# Optimal Control with Learning on the Fly: System with Unknown Drift

**Daniel Gurevich** DGUREVICH@PRINCETON.EDU
**Debdipta Goswami** GOSWAMID@PRINCETON.EDU
**Charles L. Fefferman** CF@MATH.PRINCETON.EDU
**Clarence W. Rowley** CWROWLEY@PRINCETON.EDU
*Princeton University, NJ, USA*

**Editors:** R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

## Abstract

This paper derives an optimal control strategy for a simple stochastic dynamical system with constant drift and an additive control input. Motivated by the example of a physical system with an unexpected change in its dynamics, we take the drift parameter to be unknown, so that it must be learned while controlling the system. The state of the system is observed through a linear observation model with Gaussian noise. In contrast to most previous work, which focuses on a controller's asymptotic performance over an infinite time horizon, we minimize a quadratic cost function over a finite time horizon. The performance of our control strategy is quantified by comparing its cost with the cost incurred by an optimal controller that has full knowledge of the parameters. This approach gives rise to several notions of "regret." We derive a set of control strategies that provably minimize the worst-case regret, which arise from Bayesian strategies that assume a specific fixed prior on the drift parameter. This work suggests that examining Bayesian strategies may lead to optimal or near-optimal control strategies for a much larger class of realistic dynamical models with unknown parameters.

**Keywords:** optimal control, regret, competitive ratio, adaptive control, learning

## 1. Introduction

With the proliferation of autonomous controllers in every aspect of human life, including many safety-critical systems, it is of paramount importance to ensure their robustness in the face of unexpected changes in the environment and failures of physical components. Hence, modern controllers must be able to adapt to unexpected and uncertain changes in a system's dynamics. However, such capability is currently limited. For instance, while an expert human pilot could successfully land an aircraft even after damage such as an engine failure, this is an unlikely feat for a present-day autopilot.

When a physical system undergoes a sudden change in its dynamics, forcing an adjustment of some model parameters used by the controller, there is only a limited time and amount of data available to adjust the model. Moreover, in most safety-critical scenarios, control must be applied without interruption, i.e., there is no opportunity for a dedicated learning phase. Recently developed data-driven learning methods for dynamical systems (for some examples, see Brunton et al. (2016), Cubitt et al. (2012), Hills et al. (2015), Schmidt and Lipson (2009)) are either data-intensive or require substantial *a priori* system knowledge. Ahmadi et al. (2017) uses the approach of differential

inclusions to assess the safety of a trajectory of an uncontrolled and unknown system; however, it does not suggest how to identify a control law that will guarantee this safety. A common approach for controlling an unmodeled system is to first learn the parameters in the model, and subsequently design an optimal controller from the learned model. A related approach is to divide the time horizon of the control into epochs, and after each epoch, first refine the model and then update the controller accordingly. This latter approach yields asymptotically optimal results in the limit of large time (Cohen et al., 2019). Other methods such as the *optimism-in-the-face-of-uncertainty* (OFU) paradigm first introduced by Lai and Robbins (1985) use simultaneous estimation of the system parameter and control design Abbasi-Yadkori and Szepesvári (2011) that theoretically achieves a $\tilde{O}(\sqrt{T})$ regret bound in the large time limit. However, none of the aforementioned methods are well-suited to control of a safety-critical system on the most relevant short time scales.

The present paper specifically deals with the problem of finding an optimal control on a finite time horizon with no distinct learning and control phases. In our model, we consider a stochastic dynamical system with an unknown constant drift, an additive control input, and a Wiener process noise. The system state is observed through a linear measurement that is corrupted by an independent Wiener measurement noise. Thus, both the process and observation dynamics obey linear stochastic differential equations (SDEs).

This paper builds on the prior work of Fefferman et al. (2021), which treated a similar problem with multiple special restrictions. In that work, it was assumed that the system state is one-dimensional and can be measured without any sensor noise. The present paper considers several generalizations: the state is multi-dimensional, correlations between different components of the state are taken into account, and measurements are noisy.

The paper is organized as follows. In Section 2 we formulate the Bayesian optimal control problem, in which we have a prior probability distribution for the unknown parameter, as well as the "agnostic" optimal control problem, for which we have no prior belief, and optimality is based on the notion of "regret." In Section 3 we show that Bayesian strategies are candidate optimal strategies for agnostic control. In Section 4 we present agnostic control strategies that minimize the worst-case multiplicative or additive regret. In Section 5 we summarize our findings and discuss future research directions. The proofs of the theorems can be found in the extended version of this manuscript Gurevich et al. (2022).

## 2. Problem Formulation

In the discrete version of our problem, we consider a dynamical system with noisy observations on the time interval $[0, T]$

$$\Delta\mathbf{q} = (\mathbf{a} + \mathbf{u}(t))\Delta t + \Delta\mathbf{W}, \qquad \Delta\mathbf{y} = \mathbf{q}\Delta t + \Delta\mathbf{V}, \tag{1}$$

where $\mathbf{q} \in \mathbb{R}^d$ denotes the position of a particle which is observed through a noisy measurement $\mathbf{y} \in \mathbb{R}^d$. $\Delta\mathbf{W}$ and $\Delta\mathbf{V}$ are two independent normally distributed random variables with zero mean and standard deviation $\Sigma_W\sqrt{\Delta t}$ and $\Sigma_V\sqrt{\Delta t}$ respectively. A simple scaling and rotation allow us to take $\Sigma_W = \mathbf{I}_{d\times d}$, as we will do from now on. We also take the initial conditions $\mathbf{y}(0) = 0$ and $\mathbf{q}(0)$ normally distributed with $\mathbb{E}[\mathbf{q}(0)\mathbf{q}(0)^T] = \Sigma_{\mathbf{q}_0}$. As $\Delta t \to 0$, we obtain the continuous-time version of the dynamics,

$$d\mathbf{q}(t) = (\mathbf{a} + \mathbf{u}(t))dt + d\mathbf{W}(t) \qquad d\mathbf{y}(t) = \mathbf{q}(t)dt + d\mathbf{V}(t), \tag{2}$$

where $\mathbf{W}(t)$ and $\mathbf{V}(t)$ are independent $d$-dimensional scaled Wiener processes with their generalized derivatives $d\mathbf{W}(t)$ and $d\mathbf{V}(t)$ corresponding to white noise. Given some $T_0 \in [0, T)$, we are allowed to observe the system for $t \in [0, T_0]$ without applying any control, after which we control the system on $t \in [T_0, T]$. The main focus of this paper is on the most practically challenging case arising when $T_0 = 0$, i.e., there is no dedicated learning phase and the control must be applied from the beginning. Our objective is to find the optimal control strategy $\mathbf{u}(t)$, depending on the measurements $\{y(\tau) : 0 \leq \tau < t\}$ and taking values in $\mathbb{R}^d$, that minimizes the cost function

$$J(\mathbf{q}, \mathbf{u}; \mathbf{a}) = \mathbb{E}\left[\int_{T_0}^{T} \left(\mathbf{q}^T Q \mathbf{q} + \mathbf{u}^T R \mathbf{u}\right) dt\right]. \tag{3}$$

If the drift parameter $\mathbf{a}$ is known, then the expectation in (3) is well-defined, and the task reduces to a classical optimal control problem. However, if $\mathbf{a}$ is unknown, it is not immediately clear what should be considered an optimal control strategy.

### 2.1. Notions of Optimality

When the parameter $\mathbf{a}$ is unknown, the performance of a control strategy can be quantified in two different ways.

#### 2.1.1. BAYESIAN CONTROL

In the Bayesian version of our problem, a prior belief is assumed on $\mathbf{a}$ — that is, $\mathbf{a}$ is chosen at random according to a probability measure $\mu(\mathbf{a})$. The expected cost associated with $\mathbf{u}$ can then be computed as

$$\mathcal{J}(\mathbf{q}, \mathbf{u}; \mu) = \mathbb{E}_{\mathbf{a}}\left[J(\mathbf{q}, \mathbf{u}; \mathbf{a})\right] = \int_{\mathbb{R}^d} J(\mathbf{q}, \mathbf{u}; \mathbf{a})\, d\mu(\mathbf{a}). \tag{4}$$

In this case, the objective is to find a control strategy $\mathbf{u}(t)$ that minimizes the expected value in (4) given the prior belief $\mu(a)$.

#### 2.1.2. AGNOSTIC CONTROL

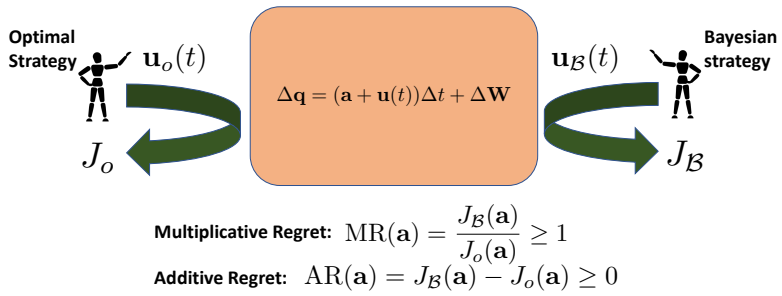



Figure 1: Different notions of "regret".

While Bayesian strategies yield a well-defined cost and performance metric, their inherent dependence on the prior belief makes them unsuitable when no prior information on the parameter is available. We wish to find a strategy that does not depend on any prior belief on $\mathbf{a}$. We call such a strategy an *agnostic strategy*, and we use the notion of *regret* to measure its performance, in particular comparing against the optimal strategy when $\mathbf{a}$ is known. Suppose the agnostic strategy $\mathcal{B}$ yields a cost $J_{\mathcal{B}}(\mathbf{a})$ according to (3), which depends on the value of $\mathbf{a}$. The optimal strategy with full knowledge of $\mathbf{a}$ instead incurs a cost $J_o(\mathbf{a})$. Clearly, $J_o(\mathbf{a}) \leq J_{\mathcal{B}}(\mathbf{a})$ for

all $\mathcal{B}$ and $\mathbf{a}$. The regret achieved by the agnostic strategy $\mathcal{B}$ can be defined in various ways. In all cases, we seek a strategy that minimizes the worst-case regret across all possible values of $\mathbf{a}$. In this paper, we focus on the following definitions of regret:

- *Additive regret*: The additive regret (often called the regret in other literature) is the difference
$$\mathrm{AR}_{\mathcal{B}}(\mathbf{a}) \triangleq J_{\mathcal{B}}(\mathbf{a}) - J_o(\mathbf{a}) \geq 0.$$
A strategy that minimizes the worst-case additive regret $\mathrm{AR}^*_{\mathcal{B}} = \sup_{\mathbf{a}} \mathrm{AR}_{\mathcal{B}}(\mathbf{a})$ is deemed optimal.

- *Multiplicative regret*: The multiplicative regret or *competitive ratio*, as it is more commonly called in the literature, is the ratio
$$\mathrm{MR}_{\mathcal{B}}(\mathbf{a}) \triangleq \frac{J_{\mathcal{B}}(\mathbf{a})}{J_o(\mathbf{a})} \geq 1.$$
Under this definition, the optimal agnostic strategy minimizes the worst-case multiplicative regret $\mathrm{MR}^*_{\mathcal{B}} = \sup_{\mathbf{a}} \mathrm{MR}_{\mathcal{B}}(\mathbf{a})$. Fig. 1 shows a schematic for different notions of regret.

### 2.2. Bayesian Strategies Are Candidate Agnostic Strategies

While a Bayesian strategy can generally be identified by variational methods, a good agnostic strategy cannot readily be constructed: the worst-case regret of a strategy is a complex nonlinear function of both the optimal cost achievable at fixed $\mathbf{a}$ and the strategy's cost, examined over all possible values of $\mathbf{a}$. However, we may search for optimal or near-optimal agnostic strategies which arise as Bayesian strategies from a fixed prior. We are assisted by the following theorem, which provides both lower and upper bounds on the optimal worst-case regret.

**Theorem 1** *Suppose that $\mathcal{A}$ is the Bayesian strategy for the prior $\mu$ and $\mathcal{B}$ is an optimal agnostic strategy in the additive or multiplicative regret setting. Then respectively*

$$\int \mathrm{AR}_{\mathcal{A}}(\mathbf{a})\, d\mu(\mathbf{a}) \leq \mathrm{AR}^*_{\mathcal{B}} \leq \mathrm{AR}^*_{\mathcal{A}} \tag{5}$$

*or*

$$\frac{\int \mathrm{MR}_{\mathcal{A}}(\mathbf{a}) J_o(\mathbf{a})\, d\mu(\mathbf{a})}{\int J_o(\mathbf{a})\, d\mu(\mathbf{a})} \leq \mathrm{MR}^*_{\mathcal{B}} \leq \mathrm{MR}^*_{\mathcal{A}}\,. \tag{6}$$

As a corollary, we obtain a simple criterion for showing that a Bayesian strategy is an optimal agnostic strategy, which was previously used in Fefferman et al. (2021).

**Corollary 2** *Suppose the Bayesian strategy $\mathcal{A}$ is optimal under the prior $\mu$, achieves constant regret $R$ (additive or multiplicative) for all $\mathbf{a}$ in the support of $\mu$, and achieves regret not exceeding $R$ for all $\mathbf{a}$ not in the support of $\mu$. Then $\mathcal{A}$ is an optimal agnostic strategy which achieves a worst-case regret of $R$.*

**Proof** Under the conditions of this corollary, both the lower and upper bounds on the optimal worst-case regret in Theorem 1 are equal to $R$, which concludes the proof. ■

In particular, any Bayesian strategy achieving constant regret for all $\mathbf{a}$ is optimal by the corollary. Thus, the plan of attack in this paper will be as follows. First, we construct the optimal Bayesian strategies for a specific set of priors $\mu$. Then we will compute the regret for those strategies for the values of $\mathbf{a}$ and solve for the specific prior $\mu$ that provides us with an optimal Bayesian strategy with constant regret, which is also an optimal agnostic strategy.

## 3. A Bayesian Strategy for Unknown $\mathbf{a}$

This section demonstrates the optimal strategy for the Bayesian problem discussed in Section 2.1.1 given a fixed prior belief $\mu(\mathbf{a})$. We restrict ourselves to Gaussian prior beliefs only, i.e., $d\mu(\mathbf{a}) = \rho(\mathbf{a})\, d\mathbf{a}$, where $\rho(\mathbf{a}) = \mathcal{N}(\mathbf{a}; 0, \Sigma)$ for some covariance matrix $\Sigma \succcurlyeq 0$. This offers the advantage that the posterior distributions of $\mathbf{a}$ and $\mathbf{q}$ at any time $t$ will remain jointly Gaussian. To approach this problem in a more traditional control-theoretic way, we write the system (1) as follows:

$$d\mathbf{x} = (F\mathbf{x} + B\mathbf{u})\, dt + G\, d\mathbf{W} \qquad d\mathbf{y} = H\mathbf{x}\, dt + d\mathbf{V}, \tag{7}$$

where $\mathbf{x} = \begin{bmatrix}\mathbf{q}\\ \mathbf{a}\end{bmatrix} \in \mathbb{R}^{2d}$, $F = \begin{bmatrix}\mathbf{0}_{d\times d} & \mathbf{I}_{d\times d}\\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d}\end{bmatrix}$, $G = B = \begin{bmatrix}\mathbf{I}_{d\times d}\\ \mathbf{0}_{d\times d}\end{bmatrix}$, and $H = \begin{bmatrix}\mathbf{I}_{d\times d} & \mathbf{0}_{d\times d}\end{bmatrix}$. The cost (4) can equivalently be written as

$$\mathcal{J}(\mathbf{x}, \mathbf{u}) = \mathbb{E}\left[\int_{T_0}^{T} (\mathbf{x}^T \tilde{Q}\mathbf{x} + \mathbf{u}^T R\mathbf{u})\, dt\right], \tag{8}$$

with $\tilde{Q} = \begin{bmatrix}Q & 0_{d\times d}\\ 0_{d\times d} & 0_{d\times d}\end{bmatrix}$. Now the problem of minimizing (8) subject to the dynamics (7) becomes a standard linear-quadratic-Gaussian optimal control problem (see, for instance, Speyer and Chung (2008)). Let $\hat{\mathbf{x}}(t) = \begin{bmatrix}\hat{\mathbf{q}}(t)\\ \hat{\mathbf{a}}(t)\end{bmatrix} = \mathbb{E}\left[\mathbf{x}(t) | \mathbf{x}(0), \mathbf{y}(\tau) : \tau \in [0, t)\right]$ be the Bayesian estimates of $\mathbf{q}(t)$ and $\mathbf{a}$ at time $t$. Then $\hat{\mathbf{q}}(t)$ and $\hat{\mathbf{a}}(t)$ satisfy the continuous-time Kalman filter equations

$$\begin{aligned} d\hat{\mathbf{q}}(t) &= (\hat{\mathbf{a}} + \mathbf{u})\, dt + P_{11}(t)\Sigma_V^{-1}\, (d\mathbf{y} - \hat{\mathbf{q}} dt)\\ d\hat{\mathbf{a}}(t) &= P_{12}^T(t)\Sigma_V^{-1}\, (d\mathbf{y} - \hat{\mathbf{q}}\, dt) \end{aligned} \tag{9}$$

with $P(t) = \begin{bmatrix}P_{11} & P_{12}\\ P_{12}^T & P_{22}\end{bmatrix} = \mathbb{E}[(\mathbf{x}(t) - \hat{\mathbf{x}}(t))(\mathbf{x}(t) - \hat{\mathbf{x}}(t))^T]$ given by the Riccati ODEs

$$\dot{P}(t) = FP + PF^T - PH^T\Sigma_V^{-1}HP + G\Sigma_W G^T, \quad P(0) = \begin{bmatrix}\Sigma_{\mathbf{q}_0} & 0\\ 0 & \Sigma\end{bmatrix}. \tag{10}$$

The optimal cost to go at time $t$ is

$$\begin{aligned} \mathcal{J}(t, \mathbf{q}, \mathbf{y}; \mu) &= \mathbb{E}\left[\int_t^T (\mathbf{x}^T \tilde{Q}\mathbf{x} + \mathbf{u}^T R\mathbf{u}) d\tau\right]\\ &= \mathbb{E}\left[\int_t^T (\hat{\mathbf{q}}^T Q\hat{\mathbf{q}} + \mathbf{u}^T R\mathbf{u}) d\tau\right] + \int_t^T \mathrm{Trace}(P(\tau)\tilde{Q}) d\tau. \end{aligned} \tag{11}$$

The optimal value of $\mathcal{J}$ solves the Hamilton-Jacobi-Bellman equation and takes the form

$$\mathcal{J}(t) = \hat{\mathbf{x}}^T S(t)\hat{\mathbf{x}} + \alpha(t) + \int_t^T \mathrm{Trace}(P(\tau)\tilde{Q}) d\tau. \tag{12}$$

In turn, the optimal control is given by

$$\mathbf{u}^*(t) = R^{-1}B^T S(t)\hat{\mathbf{x}}(t) = -R^{-1}(S_{11}\hat{\mathbf{q}}(t) + S_{12}\hat{\mathbf{a}}(t)). \tag{13}$$

The matrix $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix}$ and scalar-valued function $\alpha$ satisfy the control Riccati equations

$$\begin{aligned} -\dot{S} &= SF + F^T S + \tilde{Q} - SBR^{-1}B^T S, \quad S(T) = 0, \\ -\dot{\alpha} &= \text{Trace}\left(PH^T\Sigma_V^{-1}HPS\right), \quad \alpha(T) = 0. \end{aligned} \tag{14}$$

After some algebraic manipulation, the optimal cost-to-go at time $t$ becomes

$$\begin{aligned} \mathcal{J}(t, \mathbf{q}, \mathbf{y}; \mu) = (\hat{\mathbf{q}}^T S_{11}\hat{\mathbf{q}} + \hat{\mathbf{a}}^T S_{22}\hat{\mathbf{a}} + 2\hat{\mathbf{q}}^T S_{12}\hat{\mathbf{a}}) + \text{Trace}(P(t)S(t)) \\ + \text{Trace}\left[\int_t^T G\Sigma_W G^T S(t) + S(t)BR^{-1}B^T P(t)d\tau\right]. \end{aligned} \tag{15}$$

## 4. Towards Agnostic Control: Performance of the Bayesian Strategy

In Section 2.2, it is established that a Bayesian strategy that has constant regret independent of $\mathbf{a}$ will minimize worst-case regret. To obtain such a strategy, the performance of a general Bayesian strategy needs to be quantified for each fixed true value of the parameter $\mathbf{a}$. The total cost incurred at $\mathbf{a}$ by the Bayesian strategy arising from a prior $\mu$ is

$$\mathcal{J}(\mathbf{q}, \mathbf{a}; \mu) = \mathbb{E}\left[\int_{T_0}^T \left(\mathbf{q}^T Q\mathbf{q} + \mathbf{u}^T R\mathbf{u}\right) dt\right], \tag{16}$$

where $\mathbf{u} = R^{-1}B^t S\hat{\mathbf{x}} = R^{-1}(S_{11}\hat{\mathbf{q}} + S_{12}\hat{\mathbf{a}})$. With this control, the stochastic differential equations describing $\mathbf{q}$, $\hat{\mathbf{q}}$, and $\mathbf{a}$ become

$$\begin{aligned} d\mathbf{q} &= \left[\mathbf{a} - R^{-1}(S_{11}\hat{\mathbf{q}} + S_{12}\hat{\mathbf{a}})\right] dt + d\mathbf{W}, \\ d\hat{\mathbf{q}} &= \left[\hat{\mathbf{a}} - R^{-1}(S_{11}\hat{\mathbf{q}} + S_{12}\hat{\mathbf{a}})\right] dt + P_{11}\Sigma_V^{-1}\left[(\mathbf{q} - \hat{\mathbf{q}})dt + d\mathbf{V}\right], \\ d\hat{\mathbf{a}} &= P_{12}\Sigma_V^{-1}\left[(\mathbf{q} - \hat{\mathbf{q}})dt + d\mathbf{V}\right] \end{aligned} \tag{17}$$

Let $\bar{\mathbf{a}}(t) \triangleq \mathbb{E}[\hat{\mathbf{a}}(t)]$, $\bar{\mathbf{q}}(t) \triangleq \mathbb{E}[\hat{\mathbf{q}}(t)]$, and $\check{\mathbf{q}}(t) \triangleq \mathbb{E}[\mathbf{q}(t)]$ with the associated covariances $\Sigma_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$, $\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}}$, $\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}}$, $\Sigma_{\mathbf{q}\mathbf{q}}$, $\Sigma_{\mathbf{q}\hat{\mathbf{q}}}$, and $\Sigma_{\mathbf{q}\hat{\mathbf{a}}}$. The cost (16) can be written as

$$\begin{aligned} \mathcal{J}(\mathbf{q}, \mathbf{a}; \mu) = \mathbb{E}\left[\int_{T_0}^T \left(\mathbf{q}^T Q\mathbf{q} + \mathbf{u}^T R\mathbf{u}\right) dt\right] = \int_{T_0}^T \left[\check{\mathbf{q}}^T Q\check{\mathbf{q}} + (S_{11}\bar{\mathbf{q}} + S_{12}\bar{\mathbf{a}})^T R^{-1}(S_{11}\bar{\mathbf{q}} + S_{12}\bar{\mathbf{a}})\right] dt \\ + \int_{T_0}^T \text{Trace}\left(Q\Sigma_{\mathbf{q}\hat{\mathbf{q}}} + S_{11}^T R^{-1}S_{11}\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}} + 2S_{11}^T R^{-1}S_{12}\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}} + S_{12}^T R^{-1}S_{12}\Sigma_{\hat{\mathbf{a}}\hat{\mathbf{a}}}\right) dt. \end{aligned} \tag{18}$$

**Theorem 3** *The cost incurred by a specific Bayesian strategy can be expressed as*

$$\mathcal{J}(\mathbf{q}, \mathbf{a}; \mu) = \mathbf{a}^T X(\Sigma, \Sigma_V, T_0, T)\mathbf{a} + Y(\Sigma, \Sigma_V, T_0, T), \text{ where} \tag{19}$$

$$X(\Sigma, \Sigma_V, T_0, T) = \int_{T_0}^{T} \left[ C_1^T Q C_1 + (S_{11}C_2 + S_{12}C_3)^T R^{-1}(S_{11}C_2 + S_{12}C_3) \right] dt \text{ and}$$

$$Y(\Sigma, \Sigma_V, T_0, T) = \int_{T_0}^{T} \text{Trace}\left( Q\Sigma_{\mathbf{qq}} + S_{11}^T R^{-1} S_{11} \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}} + 2S_{11}^T R^{-1} S_{12} \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}} + S_{12}^T R^{-1} S_{12} \Sigma_{\hat{\mathbf{a}}\hat{\mathbf{a}}} \right) dt$$

*for time-varying functions $C_1(t), C_2(t)$, and $C_3(t)$. All of the quantities appearing here can be found by solving the following system of ODEs:*

$$\begin{aligned}
\dot{C}_1 &= I - R^{-1}(S_{11}C_2 - S_{12}C_3) \\
\dot{C}_2 &= C_3 - R^{-1}(S_{11}C_2 - S_{12}C_3) + P_{11}\Sigma_V^{-1}(C_1 - C_2) \\
\dot{C}_3 &= P_{12}\Sigma_V^{-1}(C_1 - C_2) \\
\dot{\Sigma}_{\hat{\mathbf{a}}\hat{\mathbf{a}}} &= (\Sigma_{\mathbf{q}\hat{\mathbf{a}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}})^T \Sigma_V^{-1} P_{12} + P_{12}\Sigma_V^{-1}(\Sigma_{\mathbf{q}\hat{\mathbf{a}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}}) + P_{12}\Sigma_V^{-1}P_{12} \\
\dot{\Sigma}_{\mathbf{qq}} &= \Sigma_W - (\Sigma_{\mathbf{q}\hat{\mathbf{q}}}S_{11} + \Sigma_{\mathbf{q}\hat{\mathbf{a}}}S_{12}^T)R^{-1} - R^{-1}(\Sigma_{\mathbf{q}\hat{\mathbf{q}}}S_{11} + \Sigma_{\mathbf{q}\hat{\mathbf{a}}}S_{12}^T)^T \\
\dot{\Sigma}_{\mathbf{q}\hat{\mathbf{q}}} &= (\Sigma_{\mathbf{qq}} - \Sigma_{\mathbf{q}\hat{\mathbf{q}}})\Sigma_V^{-1}P_{11} - \Sigma_{\mathbf{q}\hat{\mathbf{q}}}S_{11}^T R^{-1} + \Sigma_{\mathbf{q}\hat{\mathbf{a}}}(I - R^{-1}S_{12})^T - R^{-1}(S_{11}\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}} + S_{12}\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}}^T) \\
\dot{\Sigma}_{\mathbf{q}\hat{\mathbf{a}}} &= (\Sigma_{\mathbf{qq}} - \Sigma_{\mathbf{q}\hat{\mathbf{q}}})\Sigma_V^{-1}P_{12} - R^{-1}(S_{11}\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}} + S_{12}\Sigma_{\hat{\mathbf{a}}\hat{\mathbf{a}}}^T) \\
\dot{\Sigma}_{\hat{\mathbf{q}}\hat{\mathbf{q}}} &= P_{11}^T\Sigma_V^{-1}P_{11} + \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}}(I - R^{-1}S_{12})^T + (I - R^{-1}S_{12})\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}}^T + (\Sigma_{\mathbf{q}\hat{\mathbf{q}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}})^T\Sigma_V^{-1}P_{11} \\
&\quad + P_{11}\Sigma_V^{-1}(\Sigma_{\mathbf{q}\hat{\mathbf{q}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}}) - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}}S_{11}R^{-1} - R^{-1}S_{11}\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}} \\
\dot{\Sigma}_{\hat{\mathbf{q}}\hat{\mathbf{a}}} &= (\Sigma_{\mathbf{q}\hat{\mathbf{q}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{q}}})^T\Sigma_V^{-1}P_{12} + (I - R^{-1}S_{12})\Sigma_{\hat{\mathbf{a}}\hat{\mathbf{a}}} - R^{-1}S_{12}\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}} + P_{11}\Sigma_V^{-1}(\Sigma_{\mathbf{q}\hat{\mathbf{a}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}} + P_{12}) \\
\dot{\Sigma}_{\hat{\mathbf{a}}\hat{\mathbf{a}}} &= (\Sigma_{\mathbf{q}\hat{\mathbf{a}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}})^T\Sigma_V^{-1}P_{12} + P_{12}\Sigma_V^{-1}(\Sigma_{\mathbf{q}\hat{\mathbf{a}}} - \Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}}) + P_{12}\Sigma_V^{-1}P_{12}.
\end{aligned} \tag{20}$$

### 4.1. Optimal Cost for a Bayesian Strategy with Known $\mathbf{a}$

If $\mathbf{a}$ is known, then $\hat{\mathbf{a}}(t) \equiv \bar{\mathbf{a}}(t) \equiv \mathbf{a}$, $\Sigma_{\hat{\mathbf{a}}\hat{\mathbf{a}}}(t) \equiv 0$, $\Sigma_{\hat{\mathbf{q}}\hat{\mathbf{a}}}(t) \equiv 0$, and $\Sigma_{\mathbf{q}\hat{\mathbf{a}}}(t) \equiv 0$. We use similar notation as above, denoting all of the analogous quantities for this case with a superscript $*$. Hence the Riccati equation (10) becomes

$$\dot{P}_{11}^* = -P_{11}^{*T}\Sigma_V^{-1}P_{11}^* + \Sigma_W, \; P_{11}^*(0) = \Sigma_{\mathbf{q}_0}, \tag{21}$$

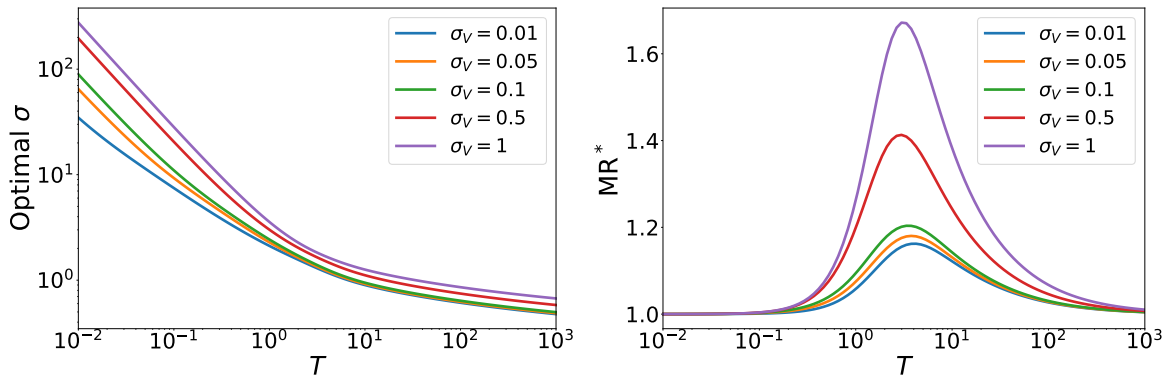
and $P_{12}^*(t) = P_{22}^*(t) \equiv 0$. The estimation equation (9) becomes

$$d\hat{\mathbf{q}}^*(t) = (\mathbf{a} + \mathbf{u})\, dt + P_{11}^*(t)\Sigma_V^{-1}\left(d\mathbf{y} - \hat{\mathbf{q}}^* dt\right). \tag{22}$$

This, in turn, yields the optimal control $\mathbf{u}^*(t) = -R^{-1}(S_{11}\hat{\mathbf{q}}^*(t) + S_{12}\mathbf{a}(t))$.

**Theorem 4** *The optimal cost for the Bayesian strategy with known* $\mathbf{a}$ *can be expressed as*

$$J(\mathbf{q}, \mathbf{a}) = \mathbf{a}^T X^*(\Sigma_V, T_0, T)\mathbf{a} + Y^*(\Sigma_V, T_0, T), \tag{23}$$




(*a*) (*b*)

Figure 2: Bayesian strategies minimizing the worst-case multiplicative regret with varying standard deviation $\sigma_V$ of the sensor noise. (a) Optimal standard deviation $\sigma$ of the prior; (b) worst-case multiplicative regret $\mathrm{MR}^*$.

$$\textit{where} \quad X^*(\Sigma_V, T_0, T) = \int_{T_0}^{T} \left[ {C_1^*}^T Q C_1^* + (S_{11}C_2^* + S_{12})^T R^{-1}(S_{11}C_2^* + S_{12}) \right] dt$$

$$\textit{and} \quad Y^*(\Sigma_V, T_0, T) = \int_{T_0}^{T} \mathrm{Trace}\left( Q\Sigma^*_{\mathbf{qq}} + S_{11}^T R^{-1} S_{11} \Sigma^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}} \right) dt,$$

*with the time-varying functions* $C_1^*(t), C_2^*(t), \Sigma^*_{\mathbf{qq}}$, *and* $\Sigma^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}}$ *described by the ODEs*

$$\begin{aligned}
\dot{C_1}^* &= I - R^{-1}(S_{11}C_2{}^* - S_{12}) \\
\dot{C_2}^* &= I - R^{-1}(S_{11}C_2{}^* - S_{12}) + P_{11}\Sigma_V^{-1}(C_1^* - C_2^*) \\
\dot{\Sigma}^*_{\mathbf{qq}} &= \Sigma_W - (\Sigma^*_{\mathbf{q}\hat{\mathbf{q}}} S_{11} R^{-1} + R^{-1} S_{11} {\Sigma^*_{\mathbf{q}\hat{\mathbf{q}}}}^T) \\
\dot{\Sigma}^*_{\mathbf{q}\hat{\mathbf{q}}} &= (\Sigma^*_{\mathbf{qq}} - \Sigma^*_{\mathbf{q}\hat{\mathbf{q}}})\Sigma_V^{-1} P_{11} - \Sigma^*_{\mathbf{q}\hat{\mathbf{q}}} S_{11}^T R^{-1} - R^{-1} S_{11} \Sigma^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}} \\
\dot{\Sigma}^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}} &= {P_{11}^*}^T \Sigma_V^{-1} P_{11} + (\Sigma^*_{\mathbf{q}\hat{\mathbf{q}}} - \Sigma^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}})^T \Sigma_V^{-1} P_{11}^* + P_{11}^* \Sigma_V^{-1} (\Sigma^*_{\mathbf{q}\hat{\mathbf{q}}} - \Sigma^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}}) \\
&\quad - \Sigma^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}} S_{11} R^{-1} - R^{-1} S_{11} \Sigma^*_{\hat{\mathbf{q}}\hat{\mathbf{q}}}.
\end{aligned} \tag{24}$$

### 4.2. Minimizing Multiplicative Regret

The optimal cost (23) with known $\mathbf{a}$ is $J(\mathbf{q}, \mathbf{a}) = \mathbf{a}^T X^*(\Sigma_V, T_0, T)\mathbf{a} + Y^*(\Sigma_V, T_0, T)$. Hence, the multiplicative regret for a Bayesian strategy with prior covariance $\Sigma$ is

$$\mathrm{MR}(\mathbf{a}) = \frac{\mathbf{a}^T X(\Sigma, \Sigma_V, T_0, T)\mathbf{a} + Y(\Sigma, \Sigma_V, T_0, T)}{\mathbf{a}^T X^*(\Sigma_V, T_0, T)\mathbf{a} + Y^*(T, T_0, \Sigma_V)}. \tag{25}$$

We want to find $\Sigma$ such that (25) becomes independent of $\mathbf{a}$. This occurs precisely when

$$X(\Sigma, \Sigma_V, T_0, T) = \frac{Y(\Sigma, \Sigma_V, T_0, T)}{Y^*(\Sigma_V, T_0, T)} X^*(\Sigma_V, T_0, T). \tag{26}$$

Eq. (26) is a matrix integral equation that can be solved for each $T$ numerically using a root-finding method. In fact, this equation has an intuitive interpretation. At optimality, the worst-case

multiplicative regret is equal to both the multiplicative regret at $\mathbf{a} = 0$, given by $\frac{Y}{Y^*}$, and the multiplicative regret in the limit as $||\mathbf{a}|| \to \infty$, which is given by the constant ratio of the $X$ and $X^*$ matrices. That is, the optimal agnostic strategy perfectly balances the regret for no drift and very large drift. Moreover, the equality of these multiplicative regrets is a sufficient condition for optimality.

For illustration, we identified the multiplicative regret-minimizing Bayesian strategy for a scalar system with dynamics (2), $\Sigma_W = 1$, $\Sigma_V = \sigma_V^2$, $Q = R = 1$, and $T_0 = 0$. The optimal prior covariance $\Sigma = \sigma^2$ was computed by solving (26) numerically using a Newton solver. The optimal prior standard deviation $\sigma$ and the worst-case multiplicative regret $\mathrm{MR}^*$ for different values of sensor noise standard deviation $\sigma_V$ are shown in Fig. 2. For small time horizons, the optimal value of $\sigma$ is quite large, reflecting the fact that there is very little time to learn the dynamics of the system and thus the prior needs to be flexible in order for the controller to adapt quickly. This effect is particularly pronounced when the sensor noise $\Sigma_V$ is large, which would otherwise cause the controller to learn the dynamics too slowly. With a large time horizon, the controller can act more conservatively as there is abundant time to stabilize the system and poor performance at small times is only weakly penalized. It can be shown that the optimal prior standard deviation $\sigma$ tends to zero as $T \to \infty$. As expected, the worst-case regret increases with the amount of sensor noise. It peaks at $T \approx 3.5$ and tends to unity as $T$ tends to 0 or $\infty$.

### 4.3. Minimizing Additive Regret

When minimizing the additive regret, we must start at a nonzero time $T_0$ for reasons to be explained shortly. The additive regret is given by

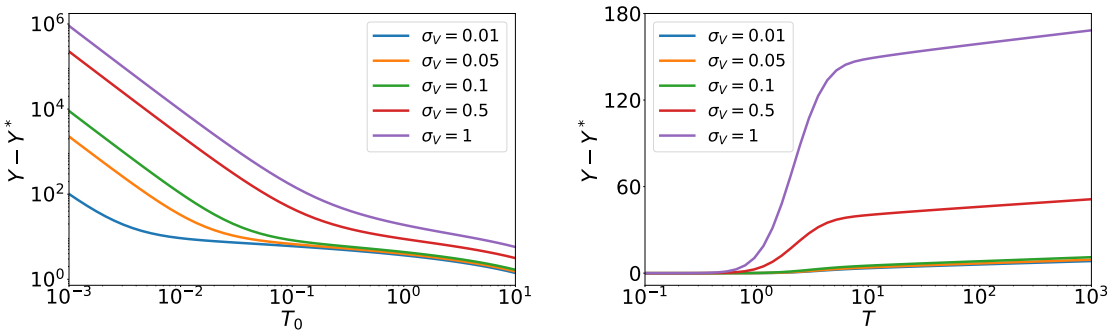
$$\mathrm{AR}(\mathbf{a}) = \mathbf{a}^T \Big[ X(\Sigma, \Sigma_V, T, T_0) - X^*(\Sigma_V, T, T_0) \Big] \mathbf{a} + \Big[ Y(\Sigma, \Sigma_V, T, T_0) - Y^*(\Sigma_V, T, T_0) \Big]. \quad (27)$$

We want to find a prior covariance $\Sigma$ such that $\mathrm{AR}(\mathbf{a})$ is independent of $\mathbf{a}$. It turns out that this occurs for $\Sigma = rI$ in the limit $r \to +\infty$.

**Theorem 5** *For* $0 \leq T_0 < T$, $\lim\limits_{\Sigma \to \infty} \mathbf{a}^T \Big[ X(\Sigma, \Sigma_V, T, T_0) - X^*(\Sigma_V, T, T_0) \Big] \mathbf{a} = 0$. *Moreover,* $\lim\limits_{\Sigma \to \infty} Y(\Sigma, \Sigma_V, T, T_0) - Y^*(\Sigma_V, T, T_0) \to \infty$ *no slower than* $O(\log \frac{1}{T_0})$ *as* $T_0 \to 0$ *for fixed* $T$.

Theorem 5 shows that the limit of Bayesian strategies with the prior covariance $\Sigma$ diverging to infinity optimizes the worst-case additive regret. Moreover, the additive regret in this limit is simply given by $\lim_{\Sigma \to \infty} Y(\Sigma, \Sigma_V, T, T_0) - Y^*(\Sigma_V, T, T_0)$. However, this quantity diverges as $T_0 \to 0$, so we must set $T_0 > 0$ to obtain a meaningful result. (Recall that we require that the control $\mathbf{u}(t)$ be set to zero for $t \in [0, T_0]$.) It remains to identify the specific strategy that arises as the limit of the Bayesian strategies. This requires knowledge of the optimal estimates $\hat{\mathbf{a}}$ and $\hat{\mathbf{q}}$ that appear in (13). The challenge is that we cannot obtain these estimates directly from the ODEs (9) as we did for the Bayesian strategies: these ODEs are singular in the limit as $\Sigma \to \infty$ and $t \to 0$. Instead, we can compute the optimal estimates from first principles. The result is

$$\hat{\mathbf{a}}(t) = 2t \left( \int_0^t s^2 \omega_0(s)\, ds + t^2 \kappa \right)^{-1} \left( \int_0^t \omega_0(s) \mathbf{y}(s)\, ds + \kappa \mathbf{y}(t) \right), \quad \hat{\mathbf{q}}(t) = t \hat{\mathbf{a}}(t) \quad (28)$$




(*a*) (*b*)

Figure 3: The optimal worst-case additive regret $Y - Y^*$ with varying standard deviation $\sigma_V$ of the sensor noise for (a) fixed $T = 100$ and varying $T_0$ and (b) fixed $T_0 = 0.1$ and varying $T$.

where the matrices $\omega_0(s)$, $\kappa$ solve the system of integral equations

$$\Sigma_V \omega_0(s) - \frac{\Sigma_W}{2T} \int_0^t \tau^2 \omega_0(\tau)\, d\tau + \Sigma_W \int_s^t (\tau - t)\omega_0(\tau)\, d\tau = I + \frac{(2s - t)}{2} \Sigma_W \kappa, \tag{29}$$

$$\forall s \in [0, t) \text{ and } \int_0^t \omega_0(\tau)\, d\tau = -\kappa.$$

We can approximate the solution of the integral equations to arbitrary precision by discretizing them with respect to time and then solving the resulting matrix equation. Fig. 3 shows plots of the additive regret for the regret-minimizing agnostic strategies corresponding to the scalar systems with parameters listed in Section 4.2. Panel (a) shows that the additive regret indeed diverges as $T_0 \to 0$ for fixed $T$, following a power law. Moreover, panel (b) indicates that the growth rate of the regret as $T \to \infty$ is logarithmic for fixed $T_0$.

## 5. Conclusions

We have identified optimal strategies for several variations of a control problem based on a stochastic dynamical system with both process and sensor noise: the classical problem where the drift $\mathbf{a}$ is known, a Bayesian problem where the prior distribution of $\mathbf{a}$ is normal, and the agnostic problem where $\mathbf{a}$ is entirely unknown and we wish to minimize the worst-case multiplicative or additive regret. This last problem is of particular practical interest, as it provides valuable intuition about how a controller should simultaneously learn and control a system without prior knowledge of the parameters. In this case, we were able to identify the optimal agnostic strategy as a Bayesian strategy or limit of Bayesian strategies arising from a Gaussian prior. Qualitatively, our results show that a wide prior is necessary to minimize the worst-case regret on a short time horizon. In fact, the worst-case additive regret is minimized as the width of the prior diverges to infinity.

As a future research direction, it is worth exploring whether it is always possible to find a Bayesian strategy which is also an optimal or near-optimal agnostic strategy. For instance, Carruth et al. (2021) treats the scalar system with the state dynamics $\Delta q = (aq + u)\Delta t + \Delta W$,. The aforementioned work provides further theoretical results supporting the hypothesis that optimal agnostic strategies for a wide variety of control problems naturally arise from Bayesian strategies.

## Acknowledgments

This work was supported by the Air Force Office of Scientific Research, award FA9550-19-1-0005.

## References

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. ISSN 0027-8424.

Toby S. Cubitt, Jens Eisert, and Michael M. Wolf. Extracting dynamical equations from experimental data is NP hard. *Physical Reviw Letters*, 108:120503, Mar 2012.

Daniel J.A. Hills, Adrian M. Grütter, and Jonathan J. Hudson. An algorithm for discovering Lagrangians automatically from data. *PeerJ Computer Science*, 1:e31, November 2015. ISSN 2376-5992.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. ISSN 0036-8075.

Mohamadreza Ahmadi, Arie Israel, and Ufuk Topcu. Safety assessemt based on physically-viable data-driven models. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 6409–6414, 2017.

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

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

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.

Charles L. Fefferman, Bernat Guillen Pegueroles, Clarence W. Rowley, and Melanie Weber. Optimal control with learning on the fly: a toy problem. *Revista Matemática Iberoamericana*, 37(1), 2021.

Daniel Gurevich, Debdipta Goswami, Charles L. Fefferman, and Clarence W. Rowley. Optimal control with learning on the fly: System with unknown drift. *arXiv preprint arXiv:2202.03620*, 2022.

Jason L Speyer and Walter H Chung. *Stochastic processes, estimation, and control*, pages 289–334. SIAM, 2008.

Jacob Carruth, Maximilian F. Eggl, Charles L. Fefferman, Clarence W. Rowley, and Melanie Weber. Controlling unknown linear dynamics with bounded multiplicative regret. *arXiv preprint arXiv:2109.06350*, 2021.