Bayesian Optimal Control of Smoothly Parameterized Systems

Yasin Abbasi-Yadkori
Queensland University of Technology

Csaba Szepesvári
University of Alberta

Abstract

We study Bayesian optimal control of a general class of smoothly parameterized Markov decision problems (MDPs). We propose a *lazy* version of the so-called posterior sampling method, a method that goes back to Thompson and Strens, more recently studied by Osband, Russo and van Roy. While Osband et al. derived a bound on the (Bayesian) regret of this method for undiscounted total cost episodic, finite state and action problems, we consider the continuing, average cost setting with no cardinality restrictions on the state or action spaces. While in the episodic setting, it is natural to switch to a new policy at the episode-ends, in the continuing average cost framework we must introduce switching points explicitly and in a principled fashion, or the regret could grow linearly. Our lazy method introduces these switching points based on monitoring the uncertainty left about the unknown parameter. To develop a suitable and easy-to-compute uncertainty measure, we introduce a new "average local smoothness" condition, which is shown to be satisfied in common examples. Under this, and some additional mild conditions, we derive rate-optimal bounds on the regret of our algorithm. Our general approach allows us to use a single algorithm and a single analysis for a wide range of problems, such as finite MDPs or linear quadratic regulation, both being instances of smoothly parameterized MDPs. The effectiveness of our method is illustrated by means of a simulated example.

1 INTRODUCTION

The topic of this paper is Bayesian optimal control, where the problem is to design a policy that achieves optimal performance on the average over control problem instances that are randomly sampled from a given distribution. This problem naturally arises when the goal is to design a controller for mass-produced systems, where production is imperfect but the errors follow a regular pattern and the goal is to maintain a good average performance over the controlled systems, rather than to achieve good performance even for the system with the largest errors.

In a Bayesian setting, the optimal policy (which exists under appropriate regularity conditions) is history dependent. Given the knowledge of the prior, the transition dynamics and costs, the problem in a Bayesian setting is to find an efficient way to *calculate* the actions that the optimal policy would take given some history. This problem was studied for finite state and action spaces by Asmuth et al. (2009) and Kolter and Ng (2009). Both works propose specific computationally efficient algorithms, which are shown to be ϵ-Bayes-optimal with probability $1 - \delta$ with the exception of $O(\text{poly}(1/\epsilon))$ many steps, where for both algorithms ϵ and δ are both part of the input. While Kolter and Ng (2009) suggest to add an exploration bonus to the rewards while using the mean estimates for the transition probabilities and considers a finite horizon setting, Asmuth et al. (2009) consider discounted total rewards and a variant of posterior sampling, originally due to Thompson (1933) and first adapted to reinforcement learning by Strens (2000). More recently, the algorithm of Strens (2000) was revisited by Osband et al. (2013) in the context of episodic, finite MDPs. An attractive feature of posterior sampling is that it requires neither the target accuracy ϵ, nor the failure probability δ as its inputs. Rather, the guarantee presented by Osband et al. (2013) is that the algorithm's (Bayesian) regret, i.e., the excess cost due to not following the optimal policy, is bounded by $\widetilde{O}(\sqrt{T})$[1] both with high probability and in expectation. The reader interested in further algorithms for Bayesian reinforcement learning (including algorithms for infinite state spaces) may consult the papers of Araya-López et al. (2012), Vlassis et al. (2012) and Guez et al. (2013), which together give an excellent overview of the literature.

[1] $\widetilde{O}(\cdot)$ hides poly-logarithmic factors.

The starting point of our paper is the work of Osband et al. (2013). In particular, just like Osband et al. (2013), we build on the posterior sampling algorithm of Strens (2000), which itself was derived from an algorithm of Thompson (1933) developed for the so-called bandit setting. Unlike Osband et al. (2013) and Strens (2000), we allow the state-action space to be infinite (subject to some regularity conditions discussed later) and we consider the *infinite horizon, continuing, average-cost setting*. As far as we known, ours is the first work deriving (Bayesian) regret bounds for any algorithms of this generality. The major assumption that we make is that *the Markov dynamics is smoothly parameterized in some unknown parameters* with known (local) "smoothness" map such that the posterior concentrates in the metric derived from this map. It is shown that this assumption is met in some common examples, such as finite MDPs, and also in linearly parameterized systems, which encompass, systems with linear dynamics.

Following a proposal of Strens (2000) who also considered the non-episodic setting, the algorithm works in phases: At the beginning of each phase, a policy is computed based on solving the optimal control problem for a random parameter vector drawn from the posterior over the parameter vectors. The algorithm keeps the policy until the parameter uncertainty is reduced by a substantial margin, when a new phase begins and the process is repeated. The idea of ending a phase when uncertainty is reduced by a significant margin goes back at least to the work of Jaksch et al. (2010).

While in the case of episodic problems the issue of how long a policy should be kept does not arise, in a continuing problem with no episodic structure, if policies are changed too often, performance will suffer (see, e.g., Example 1 of Guez et al. (2014)). To address this challenge, for non-episodic problems, Strens (2000) suggested that the lengths of phases should be adjusted to the "planning horizon" (Strens, 2000), which however, is ill-defined for the average cost setting that we consider in this paper. A major contribution of this work is that we show how the smoothness map can be used to derive the length of the phases.

The continuing setting is very common in practice; this setting is the most natural for controlled mechanical systems (e.g., CD/DVD drive control, control of manufacturing robots), or for process optimization (e.g., controlling a queuing system, resource management), where "resets" are rare or unnatural.

Under some additional technical conditions, we show that the expected (Bayesian) regret of our algorithm is $\tilde{O}(\sqrt{T} + \Sigma_T)$, where T is the number of time steps and Σ_T is controlled by the precision with which the optimal control problems are solved, thus providing an explicit bound on the cost of using imprecise calculations. In summary, the main result of the paper shows that near-optimal Bayesian optimal control is possible for a wide range of problems as long as we can efficiently sample from the parameter posteriors, the length of phases for how long the same policy is followed is carefully controlled and if we can efficiently solve the arising classical optimal control problems. Due to the lack of space, the proofs of some of our claims are given in the supplementary material.

We emphasize two contributions: (1) the invention of a class of systems which unifies many previous approaches, and permits an elegant proof. (2) the introduction of a Concentrating Posterior assumption which significantly shortens our proof compared to previous proofs and improves the bound, as we avoid the use of measure concentration arguments which were always used previously.

2 PROBLEM SETTING

We consider problems when the transition dynamics is parameterized with a matrix $\Theta_* \in \mathbb{R}^{m \times n}$, which is *randomly chosen* at time 0 (before the interaction with the learner starts) from a known prior P_0 with support $\mathcal{S} \subset \mathbb{R}^{m \times n}$. Let P_t denote the posterior of Θ_* at time t based on $x_1, a_1, \dots, a_{t-1}, x_t$. Let $\mathcal{X} \subset \mathbb{R}^n$ be the state space and $\mathcal{A} \subset \mathbb{R}^d$ be the action space, $x_t \in \mathcal{X}$ be the state at time t and $a_t \in \mathcal{A}$ be the action at time t, which is chosen based on $x_1, a_t, \dots, a_{t-1}, x_t$. It is assumed that x_1 is sampled from a fixed distribution (although, it should become clear later that this assumption is not necessary). For $M \succeq 0$ positive semidefinite, define $\|\Theta\|_M^2 = \left\|\Theta^\top M \Theta\right\|_2$, where $\|\cdot\|_2$ denotes the spectral norm of matrices (later we will drop the subindex 2). The set of positive semidefinite $m \times m$ matrices will be denoted by $\mathbb{S}^+(m)$. Our main assumption concerning the transition law is as follows:

Assumption A1 ***(Smoothly Parameterized Dynamics)*** The next state satisfies $x_{t+1} = f(x_t, a_t, \Theta_*, z_{t+1})$, where $z_{t+1} \sim U[0,1]$ is independent of the past and Θ_*. Further, there exists a (known) map $M : \mathcal{X} \times \mathcal{A} \to \mathbb{S}^+(m)$ such that for any $\Theta, \Theta' \in \mathcal{S}$, if $y = f(x, a, \Theta, z)$, $y' = f(x, a, \Theta', z)$ with $z \sim U[0,1]$, then $\mathbb{E}\left[\|y - y'\|\right] \le \|\Theta - \Theta'\|_{M(x,a)}$.

The first part of the assumption just states that given Θ_*, the dynamics is Markovian with state x_t, while the second part demands that small changes in the parameter lead to small changes in the next state. The assumption that the map M is "known" makes it possible to use M in the design of our algorithms.

Our next assumption connects the concentration of the posterior with M:

Assumption A2 ***(Concentrating Posterior)*** Let $\tilde{\mathcal{F}}_t = \sigma(x_1, a_1, \dots, a_{t-1}, x_t)$ be the σ-algebra generated by observations up to time t, $V_t = V + \sum_{s=1}^{t-1} M(x_s, a_s)$, where V is an $m \times m$ positive definite matrix. Then, there exists a positive constant C such that for any $t \ge 1$, for some $\tilde{\mathcal{F}}_t$-

measurable random variable $\widehat{\Theta}_t$, letting $\Theta'_t \sim P_t$ it holds that

$$\max\left\{\mathbb{E}\left[\|\Theta'_t - \widehat{\Theta}_t\|^2_{V_t}\right]\mathbb{E}\left[\|\Theta_* - \widehat{\Theta}_t\|^2_{V_t}\right]\right\} \le C\ .$$

The idea here is that $\widehat{\Theta}_t$ is an estimate of Θ_* based on past information available at time t, such as a maximum aposteriori (MAP) estimate (note that this estimate will *not* be needed by our algorithm). Since V_t is increasing at a linear rate, the assumption requires that $\widehat{\Theta}_t$ converges to Θ at an $O(1/\sqrt{t})$ rate. When $\Theta = \Theta_*$, this means that $\widehat{\Theta}_t$ should converge to Θ_* at this rate, which is indeed what we expect. When $\Theta = \Theta'_t$, again, we expect this to be true since Θ'_t is expected to be in the $O(1/\sqrt{t})$ vicinity of Θ_*. Note how this assumption connects M with the behavior of the posterior. One novelty of our analysis, as compared to that of Osband et al. (2013), is that while Osband et al. relies on measure-concentration, we require only the above (weaker) "variance concentration". We will show explicit examples where this variance term is easy to control using a direct calculation. Since we avoid measure-concentration, our analysis has the potential to give much tighter regret bounds for the Bayesian setting than available previously, though the study of this remains for future work. The examples we deal with include finite MDPs (where the state is represented by unit vectors) and systems with linear dynamics (i.e., when $x_{t+1} = Ax_t + Ba_t + w_{t+1}$, where $w_{t+1} \sim p_w(\cdot|x_t, a_t)$), amongst others. Explicit expressions for the map M will be given in Section 6 for these systems. In general, for systems with additive noise, finding M essentially reduces to finding a suitable local linearization of the system's dynamics.

The problem we study is to design a controller (also known as a policy) that at every time step t, based on past states $x_1, \dots, x_t$ and actions $a_1, \dots, a_{t-1}$, selects an action a_t so as to minimize the expected long-run average loss $\mathbb{E}\left[\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n \ell(x_t, a_t)\right]$. We consider any noise distribution and any loss function ℓ as long as a boundedness assumption on the variance and a smoothness assumption on the *value function* are satisfied (see Assumptions A2 and A3-ii below). It is important to note that we allow ℓ to be a nonlinear function of the last state-action pair, i.e., the framework allows one to go significantly beyond the scope of linear quadratic control as many nonlinear control problems can be transformed into a linear form (but with a nonlinear loss function) using the so-called dynamic feedback linearization techniques (Isidori, 1995).

To measure the performance of an algorithm, we use the (expected) regret R_T:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(\ell(x_t, a_t) - J(\Theta_*))\right]\ .$$

Here, $(x_t, a_t)_{t=1}^T$ denotes the state-action trajectory and $J(\Theta_*)$ is the average loss of the optimal policy given (random) parameter Θ_*. The slower the regret grows, the closer is the performance to that of an optimal policy. If the growth rate of R_T is sublinear ($R_T = o(T)$), the average loss per time step will converge to the optimal average loss as T gets large and in this sense we can say that the algorithm is asymptotically-optimal. Our main result shows that, under some conditions, the construction of such asymptotically-optimal policies can be reduced to the ability of efficiently sampling from the posterior of Θ_* and being able to solve classical (non-Bayesian) optimal-control problems. Furthermore, our main result also implies that $R_T = \widetilde{O}(\sqrt{T})$.

3 THE LAZY PSRL ALGORITHM

Our algorithm is an instance of the posterior sampling reinforcement learning (PSRL) (Osband et al., 2013). As explained beforehand, this algorithm is based on the work on Thompson (1933) and was proposed by Strens (2000). To emphasize that the algorithm keeps the current policy for a while, we call it LAZY PSRL. Our contribution is to suggest a specific schedule for updating the policy. The pseudocode of the algorithm is shown in Figure 1.

Recall that P_0 denotes the prior distribution of the parameter matrix Θ_*. Let P_t denote the posterior of Θ_* at time t based on $x_1, a_1, \dots, a_{t-1}, x_t$ and $\tau_t < t$ the last round when the algorithm chose a new policy. Further, let $V_t = V + \sum_{s=1}^{t-1} M(x_s, a_s)$, where V is some fixed, $m \times m$ positive definite matrix. Let G be a constant that controls the replanning frequency. Then, at time t, Lazy PSRL sets $\widetilde{\Theta}_t = \widetilde{\Theta}_{t-1}$ unless $\det(V_t) > G\det(V_{\tau_t})$ in which case it chooses $\widetilde{\Theta}_t$ from the posterior P_t: $\widetilde{\Theta}_t \sim P_t$. The action taken at time step t is a near-optimal action for the system whose transition dynamics is specified by $\widetilde{\Theta}_t$. We assume that a subroutine, π^*, taking the current state x_t and the parameter $\widetilde{\Theta}_t$ is available to calculate such an action. The inexact nature of calculating a near-optimal action will also be taken in our analysis.

4 RESULTS FOR BOUNDED STATE- AND FEATURE-SPACES

In this section, we study problems with a bounded state space. In particular, the number of states might be infinite, but we assume that the norm of the state vector is bounded by a constant. Before stating our main result, we state some extra assumptions.

Our first extra assumption concerns the existence of "regular" solutions to the average cost optimality equations (ACOEs), an assumption which is usually thought to be mild in the context of average-cost problems:

Assumption A3 ***(Existence of Regular ACOE Solutions)*** The following hold:

```
Inputs: $P_0$, the prior distribution of $\Theta_*$, $V$, $G$.
$V_{\text{last}} \leftarrow V$, $V_0 \leftarrow V$.
for $t \leftarrow 1, 2, \dots$ do
    if $\det(V_t) > G \det(V_{\text{last}})$ then
        Sample $\widetilde{\Theta}_t \sim P_t$.
        $V_{\text{last}} \leftarrow V_t$.
    else
        $\widetilde{\Theta}_t \leftarrow \widetilde{\Theta}_{t-1}$.
    end if
    Calculate near-optimal action $a_t \leftarrow \pi^*(x_t, \widetilde{\Theta}_t)$.
    Execute action $a_t$ and observe the new state $x_{t+1}$.
    Update $P_t$ with $(x_t, a_t, x_{t+1})$ to obtain $P_{t+1}$.
    Update $V_{t+1} \leftarrow V_t + M(x_t, a_t)$.
end for
```

Figure 1: Lazy PSRL for smoothly parameterized control problems

(i) There exists $H > 0$ such that for any $\Theta \in \mathcal{S}$, there exist a scalar $J(\Theta)$ and a function $h(\cdot, \Theta) : \mathcal{X} \to [0, H]$ that satisfy the average cost optimality equation (ACOE): for any $x \in \mathcal{X}$,

$$J(\Theta) + h(x, \Theta) = \min_{a \in \mathcal{A}} \left\{ \ell(x, a) + \int h(y, \Theta) p(dy \,|\, x, a, \Theta) \right\} , \tag{1}$$

where $p(\cdot|x, a, \Theta)$ is the next-state distribution given state x, action a and parameter Θ.

(ii) There exists $B > 0$ such that for all $\Theta \in \mathcal{S}$, and for all $x, x' \in \mathcal{X}$, $|h(x, \Theta) - h(x', \Theta)| \leq B \|x - x'\|$.

With a slight abuse of the concepts, we will call the quantity $J(\Theta)$ the average loss of the optimal policy, while function $h(\cdot, \Theta)$ will be called the value function (for the system with parameter Θ). The review paper by Arapostathis et al. (1993) gives a number of sufficient (and sometimes necessary) conditions that guarantee that a solution to ACOE exists. Lipschitz continuity usually follows from that of the transition dynamics and the losses.

Let us now discuss the condition that h should have a bounded range. A uniform lower bound on h follows, for example if the immediate cost function ℓ is lower bounded. Then, if the state space is bounded, uniform boundedness of the functions $h(\cdot, \Theta)$ follows from their uniform Lipschitzness:

Proposition 1. *Assume that the value function $h(\cdot, \Theta)$ is bounded from below ($\inf_x h(x, \Theta) > -\infty$) and is B-Lipschitz. Then, if the diameter of the state space is bounded by X (i.e., $\sup_{x, x' \in \mathcal{X}} \|x - x'\| \leq X$) then there exists a solution $h'(\cdot, \Theta)$ to* (1) *such that the range of h is included in* $[0, BX]$.

Finally, we assume that the map $M : \mathcal{X} \times \mathcal{A} \to \mathbb{S}^+(m)$ is bounded:

Assumption A4 ***(Boundedness)*** There exist $\Phi > 0$ such that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $\operatorname{trace}(M(x, a)) \leq \Phi^2$.

This assumption may be strong. In the next section we discuss an extension of the result of this section to the case when this assumption is not met.

The main theorem of this section bounds the regret of Lazy PSRL under the assumptions mentioned so far. In this result, we allow π^* to return a σ_t-suboptimal action, where $\sigma_t > 0$. By this, we mean that the action a_t satisfies

$$\ell(x_t, a_t) + \int h(y, \widetilde{\Theta}_t) p(dy|x_t, a_t, \widetilde{\Theta}_t) \leq \min_{a \in \mathcal{A}} \left\{ \ell(x_t, a) + \int h(y, \widetilde{\Theta}_t) p(dy|x_t, a, \widetilde{\Theta}_t) \right\} + \sigma_t \,. \tag{2}$$

One can control the suboptimality error in terms of the error of an approximate solution to the Bellman equation and the error of the subroutine that finds an action that minimizes the obtained approximate action values.

Theorem 2. *Assume that A1–A4 hold for some values of $C, B, X, \Phi > 0$. Consider Lazy PSRL where in time step t, the action chosen is σ_t-suboptimal. Then, for any time T, the regret of Lazy PSRL satisfies*

$$R_T = \widetilde{O}\left(\sqrt{T}\right) + \Sigma_T,$$

where $\Sigma_T = \sum_{t=1}^T \mathbb{E}[\sigma_t]$ and the constant hidden by $\widetilde{O}(\cdot)$ depends on V, C, B, X, G and Φ.

In particular, the theorem implies that Lazy PSRL is asymptotically optimal as long as $\sum_{t=1}^T \mathbb{E}[\sigma_t] = o(T)$ and it is $O(\epsilon)$-optimal if $\mathbb{E}[\sigma_t] \leq \epsilon$. The fact that the regret is bounded by the sum of suboptimality factors in solving Bellman equation is not trivial. Indeed, as actions have long term effects and we have a closed-loop system, one might suspect that the regret could blow up as a function of these errors. In this respect, the significance of our theorem is that the learner need not worry too much about each planning subproblem as the overall effect is only additive.

Due to lack of space, the proof, which combines the proof techniques of Osband et al. (2013) with that of Abbasi-Yadkori and Szepesvári (2011) in a novel fashion, is presented in the appendix.

5 FORCEFULLY STABILIZED SYSTEMS

For some applications, such as robotics, where the state can grow unbounded, the boundedness assumption (Assumption A4) is rather problematic. For such systems, it is common to use a stabilizing controller π_{stab} that is automatically turned on and is kept on as long as the state vector is "large". The stabilizing controller, however, is usually expensive (uses lots of energy), as it is designed to be robust so that it is guaranteed to drive back the state to the safe region for all possible systems under consideration. Hence a good controller should avoid relying on the stabilizing controller.

In this section, we will replace Assumption A4 with an assumption that a stabilizing controller is available. We will use this controller to override the actions coming from our algorithm as soon as the state leaves the (bounded) safe region $\mathcal{R} \subset \mathbb{R}^n$ until it returns to it. The corresponding pseudocodeis shown in Figure 2.

```
Inputs: $P_0$, the prior distribution of $\Theta_*$, $V$, the safe region $\mathcal{R} \subset \mathbb{R}^n$.
Initialize Lazy PSRL with $P_0$ and $V$, $x_1$.
for $t = 1, 2, \dots$ do
  if $x_t \in \mathcal{R}$ then
    Get action $a_t$ from Lazy PSRL
  else
    Get action $a_t$ from $\pi_{\mathrm{stab}}$
  end if
  Execute action $a_t$ and observe the new state $x_{t+1}$.
  Feed $a_t$ and $x_{t+1}$ to Lazy PSRL.
end for
```

Figure 2: Stabilized Lazy PSRL

We assume that the stabilizing controller is effective in the following sense:

Assumption A5 ***(Effective Stabilizing Controller)*** There exists $\Phi > 0$ such that the following holds: Pick any $x \in \mathcal{R}$, $a \in \mathcal{A}$ and let $x'_1, a'_1, x'_2, a'_2, \dots$ be the sequence of state-action pairs obtained when from time step two the Markovian stabilizing controller π_{stab} is applied to the controlled system whose dynamics is given by $\Theta \in \mathcal{S}$: $x'_1 = x$, $a'_1 = a$, $x'_{t+1} \sim p(\cdot|x'_t, a'_t, \Theta)$, $a'_{t+1} \sim \pi_{\mathrm{stab}}(\cdot|x'_t)$. Then, $\mathbb{E}\left[\mathrm{trace}(M(x'_t, a'_t))\right] \leq \Phi^2$ for any $t \geq 1$, where $M : \mathcal{X} \times \mathcal{A} \to \mathbb{S}^+(m)$ is the map of Assumption A1 underlying $\{p(\cdot|x, a, \Theta)\}$.

The assumption is reasonable as it only requires that the trace of $M(x'_t, a'_t)$ is bounded *in expectation*. Thus, large spikes, that no controller may prevent, can exist as long as they happen with a sufficiently low probability.

The next theorem shows that Stabilized Lazy PSRL is near Bayes-optimal for the system p' obtained from p by overwriting the action a by the action $\pi_{\mathrm{stab}}(x)$ if x is outside of the safe region $\mathcal{R} \subset \mathbb{R}^n$:

$$p'(dy|x, a, \Theta) = \begin{cases} p(dy|x, a, \Theta), & \text{if } x \in \mathcal{R}; \\ p(dy|x, \pi_{\mathrm{stab}}(x), \Theta), & \text{otherwise}. \end{cases}$$

Theorem 3. *Consider a parameterized system with the transition probability kernel family $\{p(\cdot|x, a, \Theta)\}_{\Theta \in \mathcal{S}}$ and let $\pi_{\mathrm{stab}} : \mathcal{X} \to \mathcal{A}$ be a deterministic Markovian controller. Let the smooth parameterization Assumption A1 hold for $\{p(\cdot|x, a, \Theta)\}$, the ACOE solution regularity Assumption A3 hold for $\{p'(\cdot|x, a, \Theta)\}$. Consider running the Stabilized Lazy PSRL algorithm of Figure 2 on $p(\cdot|x, a, \Theta_*)$ and let the concentration Assumption A2 hold along the trajectory obtained. Then, if in addition Assumption A5 holds then the regret of Stabilized Lazy PSRL against the Bayesian optimal controller of $\{p'(\cdot|x, a, \Theta)\}_\Theta$ with prior P_0 and immediate cost ℓ satisfies $R_T = \widetilde{O}\left(\sqrt{T}\right) + \Sigma_T$, where $\Sigma_T = \sum_{t=1}^T \mathbb{E}\left[\mathbf{1}\left\{x_t \in \mathcal{R}\right\} \sigma_t\right]$ and σ_t is the suboptimality of the action computed by Lazy PSRL at time step t.*

If the optimal controller π^* for p does not excite the condition that turns on the stabilizing controller, then this controller is also optimal for p'. In this case, Stabilized Lazy PSRL will have the same regret against π^* than what it has against the optimal controller of p' and the theorem implies that it will achieve sublinear regret in the original system, as long as Σ_T is sublinear.

6 EXAMPLES

The purpose of this section is to illustrate the results obtained. In particular, we will consider applying the results to finite MDPs and linearly parameterized controlled systems and show that for these cases all the assumptions can be satisfied and Lazy PSRL can achieve a low expected regret. We believe that our results will be applicable to many more settings, such as hybrid discrete-continuous systems where the discrete states control which continuous dynamics is used.

6.1 Finite MDPs

Consider an MDP problem with finite state and action spaces. Let the state space be $\mathcal{X} = \{1, 2, \dots, n\}$ and the action space be $\mathcal{A} = \{1, 2, \dots, d\}$. We represent the state variable by an n-dimensional binary vector x_t that has only one non-zero element at the current state and will write the dynamics in the form $x_{t+1} = \Theta_* \varphi(x_t, a_t) + \eta_t$, where Θ_* will collect the transition matrices into a single big matrix and η_t is a "Markov noise". The feature map, $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{nd}$ and the parameter matrix are defined as follows: for $1 \leq k \leq nd$,

$$\varphi_k(x, a) = \begin{cases} 1, & \text{if } k = (a-1)n + x\,; \\ 0, & \text{otherwise}\,, \end{cases} \qquad \Theta_* = \begin{pmatrix} \Theta_*^{(1)} \\ \Theta_*^{(2)} \\ \vdots \\ \Theta_*^{(d)} \end{pmatrix}.$$

Let $s \in [n]$ be a state and $a \in [d]$ be an action. The sth row of matrix $\Theta_*^{(a)}$ is a distribution over the state space that shows the transition probabilities when we take action a in state s. Thus, any row of $\Theta_*^{(a)}$ sums to one and $\mathbb{E}\left[x_{t+1}|x_t, a_t\right] = \Theta_*^\top \varphi(x_t, a_t)$.

An appropriate prior for each row is a Dirichlet distribution. Let $\alpha_1, \dots, \alpha_n$ be positive numbers and let $V' = \mathrm{diag}(\alpha_1, \dots, \alpha_n)$. Then $V = \mathrm{diag}(V', \dots, V') \in \mathbb{R}^{nd \times nd}$ is our "smoother". Let the prior for the sth row of $\Theta_*^{(a)}$ be the Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_n)$: $(P_0)_{s,:} = D(\alpha_1, \dots, \alpha_n)$. At time t, the posterior has the form

$$(P_t)_{s,:} = D(\alpha_1 + c_t(s,a,1), \dots, \alpha_n + c_t(s,a,n)),$$

where $c_t(s,a,s')$ is the number of observed transitions to state s' after taking action a in state s during the first t time steps. Matrix V_t is a diagonal matrix with diagonal elements depending only on the number of times a state-action pair is observed. In particular,

$$(V_t)_{n(a-1)+s,n(a-1)+s} = \sum_{s'} (\alpha_{s'} + c_t(s,a,s')).$$

Vector $\widehat{\Theta}_{t,(:,s')}$ is an nd-dimensional vector and its elements show the empirical frequency of transition to state s' from different state-action pairs. The mean of distribution $(P_t)_{s,:}$ is the vector $\widehat{\Theta}_{t,(n(a-1)+s,:)}$ where

$$\widehat{\Theta}_{t,(n(a-1)+s,s')} = \frac{\alpha_{s'} + c_t(s,a,s')}{\sum_{s''}(\alpha_{s''} + c_t(s,a,s''))}\,.$$

We now show that matrix-valued map M can be chosen to be $M(x,a) = (\sqrt{2}/2)\mathbb{I}$:

Proposition 4. *The above choice makes Assumptions A1 and A2 satisfied.*

Proof. Let us first show that Assumption A1 holds. Because $\mathbb{E}\left[y|x,a\right] = \Theta^\top \varphi(x,a)$, $\mathbb{E}\left[y'|x,a\right] = \Theta'^\top \varphi(x,a)$, and y and y' have only one non-zero element,

$$\begin{aligned}\mathbb{E}\left[\|y - y'\|\right] &= \sqrt{2}\mathbb{P}\left(y \neq y'\right) = \sqrt{2}\left(1 - \mathbb{P}\left(y = y'\right)\right)\\ &= \sqrt{2}\left(1 - \Theta_{(x,a),:}^\top \Theta'_{(x,a),:}\right)\\ &= \frac{\sqrt{2}}{2}\left\|\Theta_{(x,a),:} - \Theta'_{(x,a),:}\right\|^2\,,\end{aligned}$$

where the last step holds because each row of Θ and Θ' sum to one.

Let us now prove that Assumption A2 holds: Let $N = (\Theta_* - \widehat{\Theta}_t)^\top$, $\alpha_{s,a,s'} = \alpha_{s'} + c_t(s,a,s')$ and $\overline{\alpha}_{s,a} = \sum_{s'} \alpha_{s,a,s'} = V_{t,(n(a-1)+s,n(a-1)+s)}$. Let $\|.\|_F$ denote the Frobenius norm. We have that

$$\begin{aligned}\mathbb{E}\left[\left\|NV_t^{1/2}\right\|^2 \,\middle|\, \mathcal{F}_t\right] &\leq \mathbb{E}\left[\left\|NV_t^{1/2}\right\|_F^2 \,\middle|\, \mathcal{F}_t\right]\\ &= \mathbb{E}\left[\sum_{s,a} V_{t,(n(a-1)+s,n(a-1)+s)} \sum_{s'} N^2_{s',n(a-1)+s} \,\middle|\, \mathcal{F}_t\right]\\ &= \sum_{s,a} \overline{\alpha}_{s,a} \sum_{s'} \mathbb{E}\left[N^2_{s',n(a-1)+s} \,\middle|\, \mathcal{F}_t\right]\,.\end{aligned}$$

Because each row of Θ_* has a Dirichlet distribution and rows of $\widehat{\Theta}_t$ are means of these distributions, $\mathbb{E}\left[N^2_{s',n(a-1)+s} \,\middle|\, \mathcal{F}_t\right]$ is simply the variance of the corresponding Dirichlet variable. Thus,

$$\begin{aligned}\mathbb{E}\left[\left\|NV_t^{1/2}\right\|^2 \,\middle|\, \mathcal{F}_t\right] &\leq \sum_{s,a}\sum_{s'} \frac{\overline{\alpha}_{s,a}\alpha_{s,a,s'}(\overline{\alpha}_{s,a} - \alpha_{s,a,s'})}{\overline{\alpha}_{s,a}^2(1 + \overline{\alpha}_{s,a})}\\ &\leq n^2 d\,.\end{aligned}$$

□

An immediate corollary of this is that Lazy PSRL will enjoy low regret in finite MDPs:

Corollary 5. *Consider Lazy PSRL applied to a finite MDP with n states, d actions with M as above, and a Dirichlet prior as specified above. Assume that the set $\mathcal{S}$ system parameters under which Assumption A3 is satisfied is a measurable set with positive Lebesgue measure. Suppose that at time step t, the action chosen is σ_t-suboptimal. Then, for any time T, the regret of Lazy PSRL satisfies $R_T = \widetilde{O}\left(\sqrt{T}\right) + \Sigma_T$.*

Proof. The boundedness condition (Assumption A4) trivially holds, Assumption A3 holds by assumption, while Proposition 4 shows that the remaining two assumptions of Theorem 2 are satisfied. □

6.2 Linearly Parametrized Problems with Gaussian Noise

Next, we consider linearly parametrized problems with Gaussian noise:

$$x_{t+1} = \Theta_*^\top \varphi(x_t, a_t) + w_{t+1}\,, \qquad (3)$$

where w_{t+1} is a zero-mean normal random variable. The nonlinear dynamics shown in (3) shares similarities to, but allows significantly greater generality than the Linear Quadratic (LQ) problem considered by Abbasi-Yadkori and Szepesvári (2011). In particular, in the LQ problem, $\Theta_*^\top = \left(A_*, B_*\right)$ and $\varphi(x_t, a_t)^\top = \left(x_t^\top, a_t^\top\right)$. (However, Abbasi-Yadkori and Szepesvári (2011) assume only that the noise is subgaussian.)

Next, we describe a conjugate prior under the assumption that the noise is Gaussian with a known covariance matrix. Without loss of generality, we assume that

$\mathbb{E}\left[w_{t+1}w_{t+1}^\top \mid \mathcal{F}_t\right] = I$. A conjugate prior is appealing as the posterior has a compact representation that allows for computationally efficient sampling methods. Assume that the columns of matrix Θ_* are independently sampled from the following prior: for $i = 1 \dots n$,

$$P_0\left(\Theta_{*,(:,i)}\right) \propto \exp\left(\Theta_{*,(:,i)}^\top V \Theta_{*,(:,i)}\right) \mathbf{1}\left\{\Theta_{*,(:,i)} \in \mathcal{S}\right\}$$

and $\mathcal{S}$ is the set of system parameters under which Assumption A3 is satisfied, which is assumed to be a measurable set with positive Lebesgue measure. Then, by Bayes' rule, the posterior for column i of Θ_*, $P_t\left(\Theta_{*,(:,i)}\right)$, is proportional to

$$e^{\left(-0.5\left(\Theta_{*,(:,i)}-\widehat{\Theta}_{t,(:,i)}\right)^\top V_t\left(\Theta_{*,(:,i)}-\widehat{\Theta}_{t,(:,i)}\right)\right)} \mathbf{1}\left\{\Theta_{*,(:,i)} \in \mathcal{S}\right\}.$$

We now show an appropriate choice for M (which should not be surprising):

Proposition 6. *With the choice $M(x,a) = \varphi(x,a)\varphi(x,a)^\top$, Assumptions A1 and A2 are satisfied.*

Note that this choice is essentially the same as in Proposition 4.

Proof. Let us first show that Assumption A1 holds. Because $y = \Theta^\top \varphi(x,a) + w$, $y' = \Theta'^\top \varphi(x,a) + w$, we have

$$\left\|y - y'\right\|^2 = \left\|\Theta - \Theta'\right\|^2_{\varphi(x,a)\varphi(x,a)^\top} \ ,$$

which shows that this assumption is indeed satisfied with the said choice of M.

Let us now prove that Assumption A2 holds: Let Λ be a random variable with probability distribution function

$$P(\lambda) \propto \exp\left(-\frac{1}{2}\left(\lambda - \widehat{\Theta}_{t,(:,i)}\right)^\top V_t \left(\lambda - \widehat{\Theta}_{t,(:,i)}\right)\right) \ .$$

Notice that $\left(\Lambda - \widehat{\Theta}_{t,(:,i)}\right)^\top V_t^{1/2} = Z \sim \mathcal{N}(0,I)$ has the standard normal distribution. Hence $\mathbb{P}\left(|Z_j| > \alpha\right) \leq e^{-\alpha^2/2}$. Thus, since $\mathbb{P}\left(\|Z\| > \alpha\right) \leq m e^{-\alpha^2/(2m^2)}$, we have

$$\mathbb{E}\left[\left\|\left(\Theta_{*,(:,i)} - \widehat{\Theta}_{t,(:,i)}\right)^\top V_t^{1/2}\right\|^2 \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}\left[\left\|Z\right\|^2 \,\middle|\, \mathcal{F}_t\right]$$
$$= \int_0^\infty \mathbb{P}\left(\left\|Z\right\|^2 > \epsilon\right) \leq 2m^3 \ .$$

Thus,

$$\mathbb{E}\left[\left\|(\Theta_* - \widehat{\Theta}_t)^\top V_t^{1/2}\right\|^2 \,\middle|\, \mathcal{F}_t\right]$$
$$\leq \mathbb{E}\left[\left\|(\Theta_* - \widehat{\Theta}_t)^\top V_t^{1/2}\right\|_F^2 \,\middle|\, \mathcal{F}_t\right]$$
$$= \sum_{i=1}^n \mathbb{E}\left[\left\|\left(\Theta_{*,(:,i)} - \widehat{\Theta}_{t,(:,i)}\right)^\top V_t^{1/2}\right\|^2 \,\middle|\, \mathcal{F}_t\right]$$
$$\leq 2nm^3 \ .$$

This shows that Assumption A2 is satisfied, thus finishing the proof. □

An immediate corollary of this is that Lazy PSRL will enjoy low regret when applied to linearly parametrized problems with Gaussian noise. We assume an effective stabilizing controller is available. This is necessary, as the noise may make the state arbitrarily large.

Corollary 7. *Consider Stabilized Lazy PSRL applied to a linearly parametrized problem with Gaussian noise with M as in Proposition 6. Let the underlying MDP satisfy Assumption A3. Suppose in time step t, the action chosen is σ_t-suboptimal. Then, for any time T, the regret of Stabilized Lazy PSRL satisfies $R_T = \widetilde{O}\left(\sqrt{T}\right) + \Sigma_T$.*

Proof. The claim follows immediately from Proposition 6 and Theorem 3. □

7 EXPERIMENTS

In this section we illustrate the behavior of LAZY PSRL on a queueing and a web server control application.

7.1 Queuing Control Application

The queueing problem is described in (de Farias and Van Roy, 2003). The queue has a buffer size of 99. For time t, let $x_t \in \{0, 1, \dots, 99\}$ be the state. The action a_t is the departure probability or service rate and is chosen from the set $\{0.1625, 0.325, 0.4875, 0.65\}$. Let p be the (unknown) arrival rate. The dynamics is defined as follows

$$x_{t+1} = \begin{cases} x_t - 1 & \text{with probability } a_t\,; \\ x_t + 1 & \text{with probability } p\,; \\ x_t & \text{otherwise}\,. \end{cases}$$

From state $x_t = 0$, transitions to states 1 and 0 happen with probabilities p and $1-p$. From state $x_t = 99$, transitions to states 98 and 99 happen with probabilities a_t and $1 - a_t$. The loss function is $\ell(x_t, a_t) = x_t^2 + 500p^2$.

7.1.1 Numerical Results

The purpose of this experiment is to show how the LAZY PSRL algorithm can take advantage of the problem structure to obtain better performance. We compare the LAZY PSRL algorithm with UCRL (Jaksch et al., 2010). For the LAZY PSRL algorithm, we use the Beta distribution $\text{Beta}(1,1)$ as the prior for the unknown parameter p (the conditions of our theorem can be seen to be satisfied along the lines of the previous section with $M(x,a) = \text{const}$). The constant G in Figure 1 is chosen to be $G = 2$. The UCRL algorithm is an optimistic algorithm that maintains a confidence interval around each transition probability

$P(x'|x,a)$ and, in each round, finds the transition dynamics and the corresponding policy that attains the smallest average loss. Specifically, the algorithm solves the optimization problem $\widetilde{P} = \text{argmin}_P \, J(P)$, where $J(P)$ is the average loss of the optimal policy when the system dynamics is P. Then, the algorithm plays the optimal controller given the parameter $\widetilde{P}$. As we show next, the LAZY PSRL algorithm achieves lower average cost.

The time horizon in these experiments is $T = 1,000$. We repeat each experiment 10 times and report the mean and the standard deviation of the observations. Figure 3 shows average cost vs. number of rounds. Details of the implementation of the UCRL algorithm are in (Jaksch et al., 2010).

Figure 3 show the average cost of the algorithms. The LAZY PSRL algorithm outperforms the UCRL algorithm. We explain this observation by noting that the UCRL algorithm is learning components of the transition dynamics independently (400 components in total), while the LAZY PSRL algorithm takes advantage of the problem structure to speed up the learning.

7.2 Web Server Control Application

In this section we illustrate the behavior of LAZY PSRL on a simple LQR control problem. We choose an LQR control problem because it is a continuous state-action problem. Equally important is that this allowed us to compare the performance of LAZY PSRL to a competing method, the OFULQ algorithm of Abbasi-Yadkori (2012). The experiments go beyond the scope of the theory, as we did not use a stabilizing controller, though the control problem itself is such that the zero-dynamics (i.e., the dynamics under zero control) is stable, making it less likely that a stabilizing controller would be necessary for the method to work. In the next section we describe the control problem, which will be followed by the description of our results.

The problem is taken from Section 7.8.1 of the book by Hellerstein et al. (2004) (this example is also used in Section 3.4 of the book by Aström and Murray (2008)). An Apache HTTP web server processes the incoming connections that arrive on a queue. Each connection is assigned to an available process. A process drops the connection if no requests have been received in the last KEEPALIVE seconds. At any given time, there are at most MAXCLIENTS active processes. The values of the KEEPALIVE and MAXCLIENTS parameters, denoted by a_{ka} and a_{mc} respectively, are chosen by a control algorithm. Increasing a_{mc} and a_{ka} results in faster and longer services to the connections, but also increases the CPU and memory usage of the server. The state of the server is determined by the average processor load x_{cpu} and the relative memory usage x_{mem}. An *operating point of interest* of the system is given by $x_{cpu} = 0.58\,, a_{ka} = 11s\,, x_{mem} = 0.55\,, a_{mc} = 600$. A linear model around the operating point is assumed, resulting in a model of the form

$$\begin{pmatrix} x_{cpu}^{\Delta}(t+1) \\ x_{mem}^{\Delta}(t+1) \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{21} \end{pmatrix} \begin{pmatrix} x_{cpu}^{\Delta}(t) \\ x_{mem}^{\Delta}(t) \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{21} \end{pmatrix} \begin{pmatrix} a_{ka}^{\Delta}(t) \\ a_{mc}^{\Delta}(t) \end{pmatrix} + \begin{pmatrix} w_1(t+1) \\ w_2(t+1) \end{pmatrix} ,$$

where $(w_1(t+1), w_2(t+1))_t$ is an i.i.d. sequence of Gaussian random variables, with a diagonal covariance matrix $\mathbb{E}\left[w(t+1)^\top w(t+1)\right] = \sigma^2 I$. Note that these state and action variables are in fact the deviations from the operating point. We test $\sigma = 0.1$ and $\sigma = 1.0$ in our experiments. The matrices A, B, Q, R are included in the appendix.

7.2.1 Numerical Results

We compare the LAZY PSRL algorithm with OFULQ (Abbasi-Yadkori, 2012). For the LAZY PSRL algorithm, we use the standard normal distribution as the prior. The OFULQ algorithm is an optimistic algorithm that maintains a confidence ellipsoid D around the unknown parameter and, in each round, finds the parameter and the corresponding policy that attains the smallest average loss. Specifically, the algorithm solves the optimization problem

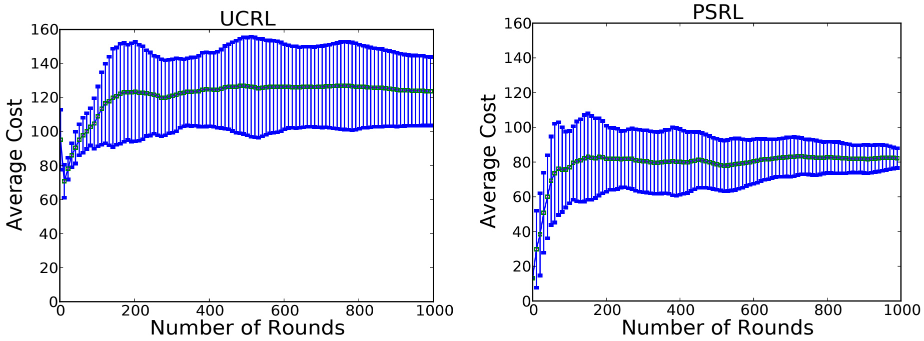
$$(\widetilde{A}, \widetilde{B}) = \underset{(A,B)\in D}{\text{argmin}} \, J(A,B)\,, \tag{4}$$

where $J(A,B)$ is the average loss of the optimal policy when the system dynamics is (A,B). Then, the algorithm plays the optimal controller given the parameter $(\widetilde{A}, \widetilde{B})$. The objective function J is not convex and thus, solving the optimistic optimization can be very time consuming. As we show next, the LAZY PSRL algorithm can have lower regret while avoiding the high computational costs of the OFULQ algorithm.

The time horizon in these experiments is $T = 1,000$. We repeat each experiment 10 times and report the mean and the standard deviation of the observations. Figure 4 shows regret vs. computation time. The horizontal axis shows the amount of time (in seconds) that the algorithm spends to process $T = 1,000$ rounds. We change the computation time by changing constant G in Figure 1, i.e. by changing how frequent an algorithm updates its policy.[2] Details of the implementation of the OFULQ algorithm are in (Abbasi-Yadkori, 2012).

The first two subfigures of Figure 4 show the regret of the algorithms when the standard deviation of the noise is $\sigma = 0.1$. The regret of the LAZY PSRL algorithm is slightly worse than what we get for the OFULQ algorithm in this case. The LAZY PSRL algorithm outperforms

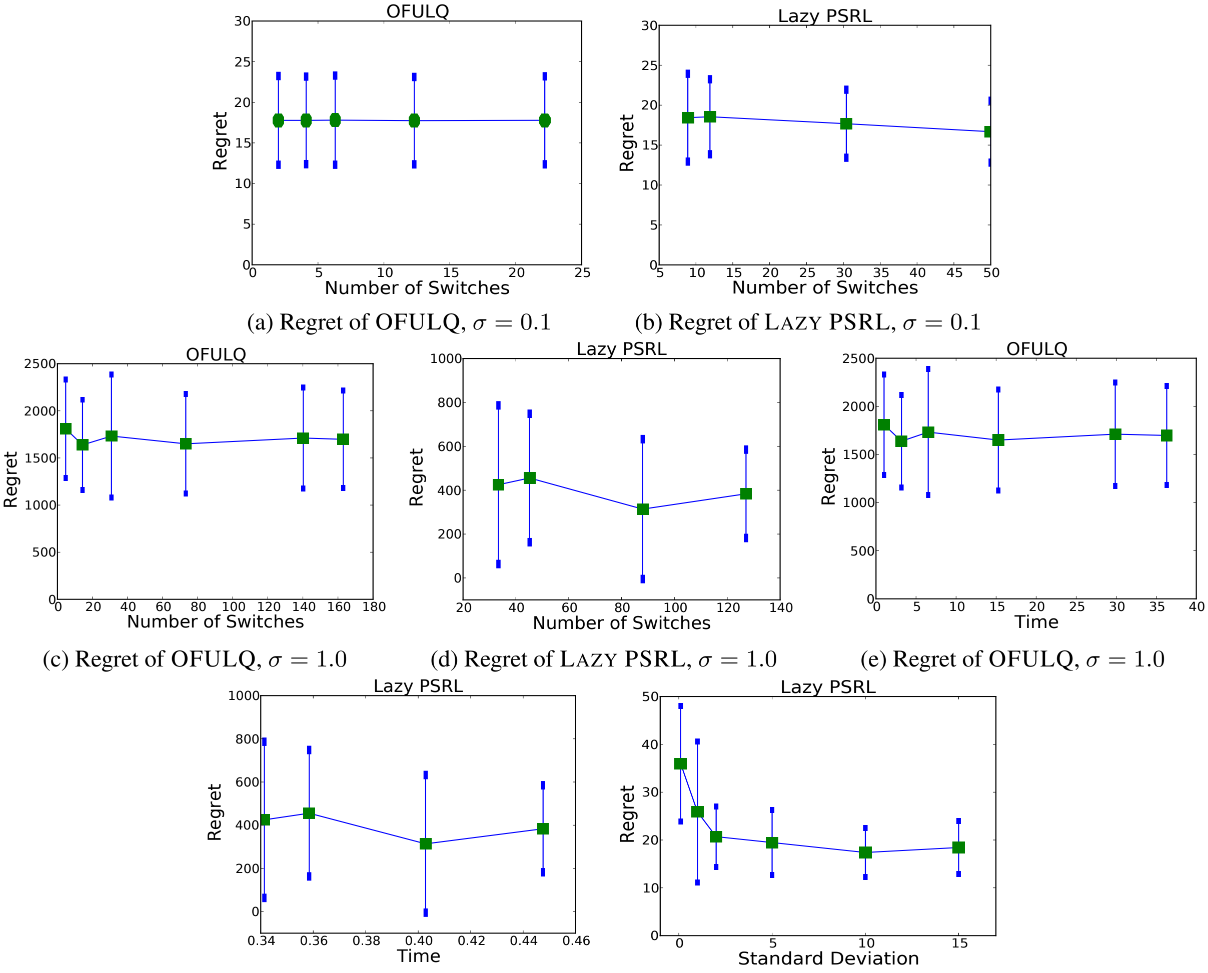
[2] For example, in Figure 4-(d), the average number of policy changes are $(33.4, 45.2, 88, 127.1)$. In Figure 4-(c) the average number of policy changes are $(5.6, 14.3, 30.8, 73.2, 140.2, 163)$.



(a) Average cost of UCRL

(b) Average cost of LAZY PSRL

Figure 3: Average cost for a queueing problem.



(a) Regret of OFULQ, $\sigma = 0.1$

(b) Regret of LAZY PSRL, $\sigma = 0.1$

(c) Regret of OFULQ, $\sigma = 1.0$

(d) Regret of LAZY PSRL, $\sigma = 1.0$

(e) Regret of OFULQ, $\sigma = 1.0$

(f) Regret of LAZY PSRL, $\sigma = 1.0$

(g) Regret of LAZY PSRL with zero mean Gaussians, changing standard deviation

Figure 4: Regret for a web server control problem.

the OFULQ algorithm when the noise variance is larger (next two subfigures). We explain this observation by noting that a larger noise variance implies larger confidence ellipsoids, which results in more difficult optimistic optimization problems (4). Finally, we performed experiments with different prior distributions. Figure 4-(e) shows regret of the LAZY PSRL algorithm when we change the prior.

References

Y. Abbasi-Yadkori. *Online Learning for Linearly Parametrized Control Problems*. PhD thesis, University of Alberta, 2012.

Y. Abbasi-Yadkori and Cs. Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, 2011.

A. Arapostathis, V.S. Borkar, E. Fernandez-Gaucherand, M.K. Ghosh, and S.I. Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31: 282–344, 1993.

M. Araya-López, V. Thomas, and O. Buffet. Near-optimal BRL using optimistic local transitions. In *ICML*, 2012.

J. Asmuth, L. Li, M. L. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *UAI*, pages 19–26, 2009.

Karl J. Aström and Richard M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.

D. P. de Farias and B. Van Roy. Approximate linear programming for average-cost dynamic programming. In *NIPS*, 2003.

A. Guez, D. Silver, and P. Dayan. Scalable and efficient Bayes-adaptive reinforcement learning based on Monte-Carlo tree search. *Journal of Artificial Intelligence Research*, 48:841–883, 2013.

A. Guez, D. Silver, and P. Dayan. Better optimism by Bayes: Adaptive planning with rich models. *CoRR*, abs/1402.1958, 2014.

Joseph L. Hellerstein, Yixin Diao, Sujay Parekh, and Dawn M. Tilbury. *Feedback Control of Computing Systems*. John Wiley & Sons, Inc., 2004.

A. Isidori. *Nonlinear Control Systems*. Springer Verlag, London, 3 edition, 1995.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563—1600, 2010.

J. Z. Kolter and A. Y Ng. Near-Bayesian exploration in polynomial time. In *ICML*, 2009.

I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, 2013.

M. Strens. A Bayesian framework for reinforcement learning. In *ICML*, 2000.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

N. Vlassis, M. Ghavamzadeh, S. Mannor, and P. Poupart. Bayesian reinforcement learning. In Marco Wieiring and Martijn van Otterlo, editors, *Reinforcement Learning: State-of-the-Art*, chapter 11, pages 359–386. Springer, 2012.

A Some Useful Lemmas

Lemma 8. *Let $V \in \mathbb{S}^+(m)$ be positive definite, $(M_t)_{t=1,2,\dots} \subset \mathbb{S}^+(m)$ be positive semidefinite matrices and define $V_t = V + \sum_{k=1}^{t-1} M_s$, $t = 1, 2, \dots$. If $\mathrm{trace}(M_t) \le L^2$ for all t, then*

$$
\begin{aligned}
\sum_{t=1}^{T} \min(1, \|V_t^{-1/2}\|_{M_t}^2) &\le 2 \left\{ \log\det(V_{T+1}) - \log\det V \right\} \\
&\le 2 \left\{ m \log\left(\frac{\mathrm{trace}(V) + TL^2}{m} \right) - \log\det V \right\} .
\end{aligned}
$$

Proof. On the one hand, we have

$$
\begin{aligned}
\det(V_T) &= \det(V_{T-1} + M_{T-1}) = \det(V_{T-1}(I + V_{T-1}^{-\frac{1}{2}} M_{T-1} V_{T-1}^{-\frac{1}{2}})) \\
&= \det(V_{T-1}) \det(I + V_{T-1}^{-\frac{1}{2}} M_{T-1} V_{T-1}^{-\frac{1}{2}}) \\
&\vdots \\
&= \det(V) \prod_{t=1}^{T-1} \det(I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}) \,.
\end{aligned}
$$

One the other hand, thanks to $x \le 2\log(1+x)$, which holds for all $x \in [0, 1]$,

$$
\begin{aligned}
\sum_{t=1}^{T} \min(1, \|V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}\|_2) &\le 2 \sum_{t=1}^{T} \log(1 + \|V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}\|_2) \\
&\le 2 \sum_{t=1}^{T} \log(\det(I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}})) \\
&= 2(\log(\det V_{T+1}) - \log(\det V)) \,,
\end{aligned}
$$

where the second inequality follows since $V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}$ is positive semidefinite, hence all eigenvalues of $I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}$ are above one and the largest eigenvalue of $I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}$ is $1 + \|V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}\|_2$, proving the first inequality. For the second inequality, note that for any positive definite matrix $S \in \mathbb{S}^+(m)$, $\log\det S \le m \log(\mathrm{trace}(S)/m)$. Applying this to V_T and using the condition that $\mathrm{trace}(M_t) \le L^2$, we get $\log\det V_T \le m \log((\mathrm{trace}(V) + TL^2)/m)$. Plugging this into the previous upper bound, we get the second part of the statement. □

Lemma 9 (Lemma 11 of Abbasi-Yadkori and Szepesvári (2011))**.** *Let $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{m \times m}$ be positive semidefinite matrices such that $A \succ B$. Then, we have*

$$
\sup_{X \neq 0} \frac{\left\| X^\top A X \right\|_2}{\left\| X^\top B X \right\|_2} \le \frac{\det(A)}{\det(B)} \,.
$$

B Proofs

Proof of Proposition 1. Note that if ACOE (1) holds for h, then for any constant C, it also holds that

$$
J(\Theta) + (h(x, \Theta) + C) = \min_{a \in \mathcal{A}} \left\{ \ell(x, a) + \int (h(y, \Theta) + C) p(dy \,|\, x, a, \Theta) \right\} .
$$

As by our assumption, the value function is bounded from below, we can choose C such that the $h'(\cdot, \Theta) = h(\cdot, \Theta) + C$ is nonnegative valued. In fact, if h assumes a minimizer x_0, by this reasoning, without loss of generality, we can assume that $h(x_0) = 0$ and so for any $x \in \mathcal{X}$, $0 \le h(x) = h(x) - h(x_0) \le B \|x - x_0\| \le BX$. The argument trivially extends to the general case when h may fail to have a minimizer over $\mathcal{X}$. □

Proof of Theorem 2. The proof follows that of the main result of Abbasi-Yadkori and Szepesvári (2011). First, we decompose the regret into a number of terms, which are then bound one by one. Define $\widetilde{x}^a_{t+1} = f(x_t, a, \widetilde{\Theta}_t, z_{t+1})$, where f is the map of Assumption A1 and let $h_t(x) = h(x, \widetilde{\Theta}_t)$ be the solution of the ACOE underlying $p(\cdot|x, a, \widetilde{\Theta}_t)$. By Assumption A3 (i), h_t exists and $h_t(x) \in [0, H]$ for any $x \in \mathcal{X}$. By Assumption A1, for any $g \in L^1(p(\cdot|x_t, a, \widetilde{\Theta}_t))$, $\int g(dy)p(dy|x_t, a, \widetilde{\Theta}_t) = \mathbb{E}\left[g(\widetilde{x}^a_{t+1})|\mathcal{F}_t, \widetilde{\Theta}_t\right]$. Hence, from (1) and (2),

$$\begin{aligned} J(\widetilde{\Theta}_t) + h_t(x_t) &= \min_{a\in\mathcal{A}} \left\{ \ell(x_t, a) + \mathbb{E}\left[h_t(\widetilde{x}^a_{t+1}) \,|\, \mathcal{F}_t, \widetilde{\Theta}_t\right]\right\} \\ &\ge \ell(x_t, a_t) + \mathbb{E}\left[h_t(\widetilde{x}^{a_t}_{t+1}) \,|\, \mathcal{F}_t, \widetilde{\Theta}_t\right] - \sigma_t \\ &= \ell(x_t, a_t) + \mathbb{E}\left[h_t(x_{t+1} + \epsilon_t) \,|\, \mathcal{F}_t, \widetilde{\Theta}_t\right] - \sigma_t \,, \end{aligned}$$

where $\epsilon_t = \widetilde{x}^{a_t}_{t+1} - x_{t+1}$. As $J(\cdot)$ is a deterministic function and conditioned on $\mathcal{F}_{\tau_t}$, $\widetilde{\Theta}_t$ and Θ_* have the same distribution,

$$\begin{aligned} R(T) &= \sum_{t=1}^T \mathbb{E}\left[\ell(x_t, a_t) - J(\Theta_*)\right] = \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\ell(x_t, a_t) - J(\Theta_*) \,|\, \mathcal{F}_{\tau_t}\right]\right] \\ &= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\ell(x_t, a_t) - J(\widetilde{\Theta}_t) \,|\, \mathcal{F}_{\tau_t}\right]\right] = \sum_{t=1}^T \mathbb{E}\left[\ell(x_t, a_t) - J(\widetilde{\Theta}_t)\right] \\ &\le \sum_{t=1}^T \mathbb{E}\left[h_t(x_t) - \mathbb{E}\left[h_t(x_{t+1} + \epsilon_t) \,|\, \mathcal{F}_t, \widetilde{\Theta}_t\right]\right] + \sum_{t=1}^T \mathbb{E}\left[\sigma_t\right] \\ &= \sum_{t=1}^T \mathbb{E}\left[h_t(x_t) - h_t(x_{t+1} + \epsilon_t)\right] + \sum_{t=1}^T \mathbb{E}\left[\sigma_t\right] \ . \end{aligned}$$

Let $\Sigma_T = \sum_{t=1}^T \mathbb{E}\left[\sigma_t\right]$ be the total error due to the approximate optimal control oracle. Thus, we can bound the regret using

$$\begin{aligned} R(T) &\le \Sigma_T + \mathbb{E}\left[h_1(x_1) - h_{T+1}(x_{T+1})\right] + \sum_{t=1}^T \mathbb{E}\left[h_{t+1}(x_{t+1}) - h_t(x_{t+1} + \epsilon_t)\right] \\ &\le \Sigma_T + H + \sum_{t=1}^T \mathbb{E}\left[h_{t+1}(x_{t+1}) - h_t(x_{t+1} + \epsilon_t)\right] \ , \end{aligned}$$

where the second inequality follows because $h_1(x_1) \le H$ and $-h_{T+1}(x_{T+1}) \le 0$. Let A_t denote the event that the algorithm has changed its policy at time t. We can write

$$\begin{aligned} R(T) - (\Sigma_T + H) &\le \sum_{t=1}^T \mathbb{E}\left[h_{t+1}(x_{t+1}) - h_t(x_{t+1} + \epsilon_t)\right] \\ &= \sum_{t=1}^T \mathbb{E}\left[h_{t+1}(x_{t+1}) - h_t(x_{t+1})\right] + \sum_{t=1}^T \mathbb{E}\left[h_t(x_{t+1}) - h_t(x_{t+1} + \epsilon_t)\right] \\ &\le 2H \sum_{t=1}^T \mathbb{E}\left[\mathbf{1}\left\{A_t\right\}\right] + B \sum_{t=1}^T \mathbb{E}\left[\|\epsilon_t\|\right] \ , \end{aligned}$$

where we used again that $0 \le h_t(x) \le H$, and also Assumption A3 (ii). Define

$$R_1 = H \sum_{t=1}^T \mathbb{E}\left[\mathbf{1}\left\{A_t\right\}\right] \,, \qquad R_2 = B \sum_{t=1}^T \mathbb{E}\left[\|\epsilon_t\|\right] \ .$$

It remains to bound R_2 and to show that the number of switches is small.

Bounding R_2 Let $\tau_t \le t$ be the last round before time step t when the policy is changed. So $\widetilde{\Theta}_t = \widetilde{\Theta}_{\tau_t}$. Letting $M_t = M(x_t, a_t)$, by Assumption A1,

$$\mathbb{E}\left[\|\epsilon_t\|\right] \le \mathbb{E}\left[\left\|\widetilde{\Theta}_t - \Theta_*\right\|_{M_t}\right].$$

Further,

$$\left\|\widetilde{\Theta}_t - \Theta_*\right\|_{M_t} \le \left\|\widetilde{\Theta}_t - \widehat{\Theta}_t\right\|_{M_t} + \left\|\widehat{\Theta}_t - \Theta_*\right\|_{M_t}.$$

For $\Theta \in \{\widetilde{\Theta}_{\tau_t}, \Theta_*\}$ we have that

$$\begin{aligned}\left\|\Theta - \widehat{\Theta}_{\tau_t}\right\|_{M_t}^2 &= \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top M_t (\Theta - \widehat{\Theta}_{\tau_t})\right\|_2 \\ &= \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{\frac{1}{2}} V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} V_t^{\frac{1}{2}} (\Theta - \widehat{\Theta}_{\tau_t})\right\|_2 \\ &\le \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{\frac{1}{2}}\right\|_2^2 \left\|V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}\right\|_2 = \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{\frac{1}{2}}\right\|_2^2 \left\|V_t^{-\frac{1}{2}}\right\|_{M_t}^2,\end{aligned}$$

where the last inequality follows because $\|\cdot\|_2$ is an induced norm and induced norms are sub-multiplicative. Hence, we have that

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}\left[\left\|\Theta - \widehat{\Theta}_{\tau_t}\right\|_{M_t}\right] &\le \mathbb{E}\left[\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2 \left\|V_t^{-1/2}\right\|_{M_t}\right] \\ &\le \mathbb{E}\left[\sqrt{\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2}\sqrt{\sum_{t=1}^T \left\|V_t^{-1/2}\right\|_{M_t}^2}\right] \\ &\le \sqrt{\mathbb{E}\left[\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2\right]}\sqrt{\mathbb{E}\left[\sum_{t=1}^T \left\|V_t^{-1/2}\right\|_{M_t}^2\right]},\end{aligned}$$

where the first inequality uses Hölder's inequality, and the last two inequalities use Cauchy-Schwarz. By Lemma 8 in Appendix A, using Assumption A4, we have that

$$\sum_{t=1}^T \min\left(1, \|V_t^{-1/2}\|_{M_t}^2\right) \le 2m \log\left(\frac{\text{trace}(V) + T\Phi^2}{m}\right).$$

Denoting by $\lambda_{\min}(V)$ the minimum eigenvalue of V, a simple argument shows $\left\|V_t^{-1/2}\right\|_{M_t}^2 \le \|M_t\|_2 / \lambda_{\min}(V) \le \Phi^2/\lambda_{\min}(V)$, where in the second inequality we used Assumption A4 again. Hence,

$$\begin{aligned}\sum_{t=1}^T \left\|V_t^{-1/2}\right\|_{M_t}^2 &\le \sum_{t=1}^T \min\left(\Phi^2/\lambda_{\min}(V), \left\|V_t^{-1/2}\right\|_{M_t}^2\right) \\ &\le \sum_{t=1}^T \max\left(1, \Phi^2/\lambda_{\min}(V)\right) \min\left(1, \left\|V_t^{-1/2}\right\|_{M_t}^2\right).\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}\left[\left\|\Theta - \widehat{\Theta}_{\tau_t}\right\|_{M_t}^2\right] &\le \sqrt{\mathbb{E}\left[2m \max\left(1, \frac{\Phi^2}{\lambda_{\min}(V)}\right) \log\left(\frac{\text{trace}(V) + T\Phi^2}{m}\right)\right]} \\ &\quad \times \sqrt{\mathbb{E}\left[\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2\right]}.\end{aligned}$$

By Lemma 9 of Appendix A and the choice of τ_t, we have that

$$\left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2 \le \sqrt{\frac{\det(V_t)}{\det(V_{\tau_t})}} \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2 \le \sqrt{2} \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2. \tag{5}$$

Thus,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left\|(\Theta-\widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2\right] &\le 2\mathbb{E}\left[\sum_{t=1}^{T}\left\|(\Theta-\widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2^2\right] && \text{(by (5))}\\
&= 2\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\left\|(\Theta-\widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2^2 \,\middle|\, \mathcal{F}_{\tau_t}\right]\right] && \text{(by the tower rule)}\\
&\le 2CT\,. && \text{(by Assumption A2)}
\end{aligned}
$$

Let $G_T = 2m\max\left(1, \frac{\Phi^2}{\lambda_{\min}(V)}\right)\log\left(\frac{\operatorname{trace}(V)+T\Phi^2}{m}\right)$. Collecting the inequalities, we get

$$
\begin{aligned}
R_2 = B\sum_{t=1}^{T}\mathbb{E}\left[\left\|(\widetilde{\Theta}_{\tau_t}-\Theta_*)^\top\varphi_t\right\|\right] &\le \sqrt{\mathbb{E}\left[G_T\right]}\sqrt{CT}\\
&\le 4B\sqrt{m\max\left(1,\frac{\Phi^2}{\lambda_{\min}(V)}\right)\log\left(\frac{\operatorname{trace}(V)+T\Phi^2}{m}\right)}\sqrt{CT}\,.
\end{aligned}
$$

Bounding R_1 If the algorithm has changed the policy K times up to time T, then we should have that $\det(V_T)\ge 2^K$. On the other hand, from Assumption A4 we have $\lambda_{\max}(V_T)\le \operatorname{trace}(V)+(T-1)\Phi^2$. Thus, it holds that $2^K \le (\operatorname{trace}(V)+\Phi^2 T)^m$. Solving for K, we get $K\le m\log_2(\operatorname{trace}(V)+\Phi^2T)$. Thus,

$$
R_1 = H\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}\left\{A_t\right\}\right] \le Hm\log_2(\operatorname{trace}(V)+\Phi^2T)\,.
$$

Putting together the bounds obtained for R_1 and R_2, we get the desired result. □

Proof of Theorem 3. First notice that Theorem 2 continues to hold if Assumption A4 is replaced by the following weaker assumption:

Assumption A6 ***(Boundedness Along Trajectories)*** There exist $\Phi>0$ such that for all $t\ge 1$, $\mathbb{E}\left[\operatorname{trace}(M(x_t,a_t))\right]\le\Phi^2$.

The reason this is true is because A4 is used only in a context where $\mathbb{E}\left[\log(\operatorname{trace}(V+\sum_{s=1}^{T}M_t))\right]$ needs to be bounded. Using that $\log$ is concave, we get

$$
\mathbb{E}\left[\log(\operatorname{trace}(V+\textstyle\sum_{s=1}^{T}M_t))\right]\le\log\left(\mathbb{E}\left[\operatorname{trace}(V+\textstyle\sum_{s=1}^{T}M_t)\right]\right)\le\log(\operatorname{trace}(V)+T\Phi^2).
$$

With this observation, the result follows from Theorem 2 applied to Lazy PSRL and $\{p'(\cdot|x,a,\Theta)\}$ as running Stabilized Lazy PSRL for t time steps in $p(\cdot|x,a,\Theta_*)$ results in the same total expected cost as running Lazy PSRL for t time steps in $p'(\cdot|x,a,\Theta_*)$ thanks to the definition of Stabilized Lazy PSRL and p'.

Hence, all what remains is to show that the conditions of Theorem 2 are satisfied when it is used with $\{p'(\cdot|x,a,\Theta)\}$. In fact, A3 and A2 hold true by our assumptions. Let us check Assumption A3 next. Defining $f'(x,a,\Theta,z)=f(x,a,\Theta,z)$ if $x\in\mathcal{R}$ and $f'(x,a,\Theta,z)=f(x,\pi_{\text{stab}}(x),\Theta,z)$ otherwise, we see that $x_{t+1}=f'(x_t,a_t,\Theta,z_{t+1})$. Further, defining $M'(x,a)=M(x,a)$ if $x\in\mathcal{R}$ and $M'(x,a)=M(x,\pi_{\text{stab}}(x))$ otherwise, we see that, thanks to the second part that of A1 applied to $p(\cdot|x,a,\Theta)$, for $y=f'(x,a,\Theta,z)$, $y'=f'(x,a,\Theta',z)$, $\mathbb{E}\left[\|y-y'\|\right]\le\mathbb{E}\left[\|\Theta-\Theta'\|_{M(x,a)}\right]$ if $x\in\mathcal{R}$ and $\mathbb{E}\left[\|y-y'\|\right]\le\mathbb{E}\left[\|\Theta-\Theta'\|_{M(x,\pi_{\text{stab}}(x))}\right]$ otherwise. Hence, $\mathbb{E}\left[\|y-y'\|\right]\le EE\|\Theta-\Theta'\|_{M'(x,a)}$, thus showing that A1 holds for $p'(\cdot|x,a,\Theta)$ when M is replaced by M'. Now, Assumption A6 follows from Assumption A5.

□

C Choice of the matrices in the web-server application

Hellerstein et al. (2004) fitted the linear model detailed earlier to an Apache HTTP server and obtained the parameters

$$A = \begin{pmatrix} 0.54 & -0.11 \\ -0.026 & 0.63 \end{pmatrix}, \qquad B = \begin{pmatrix} -85 & 4.4 \\ -2.5 & 2.8 \end{pmatrix} \times 10^{-4} \, ,$$

while the noise standard deviation was measured to be 0.1. Hellerstein et al. found that these parameters provided a reasonable fit to their data. For control purposes, the cost matrices $Q = \mathrm{diag}(5, 1)$, $R = \mathrm{diag}(1/5062, 0.1^6)$, taken from Example 6.9 of Aström and Murray (2008), were chosen.