Rev. Mat. Iberoam. 38 (2022), no. 1, 175–187
DOI 10.4171/RMI/1275

# Optimal control with learning on the fly: a toy problem

Charles L. Fefferman, Bernat Guillén Pegueroles, Clarence W. Rowley and Melanie Weber

**Abstract.** We exhibit optimal control strategies for a simple toy problem in which the underlying dynamics depend on a parameter that is initially unknown and must be learned. We consider a cost function posed over a finite time interval, in contrast to much previous work that considers asymptotics as the time horizon tends to infinity. We study several different versions of the problem, including Bayesian control, in which we assume a prior distribution on the unknown parameter; and "agnostic" control, in which we assume nothing about the unknown parameter. For the agnostic problems, we compare our performance with that of an opponent who knows the value of the parameter. This comparison gives rise to several notions of "regret", and we obtain strategies that minimize the "worst-case regret" arising from the most unfavorable choice of the unknown parameter. In every case, the optimal strategy turns out to be a Bayesian strategy or a limit of Bayesian strategies.

## 1. Motivation and introduction

We investigate control problems in which we must make decisions with little time and little data available. Our motivating example is the success of pilots learning in real time to fly and safely land an airplane after it has been severely damaged, for instance as documented in [5].

Control with learning has been considered in many different application areas; see, for example, books such as [4,7,9,11], as well as recent papers such as [1,2]. Much is known also about the closely related "multi-armed bandit" problem; see, for instance, the classic papers [3,10] and the more recent survey [6].

A standard approach to control with learning is to start by estimating the parameters in the model, and then to design a controller that is optimal for that model. A related approach is to divide the available time into epochs, and within each epoch, first refine the model, and then update the controller accordingly. In [8], it is shown that this latter approach gives results that are optimal (in some sense) in the asymptotic limit of large time. For the problem we consider, such an approach cannot lead to an optimal control strategy. We are interested in results that are optimal for a fixed time interval, rather than

*2020 Mathematics Subject Classification*: 93E20, 93E35, 49J15.
*Keywords*: Regret, competitive ratio, agnostic control, adaptive control, fuel tax regret.

for large time. Our optimal strategies do not divide into distinct phases of exploration and exploitation, but instead take past history into account at every moment.

This article analyzes a toy problem in which we apply a time-dependent control to keep the position of a moving particle close to zero. The position $q(t) \in \mathbb{R}$ is governed by Brownian motion with a drift. The drift rate is given by $a + u(t)$, where $a$ is an unknown constant, and $u(t)$ is our control at time $t$. For this simple one-dimensional toy model, we consider several different notions of optimality, and we exhibit control policies that are optimal with respect to each of these. At the end of the paper, we mention two additional toy problems to which we hope to return in later work.

## 2. The toy problem

We begin with a time interval $[0, T]$ subdivided into discrete timesteps of size $\Delta t$. From $t$ to $t + \Delta t$, the change in the position $q$ is given by

$$\Delta q = (a + u(t))\Delta t + \Delta W, \tag{2.1}$$

where $a$ is an unknown constant, $u(t)$ is our control, and $\Delta W$ is a normally distributed random variable with zero mean and standard deviation $\sigma_0(\Delta t)^{1/2}$ for a known coefficient $\sigma_0$. A simple scaling allows us to take $\sigma_0 = 1$, which we do from now on. In the limit as $\Delta t$ tends to zero, we obtain a control system in continuous time, for which $W(t)$ is Brownian motion and $dW(t)$ is white noise. Our goal is to find optimal control strategies for this continuous-time system.

From time 0 to some given time $T_0$, we are allowed only to observe the particle: i.e., we must take $u = 0$. From time $T_0$ to time $T$, we may apply any control strategy we please, provided that $u(t)$ is determined by the history $q(\tau)$ for $\tau \leq t$. We want to pick a control strategy to minimize the expected value of a cost function

$$J = \int_{T_0}^{T} (q^2 + \lambda u^2)\, dt, \tag{2.2}$$

where $\lambda > 0$ is a known coefficient. If the parameter $a$ is known, then the only randomness in the system arises from the Brownian motion, so the notion of expected value is well defined. If $a$ is unknown, the meaning of the expected value is not immediately clear. We will discuss it carefully in Sections 3.2 and 3.3.

The most interesting case arises when $T_0 = 0$. Also, a simple scaling allows us to take $\lambda = 1$ without disturbing the normalization $\sigma_0 = 1$. Unless we say otherwise, we will assume that $\lambda = 1$ and $T_0 = 0$. Furthermore, we suppose that our particle starts at position $q(0) = 0$.

## 3. Notions of optimality

In this section, we provide careful definitions of optimal strategies, first assuming that the parameter $a$ is known (classical control), then assuming a prior belief regarding $a$ (Bayesian control), and finally assuming no prior knowledge of $a$ (“agnostic” control).

### 3.1. Known parameter value *a*

If the value of $a$ is known, then we simply ask for a control strategy that minimizes the expected value of (2.2). As observed above, the expected value is well defined because $a$ is known. To calculate such an optimal control strategy is a classical problem of control theory, and we review its solution in Section 4 below.

### 3.2. Bayesian control

Next, suppose that we are given a probability measure $\mu(a)$ reflecting our prior belief about the unknown $a$. That is, the probability that $a$ lies between $\alpha$ and $\beta$ is given by

$$\int_{\alpha}^{\beta} d\mu(a).$$

Then as in the classical case, we can make sense of the expected value of $J$. We compute the expected value of $J$ assuming a given value of $a$; call that quantity $J(a)$. We then average over all $a$ according to the probability measure $\mu$:

$$E(J) = \int J(a)\, d\mu(a).$$

The goal of Bayesian control is to find a control strategy $u$ that minimizes $E(J)$. This strategy will of course depend on our prior belief $\mu$, and the optimal strategy is discussed in Section 5.

### 3.3. Agnostic control

Finally, suppose we know nothing about the parameter $a$. We hope to pick our strategy to minimize one of several notions of *regret*, which we spell out below. In all variants, we play against an opponent who knows the value of $a$ and plays perfectly, while we know nothing about $a$. Suppose we pick a control strategy $\mathcal{Q}$ without knowing $a$. Given the true value of $a$, we can compute the expected cost $J_{\mathcal{Q}}(a)$ of our strategy as in the classical case. We want to compare $J_{\mathcal{Q}}(a)$ with the expected cost $J_{\text{opponent}}(a)$ for our opponent.

We can now give precise descriptions of three problems of agnostic control:

**Additive regret.** The *additive regret*, often called simply *regret*, is the difference

$$\mathrm{AR}_{\mathcal{Q}}(a) = J_{\mathcal{Q}}(a) - J_{\text{opponent}}(a) \geq 0.$$

We can look for a control strategy $\mathcal{Q}$ that minimizes the worst-case additive regret

$$\mathrm{AR}^*_{\mathcal{Q}} = \sup_a \mathrm{AR}_{\mathcal{Q}}(a).$$

**Multiplicative regret.** The *multiplicative regret*, often called the *competitive ratio*, is the ratio

$$\mathrm{MR}_{\mathcal{Q}}(a) = \frac{J_{\mathcal{Q}}(a)}{J_{\text{opponent}}(a)} \geq 1.$$

We can look for a control strategy $\mathcal{Q}$ that minimizes the worst-case multiplicative regret

$$\mathrm{MR}^*_{\mathcal{Q}} = \sup_a \mathrm{MR}_{\mathcal{Q}}(a).$$

**Fuel tax regret.** As before, we compute our expected cost using formula (2.2) (which depends on our strategy $\mathcal{Q}$), with $\lambda = 1$. However, we now assume that our opponent incurs a cost

$$J = \int_{T_0}^{T} (q^2 + \lambda u^2)\, dt,$$

where $\lambda > 1$. Thus, our opponent pays a *fuel tax*. We can look for a control strategy $\mathcal{Q}$ such that, for any assumed value of $a$, our expected cost is at most that of our opponent. We want to find such a strategy with $\lambda$ as small as possible. We define the *fuel tax regret* to be the above minimal $\lambda$.

Solutions of the above agnostic control problems are shown in Section 7.

### 3.4. A constant-regret Bayesian strategy is optimal

Our solutions to the agnostic control problems will be Bayesian for a particular choice of prior belief $\mu$. We explain the idea for multiplicative regret; analogous ideas apply to the other agnostic control problems mentioned above.

Suppose that a particular prior belief $\mu$ gives rise to an optimal Bayesian strategy $\mathcal{B}$ whose multiplicative regret $\mathrm{MR}_{\mathcal{B}}(a)$ is *constant* (independent of $a$). Then the strategy $\mathcal{B}$ minimizes worst-case regret $\mathrm{MR}^*$. To see this, we argue as follows.

Suppose that instead of $\mathcal{B}$, we use another strategy $\mathcal{C}$. The strategy $\mathcal{C}$ cannot perform better than $\mathcal{B}$ for all values of $a$; otherwise $\mathcal{B}$ would not be optimal for the prior belief $\mu$. Thus, there is a value of $a$ for which $\mathrm{MR}_{\mathcal{C}}(a) \geq \mathrm{MR}_{\mathcal{B}}(a)$. But since the strategy $\mathcal{B}$ has constant regret, the right-hand side is independent of $a$. In other words, for some $a$, we have

$$\mathrm{MR}_{\mathcal{C}}(a) \geq \mathrm{MR}^*_{\mathcal{B}}.$$

Consequently, the worst-case regret of $\mathcal{C}$ is at least that of $\mathcal{B}$.

These ideas are clearly more general than the particular problems studied in this paper. However, note that we do not assert, for more general problems, that an optimal strategy necessarily has constant regret.

We have been told that the optimality of constant-regret strategies may be known in the context of bandit problems, although we have been unable to find a reference for this.

## 4. Optimal control for a known parameter

In this section, we review the classical control problem of minimizing the expected cost (2.2) given known $a$, via the Hamilton–Jacobi–Bellman equation [4].

Suppose we find ourselves at position $q$ at time $t$. Let $J(q, t; a)$ be the expected "cost to go":

$$J(q, t; a) = E\Big[ \int_{t}^{T} \big(q(\tau)^2 + u(\tau)^2\big)\, d\tau \Big],$$

assuming an optimal $u$. Then, considering a small time step $\Delta t$ and neglecting errors small compared with $\Delta t$, we have

$$J(q, t; a) = \min_{u} \big[ (q^2 + u^2)\Delta t + E\big(J(q + \Delta q, t + \Delta t; a)\big) \big].$$

Recall that $\Delta q = (a+u)\Delta t + \Delta W$, so (again neglecting errors $o(\Delta t)$)

$$E(\Delta q) = (a+u)\,\Delta t, \qquad E\big((\Delta q)^2\big) = \Delta t.$$

In particular, $\Delta q$ has the order of magnitude $(\Delta t)^{1/2}$. Consequently, Taylor expanding $J$ to first order in $t$ and to second order in $q$, we find that

$$0 = \partial_t J + (q^2+u^2) + (a+u)\partial_q J + \tfrac{1}{2}\partial_q^2 J. \tag{4.1a}$$

The optimal control $u$ minimizes the right-hand side of (4.1a), so

$$u = -\frac{1}{2}\partial_q J, \tag{4.1b}$$

and by definition,

$$J(q,T;a) = 0. \tag{4.1c}$$

We can guess a solution to (4.1), of the form

$$J(q,t;a) = E_2(t)q^2 + E_1(t)qa + E_0(t)a^2 + E_\sharp(t), \tag{4.2}$$

and obtain the following ordinary differential equations:

$$-\dot{E}_2 = 1 - E_2^2, \qquad E_2(T) = 0, \tag{4.3a}$$

$$-\dot{E}_1 = 2E_2 - E_1E_2, \qquad E_1(T) = 0, \tag{4.3b}$$

$$-\dot{E}_0 = E_1 - E_1^2/4, \qquad E_0(T) = 0, \tag{4.3c}$$

$$-\dot{E}_\sharp = E_2, \qquad E_\sharp(T) = 0. \tag{4.3d}$$

These may be solved exactly:

$$E_2 = \tanh(T-t), \tag{4.4a}$$

$$E_1 = 2[1-\operatorname{sech}(T-t)], \tag{4.4b}$$

$$E_0 = (T-t) - \tanh(T-t), \tag{4.4c}$$

$$E_\sharp = \log\cosh(T-t). \tag{4.4d}$$

From (4.1b) and (4.2), the optimal control is then

$$u = -E_2(t)q - \frac{E_1(t)}{2}\,a. \tag{4.5}$$

For future reference, we write down the formulas analogous to (4.2) and (4.4) without the assumption that $\lambda = 1$ in equation (2.2). In place of formula (4.2), we obtain

$$J^\lambda(q,t;a) = E_2^\lambda(t)q^2 + E_1^\lambda(t)qa + E_0^\lambda(t)a^2 + E_\sharp^\lambda(t). \tag{4.6}$$

If we define $s = (T-t)\lambda^{-1/2}$, then in place of (4.4), we obtain

$$E_2^\lambda = \lambda^{1/2}\tanh s, \tag{4.7a}$$

$$E_1^\lambda = 2\lambda(1-\operatorname{sech} s), \tag{4.7b}$$

$$E_0^\lambda = \lambda^{3/2}(s - \tanh s), \tag{4.7c}$$

$$E_\sharp^\lambda = \lambda\log\cosh s. \tag{4.7d}$$

This completes our review of the classical case, for known $a$.

## 5. A Bayesian strategy for an unknown parameter

In this section, we solve the Bayesian control problem discussed in Section 3.2. Now, the parameter $a$ is unknown, but we have a prior belief given by the probability measure $\mu$. Our problem exhibits an interesting feature not seen in general Bayesian control problems. In particular, at each time $t$, a single real number $\xi(t)$ captures all the relevant information from past history up to time $t$. To see this we argue as follows.

For ease of notation, we assume for the moment that $\mu$ is given by a probability density

$$d\mu(a) = \rho(a)\,da.$$

We first compute the posterior probability distribution for $a$, given history up to time $t$, and then use that information to find the optimal control. To this end, we compute the joint probability for the unknown $a$ and history up to time $t$, by dividing the time interval from 0 to $t$ into small steps of duration $\Delta t$. Thus we consider discrete times $\tau = 0, \Delta t, 2\Delta t, \dots, t$. The joint probability density of obtaining a particular $a$ and observing the history $q(0), q(\Delta t), \dots, q(t)$ up to time $t$ is given by

$$\rho(a) \cdot \prod_{\tau=0}^{t-\Delta t} \frac{1}{\sqrt{2\pi\Delta t}} \exp\Big(\frac{-[\Delta q(\tau) - (a + u(\tau))\Delta t]^2}{2\Delta t}\Big),$$

because $\Delta W = \Delta q - (a + u)\Delta t$ is a normal random variable (see equation (2.1)). This joint probability density has the form

$$\rho(a) \cdot \exp\Big(\Big[\sum_{\tau=0}^{t-\Delta t} (\Delta q(\tau) - u(\tau)\Delta t)\Big] a - \tfrac{1}{2} a^2 \sum_{\tau=0}^{t-\Delta t} \Delta t\Big) \cdot (\text{factor independent of } a).$$

Taking the limit as $\Delta t$ tends to zero, we obtain the joint probability density

$$\rho(a) \cdot \exp\Big(\xi a - \frac{t}{2} a^2\Big) \cdot (\text{factor independent of } a), \tag{5.1}$$

where

$$\xi(t) = q(t) - q(0) - \int_0^t u(\tau)\,d\tau. \tag{5.2}$$

Consequently, the posterior probability density for $a$, given history through time $t$, also has the form (5.1). Passing from probability densities $\rho$ to general probabilities $\mu$, we see easily that the posterior probability measure for the unknown parameter $a$, given history up to time $t$, has the form

$$d\mu_{\text{posterior}}(a) = d\mu(a) \cdot \exp\Big(\xi a - \frac{t}{2} a^2\Big) \cdot Z(\xi, t), \tag{5.3}$$

where $Z(\xi, t)$ may be computed by noting that probability measures integrate to 1.

Thus, as claimed at the beginning of this section, the posterior probability measure depends on the history $q(\tau)$ only through the single number $\xi(t)$. Furthermore, at a given time $t$, $\xi(t)$ can be computed from $q(t)$ and the history of the control $u$ up to time $t$. (Recall that we take $q(0) = 0$.) In particular, we do not need to remember the whole history of $q(\tau)$.

We can now proceed as in Section 4. Suppose we find ourselves at given values of $q$ and $\xi$ at time $t$. Let $J(q, \xi, t; \mu)$ be the expected cost to go, assuming optimal $u$. That is,

$$J(q,\xi,t;\mu) = E\Big[\int_t^T \big(q(\tau)^2 + u(\tau)^2\big)\,d\tau\Big], \tag{5.4}$$

with $u$ picked to minimize the right hand side. Proceeding as in Section 4, we may derive the following partial differential equation for $J$:

$$0 = \partial_t J + (q^2 + u^2) + (\bar{a} + u)\partial_q J + \bar{a}\partial_\xi J + \tfrac{1}{2}\partial_q^2 J + \partial_\xi \partial_q J + \tfrac{1}{2}\partial_\xi^2 J, \tag{5.5}$$

where $\bar{a}(\xi, t)$ denotes the expected value of $a$ with respect to $d\mu_{\text{posterior}}$, and the optimal $u$ is given by

$$u = -\tfrac{1}{2}\partial_q J. \tag{5.6}$$

Again we impose the boundary condition $J = 0$ at $t = T$.

Let us specialize to the case in which our prior belief $\mu$ is a normal distribution with mean zero and standard deviation $\sigma$. Then from formula (5.3), we see that our posterior belief is given by the probability density

$$\rho(a) = \exp\Big(\xi a - \frac{a^2}{2}(t + \sigma^{-2})\Big) \cdot Z(\xi, t),$$

and consequently the posterior expected value of $a$ is

$$\bar{a}(\xi, t) = \frac{\xi}{t + \sigma^{-2}}\cdot \tag{5.7}$$

As in Section 4, we guess a solution of the form

$$J(q, \xi, t) = E_2(t)q^2 + E_1(t)\bar{a}(\xi, t)q + J_0(\xi, t), \tag{5.8}$$

and obtain the following ordinary differential equations:

$$-\dot{E}_2 = 1 - E_2^2, \qquad E_2(T) = 0, \tag{5.9a}$$

$$-\dot{E}_1 = 2E_2 - E_1E_2, \qquad E_1(T) = 0, \tag{5.9b}$$

together with a partial differential equation for $J_0$:

$$0 = \partial_t J_0 + E_1\bar{a}^2 - \frac{1}{4}E_1^2\bar{a}^2 + \bar{a}\partial_\xi J_0 + E_2 + \frac{E_1}{t + \sigma^{-2}} + \tfrac{1}{2}\partial_\xi^2 J_0.$$

(It turns out that we will never need to know $J_0$, so we do not need to solve this partial differential equation.) Note that the ordinary differential equations for $E_2$ and $E_1$ are the same as those in (4.3), and thus $E_2$ and $E_1$ are again given by formulas (4.4a) and (4.4b). Thanks to formulas (5.6) and (5.8), the optimal control is now given by

$$u = -E_2(t)q - \frac{E_1(t)}{2}\,\bar{a}(\xi, t), \tag{5.10}$$

where $\bar{a}$ is given by (5.7). Note that this is the same as formula (4.5) for the optimal control for known $a$, but with $a$ (which is now unknown) replaced by $\bar{a}(\xi, t)$.

This concludes our derivation of the Bayesian strategy.

## 6. Performance of the Bayesian strategy

Now we wish to determine how well the Bayesian strategy (for unknown $a$) performs for a particular value of $a$. More precisely, we now assume that $q$ evolves according to equation (2.1), for a particular $a$, where $u$ is given by equations (5.5), (5.6) for a particular prior belief $\mu$. Let $\mathcal{J}(q,\xi,t;a,\mu)$ be the expected cost to go under the above assumptions. As in Sections 4 and 5, we can derive the following partial differential equation for $\mathcal{J}$:

$$0 = \partial_t \mathcal{J} + (a+u)\partial_q \mathcal{J} + a\partial_\xi \mathcal{J} + \tfrac{1}{2}\partial_q^2 \mathcal{J} + \partial_\xi \partial_q \mathcal{J} + \tfrac{1}{2}\partial_\xi^2 \mathcal{J} + q^2 + u^2, \tag{6.1}$$

with $u$ given by (5.5),(5.6). Again, $\mathcal{J} = 0$ when $t = T$. Let us now specialize to the case of Gaussian prior belief, as in Section 5, so that the optimal $u$ is given by (5.10).

We can guess a solution of the form

$$\mathcal{J}(q,\xi,t;a,\mu) = E_2(t)q^2 + E_1(t)qa + E_0(t)a^2 + F_0(t)(\bar{a}(\xi,t) - a)^2 + F_\sharp(t). \tag{6.2}$$

Then, thanks to (5.7), we find this solution does indeed satisfy (6.1), provided the following ODEs are satisfied:

$$-\dot{E}_2 = 1 - E_2^2, \qquad E_2(T) = 0, \tag{6.3a}$$

$$-\dot{E}_1 = 2E_2 - E_1E_2, \qquad E_1(T) = 0, \tag{6.3b}$$

$$-\dot{E}_0 = E_1 - \frac{E_1^2}{4}, \qquad E_0(T) = 0, \tag{6.3c}$$

$$-\dot{F}_0 = -\frac{2F_0}{t+\sigma^{-2}} + \frac{E_1^2}{4}, \qquad F_0(T) = 0, \tag{6.3d}$$

$$-\dot{F}_\sharp = E_2 + \frac{F_0}{(t+\sigma^{-2})^2}, \qquad F_\sharp(T) = 0. \tag{6.3e}$$

Once again, $E_2$, $E_1$, and $E_0$ are as in (4.3), and hence are given by (4.4). One readily verifies that the following satisfy equations (6.3d) and (6.3e):

$$F_0(t) = (t+\sigma^{-2})^2 \int_t^T \frac{E_1(\tau)^2}{4(\tau+\sigma^{-2})^2}\,d\tau, \tag{6.4a}$$

$$F_\sharp(t) = \int_t^T \left[E_2(\tau) + \frac{F_0(\tau)}{(\tau+\sigma^{-2})^2}\right] d\tau. \tag{6.4b}$$

This concludes our discussion of the performance of the Bayesian strategy for given $a$.

## 7. Results for agnostic control problems

In this section we determine strategies that optimize each of the three variants of agnostic control considered in Section 3.3: additive regret, multiplicative regret, and the fuel tax variant. It turns out that in each case, the optimal strategy is a Bayesian strategy in which the prior belief about $a$ is a normal distribution with mean zero and standard deviation $\sigma$. The optimal choice of $\sigma$ depends on which type of agnostic control one is considering. (Strictly speaking, for additive regret, the optimal strategy corresponds to the limit $\sigma \to \infty$.)

### 7.1. Minimizing additive regret

In this case, as we will see in a moment, we will need to consider a non-zero starting time $T_0$. The cost for the optimal control with known $a$, starting at time $T_0$, is given by (4.2):

$$J(q,T_0;a) = E_2(T_0)q^2 + E_1(T_0)qa + E_0(T_0)a^2 + E_\sharp(T_0).$$

The cost for our Bayesian strategy, for a particular $a$, also starting at time $T_0$, is given by (6.2):

$$\mathcal{J}(q,\xi,T_0;a,\sigma) = E_2(T_0)q^2 + E_1(T_0)qa + E_0(T_0)a^2 + F_0(T_0)(\bar{a}(\xi,T_0)-a)^2 + F_\sharp(T_0).$$

The difference between $\mathcal{J}$ and $J$ is therefore

$$F_0(T_0)\,(\bar{a}(\xi,T_0)-a)^2 + F_\sharp(T_0) - E_\sharp(T_0). \tag{7.1}$$

Now suppose we start at time $t=0$ and position $q=0$. Until time $T_0$ we are required to set our control $u=0$, after which we are free to pick the optimal $u$. Then $q(T_0)$ is normally distributed with mean $aT_0$ and standard deviation $T_0^{1/2}$. Moreover, $\xi(T_0)=q(T_0)$ (see equation (5.2) and recall that $q(0)=0$). Consequently, $(\bar{a}(\xi,t)-a)^2$ has expected value

$$\frac{T_0 + a^2\sigma^{-4}}{(T_0+\sigma^{-2})^2}$$

(see formula (5.7)), so from (7.1) we see that the additive regret is given by

$$\mathrm{AR}(a) = F_0(T_0)\cdot\frac{T_0 + a^2\sigma^{-4}}{(T_0+\sigma^{-2})^2} + F_\sharp(T_0) - E_\sharp(T_0). \tag{7.2}$$

As $\sigma$ tends to infinity, the control strategy becomes

$$u = -E_2(t)q - \frac{1}{2}E_1(t)\frac{\xi}{t}, \tag{7.3}$$

thanks to (5.7) and (5.10), and the additive regret is given by

$$F_0(T_0)\,T_0^{-1} + F_\sharp(T_0) - E_\sharp(T_0), \tag{7.4}$$

which is independent of $a$. Thus we have found a strategy that optimizes worst-case additive regret. In particular, the reasoning in Section 3.4 applies here even though our strategy is not a Bayesian strategy for a particular prior, but is instead a limit of such strategies.

One can check that, as $T_0$ tends to zero, the additive regret (7.4) tends to infinity (in particular, $F_\sharp$ diverges logarithmically), which is why we introduced the parameter $T_0$. In our discussions of multiplicative regret and the fuel tax variant, we will take $T_0 = 0$.

### 7.2. Minimizing multiplicative regret

Thanks to equations (4.2) and (6.2), the multiplicative regret of the optimal Bayesian strategy for normal prior belief with standard deviation $\sigma$ is given by

$$\mathrm{MR}(a) = \frac{(E_0(0)+F_0(0))\,a^2 + F_\sharp(0)}{E_0(0)\,a^2 + E_\sharp(0)}. \tag{7.5}$$

Here, $F_0$ and $F_\sharp$ depend on $\sigma$; see equations (6.4a) and (6.4b). We wish to find a value of $\sigma$ for which the expression (7.5) is independent of $a$. This occurs precisely when

$$\frac{E_0(0) + F_0(0)}{E_0(0)} = \frac{F_\sharp(0)}{E_\sharp(0)}. \tag{7.6}$$

For each $T$, (7.6) is a single equation for the one unknown $\sigma$. As a function of $T$, the solution may be computed numerically (e.g., using Newton's method). To carry out the numerics, we need to evaluate the integrals (6.4) using quadrature, or solve the ODEs for $F_0$ and $F_\sharp$ backwards, from $t = T$ to $t = 0$. The optimal $\sigma$ is shown as a function of $T$ in Figure 1. One can show that the optimal $\sigma$ tends to zero as $T \to \infty$.

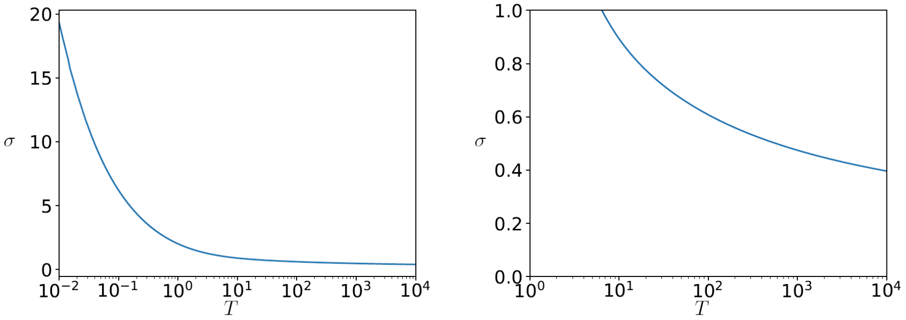

**Figure 1.** Optimum standard deviation of prior belief about $a$, to minimize multiplicative regret.

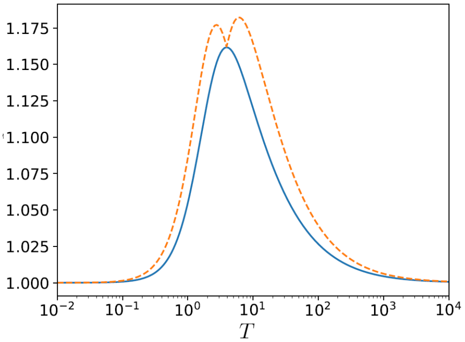

**Figure 2.** Worst-case multiplicative regret for the optimum strategy (solid); and a Bayesian strategy with a particular $\sigma$ independent of $T$ (dashed).

Figure 2 shows the worst-case regret as a function of $T$, for two different strategies. The solid curve arises from the optimum strategy: the Bayesian strategy with $\sigma$ chosen, for each $T$, to minimize worst-case multiplicative regret. The dashed curve arises from a Bayesian strategy, for a particular value of $\sigma$ chosen independently of $T$. More precisely, we pick $\sigma$ optimally for the value of $T$ for which the solid curve reaches its maximum.

Perhaps surprisingly, both strategies perform very well. The optimum strategy is never more than 17% worse than the optimum strategy for known $a$. Even the “simple” Bayesian strategy with fixed $\sigma$ is not much worse. And regardless of the choice of $\sigma$, the worst-case multiplicative regret for the Bayesian strategy goes to 1 as $T \to 0$ or $T \to \infty$.

### 7.3. Minimizing fuel tax regret

Recall from equation (2.2) the parameter $\lambda$, which represents the “cost of fuel.” Our goal here is to compare the expected cost $J_{\text{unknown } a}$ of a strategy with unknown $a$ and $\lambda = 1$ with the cost $J_{\text{known } a}(\lambda)$ of an optimal strategy with known $a$ and $\lambda > 1$ (see Section 3.3). As in our discussion of multiplicative regret, we look for a Bayesian strategy with normal prior having mean zero and standard deviation $\sigma$. For fixed $T$ and $\lambda$, we pick $\sigma$ so that the ratio $J_{\text{unknown } a} / J_{\text{known } a}(\lambda)$ is independent of $a$. This ratio is thus a function of $\lambda$, and we wish to find $\lambda$ to make the ratio equal to 1. Thanks to the discussion in Section 3.4, this value of $\lambda$ is then the fuel tax regret defined in Section 3.3. Formulas (4.6) and (4.7), together with (6.2), (4.4), and (6.4), make this a routine numerical computation.

Figure 3 shows the results of these computations. As in the case of multiplicative regret, the optimal value of $\sigma$ is a decreasing function of $T$, and we pay only a modest price for our lack of knowledge of $a$. The most challenging case arises for $T \approx 2$, with a corresponding fuel tax regret $\lambda \approx 1.3$.

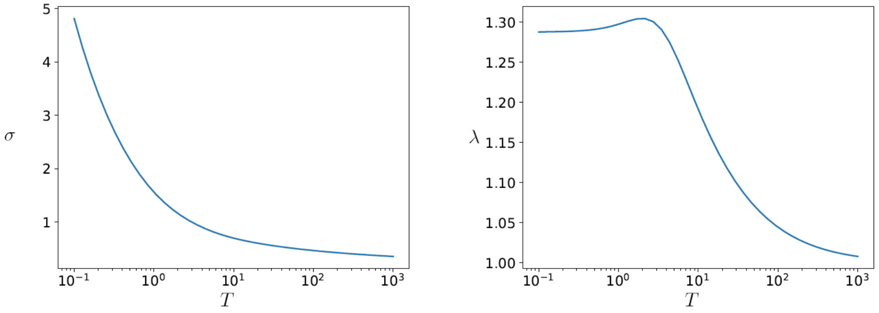

**Figure 3.** Optimal Bayesian strategies for the fuel-tax variant, and resulting fuel tax regret $\lambda$.

## 8. Conclusions and problems for further study

We have posed several optimal control problems for a toy model

$$\Delta q = (a + u(t))\Delta t + \Delta W$$

of dynamics with noise, involving a single parameter $a$. These include the classical problem in which $a$ is known, and the Bayesian problem in which we assume a prior belief regarding $a$. We posed also three “agnostic” control problems, in which nothing is assumed about $a$, and we hope to minimize some notion of regret. Here we have considered three different notions of regret: additive regret (often called simply “regret”); multiplicative regret (often called "competitive ratio"); and fuel-tax regret, which we have not seen in the literature.

Each of these optimal control problems involves minimizing a cost of the form

$$J = \int_{T_0}^{T} (q^2 + \lambda u^2)\, dt,$$

for a particular terminal time $T$. Previous studies such as [8] have produced strategies that behave well as the terminal time $T$ tends to infinity. We work in a different regime. We are concerned with a fixed value of $T$, but assume nothing about the parameter $a$.

For each of our agnostic control problems, minimum regret is achieved by a Bayesian strategy (or a limit of Bayesian strategies) with the unknown $a$ assumed to be normally distributed, with mean zero and standard deviation depending on the terminal time and the notion of regret. The optimal strategy for each of our agnostic control problems has regret independent of the actual value of $a$.

We have only begun by solving the simplest toy problem of agnostic control. Already, substantial challenges arise when we consider further toy problems, with dynamics given by

$$\Delta q = (aq + u)\Delta t + \Delta W \tag{8.1}$$

or

$$\Delta q = au\,\Delta t + \Delta W \tag{8.2}$$

in place of equation (2.1). These toy problems are the subject of ongoing work, and we hope to return to them in a future paper. Note that, if we regard the unknown $a$ as part of the state, then the dynamics are linear for our first toy problem (2.1), but not for (8.1) or (8.2). This observation may explain why toy problems (8.1) and (8.2) are more challenging than the toy problem solved here [12].

**Acknowledgements.** We are particularly grateful to Fred Leve for raising the issue of optimal control with learning on the fly, and Amir Ali Ahmadi and Elad Hazan for helpful discussions.

**Funding.** This work was supported by the Air Force Office of Scientific Research, under awards FA9550-19-1-0005 and FA9550-18-1-0069, and partially supported by the National Science Foundation, under grant DMS-1700180.

## References

[1] Agarwal, N., Bullins, B., Hazan, E., Kakade, S. M. and Singh, K.: Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning (Long Beah, California, 2019)*, 111–119. Proceedings of Machine Learning Research 97, PMLR, 2019.

[2] Agarwal, N., Hazan, E. and Singh, K.: Logarithmic regret for online control. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*, 10 pp. Advances in Neural Information Processing Systems 32, Curran Associates, 2019.

[3] Auer, P., Cesa-Bianchi, N. and Fischer, P.: Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** (2002), no. 2-3, 235–256.

[4] Bertsekas, D. P.: *Dynamic programming and optimal control, Vol. I*. Athena Scientific, Belmont, MA, 2005.

[5] Brazy, D. P.: *Group chairman's factual report of investigation*. National Transportation Safety Board Docket No. SA-532, Exhibit No. 12, 2009.

[6] Bubeck, S. and Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.* **5** (2012), no. 1, 1–122.

[7] Cesa-Bianchi, N. and Lugosi, G.: *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.

[8] Cohen, A., Koren, T. and Mansour, Y.: Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *Proceedings of the 36th International Conference on Machine Learning (Long Beah, California, 2019)*, 1300–1309. Proceedings of Machine Learning Research 97, PMLR, 2019.

[9] Hazan, E.: *Introduction to online convex optimization*. Foundation and Trends in Optimization 2, Now Publishers, 2016.

[10] Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** (1985), no. 1, 4–22.

[11] Powell, W. B.: *Approximate dynamic programming. Solving the curses of dimensionality*. Second edition, Wiley Series in Probability and Statistics, Wiley-Interscience, 2011.

[12] P. Ramadge: Personal communication, 2019.

Received February 27, 2020. Published online June 22, 2021.

**Charles L. Fefferman**
Department of Mathematics, Princeton University, Fine Hall, Princeton NJ 08544, USA;
cf@math.princeton.edu

**Bernat Guillén Pegueroles**
Program in Applied and Computational Mathematics, Princeton University, Fine Hall, Princeton NJ 08544, USA;
bernatp@math.princeton.edu

**Clarence W. Rowley**
Department of Mechanical and Aerospace Engineering, Princeton University, Engineering Quadrangle, Princeton NJ 08544, USA;
cwrowley@princeton.edu

**Melanie Weber**
Program in Applied and Computational Mathematics, Princeton University, Fine Hall, Princeton NJ 08544, USA;
mw25@math.princeton.edu