Efficient Online Linear Control with Stochastic Convex Costs and Unknown Dynamics

Asaf Cassel ACASSEL@MAIL.TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University

Alon Cohen ALONCOHEN@GOOGLE.COM
School of Electrical Engineering, Tel Aviv University, and Google Research, Tel Aviv

Tomer Koren TKOREN@TAUEX.TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University, and Google Research, Tel Aviv

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider the problem of controlling an unknown linear dynamical system under a stochastic convex cost and full feedback of both the state and cost function. We present a computationally efficient algorithm that attains an optimal $\sqrt{T}$ regret-rate compared to the best stabilizing linear controller in hindsight. In contrast to previous work, our algorithm is based on the Optimism in the Face of Uncertainty paradigm. This results in a substantially improved computational complexity and a simpler analysis.

1. Introduction

Adaptive control, the task of regulating an unknown linear dynamical system, is a classic control-theoretic problem that has been studied extensively since the 1950s (e.g., Bertsekas, 1995). Classic results on adaptive control typically pertain to the asymptotic stability and convergence to the optimal controller while contemporary research focuses on regret minimization and finite-time guarantees.

In linear control, both the state and action are vectors in Euclidean spaces. At each time step, the controller views the current state of the system, chooses an action, and the system transitions to the next state. The latter is chosen via a linear mapping from the current state and action and is perturbed by zero-mean i.i.d. noise. The controller also incurs a cost as a function of the instantaneous state and action. In classic models, such as the Linear-Quadratic Regulator (LQR), the cost function is quadratic. A fundamental result on LQR states that, when the model parameters are known, the policy that minimizes the steady-state cost takes a simple linear form; namely, that of a fixed linear transformation of the current state (see Bertsekas, 1995). In more modern formulations, the cost can be any convex Lipschitz function of the state-action pair, and the controller has a no-regret guarantee against the best fixed linear policy (e.g., Agarwal et al., 2019a,b; Simchowitz et al., 2020; Cassel and Koren, 2020).

In this paper we study linear control in a challenging setting of unknown dynamics and unknown stochastic (i.i.d.) convex costs. For the analogous scenario in tabular reinforcement learning, efficient and rate-optimal regret minimization algorithms are well-known (e.g., Auer et al., 2008). However, similar results for adaptive linear control seem significantly more difficult to obtain. Prior work in this context has established efficient $\sqrt{T}$-regret algorithms that are able to adapt to adversarially varying convex costs (Agarwal et al., 2019a), but assumed *known* dynamics. Simchowitz et al.

(2020) extended this to achieve a $T^{2/3}$-regret for unknown dynamics by using a simple explore-then-exploit strategy: in the exploration phase, the controller learns the transitions by injecting the system with random actions; in the exploitation phase, the controller runs the original algorithm using the estimated transitions. We remark that Simchowitz et al. (2020) also showed that their explore-then-exploit strategy achieves a $\sqrt{T}$ regret bound in this setting for *strongly convex* (adversarial) costs; thus demonstrating that the stringent strong convexity assumption is crucial in allowing one to circumvent the challenge of balancing exploration and exploitation.

Recently, Plevrakis and Hazan (2020) made progress in this direction. They observed that the problem of learning both stochastic transitions and stochastic convex costs under bandit feedback is reducible to an instance of stochastic bandit convex optimization for which complex, yet generic polynomial-time algorithms exist (Agarwal et al., 2011). In their case, the bandit feedback assumption requires a brute-force reduction that loses much of the structure of the problem (that would have been preserved under full-feedback access to the costs). This consequently results in a highly complicated algorithm whose running time is a high-degree polynomial in the dimension of the problem (specifically, $n^{16.5}$). Plevrakis and Hazan (2020) also give a more efficient algorithm that avoids a reduction to bandit optimization, but on the other hand assumes the cost function is known and fixed, and that the disturbances in the dynamics come from an isotropic Gaussian distribution.[1] Moreover, this algorithm still relies on computationally intensive procedures (for computing barycentric spanners) that involve running the ellipsoid method.

In this work we present a new computationally-efficient algorithm with a $\sqrt{T}$ regret guarantee for linear control with unknown dynamics and unknown stochastic convex costs under full-information feedback. Our algorithm is simple and intuitive, easily implementable, and works with any sub-Gaussian noise distribution. It is based on the "optimism in the face of uncertainty" (OFU) principle, thought previously to be computationally-infeasible to implement for general convex cost functions (see Plevrakis and Hazan, 2020). The OFU approach enables seamless integration between exploration and exploitation, simplifies both algorithm and analysis significantly, and allows for a faster running time by avoiding explicit exploration (e.g, using spanners) in high-dimensional space.

Our OFU implementation is inspired by the well-known UCB algorithm for multi-armed bandits (Auer et al., 2002). That is, we minimize a lower confidence bound that is constructed as the difference between the (convex) empirical loss and an exploration bonus term whose purpose is to draw the policy towards underexplored state-action pairs. However, since the exploration term is also convex, minimizing the lower confidence bound unfortunately results in a nonconvex optimization problem which, at first glance, can be seen as computationally-hard to solve. Using a trick borrowed from stochastic linear bandits (Dani et al., 2008), we nevertheless are able to relax the objective in such a way that allows for a polynomial-time solution, rendering our algorithm computationally-efficient overall.

Related work. The problem of adaptive LQR control with known fixed costs and unknown dynamics has had a long history. Abbasi-Yadkori and Szepesvári (2011) were the first to study this problem in a regret minimization framework. Their algorithm is also based on OFU, and while inefficient, guarantees rate-optimal $\sqrt{T}$ regret albeit with exponential dependencies on the dimensionality of the system. Since then, many works have tried improving the regret guarantee, Ibrahimi et al. (2012);

[1] Plevrakis and Hazan (2020) describe how to extend their results to more general noise distributions; however, these distributions would still need to be near-spherical since the algorithm needs to be initialized using a "warmup" period in which the dynamics are estimated uniformly.

Faradonbeh et al. (2017); Dean et al. (2018) to name a few. The latter work also presented a poly-time algorithm at a price of a $T^{2/3}$-type regret bound. Cohen et al. (2019); Mania et al. (2019) improve on this by showing how to preserve the $\sqrt{T}$ regret rate with computational efficiency. The optimality of the $\sqrt{T}$ rate was proved concurrently by Cassel et al. (2020); Simchowitz and Foster (2020). Dean et al. (2018) were the first to assume access to a stabilizing controller in order to obtain regret that is polynomial in the problem dimensions. This was later shown to be necessary by Chen and Hazan (2021).

Past work has also considered adaptive LQG control, namely LQR 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 easier than learning the full-observation controller, and, in fact, it is possible to obtain $\operatorname{poly}\log T$ regret for adaptive LQG (Lale et al., 2020).

Another line of work, initiated by Cohen et al. (2018), deals with adversarial LQR in which the transitions are fixed and known, but the cost function changes adversarially. Agarwal et al. (2019a) extended this setting to adversarial noise as well as arbitrary convex Lipschitz costs. Subsequently Agarwal et al. (2019b); Foster and Simchowitz (2020); Simchowitz (2020) provided a $\operatorname{poly}\log T$ regret guarantee for strongly-convex costs with the latter also handling fully adversarial disturbances. Cassel and Koren (2020); Gradu et al. (2020) show a $\sqrt{T}$ regret bound for bandit feedback over the cost function. Lastly, works such as Goel and Wierman (2019) bound the competitive ratio of the learning algorithm rather than its regret.

In a recent follow-up work (Cassel et al., 2022b), we provide an analogous $\sqrt{T}$ regret algorithm for the more challenging case of adversarial cost functions (and unknown dynamics). The result builds on the OFU approach introduced here and combines it with a novel and efficient online algorithm, which minimizes regret with respect to the non-convex optimistic loss functions. In both the stochastic and adversarial cases, the results strongly depend on the stochastic nature of the disturbances; this is in contrast with Simchowitz et al. (2020); Simchowitz (2020), which consider adversarial costs and disturbances. The first shows a $T^{2/3}$ regret algorithm for general convex costs, and the second gives a $\sqrt{T}$ regret algorithm for strongly-convex costs. It thus remains open whether $\sqrt{T}$ regret can be achieved for adversarial disturbances and general convex costs.

2. Problem Setup

We consider controlling an unknown linear dynamical system under stochastic convex costs and full state and cost observation. Our goal is to minimize the total control cost in the following online setting where at round t:

(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 $c_t(x_t, u_t)$;

(4) The system transitions to $x_{t+1} = A_\star x_t + B_\star u_t + w_t$, where $A_\star \in \mathbb{R}^{d_x \times d_x}$, $B_\star \in \mathbb{R}^{d_x \times d_u}$, and $w_t \in \mathbb{R}^{d_x}$.

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 $\pi \in \Pi$. Then the regret 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 all $\pi \in \Pi$.

To define the policy class Π, we use the following notion of stability due to Cohen et al. (2018), which is essentially a quantitative version of classic stability notions in linear control.

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

We consider the benchmark policy class of 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:

- **Bounded stochastic costs:** The cost functions are such that $c_t(x, u) := c(x, u; \zeta_t)$ where $(\zeta_t)_{t=1}^T$ is a sequence of i.i.d. random variables. Moreover, for all x, u, $|c(x, u; \zeta) - \mathbb{E}_{\zeta'} c(x, u; \zeta')| \leq \sigma_c$;
- **Lipschitz costs:** For any (x, u), (x', u') we have

$$|c_t(x, u) - c_t(x', u')| \leq \|(x - x', u - u')\|,$$

- **Bounded i.i.d. noise:** $(w_t)_{t=1}^T$ is a sequence of i.i.d. random variables such that $\|w_t\| \leq W$;
- **Lower-bounded covariance:** There exists some unknown $\underline{\sigma} > 0$ such that $\mathbb{E} w_t w_t^\mathsf{T} \succeq \underline{\sigma}^2 I$;
- **Stabilizing controller:** $A_\star$ is $(\kappa, \gamma)-$strongly stable, and $\|B_\star\| \leq R_B$.

Note the assumption that $A_\star$ is strongly stable is without loss of generality. Otherwise, given access to a stabilizing controller K, Cassel et al. (2022b) show a general reduction, which essentially adds Kx_t to our actions. This will replace $A_\star$ in the analysis with $A_\star + B_\star K$, which is $(\kappa, \gamma)-$strongly stable, as desired. However, this will also add the burden of adding Kx_t to our actions throughout the paper, only making for a more taxing and tiresome reading.

We also remark that the bounded noise assumption can be alleviated to sub-Gaussian noise instead, and that (sub-)Quadratic costs can also be accommodated by appropriately rescaling them. This is essentially since both sub-Gaussian noise and the state and action sequences are bounded with high probability (see Cassel and Koren, 2020; Cassel et al., 2022b for more details on these techniques).

3. Algorithm and main result

We now present our result for the general linear control problem (i.e., $A_\star \neq 0$). We begin by giving necessary preliminaries on Disturbance Action Policies, then we provide our algorithm and give a brief sketch of its regret analysis. The full details of the analysis are deferred to the full version of the paper (Cassel et al., 2022a).

3.1. Preliminaries: Disturbance Action Policies (DAP)

Following recent literature, we use the class of Disturbance Action Policies 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 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 Disturbance Action Policy 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

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

We note that there are several ways to define this class with the most common considering $\sum_{h=1}^{H} \|M^{[h]}\|$ instead of $\|M\|_F$. We chose the Frobenius norm for simplicity of the analysis and implementation, but replacing it would not change the analysis significantly.

The importance of this policy class is two-fold. First, as shown in Lemma 5.2 of Agarwal et al. (2019a), 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. Second, its parameterization preserves the convex structure of the problem, making it amenable to various online convex optimization methods. In light of the above, our regret guarantee will be given with respect to Π_{DAP}.

While the benefits of Π_{DAP} are clear, notice that it cannot be implemented under our assumptions. This 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.

Finite memory representation. As is common in recent literature, we will approximate the various problem parameters with bounded memory representations. To see this, recurse over 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 $\Psi_\star = [A_\star^{H-1} B_\star, \ldots, A_\star B_\star, B_\star, A_\star^{H-1}, \ldots, A_\star]$, and $\tilde{\rho}_t = [u_{t-H}^\mathsf{T}, \ldots, u_t^\mathsf{T}, w_{t-H}^\mathsf{T}, \ldots, w_{t-1}^\mathsf{T}]^\mathsf{T}$. Now, since $A_\star$ is strongly stable, the term $A_\star^H x_{t-H}$ quickly becomes negligible. Combining this with the DAP policy parameterization, we define the following bounded memory representations. For an arbitrary sequence of disturbances $w = \{w_t\}_{t \ge 1}$ define

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

Algorithm 1 Stochastic Linear Control Algorithm

```
1: input: memory length $H$, optimism parameter $\alpha$, regularization parameters $\lambda_\Psi, \lambda_w$.
2: set $i = j = 1, \tau_{1,1} = 1, V_1 = \lambda_\Psi I, M_1 = 0$ and $\hat{w}_t = 0, u_t = 0$ for all $t < 1$.
3: for $t = 1, 2, \ldots, T$ do
4:     play $u_t = \sum_{h=1}^{H} M_t^{[h]} \hat{w}_{t-h}$ where $M_t = M_{\tau_{i,j}}$
5:     observe $x_{t+1}$ and cost function $c_t$.
6:     calculate
           $$(A_t \; B_t) = \mathop{\arg\min}_{(A \; B) \in \mathbb{R}^{d_x \times (d_x + d_u)}} \sum_{s=1}^{t} \|(A \; B) z_s - x_{s+1}\|^2 + \lambda_w \|(A \; B)\|_F^2, \quad \text{where } z_s = \begin{pmatrix} x_s \\ u_s \end{pmatrix}.$$
7:     set $V_{t+1} = V_t + \rho_t \rho_t^\mathsf{T}$ for $\rho_t = (u_{t+1-H}^\mathsf{T}, \ldots, u_t^\mathsf{T}, \hat{w}_{t+1-H}^\mathsf{T}, \ldots, \hat{w}_{t-1}^\mathsf{T})^\mathsf{T}$ .
8:     estimate noise $\hat{w}_t = \Pi_{B_2(W)}[x_{t+1} - A_t x_t - B_t u_t]$.
9:     if $\det(V_{t+1}) > 2\det(V_{\tau_{i,1}})$ then
10:        start new epoch: $i = i + 1, j = 2, \tau_{i,1} = t + 1, \tau_{i,2} = \tau_{i,1} + 2H, M_{\tau_{i,1}} = M_{\tau_{i,2}} = 0$.
11:        estimate system parameters
               $$\Psi_{\tau_{i,1}} = \mathop{\arg\min}_{\Psi \in \mathbb{R}^{d_x \times (H d_u + (H-1) d_x)}} \left\{ \sum_{s=1}^{t} \|\Psi \rho_s - x_{s+1}\|^2 + \lambda_\Psi \|\Psi\|^2 \right\}.$$
12:    if $t + 1 - \tau_{i,1} > 2(\tau_{i,j} - \tau_{i,1})$ then
13:        start new sub-epoch: $j = j + 1, \tau_{i,j} = t + 1$.
14:        solve optimistic cost minimization ($\hat{w} = \{\hat{w}_t\}_{t \geq 1}$)
               $$M_{\tau_{i,j}} = \mathop{\arg\min}_{M \in \mathcal{M}} \sum_{s=\tau_{i,j-1}}^{\tau_{i,j}-1} \left[ c_s(x_s(M; \Psi_{\tau_{i,1}}, \hat{w}), u_s(M; \hat{w}))) - \alpha W \|V_{\tau_{i,1}}^{-1/2} P(M)\|_\infty \right].$$
```

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 our notation does not reveal this fact explicitly, it helps with both brevity and clarity.

3.2. Algorithm

Here we present Algorithm 1 for general linear systems ($A_\star \neq 0$). Notice that the system's memory as well as the use of DAP policies with the estimated noise terms ($\hat{w}_t$) can cause for cyclical probabilistic dependencies between the estimate of the model transitions, the estimate of the loss, and the estimated noise terms. To alleviate these dependencies our algorithm seldom changes its chosen policy ($\approx \log^2 T$ many times), and constructs its estimates using only observations from previous non-overlapping time intervals.

The algorithm proceeds in epochs, each starting with a least squares estimation of the unrolled model using all past observations (Line 11), and the estimate is then kept fixed throughout the epoch.

The epoch ends when the determinant of V_t is doubled (Line 9); intuitively, when the confidence of the unrolled model increases substantially.[2] An epoch is divided into subepochs of exponentially growing lengths in which the policy is kept fixed (Line 12). Each subepoch starts by minimizing an optimistic estimate of the loss (Line 14) that balances between exploration. The algorithm plays the resulting optimistic policy throughout the subepoch (Line 4). To that end we follow the technique presented in Plevrakis and Hazan (2020) to estimate the noise terms $\{w_t\}_{t\geq 1}$ on-the-fly (Line 8). Note that, for this purpose, the algorithm estimates the matrix $(A_\star \; B_\star)$ in each time step (Line 6) even though it can be derived from the estimated unrolled model. This is done to simplify our analysis, and only incurs a small price on the runtime of the algorithm.

We have the following guarantee for our algorithm:

Theorem 2. *Let $\delta \in (0,1)$ and suppose that we run Algorithm 1 with parameters $R_{\mathcal{M}}, R_B \geq 1$ and for proper choices of $H, \lambda_w, \lambda_\Theta, \alpha$. If $T \geq 64R_{\mathcal{M}}^2$ then with probability at least $1-\delta$, simultaneously for all $\pi \in \Pi_{\mathrm{DAP}}$,*

$$\mathrm{regret}_T(\pi) \leq \mathrm{poly}(\kappa, \gamma^{-1}, \underline{\sigma}^{-1}, \sigma_c, R_B, R_{\mathcal{M}}, d_x, d_u, \log(T/\delta))\sqrt{T}.$$

Efficient computation. The main hurdle towards computational efficiency is the calculation of the optimistic cost minimization step (Line 14). We compute this in polynomial-time by borrowing a trick from Dani et al. (2008): the algorithm solves $2m$ convex optimization problems with $m = d_x(2H-1)(d_x(H-1)+d_uH)$, and takes the minimum between them. To see why this is valid, observe that $\|x\|_\infty = \max_{\chi\in\{-1,1\}} \max_{k\in[m]} \chi \cdot x_k$. We can therefore write the optimistic cost minimization as

$$\min_{\chi\in\{-1,1\},k\in[m]} \; \min_{M\in\mathcal{M}} \sum_{s=\tau_{i,j-1}}^{\tau_{i,j}-1} \left[c_s(x_s(M;\Psi_{\tau_{i,1}},\hat{w}), u_s(M;\hat{w})) - \alpha W \chi \cdot \left(V_{\tau_{i,1}}^{-1/2} P(M) \right)_k \right],$$

where k is a linear index. This indeed suggests to solve for $M \in \mathcal{M}$ for each value of k and χ, then take the minimum between them. Moreover, when k and χ are fixed, the objective becomes convex in M. Consequently, As there are $2m$ such values of k and χ, this amounts to solving $2m$ convex optimization problems. We note that it suffices to solve each convex optimization problem up to an accuracy of $\approx T^{-1/2}$, which can be done using $O(T)$ gradient oracle calls.

Comparison with Plevrakis and Hazan (2020). The following compares the computational complexity of Algorithm 1 with those of Plevrakis and Hazan (2020) under a first order (value and gradient) oracle assumption on the cost functions c_t. To simplify the discussion, we denote both state and action dimensions as $d = \max\{d_x, d_u\}$, and omit logarithmic terms and $O(\cdot)$ notations from all bounds.

We show that the overall computational complexity of our algorithm is d^4T. By updating the least squares procedure in Lines 6 and 11 recursively at each time step their overall complexity is d^3T. As previously explained, in Line 4, we solve d^2 convex optimization problems, each to an accuracy of $\approx T^{-1/2}$. Since the objective is a sum of convex functions, we can do this using Stochastic Gradient Descent (SGD) with T oracle calls (in expectation). Overall, we make d^2T gradient oracle calls. For each oracle call, we further use matrix addition, and matrix vector multiplications on M, which take an additional d^2 computations. The remaining computations of the algorithm are negligible.

[2] More concretely, the volume of the confidence ellipsoid around the unrolled model decreases by a constant factor.

Now, we show that the equivalent complexity for Algorithm 2 of Plevrakis and Hazan (2020) is $d^{12}T$. The crux of their computation is finding a $2-$Barycentric spanner for their confidence set. The inherent dimension there is that of M, which is d^2. To compute the barycentric spanner, the authors explain that d^4 calls to a linear optimization oracle are required. These can in turn be implemented using the elipsoid method, whose overall computation is d^8T.

Plevrakis and Hazan (2020) also give Algorithm 5, which works with bandit feedback and uses SBCO as a black-box. Compared with their Algorithm 2 the computational complexity is higher, and the regret guarantee depends on d^{36} instead of d^3.

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 $A_\star = 0$. At the end of the section, we give an overview of the analysis for the general control setting (with $A_\star \neq 0$). The complete details in the general case are significantly more technical and thus deferred from this extended abstract (see the full version of the paper (Cassel et al., 2022a) for full details).

Concretely, following Plevrakis and Hazan (2020), suppose that $A_\star = 0$, and thus $x_{t+1} = B_\star u_t + w_t$. Next, assume that $c_t(x, u) = c_t(x)$, i.e., the costs do not depend on u. Finally, assume that we minimize the pseudo regret, i.e.,

$$\max_{u:\|u\|\le R_u} \sum_{t=1}^{T} [J(B_\star u_t) - J(B_\star u)],$$

where $J(B_\star u) = \mathbb{E}_{\zeta,w} c(B_\star u + w; \zeta)$. This setting falls under the umbrella of stochastic bandit convex optimization, making generic algorithms applicable. However, it has additional structure that we leverage to create a much simpler and more efficient algorithm. In what follows, we formally define this setting with clean notation as to avoid confusion with our general setting.

4.1. The $A_\star = 0$ case: Stochastic Convex Optimization 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. At round t the learner

(1) predicts $a_t \in \mathcal{S}$;

(2) observes cost function $\ell_t : \mathbb{R}^{d_y} \to \mathbb{R}$ and state $y_{t+1} := Q_\star a_t + w_t$;

(3) incurs cost $\ell_t(Q_\star a_t)$.

We have that $w_t \in \mathbb{R}^{d_y}$ 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 stochastic in the following sense. There exists a sequence $\zeta_1, \zeta_2, \ldots$ of i.i.d. random variables, and a function $\ell : \mathbb{R}^{d_y} \times \mathbb{R} \to \mathbb{R}$ such that $\ell_t(q) := \ell(q; \zeta_t)$. Define the expected cost $\mu(q) = \mathbb{E}_\zeta \ell(q; \zeta)$. We consider minimizing the pseudo-regret, defined as

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

Minimizing the pseudo-regret instead of the actual regret will maintain the main hardness of the problem, but will better highlight our main contributions.

Assumptions. Our assumptions in the simplified case are the following:

- $\ell(q;\zeta)$ is convex and $1-$Lipschitz in its first parameter;
- For all ζ, ζ', and any q we have $|\ell(q;\zeta) - \ell(q;\zeta')| \le \sigma_\ell$;
- There exists some known $W, R_Q \ge 0$ such that $\|w_t\| \le W$, and $\|Q_\star\| \le R_Q$.
- The diameter of $\mathcal{S}$ is $R_a = \max_{a,a' \in \mathcal{S}} \|a - a'\| < \infty$.

4.2. The simplified algorithm

Algorithm 2 SCO with hidden linear transform

```
input: optimism parameter $\alpha$, regularizer $\lambda$
set: $V_1 = \lambda I, \widehat{Q}_1 = 0$.
for $t = 1, 2, \ldots, T$ do
    play optimistic cost minimizer: $a_t \in \arg\min_{a \in \mathcal{S}} \sum_{s=1}^{t-1} [\ell_s(\widehat{Q}_t a) - \alpha \|V_t^{-1/2} a\|_\infty]$.
    observe $y_{t+1} = Q_\star a_t + w_t$ and cost function $\ell_t$, and set $V_{t+1} = V_t + a_t a_t^\mathsf{T}$.
    estimate system parameters
        $\widehat{Q}_{t+1} = \arg\min_{Q \in \mathbb{R}^{d_y \times d_a}} \sum_{s=1}^{t} \|Q a_s - y_{s+1}\|^2 + \lambda \|Q\|_F^2.$
```

Our algorithm is depicted as Algorithm 2. The algorithm maintains an estimate $\widehat{Q}_t$ of $Q_\star$. At each time step t, the algorithm plays action a_t, chosen such that it minimizes our optimistic cost function. That is, a_t is a minimizer of a lower bound on the total loss up to time t: $\sum_{s=1}^{t-1} \ell_s(Q_\star a) \approx (t-1)\mu(Q_\star a)$. This fuses together exploration and exploitation by either choosing under-explored actions, or exploiting low expected cost for already sufficiently-explored actions. We remark that the optimistic cost minimization procedure solves a non-convex optimization problem, but nevertheless show that it can be solved in polynomial-time in the sequel. Lastly, our algorithm observes y_{t+1} and uses it improve its estimate of $Q_\star$ by solving a least-squares problem.

The main hurdle in understanding why the algorithm is computationally-efficient is the calculation of the optimistic cost minimization step. Following the computational method presented in Algorithm 1, we do this by first solving $2d_a$ convex objectives and then taking their minimizer.

4.3. Analysis

We now present the main theorem for this section that 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*

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

If $T \ge \max\{\sigma_\ell, 64 D_q^2\}$ where $D_q = 3R_Q R_a + W d_y \sqrt{8 \log \frac{4T}{\delta}}$ then with probability at least $1 - \delta$,

$$\mathit{regret}_T \le 13\left[\sigma_\ell \sqrt{d_y \log \frac{3T}{\sigma_\ell \delta}} + d_a (W d_y + R_a R_Q) \log \frac{2T}{\delta}\right] \sqrt{T}.$$

At its core, the proof of Theorem 3 employs the Optimism in the Face of Uncertainty (OFU) approach. To that end, define the optimistic cost functions

$$\bar{\mu}_t(a) = \mu(\widehat{Q}_t a) - \alpha \|V_t^{-1/2} a\|_\infty,$$

where $\widehat{Q}_t, V_t$ are defined as in Algorithm 2. The following lemma shows that our optimistic loss lower bounds the true loss, and bounds the error between the two.

Lemma 4. *Suppose that $\sqrt{d_a}\|\widehat{Q}_t - Q_\star\|_{V_t} \le \alpha$. Then we have that*

$$\bar{\mu}_t(a) \le \mu(Q_\star a) \le \bar{\mu}_t(a) + 2\alpha\sqrt{a^\mathsf{T} V_t^{-1} a}.$$

Proof. We first use the Lipschitz assumption to get

$$\begin{aligned} |\mu(Q_\star a) - \mu(\widehat{Q}_t a)| &\le \|(Q_\star - \widehat{Q}_t) a\| \\ &\le \|Q_\star - \widehat{Q}_t\|_{V_t} \|V_t^{-1/2} a\| \\ &\le \frac{\alpha}{\sqrt{d_a}} \|V_t^{-1/2} a\| \\ &\le \alpha \|V_t^{-1/2} a\|_\infty, \end{aligned}$$

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,

$$\mu(Q_\star a) \ge \mu(\widehat{Q}_t a) - \alpha\|V_t^{-1/2} a\|_\infty = \bar{\mu}_t(a),$$

and on the other hand we also have

$$\mu(Q_\star a) \le \mu(\widehat{Q}_t a) + \alpha\|V_t^{-1/2} a\|_\infty = \bar{\mu}_t(a) + 2\alpha\|V_t^{-1/2} a\|_\infty \le \bar{\mu}_t(a) + 2\alpha\sqrt{a^\mathsf{T} V_t^{-1} a},$$

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

We are now ready to prove Theorem 3. The proof focuses on the main ideas, deferring some details to Section 4.4.

Proof of Theorem 3. First, notice that $|\mu(Q_\star a_t) - \mu(Q_\star a)| \le \|Q_\star\|\|a_t - a\| \le 2R_Q R_a$. Using this bound for $t = 1$, we can decompose the regret as

$$\mathrm{regret}(a) \le 2R_Q R_a + \underbrace{\sum_{t=2}^{T} \mu(Q_\star a_t) - \bar{\mu}_t(a_t)}_{R_1} + \underbrace{\sum_{t=2}^{T} \bar{\mu}_t(a_t) - \bar{\mu}_t(a)}_{R_2} + \underbrace{\sum_{t=2}^{T} \bar{\mu}_t(a) - \mu(Q_\star a)}_{R_3}.$$

We begin by bounding R_1 and R_3, which relate the true loss to its optimistic variant. To that end, we use a standard least squares estimation bound (Lemma 5) to get that Lemma 4 holds with probability at least $1 - \delta/2$. Conditioned on this event, we immediately get $R_3 \le 0$. Moreover, we get

$$R_1 \le \sum_{t=2}^{T} 2\alpha\sqrt{a_t^\mathsf{T} V_t^{-1} a_t} \le 2\alpha\sqrt{T\sum_{t=1}^{T} a_t^\mathsf{T} V_t^{-1} a_t} \le 2\alpha\sqrt{5Td_a \log T},$$

where the second inequality is due to Jensen's inequality, and the third is a standard algebraic argument (Lemma 6).

Next, we bound R_2, which is the sum of excess risk of a_t with respect to the optimistic cost. To that end, we first bound $\|\widehat{Q}_t\|$, using the least squares error bound (Lemma 5) and $V_t \geq \lambda I = R_a^2 I$ to get that

$$\|\widehat{Q}_t\| \leq \|Q_\star\| + \|\widehat{Q}_t - Q_\star\| \leq R_Q + \frac{1}{R_a}\|\widehat{Q}_t - Q_\star\|_{V_t} \leq 3R_Q + \frac{W d_y}{R_a}\sqrt{8 \log \frac{4T}{\delta}}.$$

We thus have for all $a \in \mathcal{S}$,

$$\|\widehat{Q}_t a\| \leq \|\widehat{Q}_t\|\|a\| \leq 3R_Q R_a + W d_y \sqrt{8 \log\left(\frac{4T}{\delta}\right)} = D_q.$$

Now, for all $1 \leq t \leq T$ we use a standard uniform convergence argument (Lemma 7) with $R = D_q$, and $\delta/2T$ to get that with probability at least $1 - \delta/2$ simultaneously for all $1 \leq t \leq T$

$$\begin{aligned}
\bar{\mu}_t(a_t) - \bar{\mu}_t(a) &= \mu(a_t) - \mu(a) - \alpha\left(\|V_t^{-1/2} a_t\|_\infty - \|V_t^{-1/2} a\|_\infty\right) \\
&\leq \frac{1}{t-1}\sum_{s=1}^{t-1}(\ell_s(\widehat{Q}_t a_t) - \ell_s(\widehat{Q}_t a)) - \alpha\left(\|V_t^{-1/2} a_t\|_\infty - \|V_t^{-1/2} a\|_\infty\right) + 2\sigma_\ell \sqrt{\frac{d_y \log \frac{6T^2}{\sigma_\ell \delta}}{t-1}} \\
&\leq 2\sigma_\ell \sqrt{\frac{d_y \log \frac{6T^2}{\sigma_\ell \delta}}{t-1}},
\end{aligned}$$

where the last inequality is by definition of a_t as the optimistic cost minimizer. Finally, notice that $\sum_{t=2}^{T}(t-1)^{-1/2} \leq 2\sqrt{T}$ to get that

$$R_2 = \sum_{t=2}^{T} \bar{\mu}_t(a_t) - \bar{\mu}_t(a) \leq \sum_{t=2}^{T} 2\sigma_\ell \sqrt{\frac{d_y \log \frac{6T^2}{\delta}}{t-1}} \leq 6\sigma_\ell \sqrt{T d_y \log \frac{3T}{\sigma_\ell \delta}}.$$

Finally, taking a union bound on both events and substituting for the chosen value of α completes the proof. ■

4.4. Deferred details

Here we complete the deferred details in the proof of Theorem 3. We start with 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 5 (Abbasi-Yadkori and Szepesvári, 2011). *Let $\Delta_t = Q_\star - \widehat{Q}_t$, and suppose that $\|a_t\|^2 \leq \lambda = R_a^2$, $T \geq d_y$. With probability at least $1 - \delta$, we have for all $t \geq 1$*

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

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 the full version of the paper (Cassel et al., 2022a)).

Lemma 6. *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 $\sum_{t=1}^{T} a_t^\mathsf{T} V_t^{-1} a_t \le 5 d_a \log T$.*

Finally, a standard uniform convergence result, which is used in bounding R_2 (proof in the full version of the paper (Cassel et al., 2022a)).

Lemma 7. *Let $R > 0$ and suppose that $T \ge \max\{\sigma_\ell, 64R^2\}$. Then for any $\delta \in (0,1)$ we have that with probability at least $1 - \delta$*

$$\left|\sum_{t=1}^{T} \ell_t(q) - \mu(q)\right| \le \sigma_\ell \sqrt{T d_y \log \frac{3T}{\sigma_\ell \delta}}, \quad \forall q \in \mathbb{R}^{d_y} \text{ s.t. } \|q\| \le R.$$

4.5. Extension to the general control setting

Using the assumption that $A_\star$ is strongly stable, we show it takes $2H$ time steps for one of our DAP policies to sufficiently approximate its steady-state. As a result the system may behave arbitrarily during the first $2H$ steps of each subepoch, and we bound the instantaneous regret in each of these time steps by a worst-case constant. As mentioned, though, we show that the total number of subepochs is at most $4(d_x + d_u)H \log^2 T$, making the cumulative regret during changes of policy negligible for reasonably large T.

Concretely, let $R_{\max} \ge \max_{\pi \in \Pi_{\text{DAP}} \cup \pi_{\text{alg}}, t \le T} \max\{\|(x_t^\pi, u_t^\pi)\|, \|(x_t, u_t)\|\}$ be a bound on the state action magnitude. Combining with the Lipschitz assumption, we get that

$$\left|c_t(x_t, u_t) - c_t(x_t^\pi, u_t^\pi)\right| \le 2R_{\max}.$$

Since the first two sub epochs are always at most $2H$ long, the regret decomposes as

$$\text{Regret}_T(\pi) \le 16 R_{\max} H^2 (d_x + d_u) \log^2 T + \sum_{i=1}^{N} \sum_{j=3}^{N_i} \sum_{t=\tau_{i,j}+H}^{\tau_{i,j+1}-1} c_t(x_t, u_t) - c_t(x_t^\pi, u_t^\pi),$$

where N is the number of epochs and N_i is the number of subepochs in epoch i.

We proceed by decomposing the remaining term, analyzing the regret within each subepoch. For this purpose, define an expected surrogate cost and its optimistic version

$$\begin{aligned}
f_t(M; \Psi, w, \zeta) &= c_t(x_t(M; \Psi, w), u_t(M; w), \zeta) \\
F(M; \Psi) &= \mathbb{E}_{\zeta, w} f_t(M; \Psi, w, \zeta) \\
F(M) &= F(M; \Psi_\star) \\
\bar{F}_t(M) &= F(M; \Psi_{\tau_{i(t),1}}) - \alpha W \|V_{\tau_{i(t),1}}^{-1/2} P(M)\|_\infty,
\end{aligned}$$

where $i(t) = \max\{i \;:\; \tau_{i,1} \le t\}$ is the index of the epoch to which t belongs, and $\|\cdot\|_\infty$ is the entry-wise matrix infinity norm. Letting $M_\star \in \mathcal{M}$ be the DAP approximation of $\pi \in \Pi_{\text{lin}}$, we have the following decomposition of the instantaneous regret:

$$\begin{aligned}
c_t(x_t, u_t) - c_t(x_t^\pi, u_t^\pi) &= c_t(x_t, u_t) - F(M_t) && (R_1 \text{ - Truncation + Concentration}) \\
&+ F(M_t) - \bar{F}_t(M_t) && (R_2 \text{ - Optimism}) \\
&+ \bar{F}_t(M_t) - \bar{F}_t(M_\star) && (R_3 \text{ - Excess Risk}) \\
&+ \bar{F}_t(M_\star) - F(M_\star) && (R_4 \text{ - Optimism}) \\
&+ F(M_\star) - c_t(x_t^\pi, u_t^\pi). && (R_5 \text{ - Truncation + Concentration})
\end{aligned}$$

The proof is completed by bounding each of the terms above with high probability and combining them with a union bound. Terms R_2, R_3, R_4 are similar to the ones appearing in the proof Theorem 3, while terms R_1, R_5 are new and relate the cost to that of the unrolled transition model. The latter two terms are bounded in two steps. We start by relating $c_t(x_t, u_t), c_t(x_t^\pi, u_t^\pi)$ to $f_t(M_t; \Psi_\star, w, \zeta_t), f_t(M_\star, \Psi_\star, w, \zeta_t)$, which, similarly to Agarwal et al. (2019a), uses the assumption that $A_\star$ is strongly stable, but then, also accounts for the discrepancies between w_t and $\hat{w}_t$ as was done in Plevrakis and Hazan (2020). We then conclude with a concentration argument that relates $f_t(M)$ to $F(M)$, which is its expectation with respect to w, ζ.

To bound the optimism-related terms R_2, R_4, we first show a least-squares confidence bound similar to Lemma 5. Notice that the least squares bound has to handle the fact that we use the estimated noises $\hat{w}_t$ to predict the parameters of the unrolled model. Denoting $\Delta_t = \Psi_\star - \Psi_t$, we specifically show that:

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

a comparable bound to that of Lemma 5 except for the addition of error terms e_s. Since $\hat{w}_t$ converge to the true noises w_t, we can prove that $\sum_{t=1}^{T} \|e_t\|^2 \lessapprox \log T$, i.e., the additional error terms are of a similar order to the standard estimation error and thus do not increase it significantly. Next, assuming that the above estimation error holds, we show an analogous result to Lemma 4 stating

$$0 \le F(M) - \bar{F}_t(M) \le 2\alpha W \|V_{\tau_{i(t)},1}^{-1/2} P(M)\|_F,$$

which yields that $R_4 \le 0$. To Bound R_2, we further relate $\|V_{\tau_{i(t)},1}^{-1/2} P(M)\|_F$ to $\|V_t^{-1/2} \rho_{t-1}(M; \hat{w})\|$, which is the equivalent of the harmonic term in the right hand side of Lemma 4. To that end, define the noise covariance $\Sigma = \mathbb{E} w_{t-2H:t-2} w_{t-2H:t-2}^\mathsf{T}$, and notice that our minimum eigenvalue assumption implies that $\|\Sigma^{-1/2}\| \le \underline{\sigma}^{-1}$. We thus have that

$$\begin{aligned}
\underline{\sigma}^2 \|V_{\tau_{i(t)},1}^{-1/2} P(M)\|_F^2 \le \|V_{\tau_{i(t)},1}^{-1/2} P(M) \Sigma^{1/2}\|_F^2 &= \mathrm{Tr}\left(V_{\tau_{i(t)},1}^{-1} P(M) \Sigma P(M)\right) \\
&= \mathrm{Tr}\left(V_{\tau_{i(t)},1}^{-1} P(M) \mathbb{E}[w_{t-2H:t-2} w_{t-2H:t-2}^\mathsf{T}] P(M)^\mathsf{T}\right) \\
&= \mathbb{E}_w \mathrm{Tr}\left(V_{\tau_{i(t)},1}^{-1} \rho_{t-1}(M; w) \rho_{t-1}(M; w)^\mathsf{T}\right) \qquad \text{(Eq. (2))} \\
&= \mathbb{E}_w \|V_{\tau_{i(t)},1}^{-1/2} \rho_{t-1}(M; w)\|^2.
\end{aligned}$$

Summing over t and applying several technical concentration and noise estimation arguments yields the desired term and the $O(\sqrt{T})$ bound on R_2.

Last, we deal with R_3, which is analogous the excess risk term in the proof of Theorem 3. We first show a uniform convergence property akin to Lemma 7. Here, however, the uniform convergence is done with respect to both the randomness in the loss function ζ_t and the noise terms w_t. This allows us to use observations gathered in previous subepochs to estimate the expected performance of a DAP policy in the current subepoch. Here, we once again tackle the technical difficulty that our DAP policy is defined with respect to the noise estimates $\hat{w}_t$. This adds an additional error term, proportional to the error in the noise estimates, and accumulates to $\approx \sqrt{T}$ regret overall.

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, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

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.

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*, 2022a.

Asaf Cassel, Alon Cohen, and Tomer Koren. Rate-optimal online convex optimization in adaptive linear control. *arXiv preprint arXiv:2206.01426*, 2022b.

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

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.

Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*, pages 3211–3221. PMLR, 2020.

Gautam Goel and Adam Wierman. An online algorithm for smoothed regression and lqr control. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2504–2513. PMLR, 2019.

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.

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.

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.

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.

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

Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.