Proceedings of Machine Learning Research vol 65:1–22, 2017

Fast rates for online learning in Linearly Solvable Markov Decision Processes

Gergely Neu GERGELY.NEU@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain

Vicenç Gómez VICEN.GOMEZ@UPF.EDU
Universitat Pompeu Fabra, Barcelona, Spain

Abstract

We study the problem of online learning in a class of Markov decision processes known as *linearly solvable MDPs*. In the stationary version of this problem, a learner interacts with its environment by directly controlling the state transitions, attempting to balance a fixed state-dependent cost and a certain smooth cost penalizing extreme control inputs. In the current paper, we consider an online setting where the state costs may change arbitrarily between consecutive rounds, and the learner only observes the costs at the end of each respective round. We are interested in constructing algorithms for the learner that guarantee small regret against the best stationary control policy chosen in full knowledge of the cost sequence. Our main result is showing that the smoothness of the control cost enables the simple algorithm of *following the leader* to achieve a regret of order $\log^2 T$ after T rounds, vastly improving on the best known regret bound of order $T^{3/4}$ for this setting.

Keywords: Online learning, fast rates, Markov decision processes, optimal control

1. Introduction

We consider the problem of online learning in Markov decision processes (MDPs) where a learner sequentially interacts with an environment by repeatedly taking actions that influence the future states of the environment while incurring some immediate costs. The goal of the learner is to choose its actions in a way that the accumulated costs are as small as possible. Several variants of this problem have been well-studied in the literature, primarily in the case where the costs are assumed to be independent and identically distributed (Sutton and Barto, 1998; Puterman, 1994; Bertsekas and Tsitsiklis, 1996; Szepesvári, 2010). In the current paper, we consider the case where the costs are generated by an arbitrary external process and the learner aims to minimize its total loss during the learning procedure—conforming to the learning paradigm known as *online learning* (Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012). In the online-learning framework, the performance of the learner is measured in terms of the *regret* defined as the gap between the total costs incurred by the learner and the total costs of the best comparator chosen from a pre-specified class of strategies. In the case of online learning in MDPs, a natural class of strategies is the set of all state-feedback policies: several works studied minimizing regret against this class both in the stationary-cost (Bartlett and Tewari, 2009; Jaksch et al., 2010; Abbasi-Yadkori and Szepesvári, 2011) and the non-stochastic setting (Even-Dar et al., 2009; Yu et al., 2009; Neu et al., 2010, 2012; Zimin and Neu, 2013; Dick et al., 2014; Neu et al., 2014; Abbasi-Yadkori et al., 2014). In the non-stochastic setting, most works consider MDPs with unstructured, finite state spaces and guarantee that the regret increases no faster than $O(\sqrt{T})$ as the number of interaction rounds T grows large. A notable exception is the work of Abbasi-Yadkori et al. (2014), who consider the special case of (continuous-state) linear-quadratic control with arbitrarily changing target states, and propose an algorithm that guarantees a regret bound of $O(\log^2 T)$.

In the present paper, we study another special class of MDPs that turns out to allow fast rates. Specifically, we consider the class of so-called *linearly solvable MDPs* (in short, LMDPs), first proposed and named so by Todorov (2006). This class takes its name after the special property that the Bellman optimality equations characterizing the optimal behavior policy take the form of a system of linear equations, which makes optimization remarkably straightforward in such problems. The continuous formulation (in both space and time) was discovered independently by Kappen (2005) and is known as *path integral control*. LMDPs have many interesting properties. For example, optimal control laws for LMDPs can be linearly combined to derive composite optimal control laws efficiently (Todorov, 2009). Also, the inverse optimal control problem in LMDPs can be expressed as a convex optimization problem (Dvijotham and Todorov, 2010). LMDPs generalize an existing duality between optimal control computation and Bayesian inference (Todorov, 2008). Indeed, the popular belief propagation algorithm used in dynamic probabilistic graphical models is equivalent the the power iteration method used to solve LMDPs (Kappen et al., 2012).

The LMDP framework has found applications in robotics (Matsubara et al., 2014; Ariki et al., 2016), crowdsourcing (Abbasi-Yadkori et al., 2015), and controlling the growth dynamics of complex networks (Thalmeier et al., 2017). The related path integral control framework of Kappen (2005) has been applied in several real-world tasks, including robot navigation (Kinjo et al., 2013), motor skill reinforcement learning (Theodorou et al., 2010; Rombokas et al., 2013; Gómez et al., 2014), aggressive car maneuvering (Williams et al., 2016) or autonomous flight of teams of quadrotors (Gómez et al., 2016).

In the present paper, we show that besides the aforementioned properties, the structure of LMDPs also enables constructing efficient online learning procedures with very low regret. In particular, we show that, under some mild assumptions on the structure of the LMDP, the (conceptually) simplest online learning strategy of *following the leader* guarantees a regret of order $\log^2 T$, vastly improving over the best known previous result by Guan, Raginsky, and Willett (2014), who prove a regret bound of order $T^{3/4+\epsilon}$ for arbitrarily small $\epsilon > 0$ under the same assumptions. Our approach is based on the observation that the optimal control law arising from the LMDP structure is a smooth function of the underlying cost function, enabling rapid learning without any regularization whatsoever.

The rest of the paper is organized as follows. Section 2 introduces the formalism of LMDPs and summarizes some basic facts that our technical content is going to rely on. Section 3 describes our online learning model. Our learning algorithm is described in Section 4 and analyzed in Section 5. Finally, we draw conclusions in Section 6.

Notation. We will consider several real-valued functions over a finite state-space $\mathcal{X}$, and we will often treat these functions as finite-dimensional (column) vectors endowed with the usual definitions of the ℓ_p norms. The set of probability distributions over $\mathcal{X}$ will be denoted as $\Delta(\mathcal{X})$. Indefinite sums with running variables x, y or s are understood to run through all $\mathcal{X}$.

2. Background on linearly solvable MDPs

This section serves as a quick introduction into the formalism of linearly solvable MDPs (LMDPs, Todorov (2006)). These decision processes are defined by the tuple $\{\mathcal{X}, P, c\}$, where $\mathcal{X}$ is a finite set of *states*, $P : \mathcal{X} \to \Delta(\mathcal{X})$ is a transition kernel called the *passive dynamics* (with $P(x'|x)$ being the probability of the process moving to state x' given the previous state x) and $c : \mathcal{X} \to [0, 1]$ is the *state-cost function*. Our Markov decision process is a sequential decision-making problem where the initial state X_0 is drawn from some distribution μ_0, and the following steps are repeated for an indefinite number of rounds $t = 1, 2, \ldots$:

1. The learner chooses a transition kernel $Q_t : \mathcal{X} \to \Delta(\mathcal{X})$ satisfying $\mathit{supp}\, Q_t(\cdot|x) \subseteq \mathit{supp}\, P(\cdot|x)$ for all $x \in \mathcal{X}$.

2. The learner observes $X_t \in \mathcal{X}$ and draws the next state $X_{t+1} \sim Q_t(\cdot|X_t)$.

3. The learner incurs the cost

$$\ell(X_t, Q_t) = c(X_t) + D\left(Q_t(\cdot|X_t) \| \, P(\cdot|X_t)\right),$$

where $D\left(q \| \, p\right)$ is the relative entropy (or Kullback-Leibler divergence) between the probability distributions p and q defined as $D\left(q\| \, p\right) = \sum_x q(x) \log \frac{q(x)}{p(x)}$.

The state-cost function c should be thought of as specifying the objective for the learner in the MDP, while the relative-entropy term governs the costs associated with significant deviations from the passive dynamics. Accordingly, we refer to this component as the *control cost*. A central question in the theory of Markov decision problems is finding a behavior policy that minimizes (some notion of) the long-term total costs. In this paper, we consider the problem of *minimizing the long-term average cost-per-stage* $\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \ell(X_t, Q_t)$. Assuming that the passive dynamics P is aperiodic and irreducible, this limit is minimized by a *stationary* policy Q (see, e.g., Puterman (1994, Sec. 8.4.4)). Below, we provide two distinct derivations for the optimal stationary policy that minimizes the average costs under this assumption.

2.1. The Bellman equations

We first take an approach rooted in dynamic programming (Bertsekas, 2007), following Todorov (2006). Under our assumptions, the optimal stationary policy minimizing the average cost is given by finding the solution to the Bellman optimality equation

$$v(x) = c(x) - \lambda + \min_{q \in \Delta(\mathcal{X})} \left\{ D\left(q \| \, P(\cdot|x)\right) + \sum_{x'} q(x')v(x') \right\} \tag{1}$$

for all $x \in \mathcal{X}$, where v is called the *optimal cost-to-go* and $\lambda \in \mathbb{R}$ is the average cost associated with the optimal policy[1]. Linearly solvable MDPs get their name from the fact that the Bellman optimality equation can be rewritten in a simple linear form. To see this, observe that by elementary calculations involving Lagrange multipliers, we have

$$\min_{q \in \Delta(\mathcal{X})} \left\{ D\left(q \| \, P(\cdot|x)\right) + \sum_{x'} q(x')v(x') \right\} = -\log \sum_{x'} P(x'|x)e^{-v(x')},$$

[1] This solution is guaranteed to be unique up to a constant shift of the cost-to-go: if v is a solution, then so is $v + a$ for any $a \in \mathbb{R}$. Unless stated otherwise, we will assume that v is such that $v(x_0) = 0$ holds for a fixed state $x_0 \in \mathcal{X}$.

so, after defining the exponentiated cost-to-go $z(x) = e^{-v(x)}$ for all x, plugging into Equation (1) and exponentiating both sides gives

$$z(x) = e^{\lambda - c(x)} \sum_{x'} P(x'|x) z(x'). \tag{2}$$

Rewriting the above set of equations in matrix form, we obtain the linear equations

$$e^{-\lambda} z = GPz,$$

where G is a diagonal matrix with $G_{ii} = e^{-c(i)}$. By the *Perron-Frobenius theorem* (see, e.g., Chapter 8 of Meyer (2000)) concerning positive matrices, the above system of linear equations has a unique[2] solution satisfying $z(x) \geq 0$ for all x, and this eigenvector corresponds to the largest eigenvalue $e^{-\lambda}$ of GP. Since the solution of the Bellman optimality equation (1) is unique (up to a constant shift corresponding to a constant scaling of z), we obtain that λ is the average cost of the optimal policy. In summary, the Bellman optimality equation takes the form of a *Perron–Frobenius eigenvalue problem*, which can be efficiently solved by iterative methods such as the well-known power method for finding top eigenvectors. Finally, getting back to the basic form (1) of the Bellman equations, we can conclude after simple calculations that the optimal policy can be computed for all x, x' as

$$Q(x'|x) = \frac{P(x'|x)z(x')}{\sum_y P(y|x)z(y)}.$$

2.2. The convex optimization view

We also provide an alternative (and, to our knowledge, yet unpublished) view of the optimal control problem in LMDPs, based on convex optimization. For the purposes of this paper, we find this form to be more insightful, as it enables us to study our learning problem in the framework of online convex optimization (Hazan, 2011, 2016; Shalev-Shwartz, 2012). To derive this form, observe that under our assumptions, every feasible policy Q induces a stationary distribution μ_Q over the state space $\mathcal{X}$ satisfying $\mu_Q^\mathsf{T} = \mu_Q^\mathsf{T} Q$. This stationary distribution and the policy together induce a distribution π_Q over $\mathcal{X}^2$ defined for all x, x' as $\pi_Q(x,x') = \mu_Q(x)Q(x'|x)$. We will call π_Q as the *stationary transition measure* induced by Q, which is motivated by the observation that $\pi_Q(x,x')$ corresponds to the probability of observing the transition $x \to x'$ in the equilibrium state: $\pi_Q(x,x') = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^T \mathbb{P}\left[X_t = x, X_{t+1} = x'\right]$. Notice that, with this notation, the average cost-per-stage of policy Q can be rewritten in the form

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\ell(X_t, Q)\right] &= \sum_x \mu_Q(x) \Big(c(x) + D\left(Q(\cdot|x) \| P(\cdot|x)\right) \Big) \\
&= \sum_{x,x'} \pi_Q(x,x') \left(c(x) + \log \frac{\pi_Q(x,x')}{P(x'|x) \sum_y \pi_Q(x,y)} \right) \\
&= \sum_{x,x'} \pi_Q(x,x') \log \frac{\pi_Q(x,x')}{\sum_y \pi_Q(x,y)} + \sum_{x,x'} \pi_Q(x,x') \left(c(x) - \log \left(P(x'|x) \right) \right).
\end{aligned}$$

[2] As in the case of the Bellman equations, this solution is unique up to a *scaling* of z.

The first term in the final expression above is the *negative conditional entropy* of X' relative to X, where (X, X') is a pair of random states drawn from π_Q. Since the negative conditional entropy is convex in π_Q (for a proof, see Appendix A.1) and the second term in the expression is linear in π_Q, we can see that λ is a convex function of π_Q. This suggests that we can view the optimal control problem as having to find a feasible stationary transition measure π that minimizes the expected costs. In short, defining

$$f(\pi; c) = \sum_{x,x'} \pi(x,x') \left(c(x) + \log \frac{\pi(x,x')}{P(x'|x) \sum_y \pi(x,y)} \right) \tag{3}$$

and $\Delta(M)$ as the (convex) set of feasible stationary transition measures π satisfying

$$\begin{aligned} \sum_{x'} \pi(x,x') &= \sum_{x''} \pi(x'',x) && (\forall x), \\ \sum_{x,x'} \pi(x,x') &= 1, && \\ \pi(x,x') &\geq 0 && (\forall x,x'), \\ \pi(x,x') &= 0 && (\forall x,x' : P(x'|x) = 0), \end{aligned} \tag{4}$$

the optimization problem can be succinctly written as $\min_{\pi \in \Delta(M)} f(\pi; c)$. In Appendix A.2, we provide a derivation of the optimal control given by Equation (2) starting from the formulation given above. We also remark that our analysis will heavily rely on the fact that $f(\pi; c)$ is affine in c.

3. Online learning in linearly solvable MDPs

We now present the precise learning setting that we consider in the present paper. We will study an online learning scheme where for each round $t = 1, 2, \ldots, T$, the following steps are repeated:

1. The learner chooses a transition kernel $Q_t : \mathcal{X} \to \Delta(\mathcal{X})$ satisfying $\mathit{supp}\, Q_t(\cdot|x) \subseteq \mathit{supp}\, P(\cdot|x)$ for all $x \in \mathcal{X}$.
2. The learner observes $X_t \in \mathcal{X}$ and draws the next state $X_{t+1} \sim Q_t(\cdot|X_t)$.
3. Obliviously to the learner's choice, the environment chooses state-cost function $c_t : \mathcal{X} \to [0, 1]$.
4. The learner incurs the cost

$$\ell_t(X_t, Q_t) = c_t(X_t) + D\left(Q_t(\cdot|X_t) \| \, P(\cdot|X_t)\right).$$

5. The environment reveals the state-cost function c_t.

The key change from the stationary setting described in the previous section is that the state-cost function now may *change arbitrarily* between each round, and the learner is only allowed to observe the costs *after it has made its decision*. We stress that we assume that the learner *fully knows* the passive dynamics, so the only difficulty comes from having to deal with the changing costs. As usual in the online-learning literature, our goal is to do nearly as well as the best *stationary* policy chosen in hindsight after observing the entire sequence of cost functions. To define our precise performance measure, we first define the average reward of a policy Q as

$$\mathcal{L}_T(Q) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(X'_t, Q)\right],$$

where the state trajectory X'_t is generated sequentially as $X'_t \sim Q(\cdot|X'_{t-1})$ and the expectation integrates over the randomness of the transitions. Having this definition in place, we can specify the best stationary policy[3] $Q^*_T = \arg\min_Q \mathcal{L}_T(Q)$ and define our performance measure as the (total expected) *regret* against Q^*_T:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(X_t, Q_t)\right] - \mathcal{L}_T(Q^*_T),$$

where the expectation integrates over both the randomness of the state transitions and the potential randomization used by the learning algorithm. Having access to this definition, we can now formally define the goal of the learner as having to come up with a sequence of policies $Q_1, Q_2, \ldots$ that guarantee that the total regret grows sublinearly, that is, that the average per-round regret asymptotically converges to zero.

For our analysis, it will be useful to define an idealized version of the above online optimization problem, where the learner is allowed to *immediately switch* between the stationary distributions of the chosen policies. By making use of the convex-optimization view given in Section 2.2, we define an auxiliary online convex optimization (or, in short, OCO, see, e.g., Hazan, 2011; Shalev-Shwartz, 2012) problem called the *idealized OCO problem* where in each round t, the following steps are repeated:

1. The learner chooses the stationary transition measure $\pi_t \in \Delta(M)$.
2. Obliviously to the learner's choice, the environment chooses the loss function $\widetilde{\ell}_t = f(\cdot; c_t)$.
3. The learner incurs a loss of $\widetilde{\ell}_t(\pi_t)$.
4. The environment reveals the loss function $\widetilde{\ell}_t$.

The performance of the learner in this setting is measured by the *idealized regret*

$$\overline{R}_T = \sum_{t=1}^{T} \widetilde{\ell}_t(\pi_t) - \min_{\pi \in \Delta(M)} \sum_{t=1}^{T} \widetilde{\ell}_t(\pi).$$

Throughout the paper, we will consider *oblivious environments* that choose the sequence of state-cost functions without taking into account the states visited by the learner. This assumption will enable us to simultaneously reason about the expected costs under any sequence of state distributions, and thus to make a connection between the idealized regret $\overline{R}_T$ and the true regret R_T. This technique was first used by Even-Dar et al. (2009) and was shown to be essentially inevitable by Yu et al. (2009): As discussed in their Section 3.1, no learning algorithm can avoid linear regret if the environment is not oblivious.

[3] The existence of the minimum is warranted by the fact that $\mathcal{L}_T$ is a continuous function bounded from below on its compact domain.

4. Algorithm and main result

In this section, we propose a simple algorithm for online learning in LMDPs based on the "follow-the-leader" (FTL) strategy. On a high level, the idea of this algorithm is greedily betting on the policy that seems to have been optimal for the total costs observed so far. While this strategy is known to fail catastrophically in several simple learning problems (see, e.g., Cesa-Bianchi and Lugosi 2006), it is known to perform well in several important scenarios such as sequential prediction under the logarithmic loss (Merhav and Feder, 1992) or prediction with expert advice under bounded losses, given that losses are stationary (Kotłowski, 2016) and often serves as a strong benchmark strategy (de Rooij et al., 2014; Sani et al., 2014). In our learning problem, following the leader is a very natural choice of algorithm, as the convex formulation of Section 2.2 suggests that we can effectively build on the analysis of Follow-the-Regularized-Leader-type algorithms without having to explicitly regularize the objective.

In precise terms, our algorithm computes the sequence of policies $Q_1, Q_2, \ldots, Q_T$ by running FTL *in the idealized setting*: in round t, the algorithm chooses the stationary transition measure

$$\begin{aligned}\pi_t &= \operatorname*{arg\,min}_{\pi\in\Delta(M)} \sum_{s=1}^{t-1} \widetilde{\ell}_s(\pi) = \operatorname*{arg\,min}_{\pi\in\Delta(M)} \sum_{s=1}^{t-1} f(\pi; c_s) \\ &= \operatorname*{arg\,min}_{\pi\in\Delta(M)} (t-1)\cdot f\left(\pi; \frac{1}{t-1}\sum_{s=1}^{t-1} c_s\right) = \operatorname*{arg\,min}_{\pi\in\Delta(M)} f\left(\pi; \overline{c}_t\right),\end{aligned}$$

where the third equality uses the fact that f is affine in its second argument and the last step introduces the average state-cost function $\overline{c}_t = \frac{1}{t-1}\sum_{s=1}^{t-1} c_s$. This form implies that π_t can be computed as the optimal control for the state-cost function $\overline{c}_t$, which can be done by following the procedure described in Section 2.1. Precisely, we define the diagonal matrix G_t with its i^{th} diagonal element $e^{-\overline{c}_t(i)}$, let γ_t be the largest eigenvalue of $G_t P$ and z_t be the corresponding (unit-norm) right eigenvector. Also, let $v_t = -\log z_t$ and $\lambda_t = -\log \gamma_t$, and note that $\lambda_t = f(\pi_t; \overline{c}_t)$ is the optimal average-cost-per-stage of π_t given the cost function $\overline{c}_t$. Finally, we define the policy used in round t as

$$Q_t(x'|x) = \frac{P(x'|x)z_t(x')}{\sum_y P(y|x)z_t(y)} \tag{5}$$

for all x' and x. We denote the induced stationary distribution by μ_t. The algorithm is presented as Algorithm 1.

Now we present our main result. First, we state two key assumptions about the underlying passive dynamics; both of these assumptions are also made by Guan et al. (2014).

Assumption 1 *The passive dynamics P is irreducible and aperiodic. In particular, there exists a natural number $H > 0$ such that $\left(P^n\right)(y|x) > 0$ for all $n \geq H$ and all $x, y \in \mathcal{X}$. We will refer to H as the (worst-case)* hitting time.

Assumption 2 *The passive dynamics P is ergodic in the sense that its Markov–Dobrushin ergodicity coefficient is strictly less than 1:*

$$\alpha(P) = \max_{x,y\in\mathcal{X}} \left\|P(\cdot|x) - P(\cdot|y)\right\|_1 < 1.$$

Algorithm 1 Follow The Leader in LMDPs

```
Input: Passive dynamics $P$.
Initialization: $\overline{c}_1(x) = 0$ for all $x \in \mathcal{X}$.
For $t = 1, 2, \dots, T$, repeat
    1. Construct $G_t = \left[\text{diag}(e^{-\overline{c}_t})\right]$.
    2. Find the right eigenvector $z_t$ of $G_t P$ corresponding to the largest eigenvalue.
    3. Compute the policy
       $$Q_t(x'|x) = \frac{P(x'|x) z_t(x')}{\sum_y P(y|x) z_t(y)}.$$
    4. Observe state $X_t$ and draw $X_{t+1} \sim Q_t(\cdot|X_t)$.
    5. Observe state-cost function $c_t$ and update $\overline{c}_{t+1} = \frac{(t-1)\overline{c}_t + c_t}{t}$.
```

A standard consequence (see, e.g., Seneta 2006) of Assumption 2 is that the passive dynamics mixes quickly: for any distributions $\mu, \mu' \in \Delta(\mathcal{X})$, we have

$$\left\| (\mu - \mu')^\mathsf{T} P \right\|_1 \le \alpha(P) \left\| \mu - \mu' \right\|_1 .$$

We will sometimes refer to $\tau(P) = \left(\log\left(1/\alpha(P)\right)\right)^{-1}$ as the *mixing time* associated with P. Now we are ready to state our main result:

Theorem 1 *Suppose that the passive dynamics satisfies Assumptions 1 and 2. Then, the regret of Algorithm 1 satisfies $R_T = O(\log^2 T)$.*

The asymptotic notation used in the theorem hides a number of factors that depend only on the passive dynamics P. In particular, the bound scales polynomially with the worst-case mixing time τ of any optimal policy, and shows no *explicit* dependence on the number of states.[4] We explicitly state the bound at the end of the proof presented in the next section as Equation (8), when all terms are formally defined.

5. Analysis

In this section, we provide a series of lemmas paving the way towards proving Theorem 1. The attentive reader may find some of these lemmas familiar from related work: indeed, we build on several technical results from Even-Dar et al. (2009); Neu et al. (2014) and Guan et al. (2014). Our main technical contribution is an efficient combination of these tools that enables us to go way beyond the best known bounds for our problem, proved by Guan et al. (2014). Throughout the section, we will assume that the conditions of Theorem 1 are satisfied.

Before diving into the analysis, we state some technical results that we will use several times. We defer all proofs to Appendix B. First, we present some important facts regarding LMDPs with

[4] Of course, the mixing time time does depend on the size of the state space in general.

bounded state-costs. In particular, we define $Q^*(c)$ as the optimal policy with respect to an arbitrary state-cost function c and let $\mathcal{C}$ be the set of all state-costs bounded in $[0,1]$. We define $\mathcal{Q}^*$ as the set of optimal policies induced by state-cost functions in $\mathcal{C}$: $\mathcal{Q}^* = Q^*(\mathcal{C})$. Observe that $Q_t \in \mathcal{Q}^*$ for all t, as $Q_t = Q^*(\overline{c}_t)$ and $\overline{c}_t \in \mathcal{C}$ for all t. Below, we give several useful results concerning policies in $\mathcal{Q}^*$. For stating these results, let $c \in \mathcal{C}$ and $Q = Q^*(c)$. We first note that the average cost λ of Q is bounded in $[0,1]$: By the Perron-Frobenius theorem (see, e.g., Meyer, 2000, Chapter 8), we have that the largest eigenvalue of GP is bounded by the maximal and minimal row sums of GP: $e^{-\lambda} \in [e^{-\max_x c(x)}, e^{-\min_x c(x)}]$, which translates to having $\lambda \in [0,1]$ under our assumptions. The next key result bounds the cost-to-go functions and the control costs in terms of the hitting time:

Lemma 2 *For all x, y and t, the cost-to-go satisfies $v_t(x) - v_t(y) \le H$. Furthermore, all policies $Q \in \mathcal{Q}^*$ satisfy*

$$\max_x D\left(Q(\cdot|x) \| P(\cdot|x)\right) \le H + 1.$$

The proof is loosely based on ideas from Bartlett and Tewari (2009). The second statement guarantees that the mixing time $\tau(Q) = \left(\log(1/\alpha(Q))\right)^{-1}$ is finite for all policies in $\mathcal{Q}^*$:

Lemma 3 *The Markov–Dobrushin coefficient $\alpha(Q)$ of any policy $Q \in \mathcal{Q}^*$ is bounded as*

$$\alpha(Q) \le \alpha(P) + (1 - \alpha(P))\left(1 - e^{-H-2}\right) < 1.$$

The proof builds on the previous lemma and uses standard ideas from Markov-chain theory. In what follows, we will use $\tau = \max_{Q \in \mathcal{Q}^*} \tau(Q)$ and $\alpha = \max_{Q \in \mathcal{Q}^*} \alpha(Q)$ to denote the worst-case mixing time and ergodicity coefficient, respectively. With this notation, we can state the following lemma that establishes that the cost-to-go functions are 2τ-Lipschitz with respect to the state-cost function. For pronouncing and proving the statement, it is useful to define the *span seminorm* $\|c\|_s = \max_x c(x) - \min_y c(y)$. Note that it is easy to show that $\|\cdot\|_s$ is indeed a seminorm as it satisfies all the requirements to be a norm except that it maps all constant vectors (and not just zero) to zero.

Lemma 4 *Let f and g be two state-cost functions taking values in the interval $[0,1]$ and let v_f and v_g be the corresponding optimal cost-to-go functions. Then,*

$$\|v_f - v_g\|_s \le 2\tau \|f - g\|_\infty .$$

The proof roughly follows the proof of Proposition 3 of Guan et al. (2014), with the slight difference that we make the constant factor in the bound explicit. A consequence of this result is our final key lemma in this section that actually makes our fast rates possible: a bound on the change-rate of the policies chosen by the algorithm.

Lemma 5 $\max_x \|Q_t(\cdot|x) - Q_{t+1}(\cdot|x)\|_1 \le \frac{\tau}{t}$.

The proof is based on ideas by Guan et al. (2014). As for the proof of Theorem 1, we follow the path of Even-Dar et al. (2009); Neu et al. (2014); Guan et al. (2014), and first analyze the idealized setting where the learner is allowed to directly pick stationary distributions instead of policies. Then, we show how to relate the idealized regret of FTL to its true regret in the original problem.

5.1. Regret in the idealized OCO problem

Let us now consider the idealized online convex optimization problem described at the end of Section 3. In this setting, our algorithm can be formally stated as choosing the stationary transition measure $\pi_t = \arg\min_{\pi\in\Delta(M)} f(\pi;\overline{c}_t)$. This view enables us to follow a standard proof technique for analyzing online convex optimization algorithms, going back to at least Merhav and Feder (1992). The first ingredient of our proof is the so-called "follow-the-leader/be-the-leader" lemma Cesa-Bianchi and Lugosi (2006, Lemma 3.1):

Lemma 6 $\sum_{t=1}^{T} \widetilde{\ell}_t(\pi_{t+1}) \le \min_\pi \sum_{t=1}^{T} \widetilde{\ell}_t(\pi)$.

The second step exploits the bound on the change rate of the policies to show that looking one step into the future does not buy much advantage. Note however that controlling the change rate is not sufficient by itself, as our loss functions are effectively unbounded.

Lemma 7 $\sum_{t=1}^{T} \left(\widetilde{\ell}_t(\pi_t) - \widetilde{\ell}_t(\pi_{t+1})\right) \le 2\left(\tau^2+1\right)(1+\log T)$.

In the interest of space, we only provide a proof sketch here and defer the full proof to Appendix B.5. **Proof sketch** Let us define $\Delta_t = \overline{c}_{t+1} - \overline{c}_t$. By exploiting the affinity of f in its second argument, we can start by proving $\lambda_t - \lambda_{t+1} \le \left\|\Delta_t\right\|_\infty$. Furthermore, by using the form of the optimal policy Q_t given in Eq. (5) and the form of f given in Eq. (3), we can obtain

$$\begin{aligned}\widetilde{\ell}_t(\pi_t) - \widetilde{\ell}_t(\pi_{t+1}) &= \left(\mu_t - \mu_{t+1}\right)^{\mathsf{T}}\left(c_t + \overline{c}_t\right) + \mu_{t+1}^{\mathsf{T}}\left(\overline{c}_t - \overline{c}_{t+1}\right) + \lambda_t - \lambda_{t+1}\\ &\le 2\left\|\mu_{t+1} - \mu_t\right\|_1 + 2\left\|\Delta_t\right\|_\infty .\end{aligned}$$

The first term can be bounded by a simple argument (see, e.g., Lemma 4 of Neu et al. 2014) that leads to

$$\left\|\mu_{t+1} - \mu_t\right\|_1 \le \max\left\{\tau(Q_t), \tau(Q_{t+1})\right\} \max_x \left\|Q_{t+1}(\cdot|x) - Q_t(\cdot|x)\right\|_1 .$$

Now, the first factor can be bounded by τ and the second by appealing to Lemma 5. The proof is concluded by plugging the above bounds into Equation (12), using $\left\|\Delta_t\right\|_\infty \le 1/t$, summing up both sides, and noting that $\sum_{t=1}^{T} 1/t \le 1 + \log T$. ■

Putting Lemmas 6 and 7 together, we obtain the following bound on the idealized regret of FTL:

Lemma 8 $\overline{R}_T \le 2\left(\tau^2+1\right)(1+\log T)$.

5.2. Regret in the reactive setting

We first show that the advantage of the true best policy Q_T^* over our final policy Q_{T+1} is bounded.

Lemma 9 *Let $p^* = \min_{x,x':P(x'|x)>0} P(x'|x)$ be the smallest non-zero transition probability under the passive dynamics and $B = -\log p^*$. Then, $\sum_{t=1}^{T} \overline{\ell}_t(\pi_{T+1}) - \mathcal{L}_T(Q_T^*) \le (2\tau+2)(B+1)$.*

The proof follows from applying Lemma 1 from Neu et al. (2014) and observing that $\ell_t(X_t, Q_T^*) \le B + 1$ holds for all t. It remains to relate the total cost of FTL to the total idealized cost of the algorithm. This is done in the following lemma:

Lemma 10 $\sum_{t=1}^{T} \left(\mathbb{E}\left[\ell_t(Q_t, X_t)\right] - \overline{\ell}_t(\pi_t)\right) \le (\tau+1)^3(1+\log T)^2 + 2(\tau+1)(3+\log T)$.

Proof Let $p_t(x) = \mathbb{P}\left[X_t = x\right]$. Similarly to the proof of Lemma 7, we rewrite $\overline{\ell}_t(\pi_t)$ using Equation (11) to obtain

$$\begin{aligned}\mathbb{E}\left[\ell_t(Q_t, X_t) - \overline{\ell}_t(\pi_t)\right] &= \sum_x \left(p_t(x) - \mu_t(x)\right) \left(c_t(x) + v_t(x) + \lambda_t - \overline{c}_t(x) - \sum_{x'} Q_t(x'|x) v_t(x')\right) \\ &\leq \sum_x p_t(x) \left(v_t(x) - \sum_{x'} Q_t(x'|x) v_t(x')\right) + \left\|p_t - \mu_t\right\|_1,\end{aligned}$$

where the last step uses $\sum_x \mu_t(x) Q_t(x'|x) = \mu_t(x')$ and $\left\|c_t - \overline{c}_t\right\|_\infty \leq 1$. Now, noticing that $\sum_x p_t(x) Q_t(x'|x) = p_{t+1}(x')$, we obtain

$$\begin{aligned}&\sum_{t=1}^T \mathbb{E}\left[\ell_t(Q_t, X_t) - \overline{\ell}_t(\pi_t)\right] \leq \sum_{t=1}^T \left(p_t - p_{t+1}\right)^\mathsf{T} v_t + \sum_{t=1}^T \left\|\mu_t - p_t\right\|_1 \\ &\quad= \sum_{t=1}^T p_t^\mathsf{T}\left(v_t - v_{t-1}\right) + \sum_{t=1}^T \left\|\mu_t - p_t\right\|_1 - p_{T+1}^\mathsf{T} v_T \leq \sum_{t=1}^T \frac{2\tau}{t} + \sum_{t=1}^T \left\|\mu_t - p_t\right\|_1 - p_{T+1}^\mathsf{T} v_T,\end{aligned}$$

where the last inequality uses Lemma 4 and $\left\|\overline{c}_t - \overline{c}_{t-1}\right\|_\infty \leq 1/t$ to bound the first term. By Lemma 4, this last term can be bounded by $\left\|v_T\right\|_s = \left\|v_T - v_0\right\|_s \leq 2\tau \left\|\overline{c}_T\right\|_\infty \leq 2\tau$, where v_0 is the cost-to-go corresponding to the all-zero state-cost function.

In the rest of the proof, we are going to prove the inequality

$$\left\|\mu_t - p_t\right\|_1 \leq 2e^{-(t-1)/\tau} + \frac{2(\tau+1)^3\left(1 + \log t\right)}{t}. \tag{6}$$

It is easy to see that this trivially holds for $\left(2\tau \log t\right)/t \geq 1$, so we will assume that the contrary holds in the following derivations. To prove Equation (6) for larger values of t, we can follow the proofs of Lemma 5 of Neu et al. (2014) or Lemma 5.2 of Even-Dar et al. (2009) to obtain

$$\left\|\mu_t - p_t\right\|_1 \leq 2e^{-(t-1)/\tau} + \tau\left(\tau+1\right) \sum_{n=1}^{t-1} \frac{e^{-(t-n)/\tau}}{n}. \tag{7}$$

For completeness, we include a proof in Appendix B.6. For bounding the last term, we split the sum at $B = \lfloor t - \tau \log t \rfloor$:

$$\begin{aligned}\sum_{n=1}^{t-1} \frac{e^{-(t-n)/\tau}}{n} &= \sum_{n=1}^{B} \frac{e^{-(t-n)/\tau}}{n} + \sum_{n=B+1}^{t-1} \frac{e^{-(t-n)/\tau}}{n} \\ &= e^{-(t-B)/\tau} \sum_{n=1}^{B} \frac{e^{-(B-n)/\tau}}{n} + \sum_{n=B+1}^{t} \frac{e^{-(t-n)/\tau}}{n} \\ &\leq \frac{1}{t} \cdot \frac{1}{1 - e^{-1/\tau}} + \frac{\tau \log t}{t - \tau \log t} \leq \frac{\tau}{t} + \frac{\tau \log t}{t} \cdot \frac{1}{1 - \left(\tau \log t\right)/t} \\ &\leq \frac{\tau}{t} + \frac{2\tau \log t}{t} \leq \frac{2\tau\left(1 + \log t\right)}{t},\end{aligned}$$

where the first inequality follows from bounding the $1/n$ factors by 1 and $1/B$, respectively, and bounding the sums by the full geometric sums. The second-to-last inequality follows from our assumption that $(2\tau \log t)/t \leq 1$. That is, we have successfully proved Equation (6). Now the statement of the lemma follows from summing up for all t and noting that $\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log T$ and $\sum_{t=1}^{T} e^{-(t-1)/\tau} \leq \tau + 1$. ■

Now the proof of Theorem 1 follows easily from combining the bounds of Lemmas 8–10. The result is

$$R_T \leq 2\left(\tau+1\right)^3\left(1+\log T\right)^2 + 2\left(\tau^2+\tau+2\right)\left(3+\log T\right) + \left(2\tau+2\right)\left(B+2\right). \tag{8}$$

Thus, we can see that the bound indeed demonstrates a polynomial dependence on the mixing time τ, and depends logarithmically on the smallest non-zero transition probability p^* via $B = -\log p^*$.

6. Discussion

In this paper, we have shown that, besides the well-established computational advantages, linearly solvable MDPs also admit a remarkable information-theoretic advantage: fast learnability in the online setting. In particular, we show that achieving a regret of $O(\log^2 T)$ is achievable by the simple algorithm of following the leader, thus greatly improving on the best previously known regret bounds of $O(T^{3/4})$. At first sight, our improvement may appear dramatic: in their paper, Guan et al. (2014) pose the possibility of improving their bounds to $O(\sqrt{T})$ as an important open question (Sec. VII.). In light of our results, these conjectured improvements are also grossly suboptimal. On the other hand, our new results can be also seen to complement well-known results on fast rates in online learning (see, e.g., van Erven et al. 2015 for an excellent summary). Indeed, our learning setting can be seen as a generalized variant of sequential prediction under the relative-entropy loss (see, e.g., Cesa-Bianchi and Lugosi, 2006, Sec. 3.6), which is known to be *exp-concave*. Such exp-concave losses are well-studied in the online learning literature, and are known to allow logarithmic regret bounds (Kivinen and Warmuth, 1999; Hazan et al., 2007).

Inspired by these related results, we ask the question: Is the loss function f defined in Section 2.2 exp-concave? While our derivations Appendix A.1 indicate that f has curvature in certain directions, we were not able to prove its exp-concavity. Similarly to the approach of Merhav and Feder (1992), our analysis in the current paper merely exploits the Lipschitzness of the optimal policies with respect to the cost functions, but otherwise does not explicitly make use of the curvature of f. We hope that our work presented in this paper will inspire future studies that will clarify the exact role of the LMDP structure in efficient online learnability, potentially also leading to a better understanding of policy gradient algorithms for LMDPs (Todorov, 2010).

Finally, let us comment on the tightness of our bounds. Regardless of whether the loss function f is exp-concave or not, we are almost certain that our rates can be improved to at least $O(\log T)$ by using a more sophisticated algorithm. While our focus in this paper was on improving the asymptotic regret guarantees, we also slightly improve on the results Guan et al. (2014) in that we make the leading constants more explicit. However, we expect that the dependence on these constants may also be improved in future work. Note however that the potential looseness of our bounds does not impact the performance of the algorithm itself, as it never makes use of any problem-dependent constants.

Acknowledgements This work was supported by the UPFellows Fellowship (Marie Curie COFUND program n° 600387) and the Ramon y Cajal program RYC-2015-18878 (AEI/MINEICO/FSE, UE). The authors wish to thank the three anonymous reviewers for their valuable comments that helped to improve the paper.

References

Y. Abbasi-Yadkori, P. L. Bartlett, X. Chen, and A. Malek. Large-scale Markov decision problems with KL control cost and its application to crowdsourcing. In *32nd International Conference on Machine Learning (ICML) 2015*, pages 1053–1062, 2015.

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

Y. Abbasi-Yadkori, P. Bartlett, and V. Kanade. Tracking adversarial targets. In *ICML 2014*, pages 369–377, 2014.

Y. Ariki, T. Matsubara, and S. H. Hyon. Latent Kullback-Leibler control for dynamic imitation learning of whole-body behaviors in humanoid robots. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 946–951, 2016.

P. L. Bartlett and A. Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI 2009*, 2009.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 3 edition, 2007.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

S. de Rooij, T. van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *Accepted to the Journal of Machine Learning Research*, 2014.

T. Dick, A. György, and Cs. Szepesvári. Online learning in markov decision processes with changing cost sequences. In *ICML 2014*, 2014.

K. Dvijotham and E. Todorov. Inverse optimal control with linearly-solvable mdps. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 335–342, 2010.

E. Even-Dar, S. M. Kakade, and Y. Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

V. Gómez, H. J. Kappen, J. Peters, and G. Neumann. Policy search for path integral control. *European Conference on Machine Learning and Knowledge Discovery in Databases*, 8724 LNAI (PART 1):482–497, 2014.

V. Gómez, S. Thijssen, A. C. Symington, S. Hailes, and H. J. Kappen. Real-time stochastic optimal control for multi-agent quadrotor systems. In *26th International Conference on Automated Planning and Scheduling*, 2016.

P. Guan, M. Raginsky, and R. M. Willett. Online markov decision processes with kullback–leibler control cost. *Automatic Control, IEEE Transactions on*, 59(6):1423–1438, 2014.

E. Hazan. The convex optimization approach to regret minimization. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 287–303. MIT press, 2011.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010. ISSN 1532-4435.

H. J. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95 (20):200201, 2005.

H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

K. Kinjo, E. Uchibe, and K. Doya. Evaluation of linearly solvable Markov decision process with dynamic model learning in a mobile robot navigation task. *Frontiers in Neurorobotics*, 7:1–13, 2013.

J. Kivinen and M. Warmuth. Averaging expert predictions. In *Proceedings of the Fourth European Conference on Computational Learning Theory*, pages 153–167. Lecture Notes in Artificial Intelligence, Vol. 1572. Springer, 1999.

W. Kotłowski. On minimaxity of follow the leader strategy in the stochastic setting. In *International Conference on Algorithmic Learning Theory*, pages 261–275, 2016.

T. Matsubara, V. Gómez, and H. J. Kappen. Latent Kullback Leibler control for continuous-state systems using probabilistic graphical models. *30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

N. Merhav and M. Feder. Universal sequential learning and decision from individual data sequences. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press, 1992.

C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.

G. Neu, A. György, and Cs. Szepesvári. The online loop-free stochastic shortest-path problem. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 231–243, 2010.

G. Neu, A. György, and Cs. Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *AISTATS 2012*, pages 805–813, 2012.

G. Neu, A. György, Cs. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, 2014.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, April 1994.

E. Rombokas, M. Malhotra, E. A. Theodorou, E. Todorov, and Y. Matsuoka. Reinforcement learning and synergistic control of the act hand. *IEEE/ASME Transactions on Mechatronics*, 18(2):569–577, 2013.

A. Sani, G. Neu, and A. Lazaric. Exploiting easy data in online optimization. In *NIPS-27*, pages 810–818, 2014.

E. Seneta. *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Cs. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

D. Thalmeier, V. Gómez, and H. J. Kappen. Action selection in growing state spaces: control of network structure growth. *Journal of Physics A: Mathematical and Theoretical*, 50(3):034006, 2017.

E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral control approach to reinforcement learning. *Journal of Machine Learning Research*, 11:3137–3181, 2010.

E. Todorov. Linearly-solvable Markov decision problems. In *NIPS-18*, pages 1369–1376, 2006. ISBN 0-262-23253-7.

E. Todorov. General duality between optimal control and estimation. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 4286–4292. IEEE, 2008.

E. Todorov. Compositionality of optimal control laws. In *NIPS-22*, pages 1856–1864, 2009.

E. Todorov. Policy gradients in linearly-solvable mdps. In *NIPS-23*, pages 2298–2306. CURRAN, 2010.

T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou. Aggressive driving with model predictive path integral control. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1433–1440, May 2016. doi: 10.1109/ICRA.2016.7487277.

J. Y. Yu, S. Mannor, and N. Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

A. Zimin and G. Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *NIPS-26*, pages 1583–1591, 2013.

Appendix A. The convex optimization view of optimal control in LDMPs

This section summarizes some facts regarding the convex optimization formulation of Section 2.2. We first show that the negative conditional entropy constituting the only nonlinear term in the objective $f(\pi; c)$ is convex.

A.1. The convexity of the negative conditional entropy

Let us consider the joint probability distribution π on the finite set $\mathcal{X}^2$. We denote $\mu(x) = \sum_y p(x,y)$ and $Q(y|x) = p(x,y)/\mu(x)$. We study the negative conditional entropy of $(X,Y) \sim \pi$ as a function of π:

$$R(\pi) = \sum_{x,y} \pi(x,y) \log \frac{\pi(x,y)}{\sum_{y'} \pi(x,y')} = \sum_{x,y} \pi(x,y) \log \frac{\pi(x,y)}{\mu(x)}$$

We will study the Bregman divergence B_R corresponding to R:

$$B_R\left(\pi' \middle\| \pi\right) = R(\pi') - R(\pi) - \nabla R(\pi)^{\mathsf{T}}(\pi' - \pi).$$

Our aim is to show that B_R is nonnegative, which will imply the convexity of R.

We begin by computing the partial derivative of $R(\pi)$ with respect to $\pi(x,y)$:

$$\frac{\partial R(\pi)}{\partial \pi(x,y)} = \log\left(\pi(x,y)\right) - \log\left(\mu(x)\right),$$

where we used the fact that $\frac{\partial \mu(x)}{\partial \pi(x,y)} = 1$ for all y. With this expression, we have $-H(Y|X)$:

$$\begin{aligned} R(\pi) + \nabla R(\pi)^{\mathsf{T}}(\pi' - \pi) &= \sum_{x,y} \pi(x,y) \log \frac{\pi(x,y)}{\mu(x)} + \sum_{x,y} \big(\pi'(x,y) - \pi(x,y)\big) \log \frac{\pi(x,y)}{\mu(x)} \\ &= \sum_{x,y} \pi'(x,y) \log \frac{\pi(x,y)}{\mu(x)}. \end{aligned}$$

Thus, the Bregman divergence takes the form

$$\begin{aligned} B_R\left(\pi' \middle\| \pi\right) &= \sum_{x,y} \pi'(x,y) \left(\log \frac{\pi'(x,y)}{\mu'(x)} - \log \frac{\pi(x,y)}{\mu(x)} \right) \\ &= \sum_{x,y} \pi'(x,y) \log \frac{Q'(y|x)}{Q(y|x)} = \sum_x \mu'(x) \sum_y Q'(y|x) \log \frac{Q'(y|x)}{Q(y|x)} \\ &= \sum_x \mu'(x) D\left(Q'(\cdot|x) \middle\| Q(\cdot|x)\right) \geq \frac{1}{2} \sum_x \mu'(x) \left\| Q'(\cdot|x) - Q(\cdot|x) \right\|_1^2, \end{aligned}$$

where the last step follows from Pinsker's inequality. Thus, we have shown that the Bregman divergence B_R is nonnegative on $\Delta(\mathcal{X}^2)$, proving that $R(\pi)$ is convex.

A.2. Derivation of the optimal control

Here, we give an alternative derivation of the optimal control given in Section 2.1 based on the optimization problem $\min_{\pi\in\Delta(M)} f(\mu; c)$ for an arbitrary bounded state-cost function c. As a reminder, $f(\pi; c)$ is given by

$$f(\pi; c) = \sum_{x,x'} \pi(x,x') \left(c(x) + \log \frac{\pi(x,x')}{P(x'|x) \sum_y \pi(x,y)} \right)$$

and the feasible set $\Delta(M)$ is given by the following convex constraints:

$$\begin{aligned}
\sum_{x'} \pi(x,x') &= \sum_{x''} \pi(x'',x) && (\forall x),\\
\sum_{x,x'} \pi(x,x') &= 1, &&\\
\pi(x,x') &\geq 0 && (\forall x, x'),\\
\pi(x,x') &= 0 && (\forall x, x' : P(x'|x) = 0).
\end{aligned}$$

We begin by slightly adjusting the definition of $f(\cdot; c)$ for it to become a *barrier function*: we set $f(\pi; c) = \infty$ for all π not satisfying the last two constraints. It is easy to see that this adjustment does not change the optimum of $f(\cdot; c)$, but it helps getting rid of the inequality constraints. Thus, with this form of f, we can characterize the optimum of $f(\cdot; c)$ using the technique of Lagrange multipliers[5].

Precisely, we introduce a Lagrange multiplier $v(x)$ for every x to enforce the first constraint and λ to enforce the second one, and write the Lagrangian as

$$\begin{aligned}
\mathcal{L}(\pi; v, \lambda) =& \sum_{x,x'} \pi(x,x') \left(c(x) + \log \frac{\pi(x,x')}{P(x'|x) \sum_y \pi(x,y)} \right) + \lambda \left(\sum_{x,x'} \pi(x,x') - 1 \right)\\
&+ \sum_{x,x'} v(x) \left(\pi(x,x') - \pi(x',x) \right)\\
=& \sum_{x,x'} \pi(x,x') \log \frac{\pi(x,x')}{P(x'|x) \sum_y \pi(x,y)} + \sum_{x,x'} \pi(x,x') \left(c(x) + \lambda + v(x) - v(x') \right) - \lambda
\end{aligned}$$

Let $\mu(x) = \sum_y \pi(x,y)$, noting that $\partial\mu(x)/\partial\pi(x,y) = 1$ for all y. Differentiate the Lagrangian with respect to a fixed $\pi(x,x')$:

$$\frac{\partial \mathcal{L}(\pi; v, \lambda)}{\partial \pi(x,x')} = \log \left(\pi(x,x') \right) - \log \left(\mu(x) \right) + \left(c(x) + \lambda + v(x) - v(x') - \log P(x'|x) \right).$$

Setting the gradient to zero, we obtain the following formula for $\pi(x,x')/\mu(x)$:

$$\frac{\pi(x,x')}{\mu(x)} = P(x'|x) \cdot \exp \left(-c(x) - \lambda - v(x) + v(x') \right).$$

[5] Alternatively, one could introduce KKT multipliers for all constraints and eliminate the last two by complementary slackness, which yields the same characterization.

Since $\pi(x,y) \geq 0$ for all x, y, we have $\sum_{x'} \frac{\pi(x,x')}{\sum_y \pi(x,y)} = 1$ and thus

$$\sum_{x'} P(x'|x) \exp\big(-c(x) - \lambda + v(x') - v(x)\big) = 1.$$

Introducing the variables $z(x)$ for all x, we recover the linear system of equations in Equation (2):

$$z(x) = \sum_{x'} P(x'|x) \exp\left(-c(x) - \lambda\right) z(x').$$

Plugging back into the Lagrangian, we obtain that the dual function is $\mathcal{L}(\lambda) = -\lambda$, which now needs to be maximized subject to the above constraint, implying that $\exp(-\lambda)$ is indeed the largest eigenvalue of the matrix GP. Furthermore, by strong duality, the λ maximizing the dual is indeed the minimum of the primal $f(\pi; c)$ on $\Delta(M)$.

Appendix B. Technical proofs

In this section, we prove the preliminary lemmas from Section 5.

B.1. The proof of Lemma 2

The idea of the proof is similar to the proof of Theorem 4 of Bartlett and Tewari (2009). By our Assumption 1 and the fact that all feasible control policies retain the structural properties of the passive dynamics, our MDP satisfies the conditions of Proposition 4.3.2 of Bertsekas (2007), so *value iteration* converges to the solution of the Bellman optimality equations. Let $J_n(x)$ be the total expected cost of the best n-horizon policy started output by the value iteration procedure and let q_n denote the corresponding policy. Then, consider the strategy of following the passive dynamics from x until hitting y and then switching to the optimal finite-horizon policy optimized for the remaining rounds. By the finite-horizon optimality of q_n, this strategy is clearly suboptimal: letting L denote the *random* number of steps taken for reaching y under the passive dynamics, this suboptimality can be expressed as $J_n(x) \leq \mathbb{E}\left[L + J_{n-L}(y)\right]$. Thus, by the fact that value iteration converges, we have

$$\begin{aligned} v(x) - v(y) &= \lim_{n\to\infty} \big(J_n(x) - J_n(y)\big) \\ &\leq \lim_{n\to\infty} \big(\mathbb{E}\left[L + J_{n-L}(y) - J_n(y)\right]\big) \leq H, \end{aligned}$$

where we used that $J_k(y) \leq J_n(y)$ for all $k \leq n$ and that $\mathbb{E}\left[L\right] \leq H$ by Assumption 1. This concludes the proof of the first statement. As for the second statement, note that the associated optimal policy Q can be written as

$$Q(x'|x) = P(x'|x) \exp\big(\lambda - c(x) - v(x') + v(x)\big), \tag{9}$$

so the control cost can be bounded for all x as

$$D\left(Q(\cdot|x) \| P(\cdot|x)\right) = v(x) + \lambda - c_t(x) - \sum_{x'} Q(x'|x) v(x') \leq H + 1,$$

thus concluding the proof. ■

B.2. The proof of Lemma 3

It is well known (see, e.g., Seneta (2006)) that the Markov-Dobrushin coefficient $\alpha(P)$ of a transition kernel P satisfies

$$\alpha(P) = 1 - \min_{x,y\in\mathcal{X}} \sum_s \min\left\{P(s|x), P(s|y)\right\}$$

Now, let us observe that by the expression (9), we can prove the statement of the lemma as

$$\begin{aligned}
\alpha(Q) =& 1 - \min_{x,y\in\mathcal{X}} \sum_s \min\left\{Q(s|x), Q(s|y)\right\} \\
=& 1 - \min_{x,y\in\mathcal{X}} \sum_k \min\left\{P(s|x)\exp\left(v(x) - c(x)\right), P(s|y)\exp\left(v(y) - c(y)\right)\right\}\exp\left(\lambda - v(s)\right) \\
\le& 1 - \min_{x,y\in\mathcal{X}} \sum_k \min\left\{P(s|x), P(s|y)\right\}\exp\left(\lambda - v(s) - 1 + \min_{s'} v(s')\right) \\
\le& 1 - \min_{x,y\in\mathcal{X}} \sum_k \min\left\{P(s|x), P(s|y)\right\} e^{-H-1} = e^{-1/\tau} + \left(1 - e^{-1/\tau}\right)\left(1 - e^{-H-1}\right),
\end{aligned}$$

where the last inequality follows from applying Lemma 2 to show $v(s) - v(s') \le H$ for this particular pair of states. ■

B.3. The proof of Lemma 4

The proof roughly follows the proof of Proposition 3 of Guan et al. (2014). Let us define the Bellman optimality operator $B_c : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ associated with the state-cost function c that acts on any cost-to-go v as $\left(B_c v\right)(x) = c(x) + \min_{q\in\Delta(\mathcal{X})}\left\{D\left(q\| P(\cdot|x)\right) + \sum_y q(y)v(y)\right\}$. With this operator, we can express the Bellman optimality equations for v_f as $v_f + \lambda_f = B_c v_f$. Thus, we have

$$\begin{aligned}
\left\|v_f - v_g\right\|_s &= \left\|B_f v_f - \lambda_f - B_g v_g + \lambda_g\right\|_s \\
&= \left\|B_f v_f - B_g v_g\right\|_s \le \left\|B_f v_f - B_g v_f\right\|_s + \left\|B_g v_f - B_g v_g\right\|_s,
\end{aligned} \tag{10}$$

where the second equality follows from the fact that the span seminorm is insensitive to shifting by constants and the last step follows from the triangle inequality. The first term in the above expression can be easily bounded noting that $B_f v_f - B_g v_f = f - g$ by the definition of the Bellman operator. For the second term, we can follow the argument of Guan et al. (2014) to show that

$$\left\|B_g v_f - B_g v_g\right\|_s \le \frac{1}{2}\left\|v_f - v_g\right\|_s \max_{x,y}\left\|Q_f(\cdot|x) - Q_g(\cdot|y)\right\|_1.$$

Now for controlling the last factor in the above expression, we take an arbitrary pair of states x and y and write

$$\begin{aligned}
&\frac{1}{2}\sum_s \left|Q_f(s|x) - Q_g(s|y)\right| = \frac{1}{2}\sum_s \left(Q_f(s|x) + Q_g(s|y)\right) - \sum_s \min\left\{Q_f(s|x), Q_g(s|y)\right\} \\
&= 1 - \sum_s \min\left\{P(s|x)\exp(v_f(x) - f(x) + \lambda_f - v_f(s)), P(s|y)\exp(v_g(x) - g(x) + \lambda_f - v_g(s))\right\} \\
&\le 1 - \sum_s \min\left\{P(s|x), P(s|y)\right\} e^{-H-1} = \alpha(P) + \left(1 - \alpha(P)\right)\left(1 - e^{-H-1}\right) = \alpha,
\end{aligned}$$

where the first step uses the equality $|a+b| = \frac{1}{2}|a+b| - \min\{a,b\}$ that holds for any two real numbers a, b, the second step follows from Equation (9), the inequality from Lemma 2, and the last steps use the respective definitions of $\alpha(P)$ and α. In summary, we have proved that $\left\|B_g v_f - B_g v_g\right\|_s \leq \alpha \left\|v_f - v_g\right\|_s$ holds for the worst-case ergodicity coefficient $\alpha < 1$. Plugging this bound back into Equation (10) gives $\left\|v_f - v_g\right\|_s \leq \frac{1}{1-\alpha}\left\|f-g\right\|_s$, which can be seen to imply the statement of the lemma after observing $\left\|f-g\right\|_s \leq 2\left\|f-g\right\|_\infty$ and $\frac{1}{1-\alpha} \leq \tau$. ■

B.4. The proof of Lemma 5

The proof builds on the following lemma, adapted from Proposition 4 of Guan et al. (2014):

Lemma 11 *Let f and g be two state-cost functions taking values in the interval $[0,1]$, let v_f and v_g be the corresponding optimal cost-to-go functions and Q_f and Q_g be the respective optimal policies. Then,*

$$\max_{x\in\mathcal{X}} \left\|Q_f(\cdot|x) - Q_g(\cdot|x)\right\|_1 \leq \frac{1}{2}\left\|v_f - v_g\right\|_s.$$

Proof Let us study the relative entropy between $Q_f(\cdot|x)$ and $Q_g(\cdot|x)$ for any fixed x:

$$\begin{aligned}
D\left(Q_f(\cdot|x)\middle\|\, Q_g(\cdot|x)\right) &= \sum_y Q_f(y|x) \log \frac{Q_f(y|x)}{Q_g(y|x)} \\
&= \sum_y Q_f(y|x)\left(v_g(y) - v_f(y)\right) + \log \frac{\sum_y P(y|x)e^{-v_g(y)}}{\sum_y P(y'|x)e^{-v_f(y')}} \\
&= \sum_y Q_f(y|x)\left(v_g(y) - v_f(y)\right) + \log \frac{\sum_y P(y|x)e^{-v_f(y)}}{\sum_y P(y'|x)e^{-v_f(y')}} e^{-v_g(y)+v_f(y)} \\
&\leq \frac{\left\|v_g - v_f\right\|_s^2}{8},
\end{aligned}$$

where the second equality follows from straightforward calculations and the last step follows from Hoeffding's lemma (see, e.g., Lemma A.1 in Cesa-Bianchi and Lugosi 2006). Now the statement of the lemma follows from applying Pinsker's inequality. ■

Now, we can conclude the proof of Lemma 5 by combining Lemmas 4 and 11 with the easily-seen fact $\left\|\bar{c}_{t+1} - \bar{c}_t\right\|_\infty \leq 1/t$. ■

B.5. The proof of Lemma 7

Observe that, by the definition of the algorithm and the form of Q_t given by (5), we have

$$D\left(Q_t(\cdot|x)\middle\|\, P(\cdot|x)\right) = v_t(x) + \lambda_t - \bar{c}_t(x) - \sum_{x'} Q_t(x'|x)v_t(x'),$$

which, by using $\widetilde{\ell}_t = f(\cdot; c_t)$ and the form of f given in Eq. (3) implies that

$$\begin{aligned}\widetilde{\ell}_t(\pi_t) &= \sum_x \mu_t(x) \left(c_t(x) + v_t(x) + \lambda_t - \overline{c}_t(x) - \sum_{x'} Q_t(x'|x) v_t(x') \right) \\ &= \sum_x \mu_t(x) \left(c_t(x) + \lambda_t - \overline{c}_t(x) \right),\end{aligned} \tag{11}$$

where the second equality follows from the fact that $\sum_x \mu_t(x) Q_t(x'|x) = \mu_t(x')$. Combining this with the analogous expression for $\widetilde{\ell}_t(\pi_{t+1})$, we get

$$\begin{aligned}\widetilde{\ell}_t(\pi_t) - \widetilde{\ell}_t(\pi_{t+1}) &= \left(\mu_t - \mu_{t+1}\right)^{\mathsf{T}} c_t + \mu_t^{\mathsf{T}} \overline{c}_t - \mu_{t+1}^{\mathsf{T}} \overline{c}_{t+1} + \lambda_t - \lambda_{t+1} \\ &= \left(\mu_t - \mu_{t+1}\right)^{\mathsf{T}} \left(c_t + \overline{c}_t\right) + \mu_{t+1}^{\mathsf{T}} \left(\overline{c}_t - \overline{c}_{t+1}\right) + \lambda_t - \lambda_{t+1} \\ &\le 2 \left\|\mu_{t+1} - \mu_t\right\|_1 + \left\|\overline{c}_{t+1} - \overline{c}_t\right\|_\infty + \lambda_t - \lambda_{t+1}.\end{aligned} \tag{12}$$

It remains to bound the last two terms. Defining $\Delta_t = \overline{c}_{t+1} - \overline{c}_t$, we have

$$\begin{aligned}\lambda_{t+1} = \min_\pi f(\pi; \overline{c}_{t+1}) &= \min_\pi \Big(f(\pi; \overline{c}_t) + \sum_{x,x'} \pi(x,x') \Delta_t(x) \Big) \\ &\ge \min_\pi f(\pi; \overline{c}_t) + \min_{\pi'} \sum_{x,x'} \pi'(x,x') \Delta_t(x) \ge \lambda_t - \left\|\Delta_t\right\|_\infty,\end{aligned}$$

where the second equality uses the form of f, and the last step uses the fact that π' is a probability distribution over $\mathcal{X} \times \mathcal{X}$. This gives

$$\widetilde{\ell}_t(\pi_t) - \widetilde{\ell}_t(\pi_{t+1}) \le 2 \left\|\mu_{t+1} - \mu_t\right\|_1 + 2 \left\|\Delta_t\right\|_\infty.$$

It remains to bound $\left\|\mu_{t+1} - \mu_t\right\|_1$. A simple argument (see, e.g., Lemma 4 of Neu et al. 2014) shows that

$$\left\|\mu_{t+1} - \mu_t\right\|_1 \le \max\left\{\tau(Q_t), \tau(Q_{t+1})\right\} \max_x \left\|Q_{t+1}(\cdot|x) - Q_t(\cdot|x)\right\|_1.$$

Now, the first factor can be bounded by τ and the second by appealing to Lemma 5. The proof is concluded by plugging the above bounds into Equation (12), using $\left\|\Delta_t\right\|_\infty \le 1/t$, summing up both sides, and noting that $\sum_{t=1}^T 1/t \le 1 + \log T$. ■

B.6. The proof of inequality (7)

We will now prove the inequality

$$\left\|\mu_t - p_t\right\|_1 \le 2e^{-(t-1)/\tau} + \tau\left(\tau + 1\right) \sum_{n=1}^{t-1} \frac{e^{-(t-n)/\tau}}{n}.$$

If $t = 1$, the inequality clearly holds as $\left\|\mu_1 - p_1\right\|_1 \le 2$. We let $\varepsilon_t = \max_x \left\|Q(\cdot|x) - Q_{t-1}(\cdot|x)\right\|_1$. By the triangle inequality, we have

$$\begin{aligned}\left\|p_t - \mu_t\right\|_1 &\le \left\|p_t - \mu_{t-1}\right\|_1 + \left\|\mu_{t-1} - \mu_t\right\|_1 \\ &\le e^{-1/\tau} \left\|p_{t-1} - \mu_{t-1}\right\|_1 + \left(\tau + 1\right) \varepsilon_t,\end{aligned}$$

where we used the fact that $p_t = p_{t-1}^\mathsf{T} Q_t$, and $\left\|\mu_{t-1} - \mu_t\right\|_1 \le (\tau+1)\,\varepsilon_t$, which follows from Lemma 4 of Neu et al. (2014). Continuing recursively, we obtain

$$\begin{aligned}
\left\|p_t - \mu_t\right\|_1 &\le e^{-1/\tau}\left(e^{-1/\tau}\left\|p_{t-2} - \mu_{t-2}\right\|_1 + (\tau+1)\,\varepsilon_{t-1}\right) + (\tau+1)\,\varepsilon_t \\
&\vdots \\
&\le e^{-t-1/\tau}\left\|p_1 - \mu_1\right\|_1 + (\tau+1)\sum_{n=1}^{t} \varepsilon_n e^{-(t-n)/\tau} \\
&\le e^{-t-1/\tau}\left\|p_1 - \mu_1\right\|_1 + (\tau+1)\sum_{n=1}^{t} \frac{\tau e^{-(t-n)/\tau}}{n},
\end{aligned}$$

where the last inequality follows from $\varepsilon_t \le \tau/t$, which holds by Lemma 5.