} \sqrt{H}$;
3. $\|(\rho_t(M; w)^\mathsf{T} \; w_t^\mathsf{T})^\mathsf{T}\| \leq \sqrt{2} W R_{\mathcal{M}} H$;
4. $\max\{\|x_t(M; \Psi_\star; w)\|, \|x_t^{\pi_M}\|, \|x_t\|\} \leq 2\kappa R_B W R_{\mathcal{M}} \sqrt{H}/\gamma$;

5. $\|u_t(M;w) - u_t(M;w')\| \le R_{\mathcal{M}} \|w_{t-H:t-1} - w'_{t-H:t-1}\|$;

6. $\sqrt{\|\rho_t(M;w) - \rho_t(M;w')\|^2 + \|w_t - w'_t\|^2} \le R_{\mathcal{M}} \sqrt{H} \|w_{t-2H:t-1} - w'_{t-2H:t-1}\|$.

Recall that $\rho_t = (u_{t+1-H}^\mathsf{T}, \ldots, u_t^\mathsf{T}, \hat{w}_{t+1-H}^\mathsf{T}, \ldots, \hat{w}_{t-1}^\mathsf{T})^\mathsf{T}$. The following lemma complements the previous Lemma (see proof in Appendix E).

Lemma 11. *We have that for all w such that $\|w_t\| \le W$, $M \in \mathcal{M}$, and $t \le T$:*

1. $\|\rho_t\| \le \sqrt{2} W R_{\mathcal{M}} H$;

2. $\|\rho_{t-1} - \rho_{t-1}(M_t; w)\|^2 \le 2R_{\mathcal{M}}^2 H \Big[\sum_{h=1}^{2H} \|w_{t-h} - \hat{w}_{t-h}\|^2 + \sum_{h=1}^{H} \|M_{t-h} - M_t\|_F^2\Big]$.

Surrogate and optimistic costs. We summarize the useful properties of the surrogate costs. To that end, with some abuse of notation, we extend the definition of the surrogate and optimistic cost functions to include the dependence on their various parameters:

$$\begin{aligned} f_t(M;w) &= c_t(x_t(M;\Psi_\star,w), u_t(M;w)) \\ \bar{f}_t(M;\Psi,V,w) &= c_t(x_t(M;\Psi,w), u_t(M;w)) - \alpha \|V^{-1/2} P(M) \Sigma_{2H-1}^{1/2}\|_\infty \\ \bar{f}_t(M;k,\chi,\Psi,V,w) &= c_t(x_t(M;\Psi,w), u_t(M;w)) - \alpha\chi \cdot \Big(V^{-1/2} P(M) \Sigma_{2H-1}^{1/2}\Big)_k, \end{aligned} \tag{6}$$

where $\Sigma_{2H-1} = I_{2H-1} \otimes \Sigma$ and $\otimes$ is the Kronecker product of two matrices. Recalling that $w, \tilde{w}, \hat{w}$ are the real, generated, and estimated noise sequences respectively, we use the following shorthand notations throughout:

$$\begin{aligned} f_t(M) &= f_t(M;w) \\ \bar{f}_t(M) &= \bar{f}_t(M;\Psi_{\tau_t},V_{\tau_t},w) \\ \tilde{f}_t(M) &= \bar{f}_t(M;\Psi_{\tau_t},V_{\tau_t},\tilde{w}) \\ \tilde{f}_t(M;k,\chi) &= \bar{f}_t(M;k,\chi,\Psi_{\tau_t},V_{\tau_t},\tilde{w}), \end{aligned} \tag{7}$$

where $\tau_t = \tau_{i(t-2H)} = \max\{\tau_i : \tau_i \le t - 2H\}$. The following lemma characterizes the properties of $f_t, \bar{f}_t$ as a function of the various parameters (see proof in Appendix E).

Lemma 12. *Define the functions*

$$C_f(\Psi) = 5R_{\mathcal{M}} WH \max\{\|(\Psi\ I)\|_F, \kappa\gamma^{-1} R_B\}, \qquad G_f(\Psi) = \sqrt{2} WH \|\Psi\|_F + \alpha/(R_{\mathcal{M}} \sqrt{2H}).$$

For any w, w' with $\|w_t\|, \|w'_t\| \le W$ and M, M' with $\|M\|_F, \|M'\|_F \le R_{\mathcal{M}}$, we have:

1. $|f_t(M;w) - f_t(M;w')| \le C_f(\Psi)$;

2. $|\bar{f}_t(M;\Psi,V,w) - \bar{f}_t(M;\Psi,V,w')| \le C_f(\Psi)$;

Additionally, if $V \succeq \lambda_\Psi I$ then

3. $|\bar{f}_t(M;\Psi,V,w) - \bar{f}_t(M';\Psi,V,w)| \le G_f(\Psi) \|M - M'\|_F$;

4. $|\bar{f}_t(M;k,\chi,\Psi,V,w) - \bar{f}_t(M';k,\chi,\Psi,V,w)| \le G_f(\Psi) \|M - M'\|_F$;

5. $\bar{f}_t(M;k,\chi,\Psi,V,w) \le \bar{f}_t(M;\Psi,V,w) + \alpha\sqrt{2/H}[1 + R_{\mathcal{M}}^{-1}\sqrt{d_x}]$.

Moreover, if $\|(\Psi\ I)\|_F \le 17 R_B \kappa^2 \sqrt{\gamma^{-3}(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{24T^2}{\delta}}$, then:

$$C_f(\Psi) \le 5\alpha/(H\sqrt{d_x(d_x + d_u)}), \quad \textit{and} \quad G_f(\Psi) \le \alpha\sqrt{2}/(R_{\mathcal{M}}\sqrt{H}).$$

C.2 Regret Decomposition

As seen in Eq. (1), the bounded state representation is such that it depends on the last H decisions of the algorithm. This leads to an online convex optimization with memory problem, which complicates notation significantly. While the overall analysis is the same, we avoid the "with-memory" formulation using a regret decomposition that removes the memory dependence at an early stage, replacing it with a movement cost of the predictions.

Now, the following technical lemma bounds the number of epochs N (see proof in Appendix C.6).

Lemma 13. *We have that* $N \le 2(d_x + d_u)H \log T$.

We are now ready to prove Theorem 7.

Proof of Theorem 7. Recall the surrogate cost and its optimistic version defined in Eqs. (6) and (7). Letting $M_\star \in \mathcal{M}$ be the DAP approximation of $\pi \in \Pi_{\text{lin}}$, we have the following decomposition of the regret:

$$\begin{aligned}
\mathrm{Regret}_T(\pi) = &\sum_{t=1}^T c_t(x_t, u_t) - f_t(M_t) && (R_1 \text{ - Truncation})\\
&+ \sum_{t=1}^T f_t(M_t) - \bar{f}_t(M_t) && (R_2 \text{ - Optimism})\\
&+ \sum_{t=1}^T \bar{f}_t(M_t) - \bar{f}_t(M_\star) && (R_3 \text{ - Excess Risk})\\
&+ \sum_{t=1}^T \bar{f}_t(M_\star) - f_t(M_\star) && (R_4 \text{ - Optimism})\\
&+ \sum_{t=1}^T f_t(M_\star) - c_t(x_t^\pi, u_t^\pi). && (R_5 \text{ - Truncation})
\end{aligned}$$

The proof of Theorem 2 is concluded by taking a union bound over the following lemmas, which bound each of the terms (see proofs in Appendices C.3 to C.5). The technical derivation of the final regret bound is purely algebraic and may be found in Lemma 21. ■

Lemma 14 (Truncation cost). *With probability at least* $1 - \delta/4$ *we have that*

$$R_1 + R_5 \le 24\frac{\kappa^2}{\gamma^2} W R_B^2 R_{\mathcal{M}}^2 H \sqrt{T(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2)\log\frac{4T}{\delta}} + \frac{\kappa}{\gamma^2} R_B W \sqrt{H} \sum_{t=1}^T \|M_t - M_{t-1}\|.$$

Lemma 15 (Optimism cost). *With probability at least* $1 - \delta/4$ *we have that*

$$R_2 + R_4 \le 65\alpha R_{\mathcal{M}} R_B \kappa\gamma^{-1} H \sqrt{T(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2)\log\frac{48T^2}{\delta}} + \alpha\sqrt{\frac{8H^3}{W^2}} \sum_{t=1}^T \|M_t - M_{t-1}\|_F.$$

Lemma 16 (Excess risk). *With probability at least* $1 - \delta/2$ *we have that*

$$R_3 \le 4000\alpha(d_x + d_u)\sqrt{TH^3 d_x \log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}}$$

$$\sum_{t=1}^T \|M_t - M_{t-1}\| \le 548 R_{\mathcal{M}}(d_x + d_u)H\sqrt{T\log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}}.$$

C.3 Proof of Lemma 14

Bounding R_1 and R_5 is mostly standard in recent literature. Nonetheless, we give the details for completeness. We start with the simpler R_5. Recall from Lemma 10 that for $R_{\max} = 2\kappa\gamma^{-1} R_B W R_{\mathcal{M}} \sqrt{H}$

$$\max_{M\in\mathcal{M}, \|w\|\le W, t\le T} \max\{1, \|x_t\|, \|x_t^{\pi_M}\|, \|x_t(M; w)\|\} \le R_{\max}.$$

Notice that $u_t^\pi = u_t(M_\star; w)$. We can thus use the Lipschitz assumption to get that

$$\begin{aligned} f_t(M_\star) - c_t(x_t^\pi, u_t^\pi) &= c_t(x_t(M_\star; w), u_t(M_\star; w)) - c_t(x_t^\pi, u_t^\pi) \\ &\le \|x_t(M_\star; w) - x_t^\pi\| \\ &= \|A_\star^H x_{t-H}^\pi\| && \text{(Eq. (1))} \\ &\le R_{\max}\kappa(1-\gamma)^H. && \text{(strong stability)} \end{aligned}$$

Summing over t and using that $(1-\gamma)^H \le e^{-\gamma H}$ we conclude that

$$R_5 \le R_{\max}\kappa e^{-\gamma H} T.$$

Moving to R_1, notice that $u_t = u_t(M_t; \hat{w})$. we thus have that

$$\begin{aligned} \|x_t - x_t(M_t; w)\| &= \left\| A_\star^H x_{t-H} + \sum_{h=1}^{H} A_\star^{h-1} B_\star (u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)) \right\| \\ &\le \|A_\star^H\|\|x_{t-H}\| + \sum_{h=1}^{H} \|A_\star^{h-1}\|\|B_\star\|\|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\| \\ &\le R_{\max}\kappa e^{-\gamma H} + \kappa R_B \sum_{h=1}^{H} (1-\gamma)^{h-1} \|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\|. \end{aligned}$$

Denoting $[x]_+ = \max\{0, x\}$ we further get that

$$\begin{aligned} c_t(x_t, u_t) - f_t(M_t) &= c_t(x_t, u_t) - c_t(x_t(M_t; w), u_t(M_t; w)) \\ &\le \|x_t - x_t(M_{\tau_{i,j}}; w)\| + \|u_t - u_t(M_{\tau_{i,j}}; w)\| \\ &\le \kappa \left[R_{\max} e^{-\gamma H} + R_B \sum_{h=0}^{H} (1-\gamma)^{[h-1]_+} \|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\| \right]. \end{aligned}$$

Now, we use the Lipschitz properties of u_t (see Lemma 10) to get that

$$\begin{aligned} \|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\| &\le \|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_{t-h}; w)\| \\ &\quad + \|u_{t-h}(M_{t-h}; w) - u_{t-h}(M_t; w)\| \\ &\quad R_{\mathcal{M}} \|w_{t-h-H:t-h-1} - \hat{w}_{t-h-H:t-h-1}\| + W\sqrt{H}\|M_t - M_{t-h}\|, \end{aligned}$$

and summing over t gives

$$\sum_{t=1}^{T} \|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\| \le R_{\mathcal{M}} \sqrt{TH \sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2} + hW\sqrt{H} \sum_{t=1}^{T} \|M_t - M_{t-1}\|,$$

Next, taking $H \ge \gamma^{-1} \log T$ we get that

$$\begin{aligned} R_1 + R_5 &\le \kappa \left[2R_{\max} e^{-\gamma H} T + R_B \sum_{h=0}^{H} \sum_{t=1}^{T} (1-\gamma)^{[h-1]_+} \|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\| \right] \\ &\le \kappa \left[2R_{\max} + R_B \sum_{h=0}^{H} (1-\gamma)^{[h-1]_+} \left(R_{\mathcal{M}} \sqrt{TH \sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2} + hW\sqrt{H} \sum_{t=1}^{T} \|M_t - M_{t-1}\| \right) \right] \\ &\le 2\kappa R_{\max} + 2\kappa\gamma^{-1} R_B R_{\mathcal{M}} \sqrt{TH \sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2} + \kappa\gamma^{-2} R_B W \sqrt{H} \sum_{t=1}^{T} \|M_t - M_{t-1}\|, \end{aligned}$$

Finally, suppose that Lemma 8 holds with $\delta/4$. Then we get that

$$\begin{aligned} R_1 + R_5 &\le 24\kappa^2\gamma^{-2} W R_B^2 R_{\mathcal{M}}^2 H \sqrt{T(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{4T}{\delta}} \\ &\quad + \kappa\gamma^{-2} R_B W \sqrt{H} \sum_{t=1}^{T} \|M_t - M_{t-1}\|. \end{aligned}$$

■

C.4 Proof of Lemma 15

An optimistic cost function should satisfy two properties (in expectation). On the one hand, it is a global lower bound on the true cost function. On the other, it has a small error on the realized prediction sequence. Both of these properties are established in the following lemma. Define the random variables $\Delta_t = \Psi_\star - \Psi_t$, and

$$\bar{\Delta} = \sqrt{d_x(d_x + d_u)H^2} \max_{1 \le i \le N} \|\Delta_{\tau_i} V_{\tau_i}^{1/2}\|$$
$$\tau_t = \tau_{i(t-2H)} = \max\{\tau_i : \tau_i \le t - 2H\}.$$

Notice that α is chosen such that it bounds $\bar{\Delta}$ with high probability. Let $\mathcal{F}_t$ be the filtration defined by the random variables $\{w_1, \ldots, w_{t-1}, M_1, \ldots, M_{t+2H}\}$. Notice that this is a somewhat non-standard definition that contains variables from future time steps. This is done in order to satisfy the following properties:

- Conditioning on $\mathcal{F}_{t-2H}$ does not change the distribution of $w_{t-2H:t-2}$, which are i.i.d random variables;
- $M_t, \tau_t, V_{\tau_t}, \Psi_{\tau_t}$ are $\mathcal{F}_{t-2H}$ measurable;
- $w_{1:t-1}$ is $\mathcal{F}_t$ measurable.

While the second and third requirements are trivially satisfied, the first only holds since the algorithm does not update M_t during the first $2H$ rounds of each epoch.

Lemma 17 (Optimism). *$1 \le t \le T$ and $\mathcal{F}_{t-2H}$ measurable M we have that*

$$\mathbb{E}\left[\bar{f}_t(M) - f_t(M) \ \mid\ \mathcal{F}_{t-2H}\right] \le \left(\bar{\Delta} - \alpha\right)\|V_{\tau_t}^{-1/2} P(M) \Sigma_{2H-1}^{1/2}\|_\infty$$
$$\mathbb{E}\left[f_t(M) - \bar{f}_t(M) \ \mid\ \mathcal{F}_{t-2H}\right] \le \left(\bar{\Delta} + \alpha\right)\sqrt{\mathbb{E}\left[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\|^2 \ \mid\ \mathcal{F}_{t-2H}\right]}.$$

Proof. Recall that $f_t(M) = c_t(x_t(M;\Psi_\star, w), u_t(M;w))$ and thus using the Lipschitz property we have that

$$\begin{aligned}
|f_t(M) - c_t(x_t(M;\Psi_{\tau_t}, w), u_t(M;w))| &\le \|x_t(M;\Psi_{\tau_t}, w) - x_t(M;\Psi_\star, w)\| \\
&= \|(\Psi_{\tau_t} - \Psi_\star)\rho_{t-1}(M;w)\| && \text{(Eq. (3))} \\
&= \|\Delta_{\tau_t}\rho_{t-1}(M;w)\| \\
&\le \|\Delta_{\tau_t} V_{\tau_t}^{1/2}\|\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\|. && \text{(Cauchy-Schwarz)}
\end{aligned}$$

Now, recall that $\rho_{t-1}(M;w) = P(M)w_{t-2H:t-2}$ where P is as in Eq. (2). Notice that τ_t is $\mathcal{F}_{t-2H}$ measurable. Thus for any M that is $\mathcal{F}_{t-2H}$ measurable

$$\begin{aligned}
\mathbb{E}\left[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\|^2 \ \mid\ \mathcal{F}_{t-2H}\right] &= \mathbb{E}\left[\rho_{t-1}^\mathsf{T}(M;w) V_{\tau_t}^{-1}\rho_{t-1}(M;w) \ \mid\ \mathcal{F}_{t-2H}\right] \\
&= \mathrm{Tr}\left(V_{\tau_t}^{-1}\mathbb{E}\left[\rho_{t-1}(M;w)\rho_{t-1}^\mathsf{T}(M;w) \ \mid\ \mathcal{F}_{t-2H}\right]\right) \\
&= \mathrm{Tr}\left(V_{\tau_t}^{-1}P(M)\mathbb{E}\left[w_{t-2H:t-2}w_{t-2H:t-2}^\mathsf{T} \ \mid\ \mathcal{F}_{t-2H}\right]P^\mathsf{T}(M)\right) \\
&= \mathrm{Tr}\left(V_{\tau_t}^{-1}P(M)\Sigma_{2H-1}P^\mathsf{T}(M)\right) \\
&= \|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_F^2,
\end{aligned}$$

where recall that Σ_{2H-1} is a block diagonal matrix with $2H-1$ blocks each containing Σ. Next, we use Jensen's inequality to get that

$$\begin{aligned}
\mathbb{E}\left[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\| \ \mid\ \mathcal{F}_{t-2H}\right] &\le \sqrt{\mathbb{E}\left[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\|^2 \ \mid\ \mathcal{F}_{t-2H}\right]} \\
&= \|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_F^2 \\
&\le \sqrt{2d_x(d_x+d_u)H^2}\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty
\end{aligned}$$

where $\|Q\|_\infty$ is the entry-wise infinity norm of a matrix Q and the last inequality used the fact that for $x \in \mathbb{R}^d$ we have $\|x\|_2 \le \sqrt{d}\|x\|_\infty$. Noticing that $\Psi_{\tau_t}, V_{\tau_t}$ are $\mathcal{F}_{t-2H}$ measurable, We get that

$$\begin{aligned}\mathbb{E}\big[\bar{f}_t(M) - f_t(M) \ \mid\ \mathcal{F}_{t-2H}\big] &\le \|\Delta_{\tau_t} V_{\tau_t}^{1/2}\| \cdot \mathbb{E}\Big[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\| \ \mid\ \mathcal{F}_{t-2H}\Big] \\ &\quad - \alpha\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty \\ &\le \Big(\sqrt{2d_x(d_x+d_u)H^2}\|\Delta_{\tau_t} V_{\tau_t}^{1/2}\| - \alpha\Big)\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty \\ &\le (\bar{\Delta} - \alpha)\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty,\end{aligned}$$

and on the other hand

$$\begin{aligned}\mathbb{E}\big[f_t(M_t) - \bar{f}_t(M_t) \ \mid\ \mathcal{F}_{t-2H}\big] &\le \|\Delta_{\tau_t} V_{\tau_t}^{1/2}\|\mathbb{E}\Big[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\| \ \mid\ \mathcal{F}_{t-2H}\Big] \\ &\quad + \alpha\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty \\ &\le \Big(\sqrt{2d_x(d_x+d_u)H^2}\|\Delta_{\tau_t} V_{\tau_t}^{1/2}\| + \alpha\Big)\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty \\ &\le (\bar{\Delta} + \alpha)\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_F \\ &= (\bar{\Delta} + \alpha)\sqrt{\mathbb{E}\Big[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M;w)\|^2 \ \mid\ \mathcal{F}_{t-2H}\Big]},\end{aligned}$$

as desired. ∎

Next, we need to bound the additional cost incurred by summing over the confidence bound. This cost typically takes the form of a harmonic sum, yet here we have some additional terms that arise from the expected, amortized nature of our confidence bounds. This is summarized in the following lemma (see proof in Appendix C.6).

Lemma 18. *With probability at least* $1-\delta$

$$\begin{aligned}&\sum_{t=1}^{T}\sqrt{\mathbb{E}\Big[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M_t;w)\|^2 \ \mid\ \mathcal{F}_{t-2H}\Big]} \\ &\qquad \le 13H\sqrt{T(d_x+d_u)\log\frac{4T^2}{\delta}} + \sqrt{8TW^{-2}\sum_{t=1}^{T}\|w_t - \hat{w}_t\|^2} + \sqrt{\frac{2H^3}{W^2}}\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F.\end{aligned}$$

We also need the following lemma that deals with the concentration of sums of variables that are independent when they are $2H$ apart in time (proof in Appendix E.2).

Lemma 19 (Block Concentration). *Let X_t be a sequence of random variables adapted to a filtration $\mathcal{F}_t$. Then we have the following*

- *If $|X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-2H}]| \le C_t$ where $C_t \ge 0$ are $\mathcal{F}_{t-2H}$ measurable then w.p. at least $1-\delta$*

$$\sum_{t=1}^{T}(X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-2H}]) \le 2\sqrt{\sum_{t=1}^{T}(C_t^2)H\log\frac{T}{\delta}};$$

- *If $0 \le X_t \le 1$ then with probability at least $1-\delta$*

$$\sum_{t=1}^{T}\mathbb{E}[X_t \mid \mathcal{F}_{t-2H}] \le 2\sum_{t=1}^{T}(X_t) + 8H\log\frac{2T}{\delta}.$$

We are now ready to prove Lemma 15. We begin with the slightly simpler R_4. We want to apply Lemma 19 to $\bar{f}_t(M_\star) - f_t(M_\star)$, which are $\mathcal{F}_t$ measurable. By Lemma 12 we have that for $C_f(\Psi) = 5R_{\mathcal{M}}WH\max\{\|(\Psi\ I)\|_F, \kappa\gamma^{-1}R_B\}$

$$|\bar{f}_t(M_\star) - f_t(M_\star) - \mathbb{E}\big[\bar{f}_t(M_\star) - f_t(M_\star) \ \mid\ \mathcal{F}_{t-2H}\big]| \le 2C_f(\Psi_{\tau_t})$$

We thus use Lemma 19 to get that with probability at least $1-\delta/24$

$$\begin{aligned}
&\sum_{t=1}^{T} \bar{f}_t(M_\star) - f_t(M_\star) \\
&\quad \le \sum_{t=1}^{T} \mathbb{E}\big[\bar{f}_t(M_\star) - f_t(M_\star) \;\mid\; \mathcal{F}_{t-2H}\big] + 4\sqrt{\sum_{t=1}^{T} C_f^2(\Psi_{\tau_t}) H \log\frac{24T}{\delta}} \\
&\quad \le \sum_{t=1}^{T} \big(\bar{\Delta} - \alpha\big)\|V_{\tau_t}^{-1/2} P(M_\star)\Sigma_{2H-1}^{1/2}\|_\infty + 4C_f^{\max}\sqrt{TH\log\frac{24T}{\delta}},
\end{aligned}$$

where the second transition also used Lemma 17, and the definition $C_f^{\max} = \max_{1\le i\le N} C_f(\Psi_{\tau_i})$. Next, we use a union bound on the events of Lemmas 8 and 9 each with $\delta/24$ to bound $\|\Delta_{\tau_t} V_{\tau_i}^{1/2}\|$ and $\|(\Psi_{\tau_t}\ I)\|_F$ for all $i \ge 1$. We conclude that with probability at least $1-\delta/12$

$$\begin{aligned}
\bar{\Delta} &\le \sqrt{d_x(d_x+d_u)} H \max_{t\le T}\|\Delta_t\|_{V_t} \\
&\le 21 W R_{\mathcal{M}} R_B \kappa^2 H^2 (d_x+d_u)\sqrt{\gamma^{-3} d_x (d_x^2\kappa^2 + d_u R_B^2)\log\frac{24T^2}{\delta}} = \alpha; \quad \text{and} \\
C_f^{\max} &\le 5\alpha/(H\sqrt{d_x(d_x+d_u)}). \qquad \text{(Lemma 12)}
\end{aligned}$$

Plugging this back into the above bound we conclude that on the intersection of both events we have

$$R_4 \le 20\alpha\sqrt{TH^{-1}d_x^{-1}(d_x+d_u)^{-1}\log\frac{24T}{\delta}}. \tag{8}$$

Now, for R_2 we start out similarly to R_4. Since M_t is independent of the noise terms $w_{t-2H:t-1}$,[5] we can use Lemmas 17 and 19 to get that with probability at least $1-\delta/24$

$$\begin{aligned}
R_2 &= \sum_{t=1}^{T} \bar{f}_t(M_t) - f_t(M_t) \\
&\le \sum_{t=1}^{T} \Big(\mathbb{E}\big[\bar{f}_t(M_t) - f_t(M_t) \;\mid\; \mathcal{F}_{t-2H}\big]\Big) + 20\alpha\sqrt{TH^{-1}d_x^{-1}(d_x+d_u)^{-1}\log\frac{24T}{\delta}} \\
&\le 20\alpha\sqrt{TH^{-1}d_x^{-1}(d_x+d_u)^{-1}\log\frac{24T}{\delta}} + \big(\bar{\Delta}+\alpha\big)\sum_{t=1}^{T}\sqrt{\mathbb{E}\Big[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M_t;w)\|^2 \;\mid\; \mathcal{F}_{t-2H}\Big]}.
\end{aligned} \tag{9}$$

It remains to bound the last term. Using again the events of Lemmas 8 and 9 (recall we've already taken a union bound over them when bounding R_4), we have

$$\bar{\Delta} \le \alpha, \qquad \sqrt{\sum_{t=1}^{T}\|w_t - \hat{w}_t\|^2} \le 10 W R_{\mathcal{M}} R_B \kappa\gamma^{-1}\sqrt{H(d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2)\log\frac{24T}{\delta}}.$$

Combining this with Lemma 18 with $\delta/12$ we get that with probability at least $1-\delta/12$

$$\begin{aligned}
&\sum_{t=1}^{T}\sqrt{\mathbb{E}\Big[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M_t;w)\|^2 \;\mid\; \mathcal{F}_{t-2H}\Big]} \\
&\quad\le 13H\sqrt{T(d_x+d_u)\log\frac{48T^2}{\delta}} + \sqrt{8TW^{-2}\sum_{t=1}^{T}\|w_t-\hat{w}_t\|^2} + \sqrt{\frac{2H^3}{W^2}}\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F \\
&\quad\le 30 R_{\mathcal{M}} R_B \kappa\gamma^{-1} H\sqrt{T(d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2)\log\frac{48T^2}{\delta}} + \sqrt{\frac{2H^3}{W^2}}\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F,
\end{aligned}$$

5. Indeed, $M_t = 0$ in the first $2H$ steps since the start of an epoch; afterwards V_τ is fixed.

and we conclude by combining with Eqs. (8) and (9) that

$$\begin{aligned}R_2 + R_4 &\le 40\alpha\sqrt{TH^{-1}d_x^{-1}(d_x+d_u)^{-1}\log\frac{24T}{\delta}}\\&\quad + 60\alpha R_{\mathcal{M}}R_B\kappa\gamma^{-1}H\sqrt{T(d_x+d_u)(d_x^2\kappa^2+d_uR_B^2)\log\frac{48T^2}{\delta}} + \alpha\sqrt{\frac{8H^3}{W^2}}\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F\\&\le 65\alpha R_{\mathcal{M}}R_B\kappa\gamma^{-1}H\sqrt{T(d_x+d_u)(d_x^2\kappa^2+d_uR_B^2)\log\frac{48T^2}{\delta}} + \alpha\sqrt{\frac{8H^3}{W^2}}\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F.\end{aligned}$$

■

C.5 Proof of Lemma 16

For the proof, we first need the following result (see proof in Appendix E.1).

Lemma 20. *Let $f_t(M,k,\chi)$ be a sequence of oblivious loss functions that are convex and G Lipschitz in M, and have a convex decision set S with diameter $2R$. Let $f_t(M) = \min_{k\in[d],\chi\in\{-1,1\}} f_t(M,k,\chi)$ and consider the update rule that at time t:*

1. *define loss vector ℓ_t such that $(\ell_t)_{k,\chi} = f_t(M_t(k,\chi);k,\chi)/(2GR+C)$*
2. *update experts: $M_{t+1}(k,\chi) = \Pi_{\mathcal{M}}[M_t(k,\chi) - \eta\nabla_M f_t(M_t(k,\chi);k,\chi)]$*
3. *update prediction: $(k_{t+1},\chi_{t+1}) = BFPL_\delta^*(\ell_t)$ and set $M_{t+1} = M_{t+1}(k_{t+1},\chi_{t+1})$*

where $\eta = 2R/\bar{G}\sqrt{T}$ and $C \ge 0$. Suppose that:

- *i $f_t(M;k,\chi) \le f_t(M) + C$ for all $M \in S, k\in[d], \chi\in\{\pm1\}$;*
- *ii There exist $k(M),\chi(M)$ independent of t such that $f_t(M) = f_t(M,k(M),\chi(M))$.*

Then with probability at least $1-\delta$ we have that for all $\tau \le T$

$$\sum_{t=1}^{\tau} f_t(M_t) - f_t(M) \le 151[(\bar{G} + G^2\bar{G}^{-1})R + C]\sqrt{T\log(2d)\log\frac{2T}{\delta}}$$

$$\sum_{t=1}^{\tau}\|M_t - M_{t-1}\| \le (270 + 2G\bar{G}^{-1})R\sqrt{T\log(2d)\log\frac{2T}{\delta}}.$$

We now proceed with the proof of Lemma 16. Recall the definition (Eqs. (6) and (7)):

$$\tilde{f}_t(M) = c_t(x_t(M;\Psi_{\tau_t},\tilde{w}), u_t(M;\tilde{w})) - \alpha\|V_{\tau_t}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty,$$

where $\tau_t = \max\{\tau_i : \tau_i \le t-2H\}$ is the index of the episode to which $t-2H$ belongs. These are essentially identical to $\bar{f}$ but with the actual noise sequence w replaced with our independently generated noise $\tilde{w}$. We begin by further decomposing R_3 as

$$R_3 = \underbrace{\sum_{t=1}^{T}\bar{f}_t(M_t) - \tilde{f}_t(M_t)}_{R_{3,1}} + \underbrace{\sum_{t=1}^{T}\tilde{f}_t(M_t) - \tilde{f}_t(M_\star)}_{R_{3,2}} + \underbrace{\sum_{t=1}^{T}\tilde{f}_t(M_\star) - \bar{f}_t(M_\star)}_{R_{3,3}}$$

We start with $R_{3,1}, R_{3,3}$. Let $\mathcal{G}_t$ be the filtration defined by the random variables $\{w_1,\ldots,w_{t-1}, \tilde{w}_1,\ldots,\tilde{w}_{t-1}, M_1,\ldots,M_t\}$ and recall that $\bar{f},\tilde{f}$ only depend on $w_{t-2H:t-1}, \tilde{w}_{t-2H:t-1}$ respectively and are thus $\mathcal{G}_t$ measurable. Furthermore, for any $\mathcal{G}_{t-2H}$ measurable M we have

$$\mathbb{E}[\tilde{f}_t(M) - \bar{f}_t(M) \mid \mathcal{G}_{t-2H}] = 0.$$

Recalling from Lemma 12 that $2C_f(\Psi_{\tau_t})$ bounds this sequence and are also $\mathcal{G}_{t-2H}$ measurable, we thus use Lemma 19 to get with probability at least $1-\delta/24$

$$\begin{aligned} R_{3,3} = \sum_{t=1}^{T} \tilde{f}_t(M_\star) - \bar{f}_t(M_\star) &\le 4\sqrt{\sum_{t=1}^{T} C_f^2(\Psi_{\tau_t}) H \log\frac{24T}{\delta}} \\ &\le 4C_f^{\max}\sqrt{TH\log\frac{24T}{\delta}}, \end{aligned}$$

where $C_f^{\max} = \max_{1\le i\le N} C_f(\Psi_{\tau_i})$. Next, notice that M_{t-2H} is $\mathcal{G}_{t-2H}$-measureable and thus

$$\mathbb{E}[\tilde{f}_t(M_{t-2H}) - \bar{f}_t(M_{t-2H}) \mid \mathcal{G}_{t-2H}] = 0.$$

We thus use the same set of arguments as in $R_{3,3}$ to get that with probability at least $1-\delta/24$

$$\sum_{t=1}^{T} \tilde{f}_t(M_{t-2H}) - \bar{f}_t(M_{t-2H}) \le 4C_f^{\max}\sqrt{TH\log\frac{24T}{\delta}}.$$

Now, let $G_f^{\max} = \max_{1\le i\le N} G_f(\Psi_{\tau_i})$ be the upper bound on the Lipschitz constant of $\bar{f}_t$ for all $t \le T$, as defined in Lemma 12. Then we have that

$$\begin{aligned} \left|\sum_{t=1}^{T} \bar{f}_t(M_t) - \bar{f}(M_{t-2H})\right| &\le G_f^{\max}\sum_{t=1}^{T}\|M_t - M_{t-2H}\|_F \\ &\le G_f^{\max}\sum_{t=1}^{T}\sum_{h=0}^{2H-1}\|M_{t-h} - M_{t-(h+1)}\|_F \\ &\le 2G_f^{\max} H\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F. \end{aligned}$$

Combining with the previous inequalities, we get that

$$R_{3,1} + R_{3,3} \le 8C_f^{\max}\sqrt{TH\log\frac{24T}{\delta}} + 4G_f^{\max} H\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F.$$

Moving to $R_{3,2}$, we split the analysis into epochs, and combine the results via a union bound. First, we fix some $1 \le i \le N$ and define for all $t \ge 1$, $k \in [d_\Psi] \times [(2H-1)d_x]$, $\chi \in \{\pm 1\}$ the functions

$$\begin{aligned} &\tilde{f}_t^{(i)}(M; k, \chi) \\ &\quad = c_{\tau_i+2H+t-1}(x_{\tau_i+2H+t-1}(M; \Psi_{\tau_i}, \tilde{w}), u_{\tau_i+2H+t-1}(M; \tilde{w})) - \alpha\chi\cdot\left(V_{\tau_i}^{-1/2} P(M)\Sigma_{2H-1}^{1/2}\right)_k; \\ &\tilde{f}_t^{(i)}(M) \\ &\quad = c_{\tau_i+2H+t-1}(x_{\tau_i+2H+t-1}(M; \Psi_{\tau_i}, \tilde{w}), u_{\tau_i+2H+t-1}(M; \tilde{w})) - \alpha\left\|V_{\tau_i}^{-1/2} P(M)\Sigma_{2H-1}^{1/2}\right\|_\infty. \end{aligned}$$

Let $M_t^{(i)}$ be the iterates that result from running the procedure described in Lemma 20 on the functions $\tilde{f}_t^{(i)}$ starting with an arbitrary$M_1^{(i)} \in \mathcal{M}$. Then we have the following observations:

1. $\tilde{f}_t^{(i)}$ are oblivious with respect to the iterates $M_t^{(i)}$. This is because c_t are oblivious and $\tau_i, \Psi_{\tau_i}, V_{\tau_i}, \tilde{w}$ can be determined independently of the iterates $M_t^{(i)}$;

2. $\tilde{f}_t^{(i)}(M; k, \chi)$ are convex as a sum of a linear function with a composition of a convex and affine functions;

3. $\tilde{f}_t^{(i)}(M;k,\chi)$ are $G_f(\Psi_{\tau_i})$ Lipschitz (see Lemma 12);

4. $\tilde{f}_t^{(i)}(M;k,\chi) \le \tilde{f}_t^{(i)}(M) + C$ for all $M \in \mathcal{M}, k \in [d_\Psi] \times [(2H-1)d_x], \chi \in \{\pm 1\}$ where $C = \alpha\sqrt{2/H}$ $[(1 + R_{\mathcal{M}}^{-1}\sqrt{d_x}]$ (see Lemma 12);

5. $\eta_G = 2R_{\mathcal{M}}/\bar{G}\sqrt{T}$ where $\bar{G} = \alpha\sqrt{2/H}R_{\mathcal{M}}^{-1}$.

Next, let

$$k(M), \chi(M) \in \underset{\substack{k\in[d_\Psi]\times[(2H-1)d_x],\\ \chi\in\{\pm 1\}}}{\arg\max}\ \chi \cdot \left(V_{\tau_i}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\right)_k.$$

Since the term being maximized is independent of t then so are $k(M), \chi(M)$. We thus get that for all $t \ge 1$

$$\begin{aligned}\tilde{f}_t^{(i)}(M) &= c_t(x_t(M;\Psi_{\tau_i},\tilde{w}), u_t(M;\tilde{w})) - \alpha\left\|V_{\tau_i}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\right\|_\infty \\ &= c_t(x_t(M;\Psi_{\tau_i},\tilde{w}), u_t(M;\tilde{w})) - \alpha\chi(M)\cdot\left(V_{\tau_i}^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\right)_{k(M)} \\ &= \tilde{f}_t^{(i)}(M;k(M),\chi(M)).\end{aligned}$$

We thus use Lemma 20 with $\delta/24T$ and take a union bound over the epochs to get that with probability at least $1-\delta/24$ simultaneously for all $1 \le \tau \le T, 1 \le i \le N$

$$\sum_{t=1}^{\tau} \tilde{f}_t^{(i)}(M_t^{(i)}) - \tilde{f}_t^{(i)}(M_\star) \le 151[(\bar{G} + G_f(\Psi_{\tau_i})^2\bar{G}^{-1})R_{\mathcal{M}} + C]\sqrt{T\log(6d_\Psi^2)\log\frac{48T^2}{\delta}}, \quad \text{and}$$

$$\sum_{t=1}^{\tau}\|M_t^{(i)} - M_{t-1}^{(i)}\| \le (270 + 2G_f(\Psi_{\tau_i})\bar{G}^{-1})R_{\mathcal{M}}\sqrt{T\log(6d_\Psi^2)\log\frac{48T^2}{\delta}}.$$

Next, suppose that the events of Lemmas 8 and 9 each holds with $\delta/24$. This occurs with probability at least $1-\delta/12$ and implies $\|(\Psi_{\tau_{i,1}}\ I)\|_F \le 17R_B\kappa^2\sqrt{\gamma^{-3}(d_x+d_u)(d_x^2\kappa^2 + d_uR_B^2)\log\frac{24T^2}{\delta}}$. Pluuging this into Lemma 12 we get that

$$\begin{aligned}C_f^{\max} &\le 5R_{\mathcal{M}}WH\max\{\max_{t\le T}\|(\Psi_t\ I)\|_F, \kappa\gamma^{-1}R_B\} \le 5\alpha/(H\sqrt{d_x(d_x+d_u)}) \\ G_f(\Psi_{\tau_i}) &\le G_f^{\max} \le 2WH\max_{t\le T}\|(\Psi\ I)\|_F \le \alpha\sqrt{2}/(R_{\mathcal{M}}\sqrt{H}),\end{aligned}$$

and therefore

$$\begin{aligned}[(\bar{G} + G_f(\Psi_{\tau_i})^2\bar{G}^{-1})R_{\mathcal{M}} + C] \le 2\bar{G}R_{\mathcal{M}} + C &\le 2\alpha\sqrt{2/H}R_{\mathcal{M}}^{-1}R_{\mathcal{M}} + \alpha\sqrt{2/H}[1 + R_{\mathcal{M}}^{-1}\sqrt{d_x}] \\ &\le \alpha\sqrt{2/H}[3 + R_{\mathcal{M}}^{-1}\sqrt{d_x}] \\ &\le \alpha\sqrt{8d_x/H}.\end{aligned}$$

Now, notice that for $\tau_i + 2H \le t \le \tau_{i+1} - 1$ we have that $\tilde{f}_t$ in Algorithm 1 coincide with $\tilde{f}_{t+1-(\tau_i+2H)}^{(i)}$. Moreover, $\eta_G = 2R_{\mathcal{M}}/(\bar{G}\sqrt{T})$ and the scaling factor C_M in Algorithm 1 satisfies

$$\begin{aligned}C_M(\Psi_{\tau_i}) &= \sqrt{8}WR_{\mathcal{M}}H\|\Psi\|_F + \alpha\sqrt{2/H}[(2 + R_{\mathcal{M}}^{-1}\sqrt{d_x}] \\ &= 2G_f(\Psi_{\tau_i})R_{\mathcal{M}} + \alpha\sqrt{2/H}[(1 + R_{\mathcal{M}}^{-1}\sqrt{d_x}] = 2GR + C,\end{aligned}$$

which implies that Algorithm 1 runs the same procedure as in Lemma 20 and thus $M_t = M^{(i)}_{t+1-(\tau_i+2H)}$. We conclude that

$$\begin{aligned}
R_{3,2} &\le 2G_f^{\max} R_{\mathcal{M}} N + \sum_{i=1}^{N} \sum_{t=\tau_i+2H}^{\tau_{i+1}-1} \tilde{f}_t(M_t) - \tilde{f}_t(M_\star) \\
&= 2G_f^{\max} R_{\mathcal{M}} N + \sum_{i=1}^{N} \sum_{t=1}^{\tau_{i+1}-(\tau_i+1)} \tilde{f}^{(i)}_t(M^{(i)}_t) - \tilde{f}^{(i)}_t(M_\star) \\
&\le \alpha\sqrt{8/H} N \left[1 + 151\sqrt{T d_x \log(6 d_\Psi^2) \log\frac{48T^2}{\delta}}\right] \\
&\le 860\alpha(d_x+d_u)\log(T)\sqrt{THd_x \log(6d_\Psi^2)\log\frac{48T^2}{\delta}} \\
&\le 860\alpha(d_x+d_u)\sqrt{THd_x \log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}},
\end{aligned}$$

where in the first inequality we bounded the loss at the first $2H$ rounds of each epoch using the Lipschitz property of $\tilde{f}_t$ (see Lemma 12), and in the third we bounded $N \le 2H(d_x+d_u)\log(T)$ using Lemma 13. We also get that

$$\begin{aligned}
\sum_{t=1}^{T} \|M_t - M_{t-1}\| &\le 2R_{\mathcal{M}} N + \sum_{i=1}^{N} \sum_{t=\tau_i+2H}^{\tau_{i+1}-1} \|M_t - M_{t-1}\| \\
&\le N\left[2R_{\mathcal{M}} + 272 R_{\mathcal{M}}\sqrt{T\log(6d_\Psi^2)\log\frac{48T^2}{\delta}}\right] \\
&\le 548 R_{\mathcal{M}}(d_x+d_u)H\log(T)\sqrt{T\log(6d_\Psi^2)\log\frac{48T^2}{\delta}} \\
&\le 548 R_{\mathcal{M}}(d_x+d_u)H\sqrt{T\log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}}.
\end{aligned}$$

Plugging this back into the bound for $R_{3,1} + R_{3,3}$ we get

$$\begin{aligned}
R_{3,1} + R_{3,3} &\le 8C_f^{\max}\sqrt{TH\log\frac{24T}{\delta}} + 4G_f^{\max} H \sum_{t=1}^{T}\|M_t - M_{t-1}\|_F \\
&\le 40\alpha\sqrt{TH^{-1}d_x^{-1}(d_x+d_u)^{-1}\log\frac{24T}{\delta}} + \alpha\sqrt{32H} R_{\mathcal{M}}^{-1}\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F \\
&\le 40\alpha\sqrt{TH^{-1}d_x^{-1}(d_x+d_u)^{-1}\log\frac{24T}{\delta}} + 3100\alpha(d_x+d_u)\sqrt{TH^3\log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}} \\
&\le 3140\alpha(d_x+d_u)\sqrt{TH^3\log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}}.
\end{aligned}$$

Combining the bounds for $R_{3,1}, R_{3,2}, R_{3,3}$ we conclude that

$$\begin{aligned}
R_3 &\le 3140\alpha(d_x+d_u)\sqrt{TH^3\log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}} \\
&\quad + 860\alpha(d_x+d_u)\sqrt{THd_x\log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}} \\
&\le 4000\alpha(d_x+d_u)\sqrt{TH^3 d_x\log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}}.
\end{aligned}$$

Taking a union bound over all the events throughout the lemma we have that their intersection occurs with probability at least $1-\delta/4$. ■

C.6 Side lemmas

Proof of Lemma 13. The algorithm ensures that

$$\det(V_T) \geq \det(V_{\tau_{N,1}}) \geq 2\det(V_{\tau_{N-1,1}}) \ldots \geq 2^{N-1}\det V_1,$$

and changing sides, and taking the logarithm we conclude that

$$\begin{aligned}
N &\leq 1 + \log(\det(V_T)/\det(V)) \\
&= 1 + \log\det(V^{-1/2}V_{T+1}V^{-1/2}) \\
&\leq 1 + (d_x + d_u)H\log\|V^{-1/2}V_T V^{-1/2}\| && (\det(A) \leq \|A\|^d) \\
&\leq 1 + (d_x + d_u)H\log\left(1 + \frac{1}{\lambda_\Psi}\sum_{t=1}^{T-1}\|\rho_t\|^2\right) && \text{(triangle inequality)} \\
&\leq 1 + (d_x + d_u)H\log T \\
&\leq 2(d_x + d_u)H\log T, && (T \geq 3)
\end{aligned}$$

where the second to last inequality holds since $\|\rho_t\|^2 \leq \lambda_\Psi$ by Lemma 10. ■

Proof of Lemma 18. First, we use Lemma 11 and our choice of $\lambda_\Psi = 2W^2R_\mathcal{M}^2H^2$ to get that

$$\begin{aligned}
&\sum_{t=1}^{T}\sqrt{2\mathbb{E}\left[\|V_{\tau_t}^{-1/2}(\rho_{t-1}(M_t; w) - \rho_{t-1})\|^2 \;\mid\; \mathcal{F}_{t-2H}\right]} \\
&\quad\leq \sqrt{\frac{2}{\lambda_\Psi}}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[\|\rho_{t-1}(M_t; w) - \rho_{t-1}\|^2 \;\mid\; \mathcal{F}_{t-2H}\right]} \\
&\quad\leq (WR_\mathcal{M}H)^{-1}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[2R_\mathcal{M}^2H\left[\sum_{h=1}^{2H}\|w_{t-h} - \hat{w}_{t-h}\|^2 + \sum_{h=1}^{H}\|M_{t-h} - M_t\|_F^2\right] \;\mid\; \mathcal{F}_{t-2H}\right]} \\
&\quad\leq \sqrt{\frac{2}{HW^2}}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[\sum_{h=1}^{2H}\|w_{t-h} - \hat{w}_{t-h}\|^2 \;\mid\; \mathcal{F}_{t-2H}\right] + \sum_{h=1}^{H}\|M_{t-h} - M_t\|_F^2} \\
&\quad\leq \sqrt{\frac{2}{HW^2}}\left[\sqrt{T\sum_{h=1}^{2H}\sum_{t=1}^{T}\mathbb{E}\left[\|w_{t-h} - \hat{w}_{t-h}\|^2 \mid \mathcal{F}_{t-2H}\right]} + \sum_{t=1}^{T}\sum_{h=1}^{H}\|M_{t-h} - M_t\|_F\right],
\end{aligned}$$

where the second to last transition also used the fact that for $h = 0, \ldots H$ we have M_{t-h} is $\mathcal{F}_{t-2H}$ measurable, and the last transition used both Jensen's inequality and $\|x\|_2 \leq \|x\|_1$. Now, we seek to apply Lemma 19 for the disturbances. Indeed, we have that $\|w_{t-h} - \hat{w}_{t-h}\|^2$ are non-negative, bounded by $4W^2$, and $\mathcal{F}_t$ measurable for all $1 \leq t - h \leq T$. Using Lemma 19 with $\delta/2$ we get that with probability at least $1 - \delta/2$

$$\sum_{t=1}^{T}\mathbb{E}\left[\|w_{t-h} - \hat{w}_{t-h}\|^2 \mid \mathcal{F}_{t-2H}\right] \leq 2\sum_{t=1}^{T}\|w_{t-h} - \hat{w}_{t-h}\|^2 + 32W^2H\log\frac{4T}{\delta}.$$

Next, we also have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\|M_{t-h} - M_t\|_F \leq \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{h'=0}^{h-1}\|M_{t-h'} - M_{t-(h'+1)}\|_F \leq H^2\sum_{t=1}^{T}\|M_t - M_{t-1}\|_F,$$

and thus plugging both of these into the above we get that

$$\begin{aligned}
&\sum_{t=1}^{T} \sqrt{2\mathbb{E}\left[\|V_{\tau_t}^{-1/2}(\rho_{t-1}(M_t; w) - \rho_{t-1})\|^2 \;\mid\; \mathcal{F}_{t-2H}\right]} \\
&\le \sqrt{\frac{2}{HW^2}} \left[\sqrt{T \sum_{h=1}^{2H} \left[32W^2 H \log\frac{4T}{\delta} + 2\sum_{t=1}^{T} \|w_{t-h} - \hat{w}_{t-h}\|^2\right]} + H^2 \sum_{t=1}^{T} \|M_t - M_{t-1}\|_F\right] \\
&\le \sqrt{128TH \log\frac{4T}{\delta}} + \sqrt{8TW^{-2} \sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2} + \sqrt{\frac{2H^3}{W^2}} \sum_{t=1}^{T} \|M_t - M_{t-1}\|_F.
\end{aligned}$$

Next, we seek to apply Lemma 19 to $\|V_{\tau_t}^{-1/2}\rho_{t-1}\|^2$. Indeed they are non-negative, $\mathcal{F}_t$ measurable and by Lemma 11 satisfy

$$\|V_{\tau_t}^{-1/2}\rho_{t-1}\|^2 \le \lambda_\Psi^{-1}\|o_{t-1}\|^2 \le (2W^2 R_\mathcal{M}^2 H^2)^{-1} 2W^2 R_\mathcal{M}^2 H^2 = 1$$

Applying Lemma 19 with $\delta/2$ we get that with probability at least $1 - \delta/2$

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\|V_{\tau_t}^{-1/2}\rho_{t-1}\|^2 \;\mid\; \mathcal{F}_{t-2H}\right] &\le 2\sum_{t=1}^{T} \left(\|V_{\tau_t}^{-1/2}\rho_{t-1}\|^2\right) + 8H \log\frac{4T}{\delta} \\
&\le 2\sum_{t=1}^{T} \left(\|V_{\tau_{t+2H}}^{-1/2}\rho_{t-1}\|^2\right) + 4HN + 8H \log\frac{4T}{\delta} \\
&\le 4\sum_{t=1}^{T} \left(\|V_{t-1}^{-1/2}\rho_{t-1}\|^2\right) + 4HN + 8H \log\frac{4T}{\delta} \\
&\le (20 + 8H)H(d_x + d_u)\log(T) + 8H \log\frac{4T}{\delta} \\
&\le 18H^2(d_x + d_u)\log\frac{4T^2}{\delta},
\end{aligned}$$

where the second transition used the fact that there are at most $2H$ times per epoch for which $\tau_t \neq \tau_{t+2H}$, i.e., is not the start of the current epoch, the third transition used (Cohen et al., 2019, Lemma 27), which states that for $V_1 \succeq V_2 \succeq 0$ we have $o^\mathsf{T} V_1 o \le (o^\mathsf{T} V_2 o)\det(V_1)/\det(V_2)$, and the fourth transition also used Lemma 23 to bound the harmonic sum. Using Jensen's inequality, we get that

$$\begin{aligned}
\sum_{t=1}^{T} \sqrt{\mathbb{E}\left[2\|V_{\tau_t}^{-1/2}\rho_{t-1}\|^2 \mid \mathcal{F}_{t-2H}\right]} &\le \sqrt{2T \sum_{t=1}^{T} \mathbb{E}\left[\|V_{\tau_t}^{-1/2}\rho_{t-1}\|^2 \mid \mathcal{F}_{t-2H}\right]} \\
&\le \sqrt{36TH^2(d_x + d_u)\log\frac{4T^2}{\delta}}.
\end{aligned}$$

Taking a union bound and combining the last two inequalities, we conclude that with probability at least $1 - \delta$

$$\begin{aligned}
&\sum_{t=1}^{T} \sqrt{\mathbb{E}\left[\|V_{\tau_t}^{-1/2}\rho_{t-1}(M_t; w)\|^2 \;\mid\; \mathcal{F}_{t-2H}\right]} \\
&\le \sum_{t=1}^{T} \sqrt{\mathbb{E}\left[2\|V_{\tau_t}^{-1/2}\rho_{t-1}\|^2 \mid \mathcal{F}_{t-2H}\right]} + \sum_{t=1}^{T} \sqrt{2\mathbb{E}\left[\|V_{\tau_t}^{-1/2}(\rho_{t-1}(M_t; w) - \rho_{t-1})\|^2 \;\mid\; \mathcal{F}_{t-2H}\right]} \\
&\le 13H\sqrt{T(d_x + d_u)\log\frac{4T^2}{\delta}} + \sqrt{8TW^{-2} \sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2} + \sqrt{\frac{2H^3}{W^2}} \sum_{t=1}^{T} \|M_t - M_{t-1}\|_F. \qquad \blacksquare
\end{aligned}$$

The following lemma combines the bounds in Lemmas 14 to 16 to complete the final regret bound in Theorem 7.

Lemma 21. *We have that with probability at least* $1-\delta$

$$\mathrm{Regret}_T(\pi) \le 37483 W R_{\mathcal{M}}^2 R_B^2 \kappa^3 \gamma^{-6} (d_x^2\kappa^2 + d_u R_B^2) \log^6 \frac{48T^2}{\delta} \sqrt{T(d_x+d_u)^3 \log(2d_\Psi)}.$$

Proof. Suppose that the events of Lemmas 14 to 16 hold. By a union bound, this holds with probability at least $1-\delta$. Now, we simplify each of the terms before deriving the final bound. Recall from Theorem 7 that

$$\begin{aligned}
\alpha &= 21 W R_{\mathcal{M}} R_B \kappa^2 (d_x+d_u) \sqrt{H^3 \gamma^{-3} (d_x^2\kappa^2 + d_u R_B^2) \log \frac{24T^2}{\delta}} \\
&\le 21 W R_{\mathcal{M}} R_B \kappa^2 \gamma^{-3} (d_x+d_u) \sqrt{(d_x^2\kappa^2 + d_u R_B^2) \log^4 \frac{24T^2}{\delta}}.
\end{aligned}$$

Now, plugging the movement cost bound from Lemma 16 into Lemma 14, we have

$$\begin{aligned}
R_1 + R_5 &\le 24 \frac{\kappa^2}{\gamma^2} W R_B^2 R_{\mathcal{M}}^2 H \sqrt{T(d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{4T}{\delta}} + \frac{\kappa}{\gamma^2} R_B W \sqrt{H} \sum_{t=1}^{T} \|M_t - M_{t-1}\| \\
&\le 24 \frac{\kappa^2}{\gamma^2} W R_B^2 R_{\mathcal{M}}^2 H \sqrt{T(d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{4T}{\delta}} \\
&\quad + 548 \frac{\kappa}{\gamma^2} W R_{\mathcal{M}} R_B (d_x+d_u) \sqrt{T H^3 \log(6 d_\Psi^2) \log^3 \frac{48T^2}{\delta}} \\
&\le 554 \frac{\kappa^2}{\gamma^2} W R_B^2 R_{\mathcal{M}}^2 \sqrt{T H^3 (d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2) \log(6 d_\Psi^2) \log^3 \frac{48T^2}{\delta}} \\
&\le 554 W R_B^2 R_{\mathcal{M}}^2 \kappa^2 \gamma^{-4} \log^3\left(\frac{48T^2}{\delta}\right) \sqrt{T(d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2) \log(6 d_\Psi^2)} \\
&\le W R_{\mathcal{M}}^2 R_B^2 \kappa^3 \gamma^{-8} (d_x^2\kappa^2 + d_u R_B^2) \log^6\left(\frac{48T^2}{\delta}\right) \sqrt{T d_x (d_x+d_u)^3 \log(6 d_\Psi^2)},
\end{aligned}$$

where the third inequality used the fact that $T \ge 8$ and thus $\log 48T^2 \ge 8$. Next, we do the same for R_2, R_4 to get that

$$\begin{aligned}
R_2 + R_4 &\le 65 \alpha R_{\mathcal{M}} R_B \kappa \gamma^{-1} H \sqrt{T(d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2) \log \frac{48T^2}{\delta}} + \alpha \sqrt{\frac{8H^3}{W^2}} \sum_{t=1}^{T} \|M_t - M_{t-1}\|_F \\
&\le 65 \alpha R_{\mathcal{M}} R_B \kappa \gamma^{-2} \sqrt{T(d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2) \log^3 \frac{48T^2}{\delta}} \\
&\quad + 1550 \alpha W^{-1} R_{\mathcal{M}} (d_x+d_u) \sqrt{T H^5 \log(6 d_\Psi^2) \log^3 \frac{48T^2}{\delta}} \\
&\le 1560 \alpha R_{\mathcal{M}} R_B \kappa \gamma^{-2} \sqrt{T H^5 (d_x+d_u)(d_x^2\kappa^2 + d_u R_B^2) \log(6 d_\Psi^2) \log^3 \frac{48T^2}{\delta}} \\
&\le 32760 W R_{\mathcal{M}}^2 R_B^2 \kappa^3 \gamma^{-5} (d_x^2\kappa^2 + d_u R_B^2) \sqrt{T H^5 (d_x+d_u)^3 \log(6 d_\Psi^2) \log^7 \frac{48T^2}{\delta}} \\
&\le 32760 W R_{\mathcal{M}}^2 R_B^2 \kappa^3 \gamma^{-8} (d_x^2\kappa^2 + d_u R_B^2) \log^6\left(\frac{48T^2}{\delta}\right) \sqrt{T d_x (d_x+d_u)^3 \log(6 d_\Psi^2)}.
\end{aligned}$$

Next, we plug in α into R_3 to get that

$$
\begin{aligned}
R_3 &\le 4000\alpha(d_x + d_u)\sqrt{TH^3 d_x \log(6d_\Psi^2)\log^3\frac{48T^2}{\delta}} \\
&\le 84000 W R_{\mathcal{M}} R_B \kappa^2 \gamma^{-5}(d_x + d_u)^2 \log^5\left(\frac{48T^2}{\delta}\right)\sqrt{Td_x(d_x^2\kappa^2 + d_u R_B^2)\log(6d_\Psi^2)} \\
&\le 10500 W R_{\mathcal{M}}^2 R_B^2 \kappa^3 \gamma^{-8}(d_x^2\kappa^2 + d_u R_B^2)\log^6\left(\frac{48T^2}{\delta}\right)\sqrt{Td_x(d_x + d_u)^3\log(6d_\Psi^2)}.
\end{aligned}
$$

Combining the above, we conclude that

$$
\mathrm{Regret}_T(\pi) \le 43261 W R_{\mathcal{M}}^2 R_B^2 \kappa^3 \gamma^{-8}(d_x^2\kappa^2 + d_u R_B^2)\log^6\left(\frac{48T^2}{\delta}\right)\sqrt{Td_x(d_x + d_u)^3\log(6d_\Psi^2)}. \qquad \blacksquare
$$

Appendix D. Theorem 3: Deferred details

Here we complete the deferred details in the proof of Theorem 3.

We start by proving the optimism lemma

Proof. The proof follows standard arguments (see e.g. Lemma 3 in Cassel et al. (2022)). We first use the Lipschitz assumption to get

$$
\begin{aligned}
|\ell_t(Q_\star a) - \ell_t(\widehat{Q}_{\tau_{i(t)}} a)| &\le \|(Q_\star - \widehat{Q}_{\tau_{i(t)}})a\| \\
&\le \|Q_\star - \widehat{Q}_{\tau_{i(t)}}\|_{V_{\tau_{i(t)}}} \|V_{\tau_{i(t)}}^{-1/2} a\| \\
&\le \frac{\alpha}{\sqrt{d_a}}\|V_{\tau_{i(t)}}^{-1/2} a\| \\
&\le \alpha\|V_{\tau_{i(t)}}^{-1/2} a\|_\infty,
\end{aligned}
$$

where the second and third transitions also used the estimation error and that $\|a\| \le \sqrt{d_a}\|a\|_\infty$. We thus have on one hand,

$$
\ell_t(Q_\star a) \ge \ell_t(\widehat{Q}_{\tau_{i(t)}} a) - \alpha\|V_{\tau_{i(t)}}^{-1/2} a\|_\infty = \bar{\ell}_t(a),
$$

and on the other hand we also have

$$
\ell_t(Q_\star a) \le \ell_t(\widehat{Q}_{\tau_{i(t)}} a) + \alpha\|V_{\tau_{i(t)}}^{-1/2} a\|_\infty = \bar{\ell}_t(a) + 2\alpha\|V_{\tau_{i(t)}}^{-1/2} a\|_\infty \le \bar{\ell}_t(a) + 2\alpha\sqrt{a^\mathsf{T} V_{\tau_{i(t)}}^{-1} a},
$$

where the last step also used $\|a\|_\infty \le \|a\|$. $\blacksquare$

Next, we state the following high-probability error bound for least squares estimation, that bounds the error of our estimates $\widehat{Q}_t$ of $Q_\star$, and as such also satisfies the condition of Lemma 4.

Lemma 22 (Abbasi-Yadkori and Szepesvári (2011)). *Let $\Delta_t = Q_\star - \widehat{Q}_t$, and suppose that $\|a_t\|^2 \le \lambda = R_a^2$, $T \ge d_y$. With probability at least $1-\delta$, we have for all $t \ge 1$*

$$
\|\Delta_t\|_{V_t}^2 \le \mathrm{Tr}\big(\Delta_t^\mathsf{T} V_t \Delta_t\big) \le 8W^2 d_y^2 \log\frac{T}{\delta} + 2R_a^2 R_Q^2 \le \frac{\alpha^2}{d_a}.
$$

As an immediate corollary, when Lemma 22 holds we also get that

$$
\|\widehat{Q}_t\| \le \|\widehat{Q}_t\|_F \le \lambda^{-1/2}\|\Delta_t\|_{V_t} + \|Q_\star\|_F \le \alpha(\lambda d_a)^{-1/2} + R_Q \tag{10}
$$

Next, is a well-known bound on harmonic sums (see, e.g., Cohen et al., 2019). This is used to show that the optimistic and true losses are close on the realized predictions (proof in Appendix E).

Lemma 23. *Let $a_t \in \mathbb{R}^{d_a}$ be a sequence such that $\|a_t\|^2 \le \lambda$, and define $V_t = \lambda I + \sum_{s=1}^{t-1} a_s a_s^\mathsf{T}$. Then*

$$\sum_{t=1}^{T} a_t^\mathsf{T} V_t^{-1} a_t \le 5 d_a \log T.$$

Proof. Notice that $a_t^\mathsf{T} V_t^{-1} a_t \le \|a_t\|^2 / \lambda^2 \le 1$, and so by (Cohen et al., 2019, Lemma 26) we get that $a_t V_t^{-1} a_t \le \log(\det(V_{t+1}) / \det(V_t))$. We conclude that

$$\begin{aligned}
\sum_{t=1}^{T} a_t^\mathsf{T} V^{-1} a_t &\le 2 \sum_{t=1}^{T} a_t^\mathsf{T} V_t^{-1} a_t \\
&\le 4 \sum_{t=1}^{T} \log(\det(V_{t+1}) / \det(V_t)) \\
&= 4 \log(\det(V_{T+1}) / \det(V)) && \text{(telescoping sum)} \\
&= 4 \log \det(V^{-1/2} V_{T+1} V^{-1/2}) \\
&\le 4 d_a \log \|V^{-1/2} V_{T+1} V^{-1/2}\| && (\det(A) \le \|A\|^d) \\
&\le 4 d_a \log\left(1 + \frac{1}{\lambda^2} \sum_{s=1}^{T} \|a_s\|^2\right) && \text{(triangle inequality)} \\
&\le 4 d_a \log(T+1) \\
&\le 5 d_a \log T. && (T \ge 4)
\end{aligned}$$

■

Next, the following lemma bounds the number of epochs.

Lemma 24. *We have that $N \le 2 d_a \log T$.*

Proof. The algorithm ensures that

$$\det(V_T) \ge \det(V_{\tau_N}) \ge 2 \det(V_{\tau_{N-1}}) \ldots \ge 2^{N-1} \det V_1,$$

and changing sides, and taking the logarithm we conclude that

$$\begin{aligned}
N &\le 1 + \log(\det(V_T) / \det(V)) \\
&= 1 + \log \det(V^{-1/2} V_{T+1} V^{-1/2}) \\
&\le 1 + d_a \log \|V^{-1/2} V_T V^{-1/2}\| && (\det(A) \le \|A\|^d) \\
&\le 1 + d_a \log\left(1 + \frac{1}{\lambda} \sum_{t=1}^{T-1} \|a_t\|^2\right) && \text{(triangle inequality)} \\
&\le 1 + d_a \log T \\
&\le 2 d_a \log T, && (T \ge 3)
\end{aligned}$$

where the second to last inequality holds since $\|a_t\|^2 \le R_a^2 = \lambda$. ■

Appendix E. Technical Lemmas and Proofs

E.1 OCO Results

Let $\{f_t\}_{t \ge 1}$ be a sequence of functions and $\mathcal{S} \in \mathbb{R}^d$ be a convex set. The following lemma states the regret guarantee of the Online Gradient Descent (OGD), Zinkevich (2003) update rule, given by:

$$x_{t+1} = \Pi_{\mathcal{S}}[x_t - \eta \nabla f_t(x_t)],$$

where $\Pi_{\mathcal{S}}$ is the ℓ_2 projection onto $\mathcal{S}$, and $x_0 \in \mathcal{S}$ is be chosen arbitrarily.

Lemma 25. *Running OGD with $\eta = R/(\bar{G}\sqrt{T})$ on a decision set S with diameter R, and $G-$Lipschitz convex loss functions f_t gives*

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x) \leq \frac{1}{2} R(\bar{G} + G^2\bar{G}^{-1})\sqrt{T}$$

for all $T \geq 1$, $x \in S$.

Proof. We use a standard result for OGD Zinkevich (2003) to get that for all $x \in S$

$$\sum_{t=1}^{T}[f_t(x_t) - f_t(x)] \leq \frac{R^2}{2\eta} + \frac{1}{2}\eta G^2 T = \frac{1}{2}R(\bar{G} + G^2\bar{G}^{-1})\sqrt{T}. \quad \blacksquare$$

Let $\{\ell_t\}_{t\geq 1}$ be a sequence of loss vectors in $\mathbb{R}^d$ with $\ell_{t,i}$ the $i-$th coordinate of ℓ_t. The following lemma gives a high probability regret guarantee for the Hedge algorithm (see e.g.,Chernov and Zhdanov (2010)), also known as Multiplicative Weights (MW), which draws $i_t \sim p_t$ where:

$$p_{t+1,i} \propto p_{t,i}e^{-\eta\ell_{t,i}} \propto e^{-\eta\sum_{s=1}^{t}\ell_{s,i}},$$

and p_1 is uniform.

Lemma 26. *Suppose that we play Hedge over loss vectors $\ell_t \in \mathbb{R}^d$ chosen by an oblivious adversary, and that satisfy $\|\ell_t\|_\infty \leq C$. If $\eta = \sqrt{\log(d)/4(T\bar{C}^2)}$ then with probability at least $1-\delta$*

$$\sum_{t=1}^{T} \ell_{t,i(t)} - \ell_{t,i^*} \leq \left(\bar{C} + \bar{C}^{-1}C^2\right)\sqrt{6T\log\frac{d}{\delta}}.$$

Proof. Let $\mathcal{F}_t$ be the filtration defined by all random variables up to time t, not including the randomized choice of expert. Then we have that $\mathbb{E}[\ell_{t,i(t)} \mid \mathcal{F}_t] = \sum_{i=1}^{2d} p_{t,i}\ell_{t,i}$. Moreover, $Z_t = \ell_{t,i(t)} - \mathbb{E}[\ell_{t,i(t)} \mid \mathcal{F}_t]$ is a martingale difference sequence with $|Z_t| \leq 2C$. We thus invoke the Azuma–Hoeffding inequality to get with probability at least $1-\delta$

$$\begin{aligned}
\sum_{t=1}^{T} \ell_{t,i(t)} - \ell_{t,i^*} &= \sum_{t=1}^{T} \ell_{t,i(t)} - \mathbb{E}\left[\ell_{t,i(t)}\middle|\mathcal{F}_t\right] + \sum_{t=1}^{T}\sum_{i=1}^{2d} p_{t,i}\ell_{t,i} - \ell_{t,i^*} \\
&\leq \sqrt{8TC^2\log\frac{1}{\delta}} + \sum_{t=1}^{T}\sum_{i=1}^{2d} p_{t,i}\ell_{t,i} - \ell_{t,i^*} && \text{(Azuma-Hoeffding)} \\
&\leq \sqrt{8TC^2\log\frac{1}{\delta}} + \frac{\log d}{\eta} + 4\eta C^2 T && \text{(Hedge regret bound (e.g., Chernov and Zhdanov, 2010))} \\
&= 2C\sqrt{2T\log\frac{1}{\delta}} + 2\left(\bar{C} + \bar{C}^{-1}C^2\right)\sqrt{T\log d} \\
&\leq \left(\bar{C} + \bar{C}^{-1}C^2\right)\left(\sqrt{\log\frac{1}{\delta}} + \sqrt{2\log d}\right)\sqrt{2T} && (2C \leq \left(\bar{C} + \bar{C}^{-1}C^2\right) \text{ for all } C, \bar{C} > 0) \\
&\leq \left(\bar{C} + \bar{C}^{-1}C^2\right)\sqrt{6T\log\frac{d}{\delta}}. && (\sqrt{2x} + \sqrt{y} \leq \sqrt{3(x+y)} \text{ by AM-GM inequality})
\end{aligned}$$

$\blacksquare$

The following lemma states the guarantees for the $\text{BFPL}_\delta^\star$ algorithmAltschuler and Talwar (2018), which is essentially a batched version of Follow the Perturbed Leader (FPL) Kalai and Vempala (2005), which restarts with fresh randomness every time that a set number of leader switches occur.

Lemma 27 (Altschuler and Talwar (2018)). *Suppose that we run BFPL$^\star_\delta$ with losses bounded in $[0,1]^n$. Then with probability at least $1-\delta$*

$$\sum_{t=1}^{T} \ell_{t,i(t)} - \ell_{t,i^*} \le 150\sqrt{T \log n \log \frac{2}{\delta}} \qquad \textit{and} \qquad \#\{\textit{switches}\} \le 135\sqrt{T \log n \log \frac{2}{\delta}}$$

The following result describes the guarantees of our meta-algorithm for computationally-efficient regret minimization of a particular non-convex structure.

Lemma (restatement of Lemma 20). *Let $f_t(M,k,\chi)$ be a sequence of oblivious loss functions that are convex and G Lipschitz in M, and have a convex decision set S with diameter $2R$. Let $f_t(M) = \min_{k\in[d],\chi\in\{-1,1\}} f_t(M,k,\chi)$ and consider the update rule that at time t:*

1. *define loss vector ℓ_t such that $(\ell_t)_{k,\chi} = f_t(M_t(k,\chi);k,\chi)/(2GR+C)$*
2. *update experts: $M_{t+1}(k,\chi) = \Pi_{\mathcal{M}}[M_t(k,\chi) - \eta\nabla_M f_t(M_t(k,\chi);k,\chi)]$*
3. *update prediction: $(k_{t+1},\chi_{t+1}) = BFPL^*_\delta(\ell_t)$ and set $M_{t+1} = M_{t+1}(k_{t+1},\chi_{t+1})$*

where $\eta = 2R/\bar{G}\sqrt{T}$ and $C \ge 0$. Suppose that:

- i *$f_t(M;k,\chi) \le f_t(M) + C$ for all $M \in S, k \in [d], \chi \in \{\pm 1\}$;*
- ii *There exist $k(M),\chi(M)$ independent of t such that $f_t(M) = f_t(M,k(M),\chi(M))$.*

Then with probability at least $1-\delta$ we have that for all $\tau \le T$

$$\sum_{t=1}^{\tau} f_t(M_t) - f_t(M) \le 151[(\bar{G} + G^2\bar{G}^{-1})R + C]\sqrt{T \log(2d) \log \frac{2T}{\delta}}$$
$$\sum_{t=1}^{\tau} \|M_t - M_{t-1}\| \le (270 + 2G\bar{G}^{-1})R\sqrt{T \log(2d) \log \frac{2T}{\delta}}.$$

Proof. $M_t(k,\chi)$ are exactly the iterates of running Online Gradient Descent (OGD) on the functions $f_t(\cdot;k,\chi)$, which are convex and G-Lipschitz. A classic result (see Lemma 25) then gives us that for all $M \in \mathcal{M}$ and $\tau \le T$

$$\sum_{t=1}^{\tau} f_t(M_t(k,\chi);k,\chi) - f_t(M;k,\chi) \le (\bar{G} + G^2\bar{G}^{-1})R\sqrt{T},$$

Next, we verify that ℓ_t satisfy the conditions for BFPL (Lemma 27). First, since OGD is deterministic, the iterates $M_t(k,\chi)$ and thus the losses ℓ_t are deterministic functions of the loss functions $f_1,\ldots,f_t$. Since the latter are oblivious, so are the loss vectors ℓ_t. Next, notice that BFPL is invariant to a constant shift in the loss vectors. The procedure described in the lemma is equivalent to using the losses

$$(\ell_t)_{k,\chi} = \frac{f_t(M_t(k,\chi);k,\chi) - f_t(\bar{M})}{2GR+C},$$

where $\bar{M} \in \arg\min_{M\in S} f_t(M)$. We show that $\ell_t \in [0,1]$. By definition of f_t and $\bar{M}$ we have

$$(\ell_t)_{k,\chi} = \frac{f_t(M_t(k,\chi);k,\chi) - f_t(\bar{M})}{2GR+C} \ge \frac{f_t(M_t(k,\chi)) - f_t(\bar{M})}{2GR+C} \ge 0.$$

On the other hand, using that f_t is G Lipschitz and the assumption in i we have

$$(\ell_t)_{k,\chi} = \frac{f_t(M_t(k,\chi);k,\chi) - f_t(\bar{M})}{2GR+C} \le \frac{2GR + f_t(\bar{M};k,\chi) - f_t(\bar{M})}{2GR+C} \le \frac{2GR+C}{2GR+C} \le 1.$$

We thus use Lemma 27 with δ/T, and a union bound to get that with probability at least $1-\delta$

$$\sum_{t=1}^{\tau} f_t(M_t(k_t,\chi_t);k_t,\chi_t) - f_t(M_t(k,\chi);k,\chi) \le 150(2GR+C)\sqrt{T\log(2d)\log\frac{2T}{\delta}}$$

$$\#\{\text{switches}\}_T \le 135\sqrt{T\log(2d)\log\frac{2T}{\delta}},$$

for all $k\in[d], \chi\in\{\pm1\}$, and $\tau\le T$. Now, for ease of notation denote $k^*,\chi^* = k(M),\chi(M)$ where these are taken from the lemma's assumptions. Then we conclude that with probability at least $1-\delta$ we have that for all $M\in\mathcal{M}$

$$\begin{aligned}
\sum_{t=1}^{\tau} f_t(M_t) - f_t(M) &\le \sum_{t=1}^{\tau} f_t(M_t;k_t,\chi_t) - f_t(M) && (f_t(\cdot)\le f_t(\cdot;k,\chi))\\
&= \sum_{t=1}^{\tau} f_t(M_t;k_t,\chi_t) - f_t(M;k^*,\chi^*)\\
&= \sum_{t=1}^{\tau}(f_t(M_t(k_t,\chi_t);k_t,\chi_t) - f_t(M_t(k^*,\chi^*);k^*,\chi^*))\\
&+ \sum_{t=1}^{\tau}(f_t(M_t(k^*,\chi^*);k^*,\chi^*) - f_t(M;k^*,\chi^*))\\
&\le 151[(\bar{G}+G^2\bar{G}^{-1})R+C]\sqrt{T\log(2d)\log\frac{2T}{\delta}}. && (G\le \tfrac{1}{2}(\bar{G}+G^2\bar{G}^{-1}))
\end{aligned}$$

Next, notice that if there is no expert change (switch) then $\|M_t - M_{t-1}\| \le G\eta = 2GR/(\bar{G}\sqrt{T})$, and otherwise, if there is a switch, then $\|M_t - M_{t-1}\| \le 2R$. We thus get that under the above event

$$\sum_{t=1}^{\tau}\|M_t - M_{t-1}\| \le (270+2G\bar{G}^{-1})R\sqrt{T\log(2d)\log\frac{2T}{\delta}}.$$ ■

E.2 Concentration of Measure

First, we give the following Bernstein type tail bound (see e.g., Rosenberg et al., 2020, Lemma D.4).

Lemma 28. *Let $\{X_t\}_{t\ge1}$ be a sequence of random variables with expectation adapted to a filtration $\mathcal{F}_t$. Suppose that $0\le X_t\le 1$ almost surely. Then with probability at least $1-\delta$*

$$\sum_{t=1}^{T}\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \le 2\sum_{t=1}^{T} X_t + 4\log\frac{2}{\delta}$$

Lemma (restatement of Lemma 19). *Let X_t be a sequence of random variables adapted to a filtration $\mathcal{F}_t$. Then we have the following*

- *If $|X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-2H}]| \le C_t$ where $C_t\ge 0$ are $\mathcal{F}_{t-2H}$ measurable then with probability at least $1-\delta$*

$$\sum_{t=1}^{T}(X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-2H}]) \le 2\sqrt{\sum_{t=1}^{T}(C_t^2)H\log\frac{T}{\delta}};$$

- *If $0\le X_t\le 1$ then with probability at least $1-\delta$*

$$\sum_{t=1}^{T}\mathbb{E}[X_t \mid \mathcal{F}_{t-2H}] \le 2\sum_{t=1}^{T}(X_t) + 8H\log\frac{2T}{\delta}.$$

Proof. For $h = 1, \ldots, 2H$, and $k \geq 0$ define the time indices

$$t_k^{(h)} = h + 2Hk = t_{k-1}^{(h)} + 2H,$$

and the filtration $\bar{\mathcal{F}}_k^{(h)} = \mathcal{F}_{t_k^{(h)}}$. Denoting $X_k^{(h)} = X_{t_k^{(h)}}$ we have that $X_k^{(h)}$ is $\bar{\mathcal{F}}_k^{(h)}$ measurable and that

$$|X_k^{(h)} - \mathbb{E}[X_k^{(h)} \mid \bar{\mathcal{F}}_{k-1}^{(h)}]| = |X_{t_k^{(h)}} - \mathbb{E}[X_{t_k^{(h)}} \mid \mathcal{F}_{t_k^{(h)}-2H}]| \leq C_{t_k^{(h)}}.$$

We can thus invoke the Azuma–Hoeffding inequality with a union bound over all $h = 1, \ldots, 2H$ to get that with probability at least $1 - \delta$

$$\sum_{k=1}^{K(h)} \Big(X_k^{(h)} - \mathbb{E}[X_k^{(h)} \mid \bar{\mathcal{F}}_{k-1}^{(h)}]\Big) \leq \sqrt{2 \sum_{k=1}^{K(h)} (C_{t_k^{(h)}}^2) \log \frac{2H}{\delta}} \leq \sqrt{2 \sum_{k=1}^{K(h)} (C_{t_k^{(h)}}^2) \log \frac{T}{\delta}}, \qquad (2H \leq T)$$

where we denoted $K(h) = \lfloor (T-h)/2H \rfloor$. Now, notice that

$$\{t_k : k = 1, \ldots, K(h), h = 1, \ldots, 2H\} = \{1, \ldots, T\}.$$

We conclude that

$$\begin{aligned}
\sum_{t=1}^{T} (X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-2H}]) &= \sum_{h=1}^{2H} \sum_{k=1}^{K(h)} \Big(X_k^{(h)} - \mathbb{E}[X_k^{(h)} \mid \bar{\mathcal{F}}_{k-1}^{(h)}]\Big) \\
&\leq \sum_{h=1}^{2H} \sqrt{2 \sum_{k=1}^{K(h)} (C_{t_k^{(h)}}^2) \log \frac{T}{\delta}} \\
&\leq 2 \sqrt{H \sum_{h=1}^{2H} \sum_{k=1}^{K(h)} (C_{t_k^{(h)}}^2) \log \frac{T}{\delta}} \\
&= 2 \sqrt{\sum_{t=1}^{T} (C_t^2) H \log \frac{T}{\delta}}.
\end{aligned}$$

Moving on to the second claim of the lemma, $X_k^{(h)}$ satisfies Lemma 28, and we thus invoke it with $\delta/2H$ for all $h = 1, \ldots, 2H$. Taking a union bound, we get that with probability at least $1 - \delta$ for all $h = 1, \ldots, 2H$

$$\sum_{k=1}^{K(h)} \mathbb{E}[X_k^{(h)} \mid \bar{\mathcal{F}}_{k-1}^{(h)}] \leq 2 \sum_{k=1}^{K(h)} (X_k^{(h)}) + 4 \log \frac{4H}{\delta} \leq 2 \sum_{k=1}^{K(h)} (X_k^{(h)}) + 4 \log \frac{2T}{\delta}, \qquad (2H \leq T)$$

and thus finally

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_{t-2H}] &= \sum_{h=1}^{2H} \sum_{k=1}^{K(h)} \mathbb{E}[X_k^{(h)} \mid \bar{\mathcal{F}}_{k-1}^{(h)}] \\
&\leq \sum_{h=1}^{2H} \left[2 \sum_{k=1}^{K(h)} (X_k^{(h)}) + 4 \log \frac{2T}{\delta} \right] \\
&\leq 2 \sum_{t=1}^{T} (X_t) + 8H \log \frac{2T}{\delta}.
\end{aligned}$$

■

E.3 Surrogate functions

Lemma (restatement of Lemma 11). *For all w such that $\|w_t\| \leq W$, $M \in \mathcal{M}$, and $t \leq T$, we have:*

1. $\|\rho_t\| \le \sqrt{2} W R_{\mathcal{M}} H$;
2. $\|\rho_{t-1} - \rho_{t-1}(M_t; w)\|^2 \le 2R_{\mathcal{M}}^2 H \Big[\sum_{h=1}^{2H} \|w_{t-h} - \hat{w}_{t-h}\|^2 + \sum_{h=1}^{H} \|M_{t-h} - M_t\|_F^2\Big]$.

Proof. Recall:

$$\rho_{t-1} = (u_{t-H}(M_{t-H}; \hat{w})^\mathsf{T}, \ldots u_{t-1}(M_{t-1}; \hat{w})^\mathsf{T}, \hat{w}_{t-H}^\mathsf{T}, \ldots \hat{w}_{t-2}^\mathsf{T})^\mathsf{T}, \quad \text{and}$$
$$\rho_{t-1}(M_t; w) = (u_{t-H}(M_t; w)^\mathsf{T}, \ldots u_{t-1}(M_t; w)^\mathsf{T}, w_{t-H}^\mathsf{T}, \ldots w_{t-2}^\mathsf{T})^\mathsf{T}.$$

First, $\|u_t(M; w)\| \le W R_{\mathcal{M}} \sqrt{H}$ by Lemma 10. Thus

$$\|\rho_{t-1}\| \le \sqrt{\sum_{h=1}^{H} \Big[\|u_{t-H}(M_{t-h})\|^2 + \|w_{t-h}\|^2\Big]} \le \sqrt{H[W^2 R_{\mathcal{M}}^2 H + W^2]} \le \sqrt{2} W R_{\mathcal{M}} H,$$

concluding the first part of the lemma.

For the second part, we begin by using Lemma 10 to get

$$\begin{aligned}
&\|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\|^2 \\
&\quad \le 2\|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; \hat{w})\|^2 + 2\|u_{t-h}(M_t; \hat{w}) - u_{t-h}(M_t; w)\|^2 \\
&\quad \le 2R_{\mathcal{M}}^2 \|w_{t-(h+H):t-(h+1)} - \hat{w}_{t-(h+H):t-(h+1)}\|^2 + 2R_{\mathcal{M}}^2 H \|M_{t-h} - M_t\|_F^2 \\
&\quad \le 2R_{\mathcal{M}}^2 \sum_{h'=1}^{H} \|w_{t-(h+h')} - \hat{w}_{t-(h+h')}\|^2 + 2R_{\mathcal{M}}^2 H \|M_{t-h} - M_t\|_F^2.
\end{aligned}$$

We thus get

$$\begin{aligned}
\|\rho_{t-1} - \rho_{t-1}(M_t; w)\|^2 &= \sum_{h=1}^{H} \|u_{t-h}(M_{t-h}; \hat{w}) - u_{t-h}(M_t; w)\|^2 + \sum_{h=2}^{H} \|w_{t-h} - \hat{w}_{t-h}\|^2 \\
&\le 2R_{\mathcal{M}}^2 \sum_{h=1}^{H} \sum_{h'=0}^{H} \|w_{t-(h+h')} - \hat{w}_{t-(h+h')}\|^2 + 2R_{\mathcal{M}}^2 H \sum_{h=1}^{H} \|M_{t-h} - M_t\|_F^2 \\
&\le 2R_{\mathcal{M}}^2 H \left[\sum_{h=1}^{2H} \|w_{t-h} - \hat{w}_{t-h}\|^2 + \sum_{h=1}^{H} \|M_{t-h} - M_t\|_F^2\right].
\end{aligned}$$

■

Lemma (restatement of Lemma 12). *Define the functions*

$$C_f(\Psi) = 5R_{\mathcal{M}} W H \max\{\|(\Psi\ I)\|_F, \kappa\gamma^{-1} R_B\}, \qquad G_f(\Psi) = \sqrt{2} W H \|\Psi\|_F + \alpha/(R_{\mathcal{M}} \sqrt{2H}).$$

For any w, w' with $\|w_t\|, \|w'_t\| \le W$ and M, M' with $\|M\|_F, \|M'\|_F \le R_{\mathcal{M}}$, we have:

1. $|f_t(M; w) - f_t(M; w')| \le C_f(\Psi)$;
2. $|\bar{f}_t(M; \Psi, V, w) - \bar{f}_t(M; \Psi, V, w')| \le C_f(\Psi)$;

Additionally, if $V \succeq \lambda_\Psi I$ then

3. $|\bar{f}_t(M; \Psi, V, w) - \bar{f}_t(M'; \Psi, V, w)| \le G_f(\Psi) \|M - M'\|_F$;
4. $|\bar{f}_t(M; k, \chi, \Psi, V, w) - \bar{f}_t(M'; k, \chi, \Psi, V, w)| \le G_f(\Psi) \|M - M'\|_F$;
5. $\bar{f}_t(M; k, \chi, \Psi, V, w) \le \bar{f}_t(M; \Psi, V, w) + \alpha\sqrt{2/H}[(1 + R_{\mathcal{M}}^{-1} \sqrt{d_x}]$.

Moreover, if $\|(\Psi\ I)\|_F \le 17 R_B \kappa^2 \sqrt{\gamma^{-3}(d_x + d_u)(d_x^2\kappa^2 + d_u R_B^2)\log\frac{24T^2}{\delta}}$, *then:*

$$C_f(\Psi) \le 5\alpha/(H\sqrt{d_x(d_x+d_u)}), \quad \textit{and} \;\; G_f(\Psi) \le \alpha\sqrt{2}/(R_{\mathcal{M}}\sqrt{H}).$$

Proof. First, recalling the definition of $x_t(M;\Psi,w)$ in Eq. (3), we have

$$x_t(M;\Psi,w) = \Psi_\star \rho_{t-1}(M;w) + w_{t-1} = \sum_{h=1}^{H} A_\star^{h-1}[B_\star u_{t-h}(M;w) + w_{t-h}].$$

Also noticing that $u_t(M;w)$ is $R_{\mathcal{M}}$ Lipschitz in $w_{t-H:t-1}$ (Lemma 10), we get

$$\begin{aligned}
&\|x_t(M;\Psi_\star,w) - x_t(M;\Psi_\star,w')\| = \left\|\sum_{h=1}^{H} A_\star^{h-1}\big[B_\star(u_{t-h}(M;w) - u_{t-h}(M;w')) + (w_{t-h} - w'_{t-h})\big]\right\| \\
&\le \sum_{h=1}^{H} \kappa(1-\gamma)^{h-1}\big[R_B\|u_{t-h}(M;w) - u_{t-h}(M;w')\| + \|w_{t-h} - w'_{t-h}\|\big] \\
&\le \sum_{h=1}^{H} \kappa(1-\gamma)^{h-1}\Big[R_B R_{\mathcal{M}}\|w_{t-(h+H):t-(h+1)} - w'_{t-(h+H):t-(h+1)}\| + \|w_{t-h} - w'_{t-h}\|\Big] \\
&\le \sqrt{2}\kappa \sum_{h=1}^{H} (1-\gamma)^{h-1} R_B R_{\mathcal{M}}\|w_{t-(h+H):t-h} - w'_{t-(h+H):t-h}\| \qquad (x + y \le \sqrt{2(x^2+y^2)}) \\
&\le \sqrt{2}\kappa\gamma^{-1} R_B R_{\mathcal{M}}\|w_{t-2H:t-1} - w'_{t-2H:t-1}\|,
\end{aligned}$$

where in the third inequality notice that $\|w_{1:t-1}\|^2 + \|w_t\|^2 = \|w_{1:t}\|^2$. used

Next, also using the Lipschitz assumption on c_t, and that u_t is $R_{\mathcal{M}}-$Lipschitz with respect to $w_{t-H:t-1}$ (Lemma 10) we get that

$$\begin{aligned}
|f_t(M;w) - f_t(M;w')| &= |c_t(x_t(M;\Psi_\star,w), u_t(M;w)) - c_t(x_t(M;\Psi_\star,w'), u_t(M;w'))| \\
&\le \|(x_t(M;\Psi_\star,w) - x_t(M;\Psi_\star,w'), u_t(M;w) - u_t(M;w'))\| \\
&\le \sqrt{3}\kappa\gamma^{-1} R_B R_{\mathcal{M}}\|w_{t-2H:t-1} - w'_{t-2H:t-1}\|.
\end{aligned}$$

Moreover, since $\|w_{t-2H:t-1} - w'_{t-2H:t-1}\| \le W\sqrt{8H}$ we also get that

$$|f_t(M;w) - f_t(M;w')| \le 5\kappa\gamma^{-1} R_B R_{\mathcal{M}} W\sqrt{H}.$$

Now, also recall $x_t(M;\Psi,w) = \Psi\rho_{t-1}(M;w) + w_{t-1}$, thus by Lemma 10 we have

$$\begin{aligned}
&\|x_t(M;\Psi,w)\| &&\le \sqrt{2} W R_{\mathcal{M}} H\|(\Psi\ I)\| \\
&\|(x_t(M;\Psi,w)\ u_t(M;w))\| &&\le \sqrt{3} W R_{\mathcal{M}} H\|(\Psi\ I)\| \\
&\|x_t(M;\Psi,w) - x_t(M;\Psi,w')\| &&\le \sqrt{2H} R_{\mathcal{M}}\|(\Psi\ I)\|\|w_{t-2H:t-1} - w'_{t-2H:t-1}\|.
\end{aligned}$$

Then we get

$$\begin{aligned}
&|c_t(x_t(M;\Psi,w), u_t(M;w)) - c_t(x_t(M;\Psi,w'), u_t(M;w'))| \\
&\le \|(x_t(M;\Psi,w) - x_t(M;\Psi,w'), u_t(M;w) - u_t(M;w'))\| \\
&\le \sqrt{3H} R_{\mathcal{M}}\|(\Psi\ I)\|\|w_{t-2H:t-1} - w'_{t-2H:t-1}\|,
\end{aligned}$$

and since the second term of $\bar{f}_t$ does not depend on w we also have

$$|\bar{f}_t(M;\Psi,V,w) - \bar{f}_t(M;\Psi,V,w')| \le \sqrt{3H} R_{\mathcal{M}}\|(\Psi\ I)\|\|w_{t-2H:t-1} - w'_{t-2H:t-1}\|,$$

Moreover, since $\|w_{t-2H:t-1} - w'_{t-2H:t-1}\| \le W\sqrt{8H}$ we also get

$$|\bar{f}_t(M;\Psi,V,w) - \bar{f}_t(M;\Psi,V,w')| \le 5 R_{\mathcal{M}} W H\|(\Psi\ I)\| \le C_f(\Psi).$$

This concludes the first two parts of the proof.

Next, we have that

$$
\begin{aligned}
\|u_t(M;w) - u_t(M';w)\| &= \left\| \sum_{h=1}^{H} (M^{[h]} - M'^{[h]}) w_{t-h} \right\| \\
&\le W \sum_{h=1}^{H} \left\| M^{[h]} - M'^{[h]} \right\| \\
&\le W\sqrt{H} \|M - M'\|_F, \qquad \text{(Cauchy-Schwarz)}
\end{aligned}
$$

thus we get

$$
\begin{aligned}
\|\rho_t(M;w) - \rho_t(M';w)\| &= \sqrt{\sum_{h=1}^{H} \|u_{t+1-h}(M;w) - u_{t+1-h}(M';w)\|^2} \\
&\le WH\|M - M'\|_F,
\end{aligned}
$$

and

$$
\|x_t(M;\Psi,w) - x_t(M';\Psi,w)\| \le WH\|\Psi\|\|M - M'\|_F.
$$

We thus have that

$$
\begin{aligned}
&|c_t(x_t(M;\Psi,w), u_t(M;w)) - c_t(x_t(M';\Psi,w), u_t(M';w))| \\
&\le \|(x_t(M;\Psi,w) - x_t(M';\Psi,w),\ u_t(M;w) - u_t(M';w))\| \\
&\le \sqrt{2}WH\|\Psi\|\|M - M'\|_F.
\end{aligned}
$$

Next, denote $\tilde{M} = M - M'$ and notice that

$$
\begin{aligned}
\|P(M) - P(M')\|_F &= \left\| \begin{pmatrix}
\tilde{M}^{[H]} & \tilde{M}^{[H-1]} & \dots & \tilde{M}^{[1]} & & & \\
& \tilde{M}^{[H]} & \tilde{M}^{[H-1]} & \dots & \tilde{M}^{[1]} & & \\
& & \ddots & \ddots & & \ddots & \\
& & & \tilde{M}^{[H]} & \tilde{M}^{[H-1]} & \dots & \tilde{M}^{[1]} \\
& & & & 0 & & \\
& & & & & \ddots & \\
& & & & & & 0
\end{pmatrix} \right\|_F \\
&= \sqrt{H}\|M - M'\|_F
\end{aligned}
$$

We thus have

$$
\begin{aligned}
\left| \|V^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty - \|V^{-1/2}P(M')\Sigma_{2H-1}^{1/2}\|_\infty \right| &\le \|V^{-1/2}(P(M) - P(M'))\Sigma_{2H-1}^{1/2}\|_\infty \\
&\le \|V^{-1/2}(P(M) - P(M'))\Sigma_{2H-1}^{1/2}\|_F \\
&\le \|V^{-1/2}\|\|\Sigma^{1/2}\|\|P(M) - P(M')\|_F \\
&\le \lambda_\Psi^{-1/2} W\sqrt{H}\|M - M'\|_F,
\end{aligned}
$$

which yields

$$
\begin{aligned}
|\bar{f}_t(M;\Psi,V,w) - \bar{f}_t(M';\Psi,V,w)| &\le (\sqrt{2}WH\|\Psi\| + \alpha\lambda_\Psi^{-1/2}W\sqrt{H})\|M - M'\|_F \\
&\le [\sqrt{2}WH\|\Psi\| + \alpha/(R_{\mathcal{M}}\sqrt{2H})]\|M - M'\|_F \\
&\le G_f(\Psi)\|M - M'\|_F.
\end{aligned}
$$

Identical arguments show that

$$|\bar{f}_t(M;k,\chi,\Psi,V,w) - \bar{f}_t(M';k,\chi,\Psi,V,w)| \le G_f(\Psi)\|M - M'\|_F.$$

Next, we have

$$\begin{aligned}
\bar{f}_t(M;k,\chi,\Psi,V,w) - \bar{f}_t(M;\Psi,V,w) &= \alpha\Big(\|V^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty - \chi\cdot(V^{-1/2}P(M)\Sigma_{2H-1}^{1/2})_k\Big) \\
&\le 2\alpha\|V^{-1/2}P(M)\Sigma_{2H-1}^{1/2}\|_\infty \\
&\le 2\alpha\|V^{-1/2}\|\|\Sigma^{1/2}\|\|P(M)\|_F \\
&\le 2\alpha\lambda_\Psi^{-1/2}W\sqrt{HR_{\mathcal{M}}^2 + Hd_x} \\
&\le \alpha\sqrt{(2 + 2R_{\mathcal{M}}^{-2}d_x)/H} \\
&\le \alpha\sqrt{2/H}[(1 + R_{\mathcal{M}}^{-1}\sqrt{d_x}].
\end{aligned}$$

Finally, if $\|(\Psi\ I)\|_F \le 17R_B\kappa^2\sqrt{\gamma^{-3}(d_x+d_u)(d_x^2\kappa^2 + d_uR_B^2)\log\frac{24T^2}{\delta}}$, then we have that

$$\begin{aligned}
C_f(\Psi) &\le 85WR_{\mathcal{M}}R_B\kappa^2H\sqrt{\gamma^{-3}(d_x+d_u)(d_x^2\kappa^2 + d_uR_B)\log\frac{24T^2}{\delta}} \le 5\alpha/(H\sqrt{d_x(d_x+d_u)}) \\
G_f(\Psi) &\le \frac{\sqrt{2}}{5R_{\mathcal{M}}}C_f(\Psi) + \alpha/(R_{\mathcal{M}}\sqrt{2H}) \le \alpha\sqrt{2}/(R_{\mathcal{M}}\sqrt{H}),
\end{aligned}$$

where the last transition assumed that $H \ge 2$. ■