# Task-Optimal Exploration in Linear Dynamical Systems

Andrew Wagenmaker [1] Max Simchowitz [2] Kevin Jamieson [1]

## Abstract

Exploration in unknown environments is a fundamental problem in reinforcement learning and control. In this work, we study task-guided exploration and determine what precisely an agent must learn about their environment in order to complete a particular task. Formally, we study a broad class of decision-making problems in the setting of linear dynamical systems, a class that includes the linear quadratic regulator problem. We provide instance- and task-dependent lower bounds which explicitly quantify the difficulty of completing a task of interest. Motivated by our lower bound, we propose a computationally efficient experiment-design based exploration algorithm. We show that it optimally explores the environment, collecting precisely the information needed to complete the task, and provide finite-time bounds guaranteeing that it achieves the instance- and task-optimal sample complexity, up to constant factors. Through several examples of the linear quadratic regulator problem, we show that performing task-guided exploration provably improves on exploration schemes which do not take into account the task of interest. Along the way, we establish that certainty equivalence decision making is instance- and task-optimal, and obtain the first algorithm for the linear quadratic regulator problem which is instance-optimal. We conclude with several experiments illustrating the effectiveness of our approach in practice.

## 1. Introduction

Modern reinforcement learning aims to understand how agents should best explore their environments in order to successfully complete assigned tasks. In the face of uncertainty about the environment, a naive strategy might be to explore the environment until it is uniformly understood (system identification), and then devise a plan to complete the task under this precise understanding of the environment (control). However, it is widely understood that such a two-phased approach of system identification followed by control can be wasteful since, depending on the task, some aspects of the environment ought to be estimated more accurately than others. For instance, if a task requires a precise sequence of steps to be taken in order to be completed, one need not understand all possible outcomes leading to failure after a missed early critical step. Since it may be very costly for an agent to estimate all facets of a complex or high dimensional environment to high precision, it is far preferable to direct agents' exploration only to those aspects most relevant to their tasks. Motivated by this challenge, this paper aims to answer:

*Q1. What exactly must an agent learn about its environment to carry out a particular task?*

*Q2. Given knowledge of the task, can the agent direct their exploration to speed up the process of learning this task-specific critical information?*

*Q3. Having explored its environment, how can the agent best use the information gained to complete the task of interest?*

Our work provides answers to the above questions for a family of decision-making problems in environments parameterized by a linear dynamical system, including synthesis of the linear quadratic regulator. Specifically, for Q1 we show that accomplishing a variety of tasks amounts to maximizing a task-specific linear functional of the Fisher-information matrix, a quantity of fundamental importance to optimal experimental design. Indeed, our results naturally reduce to classical linear optimal experimental design criteria (for example, $A$-optimal) in the absence of dynamics. Answering Q2 in the affirmative amounts to being able to learn just enough about the environment to drive the system to a sequence of states that maximize this task-specific function as fast as possible. We accomplish this via a sequence of experimental design problems over control inputs given a successively improving estimate of the environment. Finally, to answer Q3 we show that the *certainty equivalence*

[1] Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, Washington, USA [2] Department of Electrical Engineering and Computer Science, University of California, Berkeley, California, USA. Correspondence to: Andrew Wagenmaker <ajwagen@cs.washington.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

decision rule—choosing the policy that would optimally complete the task if the estimate of the environment was correct—is the optimal decision rule in an instance-specific sense.

### 1.1. Main Contributions

Our primary contributions are as follows:

1. We develop task- and instance-specific lower bounds which precisely quantify how parameter estimation error translates to suboptimal task performance.

2. We cast the problem of optimal exploration as a surrogate experiment design problem we call *task-optimal experiment design*. For linear dynamical systems, the task-optimal design problem can be solved efficiently by projected gradient descent. We demonstrate that the solution to the design problem yields the information-theoretically optimal exploration strategy, in a strong, instance-dependent sense.

3. The task-optimal design depends on unknown problem parameters. We therefore propose a meta-algorithm, TOPLE, which sequentially solves empirical approximations to the design objective, and demonstrate that this approach matches the performance of the optimal design given knowledge of the true system parameters. As a consequence, we obtain the first instance-optimal algorithm for the LQR problem.

4. We show through numerous mathematical examples that task-specific experiment design can perform arbitrarily better on a task of interest than uniform or task-agnostic exploration. We also rigorously prove a strong sub-optimality result for strategies with low regret for online LQR, such as optimism-under-uncertainty.

5. We show that, for *any exploration strategy* which is sufficiently non-degenerate, in a very general class of decision-making problems which includes certain classes of nonlinear dynamical systems, the certainty equivalence decision rule is instance optimal.

6. Finally, we show that our approach yields practical gains over naive exploration schemes through several numerical examples.

All our results are non-asymptotic and polynomial in terms of the natural problem parameters.

### 1.2. Task-Specific Pure Exploration

We consider linear dynamical systems of the form:

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \quad x_0 \equiv 0. \tag{1.1}$$

where $x_t, w_t \in \mathbb{R}^{d_x}, u_t \in \mathbb{R}^{d_u}$, $A_\star$ and $B_\star$ have appropriate dimensions, and where for simplicity we assume that $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$[1]. We let $\theta_\star = (A_\star, B_\star)$ capture the true dynamical system; importantly, $\theta_\star$ *is unknown to the learner*. We also define a policy $\pi$ as a mapping from past actions and states to future actions $\pi : (x_{1:t}, u_{1:t-1}) \to u_t$. We let $\mathbb{E}_{\theta,\pi}[\cdot]$ denote the expectation over trajectories induced on instance $\theta$ playing policy $\pi$. While we show in Appendix B that several of our results hold in a more general observation model which encompasses certain nonlinear systems, throughout Sections 2 to 5 we assume we are in the linear dynamical system setting.

We are interested in a general decision making problem: given some smooth loss $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ parameterized by $\theta_\star$, choose $\mathfrak{a} \in \mathbb{R}^{d_\mathfrak{a}}$ such that $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ is minimized. For every $\theta_\star$, we assume there exists some optimal decision $\mathfrak{a}_{\mathrm{opt}}(\theta_\star)$ for which $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ is minimized. We require that $\mathcal{J}_\theta(\mathfrak{a})$ and $\mathfrak{a}_{\mathrm{opt}}(\theta)$ satisfy the following assumption.

**Assumption 1** (Smooth Decision-Making, Informal)**.** *The loss $\mathcal{J}_\theta(\mathfrak{a})$ and optimal decision $\mathfrak{a}_{\mathrm{opt}}(\theta)$ are each three times differentiable within a ball around $\mathfrak{a}_{\mathrm{opt}}(\theta_\star)$ and $\theta_\star$, respectively, and their gradients can be absolutely bounded over this range. Furthermore, $\nabla_\mathfrak{a}^2 \mathcal{J}_\theta(\mathfrak{a})$ varies smoothly in $\theta$.*

Our interaction protocol is as follows.

**Task-Specific Pure Exploration Problem.** The learner's behavior is specified by an exploration policy $\pi_{\mathrm{exp}} : (x_{1:t}, u_{1:t-1}) \to u_t$ and decision rule $\mathsf{dec}$ executed in the dynamics Eq. (1.1).

1. For steps $t = 1, \ldots, T$, the learner executes $\pi_{\mathrm{exp}}$ and collects a *trajectory* $\boldsymbol{\tau} = (x_{1:T+1}, u_{1:T})$.

2. For a *budget* $\gamma^2 \geq 0$, the inputs $u_{1:T}$ must satisfy the constraint $\mathbb{E}_{\pi_{\mathrm{exp}}}[\sum_{t=1}^T \|u_t\|_2^2] \leq T\gamma^2$. [2]

3. Finally, the learner proposes a decision $\widehat{\mathfrak{a}} = \mathsf{dec}(\boldsymbol{\tau})$ as a function of $\boldsymbol{\tau}$.

The learner's performance is evaluated on the excess risk

$$\mathcal{R}(\widehat{\mathfrak{a}}; \theta_\star) := \mathcal{J}_{\theta_\star}(\widehat{\mathfrak{a}}) - \inf_{\mathfrak{a}} \mathcal{J}_{\theta_\star}(\mathfrak{a}),$$

and their goal is to choose an exploration policy $\pi_{\mathrm{exp}}$ which induces sufficient exploration to propose a decision $\widehat{\mathfrak{a}}$ with as little excess risk as possible. For simplicity, we assume that the learner only collects a single trajectory. In contrast to online control, the performance of the exploration policy

[1] See Appendix C.2 for a discussion on accommodating non-identity, possibly unknown noise covariance.

[2] The upper bounds in this paper can be easily modified to ensure that the budget constraint $\sum_{t=1}^T \|u_t\|^2 \leq T\gamma^2$ holds with high probability.

is only evaluated on its final decision $\widehat{\mathfrak{a}}$, not the trajectory generated during the learning phase. To make this setting concrete, we consider several specific examples.

### 1.3. Examples and Applications

The task-specific pure exploration problem captures many natural settings. Under the assumed linear dynamics model of (1.1) with $\theta_\star = (A_\star, B_\star)$, the unknown quantity of interest $\mathfrak{a}_{\mathrm{opt}}(\theta_\star)$ can represent any function of the environment defined by $\theta_\star$. In the simplest case of system identification, we may have $\mathfrak{a}_{\mathrm{opt}}(\theta_\star) = \theta_\star$, $\widehat{\mathfrak{a}}$ the least squares estimate of $\theta_\star$ given the trajectory $\boldsymbol{\tau} = (x_{1:T+1}, u_{1:T})$, and $\mathcal{J}_{\theta_\star}(\widehat{\mathfrak{a}})$ a measure of loss, for example the Frobenius norm: $\mathcal{J}_{\theta_\star}(\widehat{\mathfrak{a}}) = \|\widehat{\mathfrak{a}} - \mathfrak{a}_{\mathrm{opt}}(\theta_\star)\|_{\mathrm{F}}^2$. Even in this simple case, the learner can reduce $\mathcal{J}_{\theta_\star}(\widehat{\mathfrak{a}})$ far faster with a deliberate exploration policy relative to a naive policy such as playing isotropic noise. The next several examples illustrate that the task specific pure exploration framework generalizes far beyond this simple system identification task.

**Example 1.1** (Pure Exploration LQR)**.** In the LQR problem, the agent's objective is to design a policy that minimizes the infinite-horizon cumulative cost, with losses $\ell(x,u) := x^\top R_{\mathbf{x}} x + u^\top R_{\mathbf{u}} u$. The resultant cost function is

$$\mathcal{J}_{\mathrm{LQR},\theta_\star}[\pi] := \lim_{T\to\infty} \mathbb{E}_{\theta_\star,\pi}\left[\frac{1}{T}\sum_{t=1}^{T} x_t^\top R_{\mathbf{x}} x_t + u_t^\top R_{\mathbf{u}} u_t\right].$$

It is well known that the optimal policies are of the form $u_t = K x_t$ where $K \in \mathbb{R}^{d_u \times d_x}$; we denote these policies $\pi^K$, and let $\mathcal{J}_{\mathrm{LQR},\theta}(K) = \mathcal{J}_{\mathrm{LQR},\theta}[\pi^K]$. Here our decision $\mathfrak{a}$ is the controller $K$ and our loss $\mathcal{J}_\theta$ is $\mathcal{J}_{\mathrm{LQR},\theta}$. Under standard conditions, $\mathcal{J}_\theta(\cdot)$ admits a unique minimizer, which we denote $K_{\mathrm{opt}}(\theta)$. Furthermore, both $\mathcal{J}_\theta$ and $K_{\mathrm{opt}}$ are smooth functions of $K$ and $\theta$, respectively, and can be shown to satisfy Assumption 1.

**Example 1.2** (Inverse Reinforcement Learning)**.** In this setting, we assume there is some agent playing according to the control law $u_t^{\mathrm{agent}} = K^{\mathrm{agent}} x_t$ in the system $\theta_\star = (A_\star, B_\star)$, inducing the closed-loop dynamics $A_{\mathrm{cl},\star} = A_\star + B_\star K^{\mathrm{agent}}$. Furthermore, we assume that $K^{\mathrm{agent}} = K_{\mathrm{opt}}(\theta_\star; R_{\mathbf{u},\star})$ for some parameter $R_{\mathbf{u},\star} \in \mathbb{R}^{d_{\mathfrak{a}}}$ and a known map $K_{\mathrm{opt}}(\cdot;\cdot)$. $A_\star, B_\star$ and $R_{\mathbf{u},\star}$ are unknown, but we are told the value of $K^{\mathrm{agent}}$ (e.g., estimated through observation of the agent's actions). We assume we have access to the *closed-loop* system

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

and our goal is to infer the parameter, $R_{\mathbf{u},\star}$, the player is utilizing. This can be thought of as an inverse reinforcement learning problem, where we assume the agent is playing in order to minimize some cost parameterized by $R_{\mathbf{u},\star}$, and we want to determine what the cost is. In this setting our decision $\mathfrak{a}$ is the cost vector $R_{\mathbf{u}}$ and we define our loss as:

$$\mathcal{J}_{\mathrm{IRL},\theta}(R_{\mathbf{u}}) = \|R_{\mathbf{u}} - R_{\mathbf{u},\mathrm{opt}}(\theta)\|_{\mathrm{F}}^2$$

and the certainty equivalence estimate as:

$$R_{\mathbf{u},\mathrm{opt}}(\theta) = \underset{R_{\mathbf{u}}\in\mathbb{R}^{d_{\mathfrak{a}}}}{\arg\min} \|K^{\mathrm{agent}} - K_{\mathrm{opt}}(\theta; R_{\mathbf{u}})\|_{\mathrm{F}}^2.$$

Under amenable parameterizations of $K_{\mathrm{opt}}(\theta_\star; R_{\mathbf{u},\star})$, this will satisfy Assumption 1.

**Example 1.3** (System Identification with Parametric Uncertainty)**.** Consider the system identification problem where we only care about estimating particular entries of $(A_\star, B_\star)$—for example, the gain of a particular actuator or the friction coefficient of a surface. In this setting, we choose our loss to be:

$$\mathcal{J}_{\mathrm{SID},\theta}(\widehat{\theta}) = \|\widehat{\theta} - \theta_\star\|_M^2 := \mathrm{vec}(\widehat{\theta} - \theta_\star)^\top M \mathrm{vec}(\widehat{\theta} - \theta_\star)$$

where $M \succeq 0$ has a value of 0 at coordinates which correspond to the known entries of $(A_\star, B_\star)$ and a value of 1 at coordinates which correspond to the unknown entries of $(A_\star, B_\star)$. Our decision, $\widehat{\mathfrak{a}}$, is the least squares estimate of $\theta_\star$.

**Example 1.4** (Linear Experimental Design)**.** If $A_\star = 0$, $B_\star^\top = \phi_\star \in \mathbb{R}^{d_u}$, (1.1) reduces to

$$y_t = \phi_\star^\top u_t + w_t \tag{1.2}$$

for $y_t, w_t \in \mathbb{R}$, $u_t \in \mathbb{R}^{d_u}$. This is the standard linear regression setting, and our framework therefore encompasses optimal linear experiment design in arbitrary smooth losses (Pukelsheim, 2006). For example, we may consider the $A$-optimal objective $\mathcal{J}_{\mathrm{LED},\phi_\star}(\phi) = \|\phi - \phi_\star\|_2^2$. Alternatively, we could minimize the negative log-likelihood relative to some reference distribution $\nu$ so that $\mathcal{J}_{\mathrm{LED},\phi_\star}(\phi) = \mathbb{E}_{U\sim\nu, Y\sim p(\cdot|U,\phi_\star)}[-\log(p(Y|U,\phi))]$ where $P(Y|U,\phi_\star)$ is the likelihood of observations such that $y_t \sim p(\cdot|u_t,\phi_\star)$ (Chaudhuri & Mykland, 1993; Chaudhuri et al., 2015; Pronzato & Pázman, 2013). Non-smooth $G$-optimal-like objectives such as $J_{\phi_\star}(\phi) = \max_{x\in\mathcal{X}}\langle x, \widehat{\phi} - \phi_\star\rangle^2$ for some finite set $\mathcal{X} \subset \mathbb{R}^{d_u}$ can be captured in our framework by using an approximate smoothed objective such as $\mathcal{J}_{\mathrm{LED},\phi_\star}(\phi) = \frac{1}{\lambda}\log\left(\sum_{x\in\mathcal{X}} e^{\lambda\langle x,\phi-\phi_\star\rangle^2}\right)$ for large $\lambda$.

Many other examples exist—from more general control problems, to incentive design, and beyond. As we will show, there is a provable gain to performing task-guided exploration on examples such as these. We present our results for general loss functions $\mathcal{J}$, but consider several of the examples stated here in more detail in Section 3.

### 1.4. Related Works

**Experiment Design and Control.** Experiment design has over a century-old history in statistics, and numerous surveys have been written addressing its classical results (see e.g. (Pukelsheim, 2006; Pronzato & Pázman, 2013)). More recently, (Chaudhuri et al., 2015) gives a non-asymptotic active learning procedure for adaptive maximum likelihood estimation, again adapting the design to the unknown parameter of interest; unlike our work, their setting does not address dynamical systems.

In the controls literature, there has been significant attention devoted to optimally exciting dynamical systems (Mehra, 1976; Goodwin & Payne, 1977; Jansson & Hjalmarsson, 2005; Gevers et al., 2009; Manchester, 2010; Hägg et al., 2013) to optimize classical design criteria for system identification. More recent works (Hjalmarsson et al., 1996; Hildebrand & Gevers, 2002; Katselis et al., 2012) have focused on designing inputs to meet certain task-specific objectives, as is the focus of this work. In control, the optimal design depends on the unknown parameters of the system, and prior work rely on either robust experiment design (Rojas et al., 2007; 2011; Larsson et al., 2012; Hägg et al., 2013) or adaptive experimental design (Lindqvist & Hjalmarsson, 2001; Gerencsér & Hjalmarsson, 2005; Barenthin et al., 2005; Gerencsér et al., 2007; 2009), the method of choice in this work, to address this challenge. Past results were often heuristic, and rigorous bounds are asymptotic in nature (Gerencsér et al., 2007; 2009). In contrast, we provide finite sample upper bounds, and unconditional *information-theoretic* lower bounds which validate the optimality of our approach. Our adaptive algorithm also admits an efficient implementation via projected gradient descent, whereas past designs require the solution of semi-definite programs, which may be prohibitive in high dimensions.

More recently, (Wagenmaker & Jamieson, 2020) provided a finite sample analysis of system identification in the operator norm. Our work shows that designs which optimize operator norm recovery can fare arbitrarily worse for control tasks compared to task-optimal designs. Moreover, the techniques in this work translate to providing an efficient implementation of the computationally inefficient procedure proposed by (Wagenmaker & Jamieson, 2020). In addition, our lower bounds consider a more realistic "moderate $\delta$" regime (see Remark D.1 for comparison).

**Non-Asymptotic Learning for Control.** While the adaptive control problem has been extensively studied within the controls community (Åström & Wittenmark, 2013), machine learning has produced considerable recent interest in finite-time performance guarantees for system identification and control, guarantees which the classical adaptive control literature lacked. In the control setting, results have focused on finite time regret bounds for the LQR problem with unknown dynamics (Abbasi-Yadkori & Szepesvári, 2011; Dean et al., 2017; 2018; Mania et al., 2019; Dean et al., 2019; Cohen et al., 2019), with (Simchowitz & Foster, 2020) ultimately settling the minimax optimal regret in terms of dimension and time horizon. Others have considered regret in online adversarial settings (Agarwal et al., 2019; Simchowitz et al., 2020). Recent results in system identification have focused on obtaining finite time high probability bounds on the estimation error of the system's parameters when observing the evolution over time (Tu et al., 2017; Faradonbeh et al., 2018; Hazan et al., 2018; Hardt et al., 2018; Simchowitz et al., 2018; Sarkar & Rakhlin, 2018; Oymak & Ozay, 2019; Simchowitz et al., 2019; Sarkar et al., 2019; Tsiamis & Pappas, 2019). Existing results rely on excitation from random noise to guarantee learning and do not consider the problem of learning with arbitrary sequences of inputs or optimally choosing inputs for excitation. Recent work has begun to consider instance-optimal bounds with more targeted excitation (Wagenmaker & Jamieson, 2020; Ziemann & Sandberg, 2020); the former is discussed above. The latter presents an asymptotic, instance-dependent lower bound for the online LQR problem. Our results, in contrast, consider offline pure-exploration for a class of tasks much more general than LQR, and are finite-time. Furthermore, (Ziemann & Sandberg, 2020) do not provide a matching upper bound for their lower bound.

**Reinforcement Learning.** Viewing linear dynamical systems as a particular class of Markov Decision Processes (MDPs), our work can also be seen as studying PAC reinforcement learning (RL), where the goal is to find an $\epsilon$-good policy with probability $1 - \delta$ on a fixed MDP and reward function. Existing literature on PAC RL has tended to focus on obtaining coarse, worst-case bounds (Dann & Brunskill, 2015; Dann et al., 2017; 2019; Ménard et al., 2020). Only recently has progress been made in obtaining instance-dependent bounds, and here the results are either restricted to the much simpler generative model setting (Zanette et al., 2019; Marjani & Proutiere, 2020), or are asymptotic in nature and only apply to finding the *optimal* policy (Marjani et al., 2021). In contrast, our work provides tight, non-asymptotic, and instance-dependent upper and lower bounds for finding $\epsilon$-good policies, albeit in a restricted class of continuous RL problems.

### 1.5. Organization

The remainder of this paper is organized as follows. In Section 2 we provide an overview of our results, state an informal version of TOPLE, and introduce the essential quantities used in our analysis. Section 3 states several corollaries of our main result in specific settings and gives our bound for the LQR problem. Section 4 provides explicit examples where task-guided exploration yields provable gains over task-agnostic exploration, and Section 5 presents numerical experiments demonstrating that this improvement occurs in practice as well. We close in Section 6 with several interesting questions motivated by this work.

## 2. Summary of Results

We now turn to the presentation of our results. We assume we are in the setting described in Section 1.2. Throughout, we let $\mathcal{O}^\star(\cdot)$ suppress terms polynomial in problem parameters, $\log \frac{1}{\delta}$, and $\log \log T$; we let $a \lesssim b$ if $a \leq C \cdot b$ for a universal constant $C > 0$.

### 2.1. Optimality of Certainty Equivalence

Before describing the optimal exploration policy for collecting data, we resolve the optimal procedure for synthesizing a decision, and its sample complexity. Given a trajectory $\boldsymbol{\tau} = (x_{1:T+1}, u_{1:T})$, the *least squares* estimator of $\theta_\star$ is

$$\widehat{\theta}_{\mathrm{ls}}(\boldsymbol{\tau}) := \arg\min_{A,B} \sum_{t=1}^{T} \|x_{t+1} - Ax_t - Bu_t\|_2^2.$$

Note that $\widehat{\theta}_{\mathrm{ls}}$ is the maximum-likelihood estimator of $\theta_\star$. For our upper bounds, we propose the *certainty-equivalent* decision rule:

**Definition 2.1.** The *certainty equivalence* decision rule selects the optimal control policy for the least-squares estimate of the dynamics; $\mathsf{ce}(\boldsymbol{\tau}) := \mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_{\mathrm{ls}}(\boldsymbol{\tau}))$.

Certainty Equivalence has a long history in controller design (Theil, 1957; Simon, 1956). To analyze this strategy, we quantify both the error in our least squares estimator, and how it translates into uncertainty about the control synthesis. The former is quantified in terms of the expected covariance matrices under exploration policies:

$$\Gamma_T(\pi;\theta) := \frac{1}{T}\mathbb{E}_{\theta,\pi}\left[\sum_{t=1}^{T}\begin{bmatrix} x_t \\ u_t\end{bmatrix}\begin{bmatrix} x_t \\ u_t\end{bmatrix}^\top\right],$$
$$\boldsymbol{\Gamma}_T(\pi;\theta) := I_{d_x} \otimes \Gamma_T(\pi;\theta).$$

where $\otimes$ denotes the Kronecker product. The latter requires that we measure how uncertainty in $\theta$ translates into uncertainty about the optimal decision rule for the task of interest:

**Definition 2.2** (Model-Task Hessian and Idealized Risk)**.** We define the *model-task Hessian* as

$$\mathcal{H}(\theta_\star) := \nabla_\theta^2 \mathcal{R}_{\theta_\star}(\mathfrak{a}_{\mathrm{opt}}(\theta))\big|_{\theta=\theta_\star},$$

and the idealized risk

$$\Phi_T(\pi;\theta_\star) := \mathrm{tr}(\mathcal{H}(\theta_\star)\boldsymbol{\Gamma}_T(\pi;\theta_\star)^{-1}).$$

Intuiviely, the model-task Hessian measures the local curvature of $\mathcal{J}_{\theta_\star}(\mathfrak{a})$, as the decision $\mathfrak{a}$ varies along the directions of optimal policies $\mathfrak{a}_{\mathrm{opt}}(\theta)$ for parameters $\theta$ in a neighborhood of $\theta_\star$. The idealized risk capture how the least-squares error propagates through this uncertainty.

Our results will show that $\Phi$ characterizes the instance-optimal sample complexity for decision making, and consequently, by optimizing over $\pi$, of pure exploration. To formalize this, we require a notion of *local minimax risk*:

**Definition 2.3** (Decision-Making Local Minimax Risk)**.** Let $\mathcal{B} \subset \Theta$ denote a subset of instances. The $T$-sample local minimax decision risk on $\mathcal{B}$ under exploration policy $\pi_{\mathrm{exp}}$ is

$$\mathfrak{M}_{\pi_{\mathrm{exp}}}(\mathcal{R};\mathcal{B}) := \min_{\mathsf{dec}} \max_{\theta\in\mathcal{B}} \mathbb{E}_{\boldsymbol{\tau}\sim\theta,\pi_{\mathrm{exp}}}\left[\mathcal{R}_\theta(\mathsf{dec}(\boldsymbol{\tau}))\right]$$

where the minimization is over all maps from trajectories to decisions. Typically, we shall let $\mathcal{B}$ take the form $\mathcal{B}_{\mathrm{F}}(r;\theta_\star) := \{\theta : \|\theta - \theta_\star\|_{\mathrm{F}} \leq r\}$.

By choosing $\mathcal{B}$ to contain only instances close to $\theta_\star$, the local minimax risk captures the difficulty of completing our task on the specific instance $\theta_\star$, yielding an effectively instance-specific lower bound. Finally, we make the following assumption on the system dynamics.

**Assumption 2.** *Let $\Theta$ denote the set of all stable $\theta$: $\Theta := \{\theta = (A,B) : \rho(A) < 1\}$, where $\rho(A)$ denotes the spectral radius of $A$. We assume that $\theta_\star \in \Theta$.*

While this assumption restricts our results to stable systems, similar assumptions are standard in much of the recent literature. Appendix C discusses generalization to unstable systems. Under this assumption we have the following result.

**Theorem 2.1** (Optimality of Certainty Equivalence)**.** *Let $\pi_{\mathrm{exp}}$ be any sufficiently regular policy, and consider $\mathcal{J}_\theta(\mathfrak{a})$ and $\mathfrak{a}_{\mathrm{opt}}(\theta)$ satisfying Assumption 1 and $\theta_\star$ satisfying Assumption 2. Then for all $\delta \in (0, 1/3)$ and all $T$ sufficiently large, a trajectory $\boldsymbol{\tau}$ generated by $\pi_{\mathrm{exp}}$ and $\theta_\star$ satisfies the following with probability $1-\delta$,*

$$\mathcal{R}_{\theta_\star}(\mathsf{ce}(\boldsymbol{\tau})) \lesssim \sigma_w^2 \log(\tfrac{d_x}{\delta}) \cdot \frac{\Phi_T(\pi_{\mathrm{exp}};\theta_\star)}{T} + \mathcal{O}^\star\left(\tfrac{1}{T^{3/2}}\right).$$

*Moreover, for some $r = \Omega(1/T^{5/12})$, and $\mathcal{B} = \mathcal{B}_{\mathrm{F}}(r;\theta_\star)$, the synthesis minimax risk is lower bounded as*

$$\mathfrak{M}_{\pi_{\mathrm{exp}}}(\mathcal{R};\mathcal{B}) \geq \frac{\sigma_w^2}{3} \cdot \frac{\Phi_T(\pi_{\mathrm{exp}};\theta_\star)}{T} - \mathcal{O}^\star\left(\tfrac{1}{T^{5/4}}\right).$$

In Appendix B we state the full version of this result, which holds in a more general *martingale decision making* setting encompassing certain instances of the nonlinear system formulation considered in (Mania et al., 2020). This result establishes that, for a *given* exploration policy $\pi_{\mathrm{exp}}$ and for $T$ sufficiently large, the certainty equivalence decision $\mathsf{ce}(\boldsymbol{\tau})$ is the locally minimax optimal synthesis rule—there does not exist a more efficient way to utilize the acquired information to produce a decision. Note that, under some reasonable assumptions, an expectation bound can be obtained from the high probability bound. We precisely quantify what it means for a policy to be sufficiently regular in Appendix B. In short, it entails that the policy sufficiently excites the system, and that the covariates concentrate to their mean. Lastly, note that our lower bound differs substantively from the $\delta \to 0$ lower bounds common in the adaptive estimation literature (see Remark D.1). Appendix C provides a more thorough discussion of these points.

*Proof Sketch of Theorem 2.1.* For the proof of the lower bound, we first show that for any decision $\widehat{\mathfrak{a}} = \mathsf{dec}(\boldsymbol{\tau})$ for which $\mathcal{R}_\theta(\widehat{\mathfrak{a}})$ is small, we can infer an instance $\widehat{\theta}(\widehat{\mathfrak{a}})$ such that $\|\widehat{\theta}(\widehat{\mathfrak{a}}) - \theta\|_{\mathcal{H}(\theta_\star)}^2$ is also small (see Appendix E). This equivalence reduces our problem to that of estimating $\theta$ in the $\mathcal{H}(\theta_\star)$ norm. We then show a lower bound on a Gaussian martingale regression problem with general quadratic losses via a careful though elementary Bayesian MMSE computation. Unlike vanilla Cramer-Rao, this approach allows us to obtain a lower bound which holds for any estimator, not simply unbiased estimators (see Appendix D). Combining these results gives the stated lower bound. The proof of our upper bound mirrors this: we approximate $\mathcal{R}_{\theta_\star}(\mathsf{ce}(\boldsymbol{\tau}))$ as a quadratic, $\|\widehat{\theta}(\boldsymbol{\tau}) - \theta_\star\|_{\mathcal{H}(\theta_\star)}^2$, and prove an upper bound on the error of the least squares estimator for martingale regression in general norms (see Appendix F). □

### 2.2. Task-Optimal Experiment Design

Given that the optimal risk for a fixed exploration policy $\pi_{\mathrm{exp}}$ is governed by

$$\Phi_T(\pi_{\mathrm{exp}}; \theta_\star) = \mathrm{tr}(\mathcal{H}(\theta_\star)\boldsymbol{\Gamma}_T(\pi_{\mathrm{exp}}; \theta_\star)^{-1}),$$

it stands to reason that the optimal design procedure seeks to minimize this quantity. To this end, we introduce several quantities describing the optimality properties.

**Definition 2.4** (Power-Constrained Policies)**.** Let $\Pi_{\gamma^2}$ denote the set of causal polices that have expected average power bounded as $\gamma^2$. That is, for any $\pi \in \Pi_{\gamma^2}$, we will have $\mathbb{E}_{\theta,\pi}[\sum_{t=1}^T \|u_t\|_2^2] \le T\gamma^2$ for all $\theta$.

**Definition 2.5** (Optimal Risk)**.** We define:

$$\Phi_{\mathrm{opt}}(\gamma^2; \theta_\star) := \liminf_{T \to \infty} \inf_{\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}} \Phi_T(\pi_{\mathrm{exp}}; \theta_\star),$$

the risk obtained by the policy minimizing the complexity $\Phi_T(\pi_{\mathrm{exp}}; \theta_\star)$.

**Definition 2.6** (Exploration Local Minimax Risk)**.** Let $\mathcal{B} \subset \Theta$ denote a subset of instances. The $T$-sample local minimax exploration risk on $\mathcal{B}$ with budget $\gamma^2$ is

$$\mathfrak{M}_{\gamma^2}(\mathcal{R}; \mathcal{B}) := \min_{\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}} \min_{\mathsf{dec}} \max_{\theta \in \mathcal{B}} \mathbb{E}_{\boldsymbol{\tau} \sim \theta, \pi_{\mathrm{exp}}}[\mathcal{R}_\theta(\mathsf{dec}(\boldsymbol{\tau}))].$$

**Algorithm Sketch.** We are now ready to state our algorithm. TOPLE proceeds in epochs. At each epoch it chooses a policy $\pi$ that minimize the certainty-equivalence design objective, $\Phi_T(\pi; \widehat{\theta})$, based on the estimate of the system's parameters produced in the previous epoch. As the estimate of $\theta_\star$ is refined, the exploration policy is improved, and ultimately achieves near-optimal excitation of the system for the task of interest.

In the policy optimization step on Line 6, we optimize over a restricted class of policies, $\Pi_{\gamma^2}^{\mathrm{P}}$, which contains only *periodic* signals. As we show, this restriction is expressive enough to contain a near-optimal policy, while allowing us to represent $\boldsymbol{\Gamma}_T(\pi; \widehat{\theta}_i)$ in a convenient frequency-domain form. We then adopt a (sharp) *convex relaxation* of these policies that transforms the experiment design into a convex program, admitting a simple, efficient projected gradient descent implementation. A formal definition of TOPLE and detailed explanation of these points is given in Appendix B.

**Algorithm 1 T**ask-**OP**tima**L E**xperiment Design (TOPLE), Informal

```
1: **Input:** Initial epoch length $T_0$, budget $\gamma^2$
2: $\pi_0 \leftarrow \mathcal{N}(0, \gamma^2/d_u \cdot I)$.
3: **for** phase $i = 0, 1, 2, \ldots$ **do**
4:     Run system for $T_0 2^i$ steps, playing input
           $u_t = \pi_i(x_{1:t}, u_{1:t-1})$
5:     Compute least squares estimate
           $\widehat{\theta}_i \in \arg\min_{A,B} \sum_{t=1}^T \|x_{t+1} - Ax_t - Bu_t\|_2^2$
6:     Select policy for epoch $i+1$,
           $\pi_{i+1} \leftarrow \arg\min_{\Pi_{\gamma^2}^{\mathrm{P}}} \mathrm{tr}(\mathcal{H}(\widehat{\theta}_i)\boldsymbol{\Gamma}_T(\pi; \widehat{\theta}_i)^{-1})$
```

**Theorem 2.2** (Task-Optimal Experiment Design)**.** *Consider $\mathcal{J}_\theta(\mathfrak{a})$ and $\mathfrak{a}_{\mathrm{opt}}(\theta)$ satisfying Assumption 1 and $\theta_\star$ satisfying Assumption 2. For sufficiently large $T$, the trajectory $\boldsymbol{\tau}$ generated by Algorithm 1 enjoys the following guarantee with probability at least $1 - \delta$:*

$$\mathcal{R}_{\theta_\star}(\mathsf{ce}(\boldsymbol{\tau})) \lesssim \sigma_w^2 \log(\tfrac{d_x}{\delta}) \cdot \frac{\Phi_{\mathrm{opt}}(\gamma^2; \theta_\star)}{T} + \mathcal{O}^\star\left(\tfrac{1}{T^{3/2}}\right).$$

*Moreover, it produces inputs satisfying $\mathbb{E}_{\theta_\star,\mathrm{TOPLE}}[\sum_{t=1}^T \|u_t\|_2^2] \le T\gamma^2$, and can be implemented in polynomial time. Finally, for $r = \mathcal{O}^\star(1/T^{5/12})$ and $\mathcal{B} = \mathcal{B}_{\mathrm{F}}(r; \theta_\star)$, the local minimax risk is lower bounded by*

$$\mathfrak{M}_{\gamma^2}(\mathcal{R}; \mathcal{B}) \ge \frac{\sigma_w^2}{64} \cdot \frac{\Phi_{\mathrm{opt}}(\gamma^2; \theta_\star)}{T} - \mathcal{O}^\star\left(\tfrac{1}{T^{5/4}}\right).$$

We emphasize that the only assumptions needed for Theorem 2.2 to hold are that our system, $\theta_\star$, is stable, and that the loss we are considering, $\mathcal{R}_{\theta_\star}(\mathfrak{a})$, is sufficiently smooth. For *any* system and *any* loss satisfying these minimal assumptions, including those stated in Section 1.3, Theorem

2.2 shows that certainty equivalence decision making is instance-wise optimal, and that TOPLE hits this optimal rate. Furthermore, while TOPLE relies on experiment design, its sample complexity is also optimal over algorithms which incorporate feedback. We precisely quantify the lower order terms and burn-in times necessary for this result to hold in Appendix B, and consider relaxations to our assumptions in Appendix C.

*Proof Sketch of Theorem 2.2.* The key technical difficulty lies in proving that our restricted class of policies, $\Pi^{\mathrm{p}}_{\gamma^2}$, contains a near-optimal policy. We show this in Appendix H by a careful truncation argument and application of Caratheodory's Theorem. Given this, the lower bound follows by a similar argument as in Theorem 2.1. For the upper bound, we show that once $\theta_\star$ has been estimated well enough, the certainty equivalence experiment design on Line 6 achieves the near-optimal rate (see Appendix J). □

## 3. Interpreting the Results

To make our results more concrete, we return to the examples introduced in Section 1.3, and show how TOPLE applies in these settings.

### 3.1. Instance-Optimal LQR Synthesis

Consider the pure exploration LQR problem stated in Example 1.1. We define

$$\mathcal{R}_{\mathrm{LQR},\theta_\star}(K) := \mathcal{J}_{\mathrm{LQR},\theta_\star}(K) - \min_K \mathcal{J}_{\mathrm{LQR},\theta_\star}(K)$$

where $\mathcal{J}_{\mathrm{LQR},\theta_\star}(K)$ is given in Example 1.1. Recall the *discrete algebraic Ricatti equation*, defined for some $(A, B)$:

$$P = A^\top P A - A^\top P B(R_{\mathbf{u}} + B^\top P B)B^\top P A + R_{\mathbf{x}}$$

If $\theta$ is stabilizable and $R_{\mathbf{x}}, R_{\mathbf{u}} \succ 0$, it is a well-known fact that this has a unique solution, $P \succeq 0$. We denote the solution for the instance $\theta_\star = (A_\star, B_\star)$ by $P_\star$. We also recall the definition of the $\mathcal{H}$-infinity norm of a system:

$$\|A_\star\|_{\mathcal{H}_\infty} = \max_{\omega \in [0, 2\pi]} \|(e^{\iota\omega} I - A_\star)^{-1}\|_{\mathrm{op}}$$

Finally, we let $\Phi_{\mathrm{LQR}}(\gamma^2; \theta_\star) := \Phi_{\mathrm{opt}}(\gamma^2; \theta_\star)$ in the case when our loss is the LQR loss, $\mathcal{J}_{\theta_\star} = \mathcal{J}_{\mathrm{LQR},\theta_\star}$. Given these definitions, the following corollary shows the performance of TOPLE on the pure exploration LQR problem, and that relevant quantities can be expressed in terms of the problem-dependent constants $\|P_\star\|_{\mathrm{op}}$, $\|B_\star\|_{\mathrm{op}}$, and $\|A_\star\|_{\mathcal{H}_\infty}$.

**Corollary 1.** *As long as $T \geq C_{\mathrm{LQR}}(d_x \log^2 T + d_x^2)$, with probability at least $1 - \delta$, TOPLE achieves the following rate for the LQR problem:*

$$\mathcal{R}_{\mathrm{LQR},\theta_\star}(\mathsf{ce}(\boldsymbol{\tau})) \lesssim \sigma_w^2 \log(\tfrac{d_x}{\delta}) \cdot \frac{\Phi_{\mathrm{LQR}}(\gamma^2; \theta_\star)}{T} + \tfrac{C_{\mathrm{LQR}} d_x^5}{T^{3/2}}.$$

*Furthermore, any algorithm must incur the following loss:*

$$\mathfrak{M}_{\gamma^2}(\mathcal{R}_{\mathrm{LQR}}; \mathcal{B}) \geq \frac{\sigma_w^2}{64} \cdot \frac{\Phi_{\mathrm{LQR}}(\gamma^2; \theta_\star)}{T} - \tfrac{C_{\mathrm{LQR}} d_x^5}{T^{5/4}}$$

*where $\mathcal{B}$ is as in Theorem 2.2 and $C_{\mathrm{LQR}} = C'_{\mathrm{LQR}} / \min\{\sigma_w^6, \gamma^6/d_u^3, 1\}$ for $C'_{\mathrm{LQR}}$ polynomial in $\|P_\star\|_{\mathrm{op}}$, $\|B_\star\|_{\mathrm{op}}$, $\|B_\star\|_{\mathrm{op}}^{-1}, \|A_\star\|_{\mathcal{H}_\infty}, \|R_{\mathbf{u}}\|_{\mathrm{op}}, \gamma^2, \sigma_w^2, d_u$, $\log\log T$, and $\log \frac{1}{\delta}$.*

As this result shows, TOPLE is instance-optimal for the LQR problem, with sample complexity governed by the constant $\Phi_{\mathrm{LQR}}(\gamma^2; \theta_\star)$. To the best of our knowledge, this is the first algorithm provably instance-optimal for LQR—albeit in the offline LQR setting.

### 3.2. System Identification in Arbitrary Norms

Next, we consider the case of system identification in arbitrary norms outlined in Example 1.3. In this setting our loss is $\mathcal{R}_{\mathrm{SID},\theta_\star}(\widehat{\theta}) := \mathcal{J}_{\mathrm{SID},\theta_\star}(\widehat{\theta}) = \|\widehat{\theta} - \theta_\star\|_M^2$, and it can be shown our idealized risk is $\Phi_T(\pi_{\mathrm{exp}}; \theta_\star) = \mathrm{tr}(M \boldsymbol{\Gamma}_T(\pi_{\mathrm{exp}}; \theta_\star)^{-1})$. Defining

$$\Phi_{\mathrm{SID}}(\gamma^2; \theta_\star) := \liminf_{T \to \infty} \inf_{\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}} \mathrm{tr}(M \boldsymbol{\Gamma}_T(\pi_{\mathrm{exp}}; \theta_\star)^{-1})$$

Theorem 2.2 implies that

$$\mathcal{R}_{\mathrm{SID},\theta_\star}(\mathsf{ce}(\boldsymbol{\tau})) \lesssim \sigma_w^2 \log(\tfrac{d_x}{\delta}) \cdot \frac{\Phi_{\mathrm{SID}}(\gamma^2; \theta_\star)}{T} + \tfrac{C_{\mathrm{SID}} \mathrm{tr}(M) d_x^3}{T^{3/2}}$$

and that this rate is instance-optimal, for some constant $C_{\mathrm{SID}}$ polynomial in $\|B_\star\|_{\mathrm{op}}$, $\|A_\star\|_{\mathcal{H}_\infty}$, $\gamma^2, \sigma_w^2, d_u$, $\log \frac{1}{\delta}$, and $\log\log T$. In particular, if $M = I$ our loss $\mathcal{R}_{\mathrm{SID},\theta_\star}(\widehat{\theta})$ reduces to the Frobenius norm, implying that TOPLE is the optimal Frobenius norm identification algorithm.

## 4. Task-Guided Exploration yields Provable Gains

We turn now to several examples which illustrate that taking into account the task of interest when performing exploration yields provable gains over task-agnostic exploration schemes. We focus on the LQR setting and compare against the following natural exploration baselines:

- **System Identification in Operator Norm** (Wagenmaker & Jamieson, 2020): Let $\pi_{\mathrm{op}}$ denote the exploration policy that is optimal for estimating $\theta_\star = (A_\star, B_\star)$ under the operator norm $\|\widehat{\theta} - \theta_\star\|_{\mathrm{op}} = \sup_{u : \|u\|_2 \leq 1} \|(\widehat{\theta} - \theta_\star)u\|_2$. Explicitly, $\pi_{\mathrm{op}} = \arg\max_{\pi \in \Pi_{\gamma^2}} \lambda_{\min}(\boldsymbol{\Gamma}_T(\pi; \theta_\star))$.
- **System Identification in Frobenius Norm**: Let $\pi_{\mathrm{fro}}$ denote the exploration policy that is optimal for estimating $\theta_\star = (A_\star, B_\star)$ under the Frobenius norm: $\pi_{\mathrm{fro}} = \arg\min_{\pi \in \Pi_{\gamma^2}} \mathrm{tr}(\boldsymbol{\Gamma}_T(\pi; \theta_\star)^{-1})$.

- **Task-Optimal Gaussian Noise**: Let $\pi_{\text{noise}}$ denote the exploration policy such that $\pi_{\text{noise}} =: \pi_{\text{noise}}(\Lambda_\star)$ where $\pi_{\text{noise}}(\Lambda_\star)$ plays the inputs $u_t \sim \mathcal{N}(0, \Lambda_\star)$ and $\Lambda_\star = \arg\min_{\Lambda : \mathrm{tr}(\Lambda) \leq \gamma^2} \mathrm{tr}(\mathcal{H}(\theta_\star) \mathbf{\Gamma}_T(\pi_{\text{noise}}(\Lambda); \theta_\star)^{-1})$.

In stating our results, we overload notation and let $\mathcal{R}_{\text{LQR},\theta_\star}(\pi_{\text{exp}}) = \mathcal{R}_{\text{LQR},\theta_\star}(\mathsf{ce}(\boldsymbol{\tau}))$ for $\boldsymbol{\tau} \sim \pi_{\text{exp}}, \theta_\star$. We are concerned primarily in how the complexity scales with the dimension, $d_x$, and $\frac{1}{1-\rho}$ where $\rho$ is the spectral radius of the system, and use $\Theta(\cdot)$ and $\mathcal{O}(\cdot)$ to suppress lower order dependence on these terms. Our first example shows that, if $(A_\star, B_\star)$ is properly structured, TOPLE achieves a tighter scaling in $\frac{1}{1-\rho}$ than all naive exploration approaches.

**Proposition 4.1.** *Consider the system $A_\star = \rho \mathbf{e}_1 \mathbf{e}_1^\top$, $B_\star = bI$, $R_{\mathbf{x}} = \kappa I$, and $R_{\mathbf{u}} = \mu I$. There exist values of $b, \kappa, \mu$, and $\sigma_w$ such that the loss of* TOPLE*, optimal operator norm identification (Wagenmaker & Jamieson, 2020), optimal Frobenius norm identification, and optimally exciting Gaussian noise have the following scalings:*

$$\mathcal{R}_{\text{LQR},\theta_\star}(\text{TOPLE}) = \mathcal{O}\left(\frac{d_x^2}{(1-\rho)^2} \frac{\sigma_w^2}{\gamma^2 T}\right)$$

$$\mathcal{R}_{\text{LQR},\theta_\star}(\pi_{\text{fro}}) = \Theta\left(\left(\frac{d_x^2}{(1-\rho)^2} + \frac{d_x}{(1-\rho)^{5/2}}\right) \frac{\sigma_w^2}{\gamma^2 T}\right)$$

$$\mathcal{R}_{\text{LQR},\theta_\star}(\pi_{\text{op}}) = \Theta\left(\left(\frac{d_x^2}{(1-\rho)^2} + \frac{d_x}{(1-\rho)^3}\right) \frac{\sigma_w^2}{\gamma^2 T}\right)$$

$$\mathcal{R}_{\text{LQR},\theta_\star}(\pi_{\text{noise}}) = \Theta\left(\left(\frac{(1-\rho)^{-1} + d_x^4(1-\rho)}{(1-\rho)^2}\right) \frac{\sigma_w^2}{\gamma^2 T}\right).$$

TOPLE achieves the optimal scaling in $\frac{1}{1-\rho}$ and, as $\rho \to 1$, will outperform other approaches by an arbitrarily large factor. In addition, we note that Frobenius norm identification outperforms operator norm identification for this task. A key ingredient in the proof of this result is our convex relaxation of the optimal policy computation. Intuitively, on this instance, the first coordinate is easily excited and $\pi_{\text{op}}$ and $\pi_{\text{fro}}$ will therefore devote the majority of their energy to reducing the uncertainty in the remaining coordinates. However, the LQR cost will primarily be incurred in the first coordinate due to the same effect—this coordinate is easily excited and therefore the first coordinate of the state grows at a much faster rate. As such, the task-optimal allocation does the opposite of $\pi_{\text{op}}$ and $\pi_{\text{fro}}$ and seeks to learn the first coordinate more precisely than the remaining coordinates so as to mitigate this growth.

In our next example, our system behaves isotropically but our costs are non-isotropic. As a result, certain directions incur greater cost than others, and the task-optimal allocation seeks to primarily reduce uncertainty in these directions.

**Proposition 4.2.** *Consider the system $A_\star = \rho I, B_\star = I, R_{\mathbf{x}} = I + \kappa e_1 e_1^\top$ and $R_{\mathbf{u}} = \mu I$. Then there exists a choice of $\mu, \kappa$, and $\sigma_w$ such that*

$$\mathcal{R}_{\text{LQR},\theta_\star}(\text{TOPLE}) = \mathcal{O}\left(\frac{1}{(1-\rho)^4} \frac{\sigma_w^2}{\gamma^2 T}\right)$$

$$\mathcal{R}_{\text{LQR},\theta_\star}(\pi_{\text{op}}) = \mathcal{R}_{\text{LQR},\theta_\star}(\pi_{\text{fro}}) = \Theta\left(\frac{d_x}{(1-\rho)^4} \frac{\sigma_w^2}{\gamma^2 T}\right)$$

$$\mathcal{R}_{\text{LQR},\theta_\star}(\pi_{\text{noise}}) = \Theta\left(\frac{d_x^2}{(1-\rho)^4} \frac{\sigma_w^2}{\gamma^2 T}\right).$$

We note that TOPLE improves on task-agnostic exploration by a factor of at least the dimensionality. These examples make clear that, in the setting of a linear dynamical system, when our goal is to perform a specific task, exploration agnostic to this task can be arbitrarily suboptimal.

### 4.1. Suboptimality of Low-Regret Algorithms

In contrast to our pure-exploration setting, where we do not incur cost during exploration, a significant body of work exists on regret-minimization for the *online* LQR problem with unknown $A_\star, B_\star$. Here the goal is to choose a *low regret* policy $\pi_{\text{lr}}$ so as to minimize

$$\text{Reg}_T := \mathbb{E}_{\theta_\star, \pi_{\text{lr}}}\left[\sum_{t=1}^T \ell(x_t, u_t)\right] - T \min_K J_{\text{LQR},\theta_\star}(K)$$

for $\ell(x_t, u_t)$ as defined in Example 1.1. While our objectives differ, it would seem a natural strategy to run a low-regret algorithm for $T$ steps to obtain a controller $K_{\text{lr}}$, and then evaluate the cost $\mathcal{J}_{\text{LQR},\theta_\star}$ on this $K_{\text{lr}}$. The following result shows that there is a fundamental tradeoff between regret and estimation; in particular, the optimal $\Theta_\star(\sqrt{T})$ (see (Simchowitz & Foster, 2020)) regret translates to a (very suboptimal) $\Omega_\star(1/\sqrt{T})$ excess risk $\mathcal{R}_{\theta_\star,\text{LQR}}(K_{\text{lr}})$.

**Proposition 4.3** (Suboptimality of Low Regret, Informal)**.** *For any sufficiently large $T$ and any regret bound $R \in [\sqrt{T}, T]$, any policy $\pi_{\text{lr}}$ with regret $\mathbb{E}_{\pi_{\text{lr}},\theta_\star}[\text{Reg}_T] \leq R$ which returns a controller $K_{\text{lr}}$ as a function of its trajectory must have $\mathbb{E}_{\theta_\star,\pi_{\text{lr}}}[\mathcal{R}_{\text{LQR},\theta_\star}(K_{\text{lr}})] = \Omega_\star(\frac{d_u^2 d_x}{R})$.*

In particular, Proposition 4.3 implies that popular low-regret strategies, such as optimism-in-the-face-of-uncertainty (Abbasi-Yadkori & Szepesvári, 2011; Abeille & Lazaric, 2020), are highly suboptimal in our setting. The key intuition behind the proof is that low regret algorithms converge to inputs $\mathbf{u}_t \approx K_\star \mathbf{x}_t$ approaching the optimal control policy; in doing so, they under-explore directions *perpendicular* to the hyperplane $\{(x, u) : u = K_\star x\}$, which are necessary for identifying the optimal control policy. We formally state and prove this result in Appendix C.7.

## 5. Numerical Experiments

Finally, we show that task-guided exploration yields practical gains. Figures 1, 2, and 3 illustrate the performance of TOPLE on several instances of the pure-exploration LQR problem. We compare against the baselines presented in Section 4 and the oracle task-optimal algorithm (which we refer to as "TOPLE Oracle"). For all baselines, we compute the inputs in an oracle, offline manner, using knowledge of $A_\star$

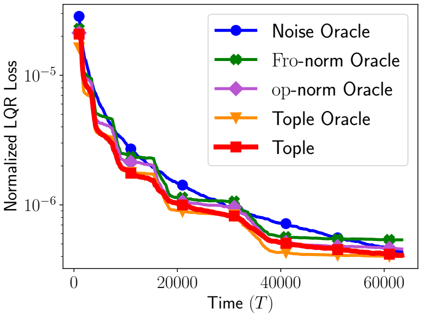

*Figure 1.* LQR loss vs time on $A_\star$ a Jordan block and $B_\star$, $R_\mathbf{x}$, $R_\mathbf{u}$ randomly generated.

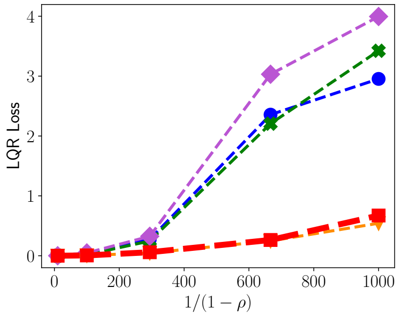

*Figure 2.* LQR loss when varying $\rho$ on example stated in Proposition 4.1.

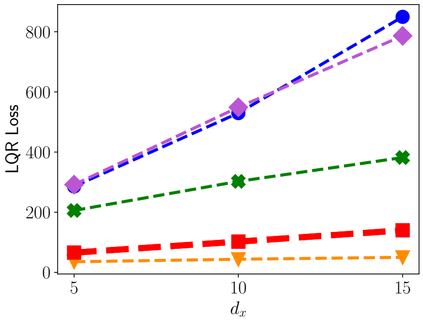

*Figure 3.* LQR loss when varying $d_x$ on example stated in Proposition 4.2.

and $B_\star$, and play them for the entire trajectory. Our implementation of TOPLE follows precisely the formal statement of the algorithm given in Appendix B, and we rely on the aforementioned convex relaxation and a projected gradient descent solution to efficiently solve the experiment design problem. This convex relaxation can, in fact, also be applied to the optimal operator norm identification algorithm, rendering the algorithm from (Wagenmaker & Jamieson, 2020) computationally efficient. We therefore rely on this relaxation and a projected subgradient descent method in our implementation of the operator norm identification algorithm. All data points correspond to averaging over at least 50 runs of the algorithm. Additional details and plots with error bars are provided in Appendix M.

Figures 2 and 3 illustrate performance on the instances stated in Proposition 4.1 and Proposition 4.2, respectively. Every point in the plot corresponds to the LQR loss obtained after $T = 60000$ steps. As these plots clearly illustrate, the theoretical gains stated in Proposition 4.1 and Proposition 4.2 appear in practice as well—there is a clear improvement in terms of the scaling in $\rho$ and $d_x$ when performing task-guided exploration, even over moderate time regimes. Figure 1 illustrates the performance of TOPLE on a more "typical" problem instance: $A_\star$ a single Jordan block and $B_\star, R_\mathbf{x}$, and $R_\mathbf{u}$ randomly generated. Figure 1 gives the average loss versus time obtained by averaging the performance over 15 different realizations of $B_\star, R_\mathbf{x}, R_\mathbf{u}$. As in the previous examples, TOPLE outperforms all other approaches.

## 6. Conclusion

In this work, we have shown that task-guided exploration of an unknown environment yields significant improvements over task-agnostic exploration. Furthermore, we have derived an instance- and task-optimal exploration algorithm which applies to a wide range of decision making problems, and derived corresponding instance- and task-dependent lower bounds. Our results also establish that certainty equivalence decision making is optimal, and we obtain the first instance-optimal algorithm for the LQR problem. This work raises several interesting questions:

- While our martingale decision making setting encompasses certain classes of nonlinear systems, all our results fundamentally rely on linear observations of the parameter of interest, $\theta_\star$. Task-optimal exploration remains an open question for general nonlinear systems, and is an interesting future direction.

- We show that the smoothness conditions on our loss are met by a wide range of decision making problems. However, it remains an interesting future direction to obtain an optimal algorithm that holds without these smoothness assumptions. As (Wagenmaker & Jamieson, 2020) shows, when the loss is the operator norm—which we note does not satisfy our smoothness assumption—the optimal algorithm takes a form very similar to TOPLE. Does a general algorithm and analysis exist for both smooth and non-smooth losses?

- Our work focuses on the offline, pure-exploration setting. Extending our analysis to obtain instance- and task-optimal rates in the *online* setting is an interesting direction of future work. For the online LQR problem, (Simchowitz & Foster, 2020) obtain the optimal scaling in terms of dimension but their rates are suboptimal in terms of other problem-dependent constants. On the lower bound side, (Ziemann & Sandberg, 2020) provide an instance-dependent lower bound but give no upper bound. Solving this problem may require new algorithmic ideas, and we leave this for future work.

### ACKNOWLEDGEMENTS

The work of AW is supported by an NSF GFRP Fellowship DGE-1762114. MS is generously supported by an Open Philanthropy AI Fellowship. The work of KJ is supported in part by grants NSF RI 1907907 and NSF CCF 2007036.

## References

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

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Abeille, M. and Lazaric, A. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *International Conference on Machine Learning*, pp. 23–31. PMLR, 2020.

Agarwal, N., Bullins, B., Hazan, E., Kakade, S., and Singh, K. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pp. 111–119. PMLR, 2019.

Anderson, J., Doyle, J. C., Low, S. H., and Matni, N. System level synthesis. *Annual Reviews in Control*, 47:364–393, 2019.

Arias-Castro, E., Candes, E. J., and Davenport, M. A. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Åström, K. J. and Wittenmark, B. *Adaptive control*. Courier Corporation, 2013.

Barenthin, M., Jansson, H., and Hjalmarsson, H. Applications of mixed h2 and hinfin; input design in identification. *IFAC Proceedings Volumes*, 38(1):458–463, 2005.

Chaudhuri, K., Kakade, S., Netrapalli, P., and Sanghavi, S. Convergence rates of active learning for maximum likelihood estimation. *arXiv preprint arXiv:1506.02348*, 2015.

Chaudhuri, P. and Mykland, P. A. Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.

Cohen, A., Koren, T., and Mansour, Y. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. *arXiv preprint arXiv:1510.08906*, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pp. 4188–4197, 2018.

Dean, S., Tu, S., Matni, N., and Recht, B. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, pp. 5582–5588. IEEE, 2019.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Gerencsér, L. and Hjalmarsson, H. Adaptive input design in system identification. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 4988–4993. IEEE, 2005.

Gerencsér, L., Mårtensson, J., and Hjalmarsson, H. Adaptive input design for arx systems. In *2007 European Control Conference (ECC)*, pp. 5707–5714. IEEE, 2007.

Gerencsér, L., Hjalmarsson, H., and Mårtensson, J. Identification of arx systems with non-stationary inputs—asymptotic analysis with application to adaptive input design. *Automatica*, 45(3):623–633, 2009.

Gevers, M., Bazanella, A. S., Bombois, X., and Miskovic, L. Identification and the information matrix: how to get just sufficiently rich? *IEEE Transactions on Automatic Control*, 54(ARTICLE):2828–2840, 2009.

Gill, R. D., Levit, B. Y., et al. Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, 1 (1-2):59–79, 1995.

Goodwin, G. C. and Payne, R. L. *Dynamic system identification: experiment design and data analysis*. Academic press, 1977.

Hägg, P., Larsson, C. A., and Hjalmarsson, H. Robust and adaptive excitation signal generation for input and output constrained systems. In *2013 European Control Conference (ECC)*, pp. 1416–1421. IEEE, 2013.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pp. 4634–4643, 2018.

Hildebrand, R. and Gevers, M. Identification for control: optimal input design with respect to a worst-case $\nu$-gap cost function. *SIAM Journal on Control and optimization*, 41(5):1586–1608, 2002.

Hjalmarsson, H., Gevers, M., and De Bruyne, F. For model-based control design, closed-loop identification gives better performance. *Automatica*, 32(12):1659–1673, 1996.

Hsu, D., Kakade, S., Zhang, T., et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Jansson, H. and Hjalmarsson, H. Input design via lmis admitting frequency-wise model specifications in confidence regions. *IEEE transactions on Automatic Control*, 50(10):1534–1549, 2005.

Jedra, Y. and Proutiere, A. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 2676–2681. IEEE, 2019.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

Katselis, D., Rojas, C. R., Hjalmarsson, H., and Bengtsson, M. Application-oriented finite sample experiment design: A semidefinite relaxation approach. *IFAC Proceedings Volumes*, 45(16):1635–1640, 2012.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Larsson, C., Geerardyn, E., and Schoukens, J. Robust input design for resonant systems under limited a priori information. *IFAC Proceedings Volumes*, 45(16):1611–1616, 2012.

Lindqvist, K. and Hjalmarsson, H. Identification for control: Adaptive input design using convex optimization. In *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No. 01CH37228)*, volume 5, pp. 4326–4331. IEEE, 2001.

Manchester, I. R. Input design for system identification via convex relaxation. In *49th IEEE Conference on Decision and Control (CDC)*, pp. 2041–2046. IEEE, 2010.

Mania, H., Tu, S., and Recht, B. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Marjani, A. A. and Proutiere, A. Best policy identification in discounted mdps: Problem-specific sample complexity. *arXiv preprint arXiv:2009.13405*, 2020.

Marjani, A. A., Garivier, A., and Proutiere, A. Navigating to the best policy in markov decision processes. *arXiv preprint arXiv:2106.02847*, 2021.

Mehra, R. K. Synthesis of optimal inputs for multiinput-multioutput (mimo) systems with process noise part i: Frequenc y-domain synthesis part ii: Time-domain synthesis. In *Mathematics in Science and Engineering*, volume 126, pp. 211–249. Elsevier, 1976.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442*, 2020.

Ok, J., Proutiere, A., and Tranos, D. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 8874–8882, 2018.

Oymak, S. and Ozay, N. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American Control Conference (ACC)*, pp. 5655–5661. IEEE, 2019.

Pronzato, L. and Pázman, A. Design of experiments in nonlinear models. *Lecture notes in statistics*, 212, 2013.

Pukelsheim, F. *Optimal design of experiments*. SIAM, 2006.

Rojas, C. R., Welsh, J. S., Goodwin, G. C., and Feuer, A. Robust optimal experiment design for system identification. *Automatica*, 43(6):993–1008, 2007.

Rojas, C. R., Aguero, J.-C., Welsh, J. S., Goodwin, G. C., and Feuer, A. Robustness in experiment design. *IEEE Transactions on Automatic Control*, 57(4):860–874, 2011.

Sarkar, T. and Rakhlin, A. How fast can linear dynamical systems be learned? *arXiv preprint arXiv:1812.01251*, 2018.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pp. 5610–5618. PMLR, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Simchowitz, M. and Foster, D. J. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.

Simchowitz, M., Jamieson, K., and Recht, B. The simulator: Understanding adaptive sampling in the moderate-confidence regime. *arXiv preprint arXiv:1702.05186*, 2017.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Simchowitz, M., Boczar, R., and Recht, B. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Simchowitz, M., Singh, K., and Hazan, E. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pp. 3320–3436. PMLR, 2020.

Simon, H. A. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica, Journal of the Econometric Society*, pp. 74–81, 1956.

Theil, H. A note on certainty equivalence in dynamic planning. *Econometrica: Journal of the Econometric Society*, pp. 346–349, 1957.

Tsiamis, A. and Pappas, G. J. Finite sample analysis of stochastic system identification. *arXiv preprint arXiv:1903.09122*, 2019.

Tu, S., Boczar, R., Packard, A., and Recht, B. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

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

Wagenmaker, A. and Jamieson, K. Active learning for identification of linear dynamical systems. *arXiv preprint arXiv:2002.00495*, 2020.

Zanette, A., Kochenderfer, M., and Brunskill, E. Almost horizon-free structure-aware best policy identification with a generative model. 2019.

Ziemann, I. and Sandberg, H. On uninformative optimal policies in adaptive lqr with unknown b-matrix. *arXiv preprint arXiv:2011.09288*, 2020.