
Regret-optimal control in dynamic environments

Gautam Goel

Caltech

Babak Hassibi

Abstract

We consider the control of linear time-varying dynamical systems from the perspective of regret minimization. Unlike most prior work in this area, we focus on the problem of designing an online controller which competes with the best dynamic sequence of control actions selected in hindsight, instead of the best controller in some specific class of controllers. This formulation is attractive when the environment changes over time and no single controller achieves good performance over the entire time horizon. We derive the structure of the regret-optimal online controller via a novel reduction to H_∞ control and present a clean data-dependent bound on its regret. We also present numerical simulations which confirm that our regret-optimal controller significantly outperforms the H_2 and H_∞ controllers in dynamic environments.

1 Introduction

The central question in control theory is how to regulate the behavior of an evolving system with state x that is perturbed by a disturbance w by dynamically adjusting a control action u . Traditionally, this question has been studied in two distinct settings: in the H_2 setting, we assume that the disturbance w is generated by a stochastic process and seek to select the control u so as to minimize the *expected* control cost, whereas in the H_∞ setting we assume the noise is selected adversarially and instead seek to minimize the *worst-case* control cost.

Both H_2 and H_∞ controllers suffer from an obvious drawback: they are designed with respect to a specific

class of disturbances, and if the true disturbances fall outside of this class, the performance of the controller may be poor. Indeed, the loss in performance can be arbitrarily large if the disturbances are carefully chosen [9].

This observation naturally motivates the design of *adaptive* controllers, which dynamically adjust their control strategy as they sequentially observe the disturbances instead of blindly following a prescribed strategy. This problem has attracted much recent attention in machine learning (e.g. [2, 3, 6, 7, 11, 14, 18]), mostly from the perspective of regret minimization. In this framework, the online controller is chosen so as to minimize the difference between its cost and the best cost achievable in hindsight by a controller from some fixed class of controllers. The resulting controllers are adaptive in the sense that they seek to minimize cost irrespective of how the disturbances are generated.

In this paper, we take a somewhat different approach to the design of adaptive controllers. Instead of designing a controller to minimize regret against the best controller selected in hindsight from some specific class, we instead focus on designing a controller which minimizes regret against the best *dynamic sequence* of control actions selected in hindsight. By best dynamic sequence, we simply mean the globally optimal sequence of control actions

$$\operatorname{argmin}_{u_0, \dots, u_{T-1}} \operatorname{cost}(w_0, \dots, w_{T-1}, u_0, \dots, u_{T-1}),$$

where $\operatorname{cost}(u, w)$ is the control cost incurred by the control action u and the disturbance w .

We believe that this formulation of regret minimization in control is more attractive than the standard formulation, where the controller learns the best fixed controller in some specific class, for two fundamental reasons. Firstly, it is more general: instead of imposing *a priori* some specific structure on the controller we learn (e.g. state feedback, LTI controllers, etc), which may or may not be appropriate for the given control task, we instead try to compete with the globally optimal dynamic sequence of control ac-

tions, with no artificial constraints. Secondly, and perhaps more importantly, the controllers we obtain are more robust to changes in the environment. Consider, for example, a scenario in which an online controller is trying to compete with the best state feedback or LTI controller selected in hindsight, and suppose the disturbances alternate between being generated by a stochastic process and being generated adversarially. When the disturbances are stochastic, an optimistic controller (such as the H_2 controller) will perform well; conversely, when the disturbances are adversarial, a more conservative, pessimistic controller (such as an H_∞ controller) will perform well. No single controller will perform well over the entire time horizon, and hence any online algorithm which tries to learn the best static controller will incur high cumulative cost. A controller which minimizes regret against the best dynamic sequence, however, is not constrained to converge to any static controller, and hence can potentially outperform standard regret-minimizing control algorithms when the environment is dynamic.

Our approach to control is very similar in outlook to a series of works in online learning (e.g. [19, 20]), which seek to design algorithms with low *adaptive regret*, i.e. algorithms which compete with a dynamic sequence of actions instead of the best fixed action selected in hindsight. This formulation is natural when the reward distribution encountered by the online algorithm varies over time. We refer the reader to Ch. 12 of [17] for a survey of this area.

1.1 Contributions of this paper

We focus on the Linear Quadratic Regulator (LQR) setting where the costs are quadratic and the underlying dynamical system is linear; we allow the dynamics and costs to be time-varying. Our main contribution is to derive the structure of the controller which minimizes regret against the best dynamic sequence of control actions; we call this controller the “regret-optimal” controller. We first derive the structure of the optimal noncausal controller, which selects the globally optimal control u with full knowledge of the disturbance w . Given this noncausal controller, we show that the design of the regret-optimal controller can be recovered via a novel reduction to H_∞ control. Because the optimal H_∞ controller is already known, a nice bonus of this reduction is that the regret-optimal controller we derive is truly optimal; no controller can achieve lower regret.

We show that the regret-optimal and optimal noncausal controllers enjoy a surprising symmetry: the optimal noncausal controller is the optimal H_2 controller plus a correction that depends on current and future disturbances, whereas the regret-optimal con-

troller is the optimal H_2 controller plus a correction that depends on current and past disturbances. We present numerical simulations which confirm that our adaptive, regret-optimal controller captures almost all the performance of a fixed controller when the environment is static and can significantly outperform a fixed controller when the environment changes over time.

1.2 Related work

Regret minimization in control. There has been a surge of interest in regret minimization in control in the past few years, so much so that we are able to survey only a tiny fraction of the papers in this area. One of the first works in this area was [2], which focused on regret minimization when the noise is stochastic; a more general setting where the noise is stochastic but the costs are adversarial was considered in [6]. A series of more recent papers (e.g. [3, 7, 11, 18]) consider the setting where the noise is adversarial. In all of these works, the online learner is trying to minimize static regret against a fixed benchmark controller, often taken to be a state feedback or LTI controller.

Competitive analysis. A key distinction between this paper and most previous work at the intersection of machine learning and control is that we focus on designing an online controller which competes against a globally optimal noncausal controller. This problem was also studied in [14] (albeit through the lens of *competitive ratio* rather than regret) where it was shown that the Online Balanced Descent algorithm introduced in [5] could be used to give some performance guarantees in the LQR setting; this result was improved in [15]. We note that the reduction in those works relied crucially on very strong assumptions about the structure of the dynamics, such as invertibility of the control matrix. Similar assumptions were made in [12], which considered the problem of timescale separation in LQR control.

2 Preliminaries

We consider a linear dynamical system governed by the following evolution equation:

$$x_{t+1} = A_t x_t + B_{u,t} u_t + B_{w,t} w_t.$$

Here $x_t \in \mathbb{R}^n$ is a state variable we are interested in regulating, $u_t \in \mathbb{R}^m$ is a control variable which we can dynamically adjust to influence the evolution of the system, and $w_t \in \mathbb{R}^n$ is a disturbance. We assume without any real loss of generality that the initial state is $x_0 = 0$. We consider the evolution of this system over a finite time horizon $t = 0, \dots, T-1$ and often use the notation $u = (u_0, \dots, u_{T-1})$, $x = (x_0, \dots, x_{T-1})$,

$w = (w_0, \dots, w_{T-1})$. We formulate the problem of regulating the system over a finite time horizon $t = 0 \dots T-1$ as an optimization problem, where the goal is to select the control actions so as to minimize the LQR cost

$$\text{cost}(w, u) = x_T^\top Q_T x_T + \sum_{t=0}^{T-1} (x_t^\top Q_t x_t + u_t^\top R_t u_t),$$

where $Q_t, R_t \succ 0$ for $t = 0, \dots, T-1$ and $Q_T \succ 0$ is a terminal state cost. We say that a controller is *causal* if the control action at time t depends only on the previous disturbances up to time t ; otherwise we say the controller is *noncausal*. Define $A = (A_0, \dots, A_{T-1})$ and define B_u, B_w, Q, R analogously. We assume that A, B_u, B_w, Q, R are known, so the only uncertainty in the evolution of the system comes from the disturbance w . We will assume in each timestep that $R_t = I$; we emphasize that this imposes no real restriction, since for all $R_t \succ 0$ we can always rescale u_t so that $R_t = I$.

In a seminal paper [22], Kalman considered the setting where the noise is stochastic and derived the structure of the causal controller which minimizes the expected cost, i.e. the H_2 optimal controller:

Theorem 1 (Kalman, 1960). *The optimal H_2 controller has the form*

$$u_t = -H_t^{-1} B_{u,t}^\top P_{t+1} (A_t x_t + B_{w,t} w_t),$$

where $H_t = (R_t + B_{u,t}^\top P_{t+1} B_{u,t})$ and P_t is the solution of the backwards Riccati recursion

$$P_t = Q_t + A_t^\top P_{t+1} A - A_t^\top P_{t+1} B_t H_t^{-1} P_{t+1} A_t \quad (1)$$

where we initialize $P_T = Q_T$.

One can also consider a setting with adversarial noise:

Problem 1 (Optimal H_∞ control problem). *Find a causal control strategy $\hat{u}_t = \mathcal{F}(w_0, \dots, w_t, u_0, \dots, u_{t-1})$ that minimizes*

$$\sup_w \frac{\text{cost}(w, u)}{\sum_{t=0}^{T-1} \|w_t\|_2^2}.$$

We let γ_{opt}^2 denote the optimal value of this ratio. In general, it is not known how to derive a closed-form for γ_{opt}^2 or the optimal H_∞ controller, so instead it is common to consider a relaxation:

Problem 2 (Suboptimal H_∞ control problem). *Given a performance level $\gamma > 0$, find a causal control strategy $\hat{u}_t = \mathcal{F}(w_0, \dots, w_t, u_0, \dots, u_{t-1})$ such that*

$$\sup_{x_0, w \in \ell_2} \frac{\text{cost}(w, u)}{\sum_{t=0}^{T-1} \|w_t\|_2^2} < \gamma^2,$$

or determine whether no such policy exists.

This problem has a well-known solution:

Theorem 2 (Theorem 9.5.1 in [16]). *For any performance level $\gamma^2 > 0$, define Δ_t as*

$$-\gamma^2 I + B_{w,t}^\top P_{t+1} B_{w,t} - B_{w,t}^\top P_{t+1} B_{u,t} H_t^{-1} B_{u,t}^\top P_{t+1} B_{w,t},$$

where

$$H_t = (R_t + B_{u,t}^\top P_{t+1} B_{u,t})$$

P_t is the solution of the backwards-time Riccati equation

$$P_t = Q_t + A_t^\top P_{t+1} A_t - A_t^\top P_{t+1} \hat{B}_t \hat{H}_t^{-1} \hat{B}_t^\top P_{t+1} A_t$$

with initialization $P_T = Q_T$ and we define

$$\hat{B}_t = [B_{u,t} \quad B_{w,t}],$$

$$\hat{H}_t = \begin{bmatrix} R_t & 0 \\ 0 & -\gamma^2 I \end{bmatrix} + \hat{B}_t^\top P_{t+1} \hat{B}_t.$$

A suboptimal H_∞ controller exists at performance level γ^2 if and only if $\Delta_t \prec 0$ for $t = 0, \dots, T-1$.

In this case, the suboptimal H_∞ controller has the form

$$u_t = -H_t^{-1} B_{u,t}^\top P_{t+1} (A_t x_t + B_{w,t} w_t).$$

Note that the optimal H_∞ controller is easily obtained from the solution of the suboptimal H_∞ problem by bisection on γ^2 ; we iteratively reduce the value of γ^2 until we find the smallest value of γ^2 such that the constraints on Δ_t are satisfied. This value of γ^2 is precisely γ_{opt}^2 and the corresponding controller is the optimal H_∞ controller [8, 23–25].

In this paper, instead of minimizing the worst-case cost, our goal is to minimize the worst-case *regret*; this problem has a natural analog of the H_∞ problem:

Problem 3 (Regret-optimal control problem). *Find a causal control strategy $\hat{u}_t = \mathcal{F}(w_0, \dots, w_t, u_0, \dots, u_{t-1})$ that minimizes*

$$\sup_w \frac{\text{cost}(w, u) - \min_u \text{cost}(w, u)}{\sum_{t=0}^{T-1} \|w_t\|_2^2}.$$

As in the H_∞ setting, we consider the relaxation:

Problem 4 (Regret-suboptimal control problem). *Given a performance level $\gamma > 0$, find a causal control strategy $\hat{u}_t = \mathcal{F}(w_0, \dots, w_t, u_0, \dots, u_{t-1})$ such that*

$$\frac{\text{cost}(w, u) - \min_u \text{cost}(w, u)}{\sum_{t=0}^{T-1} \|w_t\|_2^2} < \gamma^2$$

for all disturbances w , or determine whether no such policy exists.

We emphasize that, as in the H_∞ setting, if we can solve the regret-suboptimal problem, we can easily recover the solution to the regret-optimal problem by performing bisection on γ^2 until convergence to γ_{opt}^2 .

3 Derivation of the optimal noncausal controller

We first derive the structure of the optimal noncausal controller before turning to the design of the regret-optimal controller. Let $s_t = Q_t^{\frac{1}{2}}x_t$, and define $s = (s_0, \dots, s_{T-1})$. We can write $s = Fu + Gw$, where F and G are causal (i.e. lower triangular) operators encoding the dynamics (A, B_u, B_w) and the costs (Q, R) . Fix any noise realization w . The cost-minimizing noncausal sequence of control actions with respect to w is the solution of

$$\min_u \|Fu + Gw\|_2^2 + \|u\|_2^2.$$

Notice that we allow u to be arbitrary; in particular we do not assume *a priori* that u is a linear function of w . Solving for the optimal u via completion-of-squares, we obtain

$$u^* = -(I + F^\top F)^{-1}F^\top Gw. \quad (2)$$

Notice that this policy is a fixed linear function of w ; we have shown that the optimal noncausal policy is to always select the control sequence $u^* = K^*w$ where

$$K^* = -(I + F^\top F)^{-1}F^\top G.$$

In light of (2), it is easy to compute the optimal noncausal cost:

$$\text{cost}(w; K^*w) = w^\top G^\top (I + FF^\top)^{-1}Gw. \quad (3)$$

While these calculations describe the optimal noncausal controller in “operator” form, it is often more computationally useful to have a state-space description of the controller. We can use dynamic programming to recursively compute the optimal noncausal control actions, starting from the last time step and moving backwards in time; this gives the desired state-space description. For this result we drop the assumption that $R_t = I$.

For any fixed sequence of noise increments $w = (w_0, \dots, w_{T-1})$, define the “offline cost-to-go” function $V_t^w(x)$ as

$$\min_u [x^\top Q_t x + u^\top R_t u + V_{t+1}^w(A_t x + B_{u,t}u + B_{w,t}w_t)]$$

for $t = 1 \dots T - 1$, with $V_T(x) = x^\top Q_T x$. This function measures the aggregate cost over the future time horizon starting at the state x at time t , under the assumption that in each time step, the offline controller picks the control action which minimizes the future cost given the current state and the realizations $w_t \dots w_{T-1}$.

We will show that $V_t^w(x)$ can be written as $x^\top P_t x + v_t^\top x_t + q_t$ for all $t \in [1 \dots T]$, where P_t is the solution of a backwards Riccati recursion. The claim is

clearly true for $t = T$, since we can take $(P_T, v_T, q_T) = (Q_T, 0, 0)$. Proceeding by backwards induction, suppose $V_{t+1}^w(x) = x^\top P_{t+1}x + v_{t+1}^\top x + q_{t+1}$ for some v_T, q_T . We can therefore write $V_t^w(x)$ as

$$\begin{aligned} & \min_u [x^\top Q_t x + u^\top R_t u \\ & + (A_t x + B_{u,t}u + B_{w,t}w_t)^\top P_{t+1}(A_t x + B_{u,t}u + B_{w,t}w_t) \\ & + v_{t+1}^\top (A_t x + B_{u,t}u + B_{w,t}w_t) + q_{t+1}]. \end{aligned}$$

Solving for the minimizing u , we see that the offline optimal control action is

$$u_t^* = -H_t^{-1}B_{u,t}^\top \left(P_{t+1}A_t x_t + P_{t+1}B_{w,t}w_t + \frac{1}{2}v_{t+1} \right)$$

where we define $H_t = (R_t + B_{u,t}^\top P_{t+1}B_{u,t})$. In light of Theorem 1, we have shown that *the optimal noncausal control action is the sum of the optimal H_2 control action and a term which depends only on current and future disturbances.*

Plugging this choice of u_t^* into our expression for $V_t^w(x)$ and collecting terms, we see that $V_t^w(x) = x^\top P_t x + v_t^\top x + q_t$ where P_t is the solution of the discrete time Riccati equation obtained by Kalman, and v_t and q_t satisfy the recurrences

$$v_t = 2A_t^\top S_t B_{w,t}w_t + A_t^\top S_t P_{t+1}^{-1}v_{t+1}, \quad (4)$$

$$\begin{aligned} q_t &= w_t^\top B_{w,t}^\top S_{t+1} B_{w,t}w_t + v_{t+1}^\top P_{t+1}^{-1} S_t B_{w,t}w_t \\ &\quad - \frac{1}{4}v_{t+1}^\top B_{u,t}H_t^{-1}B_{u,t}^\top v_{t+1} + q_{t+1}, \end{aligned} \quad (5)$$

and we define

$$S_t = P_{t+1} - P_{t+1}B_{u,t}H_t^{-1}B_{u,t}^\top P_{t+1}. \quad (6)$$

We have proven:

Theorem 3. *The optimal noncausal controller has the form*

$$u_t^* = -H_t^{-1}B_{u,t}^\top \left(P_{t+1}A_t x_t + P_{t+1}B_{w,t}w_t + \frac{1}{2}v_{t+1} \right),$$

where we define

$$H_t = R_t + B_{u,t}^\top P_{t+1}B_{u,t},$$

P_t is the solution of the discrete-time Riccati recurrence 1, and v_t satisfies the recurrence (4).

We note that this result generalizes one recently obtained in [13], which considered the special case where the dynamics and costs are time-invariant.

4 Derivation of the regret-optimal controller

We now turn to the problem of deriving the regret-optimal controller. Our approach is to reduce the regret-suboptimal control problem to the suboptimal H_∞ problem. As in the H_∞ setting, once the regret-suboptimal controller is found, the regret-optimal controller is easily obtained by bisection.

Let u^* be the optimal noncausal sequence of control actions derived in (2) and let $s^* = Fu^* + Gw$ be the corresponding sequence of state costs. Then, (assuming $R = I$), the offline cost is $\|s^*\|_2^2 + \|u^*\|_2^2$. Recall that the regret-suboptimal problem (Problem 4) with performance level γ^2 is to find, if possible, a causal strategy $u_t = \mathcal{F}(x_0, w_0, \dots, w_t, u_0, \dots, u_{t-1})$, such that for all w ,

$$\|u\|_2^2 + \|s\|_2^2 < \gamma^2 \|w\|_2^2 + \|u^*\|_2^2 + \|s^*\|_2^2, \quad (7)$$

where we define $s = Fu + Gw$ as in Section 3.

Our approach shall be to find a change of variables such that this problem takes the form of the suboptimal H_∞ control problem (Problem 2). Recall from equation (3) that

$$\|u^*\|_2^2 + \|s^*\|_2^2 = w^\top G^\top (I + FF^\top)^{-1} Gw,$$

hence

$$\gamma^2 \|w\|_2^2 + \|u^*\|_2^2 + \|s^*\|_2^2 = w^\top (\gamma^2 I + G^\top (I + FF^\top)^{-1} G)w.$$

Suppose we can find a causal, invertible matrix L such that

$$\gamma^2 I + G^\top (I + FF^\top)^{-1} G = L^\top L.$$

In this case,

$$\gamma^2 \|w\|_2^2 + \|u^*\|_2^2 + \|s^*\|_2^2 = \|Lw\|_2^2.$$

Letting $z = Lw$ and $G' = GL^{-1}$, we have $s = Fu + G'z$. With this change of variables, the regret-suboptimal problem (7) takes the form of finding a causal strategy $u_t = \mathcal{F}(x_0, w_0, \dots, w_t, u_0, \dots, u_{t-1})$, such that for all w ,

$$\|u\|_2^2 + \|s\|_2^2 < \|z\|_2^2.$$

In other words, the regret-suboptimal problem with performance level γ^2 on the system with disturbance w and dynamics given by F and G is precisely equivalent to the suboptimal H_∞ problem with performance level 1, disturbance z , and dynamics given by F and G' .

The main technical challenge is to factor $\gamma^2 I + G^\top (I + FF^\top)^{-1} G$ as $L^\top L$. Our approach is centered around

repeated use of the celebrated Kalman filter, which, given a covariance matrix Σ , computes a Cholesky factorization $\Sigma = M^\top M$ where M is causal; we refer the reader to [21] for more background on Kalman filters and state-space models. We first construct a random variable with covariance $I + FF^\top$; we then use the Kalman filter to factor this operator as $\Delta\Delta^\top$ where Δ is causal. We subsequently construct a new random variable with covariance $\gamma^2 I + G^\top (I + FF^\top)G = \gamma^2 I + (\Delta^{-1}G)^\top (\Delta^{-1}G)$; again we use the Kalman filter to factor this operator as $L^\top L$ where L is causal. Once we have obtained L , it is straightforward to recover the regret-optimal controller.

We now state our main result:

Theorem 4. *The regret-suboptimal controller at performance level γ^2 has the form*

$$u_t = -\hat{H}_t^{-1} \hat{B}_{u,t}^\top \hat{P}_{t+1} \left(\hat{A}_t \begin{bmatrix} \zeta_t \\ \hat{\nu}_t \end{bmatrix} + \hat{B}_{w,t} z_t \right),$$

where we define

$$\hat{A}_t = \begin{bmatrix} A_t & -B_{w,t}(K_{l,t}^b)^\top \\ 0 & \tilde{A}_t - B_{w,t}(K_{l,t}^b)^\top \end{bmatrix},$$

$$\hat{B}_{u,t} = \begin{bmatrix} B_{u,t} \\ 0 \end{bmatrix},$$

$$\hat{B}_{w,t} = \begin{bmatrix} B_{w,t}(R_{e,t}^b)^{-\frac{1}{2}} \\ B_{w,t}(R_{e,t}^b)^{-\frac{1}{2}} \end{bmatrix},$$

$$\hat{Q}_t = \begin{bmatrix} Q_t & 0 \\ 0 & 0 \end{bmatrix},$$

$$\hat{H}_t = I + \hat{B}_{u,t}^\top \hat{P}_{t+1} \hat{B}_{u,t},$$

$$\tilde{A}_t = A_t - K_{p,t} Q_t^{\frac{1}{2}},$$

$$K_{p,t} = A_t P_t Q_t^{\frac{1}{2}} R_{e,t}^{-1},$$

$$R_{e,t} = I + Q_t^{\frac{1}{2}} P_t Q_t^{\frac{1}{2}},$$

$$K_{l,t}^b = \tilde{A}_t^\top P_t B_{w,t} (R_{e,t}^b)^{-1},$$

$$R_{e,t}^b = \gamma^2 I + B_{w,t}^\top P_t B_{w,t},$$

$$\delta_{t+1} = \tilde{A}_t \delta_t + B_{w,t} w_t,$$

$$z_t = (R_{e,t}^b)^{\frac{1}{2}} (K_{l,t}^b)^\top \delta_t + (R_{e,t}^b)^{\frac{1}{2}} w_t,$$

and we initialize $\delta_0 = 0$. The state variables ζ_t and $\hat{\nu}_t$ evolve according to the dynamics

$$\begin{bmatrix} \zeta_{t+1} \\ \hat{\nu}_{t+1} \end{bmatrix} = \hat{A}_t \begin{bmatrix} \zeta_t \\ \hat{\nu}_t \end{bmatrix} + \hat{B}_{u,t} u_t + \hat{B}_{w,t} z_t,$$

where we initialize $\zeta_0, \hat{\nu}_0 = 0$. We define P_t to be the solution of the forwards Riccati recursion

$$P_{t+1} = A_t P_t A_t^\top + B_{u,t} B_{u,t}^\top - K_{p,t} R_{e,t} K_{p,t}^\top,$$

where we initialize $P_0 = 0$, and \hat{P}_t, P_t^b to be the solutions of the backwards Ricatti recursions

$$P_{t-1}^b = \tilde{A}_t^\top P_t^b \tilde{A}_t + Q_t^{\frac{1}{2}} (R_{e,t})^{-1} Q_t^{\frac{1}{2}} - K_{l,t}^b R_{e,t}^b (K_{l,t}^b)^\top,$$

$$\hat{P}_t = \hat{Q}_t + \hat{A}_t^\top \hat{P}_{t+1} \hat{A}_t - \hat{A}_t^\top \hat{P}_{t+1} \hat{B}_{u,t} \hat{H}_t^{-1} \hat{B}_{u,t}^\top \hat{P}_{t+1} \hat{A}_t,$$

where we initialize $P_T^b = 0, \hat{P}_T = \hat{Q}_T$.

The regret-optimal controller is the regret-suboptimal controller at performance level γ_{opt}^2 , where γ_{opt}^2 is the smallest value of γ^2 which satisfies $\hat{\Delta}_t \prec 0$ for $t = 0, \dots, T-1$ and we define $\hat{\Delta}_t$ as

$$-\gamma^2 I + \hat{B}_{w,t}^\top \hat{P}_{t+1} \hat{B}_{w,t} - \hat{B}_{w,t}^\top \hat{P}_{t+1} \hat{B}_{u,t} \hat{H}_t^{-1} \hat{B}_{u,t}^\top \hat{P}_{t+1} \hat{B}_{w,t}.$$

Furthermore, the regret incurred by the regret-optimal controller on any specific noise sequence $w = (w_0, \dots, w_{T-1})$ is at most $\gamma_{opt}^2 \sum_{t=1}^T \|w_t\|_2^2$.

The regret-optimal controller described in Theorem 4 may appear mysterious, so we take a brief detour to better understand its structure before presenting the proof of Theorem 4. We introduce the block decomposition

$$\hat{P}_t = \begin{bmatrix} P_{11,t} & P_{12,t} \\ P_{21,t} & P_{22,t} \end{bmatrix},$$

where each of the submatrices has size $n \times n$. With this notation, we see that

$$\hat{H}_t = I + B_{u,t}^\top P_{11,t} B_{u,t}$$

and

$$\begin{aligned} u_t &= -\hat{H}_t^{-1} B_{u,t}^\top [P_{11,t} \quad P_{12,t}] \left(\hat{A}_t \begin{bmatrix} \zeta_t \\ \hat{\nu}_t \end{bmatrix} + \hat{B}_{w,t} z_t \right) \\ &= -\hat{H}_t^{-1} B_{u,t}^\top P_{11,t} (A_t \zeta_t + B_{w,t} (R_{l,t}^b)^{-\frac{1}{2}} z_t) \\ &\quad + \hat{H}_t^{-1} B_{u,t}^\top P_{11,t} B_{w,t} (K_{l,t}^b)^\top \hat{\nu}_t \\ &\quad - \hat{H}_t^{-1} B_{u,t}^\top P_{12,t} (\tilde{A}_t - B_{w,t} (K_{l,t}^b)^\top) \hat{\nu}_t. \end{aligned}$$

After simplifying the backwards Ricatti recursion for \hat{P}_t , we see that $P_{11,t}$ satisfies the recursion

$$P_{11,t} = Q_t + A_t^\top P_{11,t+1} A - A^\top P_{11,t+1} B_t \hat{H}_t^{-1} P_{11,t+1} A_t,$$

where $P_{11,T} = Q$. Notice that this is precisely the recursion that appears in the H_2 controller! We can hence recognize the term $-\hat{H}_t^{-1} B_{u,t}^\top P_{11,t} A_t \zeta_t$ as the optimal H_2 control action described in Theorem 1. It is worth emphasizing this result: *The regret-optimal control action is the sum of the H_2 optimal control action and a term which depends only on current and past disturbances.* This is especially interesting in light of Theorem 3, which shows that the optimal noncausal controller is the H_2 optimal controller, plus an additional term that depends only on current and future disturbances.

We now return to the proof of Theorem 4.

Proof. Consider the state-space model

$$\xi_{t+1} = A_t \xi_t + B_{u,t} u_t, \quad y_t = Q_t^{\frac{1}{2}} \xi_t + v_t,$$

where u_t, v_t are zero mean noise variables such that $\mathbb{E}[u_t u_t^\top] = \mathbb{E}[v_t v_t^\top] = I$ and $\mathbb{E}[u_t v_t^\top] = 0$. Let $y = (y_0, \dots, y_{T-1})$, $u = (u_0, \dots, u_{T-1})$, and $v = (v_0, \dots, v_{T-1})$. Notice that $y = Fu + v$ and $\mathbb{E}[yy^\top] = I + FF^\top$. Suppose we can find a causal matrix Δ such that $y = \Delta e$ where e is a zero-mean random variable such that $\mathbb{E}[ee^\top] = I$. Then $\mathbb{E}[yy^\top] = \Delta \Delta^\top$, so $I + FF^\top = \Delta \Delta^\top$ as desired.

Using the Kalman filter, we obtain a state-space model for Δ given by

$$\hat{\xi}_{t+1} = A_t \hat{\xi}_t + K_{p,t} R_{e,t}^{\frac{1}{2}} e_t, \quad y_t = Q_t^{\frac{1}{2}} \hat{\xi}_t + R_{e,t}^{\frac{1}{2}} e_t,$$

where we define $K_{p,t} = A_t P_t Q_t^{\frac{1}{2}} R_{e,t}^{-1}$ and $R_{e,t} = I + Q_t^{\frac{1}{2}} P_t Q_t^{\frac{1}{2}}$ and P_t is defined recursively as

$$P_{t+1} = A_t P_t A_t^\top + B_{u,t} B_{u,t}^\top - K_{p,t} R_{e,t} K_{p,t}^\top$$

and $P_0 = 0$. A state-space model for Δ^{-1} is hence given by

$$\hat{\xi}_{t+1} = A_t \hat{\xi}_t + K_{p,t} (y_t - Q_t^{\frac{1}{2}} \hat{\xi}_t), \quad e_t = R_{e,t}^{-\frac{1}{2}} (y_t - Q_t^{\frac{1}{2}} \hat{\xi}_t).$$

We have hence factored $I + FF^\top$ as $\Delta \Delta^\top$, so $\gamma^2 I + G^\top (I + FF^\top)^{-1} G = \gamma^2 I + (\Delta^{-1} G)^\top (\Delta^{-1} G)$. Notice that $\Delta^{-1} G$ is strictly causal, since Δ^{-1} is causal and G is strictly causal. A state-space model for G is given by

$$\eta_{t+1} = A_t \eta_t + B_{w,t} w_t, \quad s_t = Q_t^{\frac{1}{2}} \eta_t.$$

Equating s and y , we see that a state-space model for $\Delta^{-1} G$ is given by

$$\begin{bmatrix} \hat{\xi}_{t+1} \\ \eta_{t+1} \end{bmatrix} = \begin{bmatrix} \tilde{A}_t & K_{p,t} Q_t^{\frac{1}{2}} \\ 0 & A_t \end{bmatrix} \begin{bmatrix} \hat{\xi}_t \\ \eta_t \end{bmatrix} + \begin{bmatrix} 0 \\ B_{w,t} \end{bmatrix} w_t,$$

$$e_t = R_{e,t}^{-\frac{1}{2}} Q_t^{\frac{1}{2}} (\eta_t - \hat{\xi}_t).$$

Setting $\nu_t = \eta_t - \hat{\xi}_t$ and simplifying, we see that a minimal representation for a state-space model for $\Delta^{-1} G$ is given by

$$\nu_{t+1} = \tilde{A}_t \nu_t + B_{w,t} w_t, \quad e_t = R_{e,t}^{-\frac{1}{2}} Q_t^{\frac{1}{2}} \nu_t,$$

where we defined $\tilde{A}_t = A_t - K_{p,t} Q_t^{\frac{1}{2}}$. It follows that a minimal representation for a state-space model for $(\Delta^{-1} G)^\top$ is given by

$$\nu_{t-1} = \tilde{A}_t^\top \nu_t + Q_t^{\frac{1}{2}} R_{e,t}^{-\frac{1}{2}} w_t, \quad e_t = B_{w,t}^\top \nu_t.$$

Recall that our original goal was to obtain a factorization $\gamma^2 I + G^\top(I + FF^\top)^{-1}G = L^\top L$, where L is causal and invertible. Define $z = (\Delta^{-1}G)^\top a + b$, where a and b are zero-mean random variables such that $\mathbb{E}[aa^\top] = I, \mathbb{E}[ab^\top] = 0$, and $\mathbb{E}[bb^\top] = \gamma^2 I$. Suppose that we can find a causal matrix L such that $z = L^\top f$, where f is a zero-mean random variable such that $\mathbb{E}[ff^\top] = I$. Notice that $\mathbb{E}[zz^\top] = \gamma^2 I + (\Delta^{-1}G)^\top(\Delta^{-1}G) = \gamma^2 I + G^\top(I + FF^\top)^{-1}G$; on the other hand $\mathbb{E}[zz^\top] = L^\top L$ as desired.

A backwards-time state-space model for z is given by

$$\begin{aligned}\nu_{t-1} &= \tilde{A}_t^\top \nu_t - Q_t^{\frac{1}{2}} R_{e,t}^{-\frac{1}{2}} a_t, \\ z_t &= B_{w,t}^\top \nu_t + b_t.\end{aligned}$$

Using the (backwards time) Kalman filter, we see that a state-space model for L^\top is given by

$$\begin{aligned}\hat{\nu}_{t-1} &= \tilde{A}_t^\top \hat{\nu}_t + K_{l,t}^b (R_{e,t}^b)^{\frac{1}{2}} f_t, \\ z_t &= B_{w,t}^\top \hat{\nu}_t + (R_{e,t}^b)^{\frac{1}{2}} f_t,\end{aligned}$$

where we define $K_{l,t}^b = (A_t - K_{p,t} P_t^b Q_t^{\frac{1}{2}})^\top P_t^b B_{w,t} (R_{e,t}^b)^{-1}$ and $R_{e,t}^b = \gamma^2 I + B_{w,t}^\top P_t^b B_{w,t}$, and P_t^b is the solution to the backwards Ricatti recursion

$$\begin{aligned}P_{t-1}^b &= \tilde{A}_t^\top P_t^b \tilde{A}_t \\ &\quad + Q_t^{\frac{1}{2}} (R_{e,t}^b)^{-1} Q_t^{\frac{1}{2}} - K_{l,t}^b R_{e,t}^b (K_{l,t}^b)^\top,\end{aligned}$$

and $P_T^b = 0$. It follows that a state-space model for L is given by

$$\begin{aligned}\hat{\nu}_{t+1} &= \tilde{A}_t \hat{\nu}_t + B_{w,t} f_t, \\ z_t &= (R_{e,t}^b)^{\frac{1}{2}} (K_{l,t}^b)^\top \hat{\nu}_t + (R_{e,t}^b)^{\frac{1}{2}} f_t.\end{aligned}$$

Therefore a state-space model for L^{-1} is

$$\begin{aligned}\hat{\nu}_{t+1} &= (\tilde{A}_t - B_{w,t} (K_{l,t}^b)^\top) \hat{\nu}_t + B_{w,t} (R_{e,t}^b)^{-\frac{1}{2}} z_t, \\ f_t &= -(K_{l,t}^b)^\top \hat{\nu}_t + (R_{e,t}^b)^{-\frac{1}{2}} z_t.\end{aligned}$$

Recall that a state-space model for G is given by

$$\eta_{t+1} = A_t \eta_t + B_{w,t} w_t, \quad s_t = Q_t^{\frac{1}{2}} \eta_t.$$

Equating f_t and w_t , we see that a state-space model for GL^{-1} is

$$\begin{aligned}\hat{\nu}_{t+1} &= (\tilde{A}_t - B_{w,t} (K_{l,t}^b)^\top) \hat{\nu}_t + B_{w,t} (R_{e,t}^b)^{-\frac{1}{2}} z_t, \\ \eta_{t+1} &= A_t \eta_t - B_{w,t} (K_{l,t}^b)^\top \hat{\nu}_t + B_{w,t} (R_{e,t}^b)^{-\frac{1}{2}} z_t, \\ s_t &= Q_t^{\frac{1}{2}} \eta_t.\end{aligned}$$

A model for F is given by

$$\psi_{t+1} = A_t \psi_t + B_{u,t} u_t, \quad s_t = Q_t^{\frac{1}{2}} \psi_t.$$

Letting $\zeta_t = \eta_t + \psi_t$, we see that a state-space model for the overall system is given by

$$\begin{aligned}\begin{bmatrix} \zeta_{t+1} \\ \hat{\nu}_{t+1} \end{bmatrix} &= \begin{bmatrix} A_t & -B_{w,t} (K_{l,t}^b)^\top \\ 0 & \tilde{A}_t - B_{w,t} (K_{l,t}^b)^\top \end{bmatrix} \begin{bmatrix} \zeta_t \\ \hat{\nu}_t \end{bmatrix} \\ &\quad + \begin{bmatrix} B_{u,t} \\ 0 \end{bmatrix} u_t + \begin{bmatrix} B_{w,t} (R_{e,t}^b)^{-\frac{1}{2}} \\ B_{w,t} (R_{e,t}^b)^{-\frac{1}{2}} \end{bmatrix} z_t, \\ s_t &= \begin{bmatrix} Q_t^{\frac{1}{2}} & 0 \end{bmatrix} \begin{bmatrix} \zeta_t \\ \hat{\nu}_t \end{bmatrix}.\end{aligned}$$

To derive the regret-optimal controller, we can plug this state-space model into the formula for the optimal H_∞ controller given in Theorem 2. We see that the regret-optimal controller is given by

$$u_t = -\hat{H}_t^{-1} \hat{B}_{u,t}^\top \hat{P}_{t+1} \left(\hat{A}_t \begin{bmatrix} \zeta_t \\ \hat{\nu}_t \end{bmatrix} + \hat{B}_{w,t} z_t \right),$$

where we use the notation $\hat{\cdot}$ to represent a variable appearing in the H_∞ controller and define

$$\begin{aligned}\hat{A}_t &= \begin{bmatrix} A_t & -B_{w,t} (K_{l,t}^b)^\top \\ 0 & \tilde{A}_t - B_{w,t} (K_{l,t}^b)^\top \end{bmatrix}, \\ \hat{B}_{u,t} &= \begin{bmatrix} B_{u,t} \\ 0 \end{bmatrix}, \\ \hat{B}_{w,t} &= \begin{bmatrix} B_{w,t} (R_{e,t}^b)^{-\frac{1}{2}} \\ B_{w,t} (R_{e,t}^b)^{-\frac{1}{2}} \end{bmatrix}, \\ \hat{Q}_t &= \begin{bmatrix} Q_t & 0 \\ 0 & 0 \end{bmatrix}, \\ \hat{H}_t &= I + \hat{B}_{u,t}^\top \hat{P}_{t+1} \hat{B}_{u,t},\end{aligned}$$

and \hat{P}_t is the solution of the backwards Ricatti recursion

$$\hat{P}_t = \hat{Q}_t + \hat{A}_t^\top \hat{P}_{t+1} \hat{A}_t - \hat{A}_t^\top \hat{P}_{t+1} \hat{B}_{u,t} \hat{H}_t^{-1} \hat{B}_{u,t}^\top \hat{P}_{t+1} \hat{A}_t.$$

We emphasize that the driving disturbance in this system is not w , but rather $z = Lw$. We have already computed L , so it is easy to see that a state-space model for z is given by

$$\begin{aligned}\delta_{t+1} &= \tilde{A}_t \delta_t + B_{w,t} w_t, \\ z_t &= (R_{e,t}^b)^{\frac{1}{2}} (K_{l,t}^b)^\top \delta_t + (R_{e,t}^b)^{\frac{1}{2}} w_t,\end{aligned}$$

where we initialize $\delta_0 = 0$.

The regret bound stated in Theorem 4 is immediate: by definition the regret-suboptimal controller at performance level γ^2 has regret at most $\gamma^2 \|w\|_2^2$. To find the regret-optimal controller, we minimize γ^2 subject to the constraint that the matrices $\hat{\Delta}_t$ (defined in Theorem 2) satisfy $\hat{\Delta}_t \prec 0$ for $t = 0, \dots, T-1$. \square

5 Numerical experiments

We benchmark the H_2 , H_∞ , and regret-optimal controllers in the context of the inverted pendulum, a classic control system which has been widely studied in works at the intersection of machine learning and control (e.g. [4, 10]).

We refer the reader to Examples 2.2 and 4.4 in [1] for the details on the physics of the inverted pendulum; for our purposes it suffices to know that the continuous-time closed-loop dynamics of the inverted pendulum is given by the nonlinear differential equation

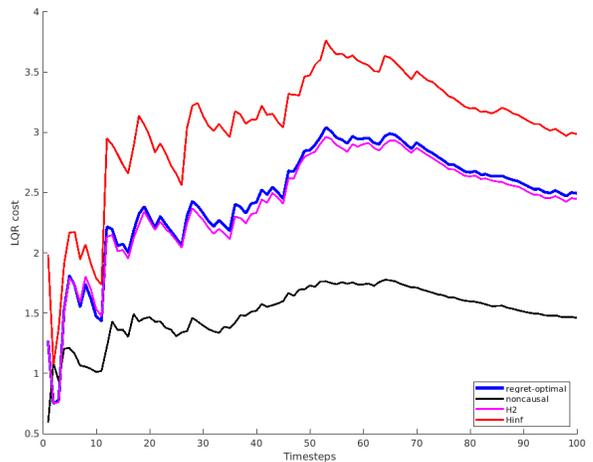
$$\frac{dx}{dt} = \begin{bmatrix} x_2 \\ \sin x_1 - cx_2 + u \cos x_1 \end{bmatrix},$$

where u is a scalar control input, and c is a physical parameter of the system. Using the linearization $\sin x \approx x, \cos x \approx 1$ for small x and adding a disturbance term, we can capture these dynamics by the state-space model

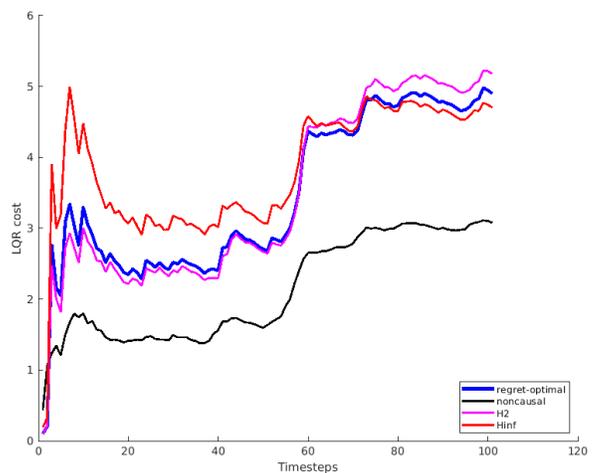
$$x_{t+1} = \begin{bmatrix} 0 & 1 \\ 1 & -c \end{bmatrix} x_t + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_t + w_t.$$

In our experiments we set $c = 0.05$, $Q, R = I$.

We measure the performance of the H_2, H_∞ , and regret-optimal controllers in two settings. In the first setting (Figure 1a), the components of the noise are drawn i.i.d from a standard Gaussian distribution. As expected, the H_2 controller easily outperforms the more conservative H_∞ controller. The regret-optimal controller is still able to capture most of the performance of the H_2 controller, despite not being specifically tuned for i.i.d zero mean noise. In the second setting (Figure 1b), we consider a more challenging scenario where the noise distribution changes over time instead of being fixed over the whole time horizon. In the first phase, the noise is again drawn from a standard Gaussian distribution, favoring the H_2 controller; but in the second phase each of the two components of the noise is drawn from $\mathcal{N}(1, 1)$. This favors the H_∞ controller, which does not assume that the noise will be zero mean. In this phase, the cost incurred by the H_2 controller increases rapidly, eventually surpassing that of the H_∞ controller. The regret-optimal controller consistently incurs the lowest or near-lowest cost over the full time horizon. Our simulations highlight that the regret-optimal controller captures almost all the performance of a fixed controller when the environment is static and can significantly outperform a fixed controller when the environment changes over time.



(a)



(b)

Figure 1: We plot the time-averaged costs of the LQR controllers. Figure 1a shows that when the noise is drawn from $\mathcal{N}(0, 1)$ the regret-optimal controller nearly matches the optimal H_2 controller’s performance. Figure 1b shows how the controllers perform when the process generating the disturbance varies over time. From $t = 1 \dots 50$ each component of the noise is drawn from $\mathcal{N}(0, 1)$, favoring the H_2 controller, but from $t = 51 \dots 100$ each component of the noise is drawn from $\mathcal{N}(1, 1)$, favoring the H_∞ controller. The regret-optimal controller achieves near-lowest cost in both regimes.

References

- [1] Karl Johan AÅstrom and Richard M Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2010.
- [2] Yasin Abbasi-Yadkori and Csaba Szepesvari. “Regret bounds for the adaptive control of linear quadratic systems”. In: *Proceedings of the 24th Annual Conference on Learning Theory*. 2011, pp. 1–26.
- [3] Naman Agarwal et al. “Online control with adversarial disturbances”. In: *arXiv preprint arXiv:1902.08721* (2019).
- [4] Charles W Anderson. “Learning to control an inverted pendulum using neural networks”. In: *IEEE Control Systems Magazine* 9.3 (1989), pp. 31–37.
- [5] Niangjun Chen, Gautam Goel, and Adam Wierman. “Smoothed online convex optimization in high dimensions via online balanced descent”. In: *arXiv preprint arXiv:1803.10366* (2018).
- [6] Alon Cohen, Tomer Koren, and Yishay Mansour. “Learning Linear-Quadratic Regulators Efficiently with only \sqrt{T} Regret”. In: *arXiv preprint arXiv:1902.06223* (2019).
- [7] Sarah Dean et al. “Regret bounds for robust adaptive control of the linear quadratic regulator”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 4188–4197.
- [8] John Doyle et al. “State-space solutions to standard H 2 and H-infinity control problems”. In: *1988 American Control Conference*. IEEE. 1988, pp. 1691–1696.
- [9] John C Doyle. “Guaranteed margins for LQG regulators”. In: *IEEE Transactions on automatic Control* 23.4 (1978), pp. 756–757.
- [10] Rafael Figueroa et al. “Reinforcement learning for balancing a flying inverted pendulum”. In: *Proceeding of the 11th World Congress on Intelligent Control and Automation*. IEEE. 2014, pp. 1787–1793.
- [11] Dylan J Foster and Max Simchowitz. “Logarithmic regret for adversarial online control”. In: *arXiv preprint arXiv:2003.00189* (2020).
- [12] Gautam Goel, Niangjun Chen, and Adam Wierman. “Thinking fast and slow: Optimization decomposition across timescales”. In: *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE. 2017, pp. 1291–1298.
- [13] Gautam Goel and Babak Hassibi. “The Power of Linear Controllers in LQR Control”. In: *arXiv preprint arXiv:2002.02574* (2020).
- [14] Gautam Goel and Adam Wierman. “An online algorithm for smoothed regression and lqr control”. In: *Proceedings of Machine Learning Research* 89 (2019), pp. 2504–2513.
- [15] Gautam Goel et al. “Beyond online balanced descent: An optimal algorithm for smoothed online optimization”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 1875–1885.
- [16] Babak Hassibi, Ali H Sayed, and Thomas Kailath. *Indefinite-quadratic estimation and control: a unified approach to H 2 and H-infinity theories*. SIAM, 1999.
- [17] Elad Hazan. “Introduction to online convex optimization”. In: *arXiv preprint arXiv:1909.05207* (2019).
- [18] Elad Hazan, Sham Kakade, and Karan Singh. “The nonstochastic control problem”. In: *Algorithmic Learning Theory*. PMLR. 2020, pp. 408–421.
- [19] Elad Hazan and Comandur Seshadhri. “Efficient learning algorithms for changing environments”. In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 393–400.
- [20] Mark Herbster and Manfred K Warmuth. “Tracking the best expert”. In: *Machine learning* 32.2 (1998), pp. 151–178.
- [21] Thomas Kailath, Ali H Sayed, and Babak Hassibi. *Linear estimation*. Prentice Hall, 2000.
- [22] Rudolf Emil Kalman et al. “Contributions to the theory of optimal control”. In: *Bol. soc. mat. mexicana* 5.2 (1960), pp. 102–119.
- [23] DJN Limebeer, Michael Green, and D Walker. “Discrete-time H/sup infinity/control”. In: *Proceedings of the 28th IEEE Conference on Decision and Control*, IEEE. 1989, pp. 392–396.
- [24] Krishan M Nagpal and Pramod P Khargonekar. “Filtering and smoothing in an H/sup infinity/setting”. In: *IEEE Transactions on Automatic Control* 36.2 (1991), pp. 152–166.
- [25] Kenko Uchida and Mayasuki Fujita. “Finite horizon H/sup infinity/control problems with terminal penalties”. In: *29th IEEE Conference on Decision and Control*. IEEE. 1990, pp. 1808–1813.