Topics in Online Markov Decision Processes

by

Peng Guan

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Maxim Raginsky, Supervisor

_____
Rebecca Willett, Co-Chair

_____
Henry Pfister, Co-Chair

_____
Mauro Maggioni

_____
Vincent Conitzer

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University
2015

Abstract

Topics in Online Markov Decision Processes

by

Peng Guan

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Maxim Raginsky, Supervisor

_____
Rebecca Willett, Co-Chair

_____
Henry Pfister, Co-Chair

_____
Mauro Maggioni

_____
Vincent Conitzer

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2015

Abstract

This dissertation describes sequential decision making problems in non-stationary environments. Online learning algorithms deal with non-stationary environments, but generally there is no notion of a dynamic state to model future impacts of past actions. State-based models are common in stochastic control settings, but well-known frameworks such as Markov decision processes (MDPs) assume a known stationary environment. In recent years, there has been a growing interest in fusing the above two important learning frameworks and considering an MDP setting in which the cost function is allowed to change arbitrarily over time. A number of online MDP algorithms have been designed to work under various assumptions about the dynamics of state transitions so far and provide performance guarantees, i.e. bounds on the regret defined as the performance gap between the total cost incurred by the learner and the total cost of the best available stationary policy that could have been chosen in hindsight. However, most of the work in this area has been algorithmic: given a problem, one would develop an algorithm almost from scratch and prove the performance guarantees on a case-by-case basis. Moreover, the presence of the state and the assumption of an arbitrarily varying environment complicate both the theoretical analysis and the development of computationally efficient methods. Another potential issue is that, by removing distributional assumptions about the mechanism generating the cost sequences, the existing methods have to consider the worst-case scenario, which may render their solutions too conservative in situations

where the environment exhibits some degree of predictability.

This dissertation contributes several novel techniques to address the above challenges of the online MDP framework and opens up new research directions for online MDPs. Our proposed general framework for deriving algorithms in the online MDP setting leads to a unifying view of existing methods and provides a general procedure for constructing new ones. Several new algorithms are developed and analyzed using this framework. We develop convex-analytical algorithms that take advantage of possible regularity of observed sequences, yet maintain the worst case performance guarantees. To further study the convex-analytic methods we applied above, we take a step back to consider the traditional MDP problem and extend the LP approach to MDPs by adding a relative entropy regularization term. A computationally efficient algorithm for this class of MDPs is constructed under mild assumptions on the state transition models. Two-player zero-sum stochastic games are also investigated in this dissertation as an important extension of the online MDP setting. In short, this dissertation provides in-depth analysis of the online MDP problem and answers several important questions in this field.

Contents

Abstract **iv**

List of Tables **xi**

List of Figures **xii**

List of Abbreviations and Symbols **xiii**

Acknowledgements **xv**

1 Introduction **1**

1.1 Review of existing learning models 6

1.2 Contributions of the dissertation 9

1.3 Notation 11

2 From minimax value to low-regret algorithms **15**

2.1 Introduction 16

2.1.1 A summary of results 19

2.2 Problem formulation 21

2.2.1 Minimax value 24

2.2.2 Major challenges 29

2.3 The general framework for constructing algorithms in online MDPs . 30

2.3.1 Stationarization 31

2.3.2 The value-function approach 34

2.3.3 The convex-analytic approach 42

2.4 Example derivations of explicit algorithms 46
2.4.1 The value-function approach 47
2.4.2 The convex-analytic approach 65
2.5 Conclusions 74
3 Online MDPs with predictable sequences 75
3.1 Introduction 75
3.2 Problem setup 76
3.3 The proposed strategy 78
3.3.1 Algorithm description 79
3.3.2 Main result 80
3.4 Conclusion 84
4 Information projection onto the state-action polytope 85
4.1 Introduction 86
4.2 Entropy-regularized MDPs 88
4.3 Problem formulation 89
4.3.1 Motivation 92
4.4 Main results on the Gibbs policy 94
4.4.1 Embedding into an exponential family 94
4.4.2 Basic properties of the Gibbs policy 95
4.4.3 Sensitivity to perturbations in f and β 98
4.4.4 Canonical parameters and the relative value function 100
4.5 Conclusion 104
5 Online MDPs with Kullback-Leibler control cost 105
5.1 Introduction 105
5.1.1 Our contributions and comparison with relevant literature . . 110

5.1.2 Organization of the chapter . . . 113
5.1.3 Notation . . . 113
5.2 Problem formulation and the main result . . . 114
5.2.1 The model . . . 114
5.2.2 Strategies and regret . . . 115
5.2.3 The main result . . . 116
5.3 Preliminaries . . . 118
5.3.1 Linearly solvable MDPs . . . 119
5.3.2 Some properties of Todorov's optimal policy . . . 121
5.4 The proposed strategy . . . 132
5.5 Proof of Theorem 21 . . . 133
5.5.1 The main idea . . . 133
5.5.2 Preliminary lemmas . . . 134
5.5.3 Details . . . 137
5.6 Simulations . . . 143
5.7 Conclusion and future work . . . 147
6 No-regret algorithms for zero-sum stochastic games 149
6.1 Problem formulation . . . 152
6.1.1 Preliminaries . . . 153
6.1.2 From online linear optimization to a single-stage game . . . 155
6.2 Regret minimization in stochastic games . . . 157
6.2.1 Stationary strategies for finite-horizon stochastic games . . . 157
6.2.2 Stationary strategies for infinite-horizon stochastic games . . . 162
6.3 Conclusion . . . 163

7 Conclusion **164**

7.1 Summary of contributions 164

7.2 Future work 166

Bibliography **168**

Biography **175**

List of Tables

1.1 Advantages of the online MDP setting over existing methods. 9

List of Figures

5.1 Regret versus time. The red curve shows the average of the regret (the difference between the total cost of our strategy up to each time t and the total cost of the best stationary policy up to that time) over 100 independent realizations of the simulation. At each time t, the height of the gray region corresponds to one sample standard deviation. 145

5.2 Comparison of our proposed strategy to the best stationary policy in a set of 10^5 randomly sampled policies. The red curve shows the average of the regret (the difference between the total cost of our strategy up to each time t and the total cost of the best stationary policy up to that time) over 100 independent realizations of the simulation. At each time t, the height of the gray area corresponds to one sample standard deviation. 147

List of Abbreviations and Symbols

Symbols

Here is a quick reference for the notation used throughout this document. Some notations will be introduced in more details later on.

X	Finite state space of a Markov decision process
x	A specific state
U	Action space of a Markov decision process
u	A specific action
$\mathcal{P}(A)$	Set of all probability distributions on set A
P	A Markov matrix or policy
$\mathcal{M}$	Set of Markov matrices
K	Stochastic transition kernel of a Markov decision process
T	Time horizon
f	Real-valued function
h	Relative value function
$\mathcal{F}$	Set of functions
μ, ν	Probability measures
π	Invariant state distribution
$\boldsymbol{\gamma}$	Behavioral strategy
R	Regret

V	Value of a game
ϵ	Rademacher random variable with $\Pr(\varepsilon = \pm 1) = 1/2$.
$\mathcal{R}$	Rademacher complexity
$\widehat{V}, \widehat{W}$	Relaxation of conditional value of a game
β	Learning rate
c	Cost function
Ψ	Comparator term
Φ	Regularization
$\mathcal{G}$	State-action polytope
P^*	Passive dynamics
$\check{P}$	Twisted kernel
α	Dobrushin ergodicity coefficient
$\mathbb{T}$	Dynamic programming operator

Abbreviations

ACOE	Average cost optimality equation
FPE	Frobenius–Perron eigenvalue
FRL	Follow the regularized leader
KL	Kullback–Leibler
MD	Mirror descent
MDP	Markov decision process
MPE	Multiplicative Poisson equation
RL	Reinforcement learning
RWM	Randomized Weighted Majority
SRC	Sequential Rademacher complexity

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my advisors Prof. Rebecca Willett and Prof. Maxim Raginsky. Without your continuous support, inspiration and motivation, I would not have come this far. Thank you for welcoming me to your labs and introducing me to the exciting fields of learning and control theory. Thank you for always being patient and helpful when I get stuck on a problem. Thank you for showing me what it takes to be a well-rounded researcher. Thank you for teaching me to think critically and independently, to write and present clearly and concisely. Thank you for the kind encouragement when our paper got rejected. Thank you for always being able to drag me out of dead ends with your immense knowledge. Thank you, Becca, Max, for being the best mentors ever.

I would also like to thank my dissertation committee members, Prof. Vincent Conitzer, Prof. Mauro Maggioni and Prof. Henry Pfister for your kind support and insightful comments. I am also very grateful to Prof. Robert Calderbank, Prof. Lawrence Carin, Prof. Michael Gustafson, Prof. Jeffery Krolik and Prof. Robert Wolpert. You have kindly provided me with guidance and help through committee meetings and classes. I would also like to acknowledge support for this work provided by NSF grant CCF-1017564 and by AFOSR grant FA9550-10-1-0390.

I am very lucky to have met and learned from many researchers and fellow labmates during my time at Duke and UIUC: Dr. Zachary Harmany, Dr. Kalyani Krishnamurthy, Dr. Jorge Silva, Dr. Nooshin Kiarashi, Eric Hall, Shirisha Reddy,

Rubenka Bandyopadhyay, Viola Gomes, Prof. Joseph Salmon, Prof. Yao Xie, Xin Jiang, Jiaji Huang, Albert Oh, Chi Chen, Juan Ramirez, Aolin Xu, Ehsan Shafieepoorfard, Xiaoyu Guang, Jaeho Lee, Aray Kaliyeva and Dr. Daphney-Stavroula Zois. I thank all of you for your friendship, for all the time we have spent together, and all your help and guidance. I would like to thank other graduate students at Duke: Dr. Zhengming Xing, Dr. Miao Liu, and Wenzhao Lian. I am especially thankful to Shaobo Han and his girlfriend Kun Li, for always being great friends and all the time we have spent together at Wilson center and during our way home. Apologize to those that I thank not in ink but only in spirit.

Finally, I want to thank my family for always be there for me. Thank you, Mom and Dad, for your endless love and care during my whole life. I am sorry that I cannot always be there for you. Thank you, Shenyu, for being my lover, my best friend and my inspiration all the time.

1

Introduction

In this dissertation, we study sequential decision making problems. The goal is to develop algorithms for agents to make good decisions in non-stationary environments based on past observations. Sequential decision making problems come in a variety of different forms. Classical statistical learning (Vapnik (1998); Anthony and Bartlett (1999); Bousquet et al. (2004)) assumes the agents work in a stationary world in the sense that each element of the observed sequence is generated independently from a fixed distribution. To better deal with non-stationary practical problems, in the past decade, learning theorists have focused on relaxing or removing distributional assumptions on the observed sequences, and numerous online learning methods have been developed (cf. Cesa-Bianchi and Lugosi (2006)). In the online learning setting, the agent faces an identical decision-making problem at each time step, and the action of the agent doesn't affect future costs. However, in many cases, it does not suffice to assume that the decision-making agent is state-less, i.e., there are situations where the agent's action has a future impact. Markov decision processes (MDPs) (Puterman (1994); Hernández-Lerma and Lasserre (1996); Arapostathis et al. (1993)) and reinforcement learning (RL) methods (Watkins and Dayan (1992); Tsitsiklis (1994);

Kaelbling et al. (1996)) all provide frameworks for sequential decision-making in a random dynamic environment where the agent has a state. In those settings, the agent's action affects future states and costs through the evolution of the state variable. However, the key assumption underlying those settings is that the agent is operating in a known or stochastically stable environment, which limits the applicability of those methods. To address this shortcoming, recent research combines the MDP and the online learning frameworks into what is called *online MDPs*. Such a setting arises as an extension to the online learning setting, where the agent is making decisions in a dynamic and non-stationary environment and the action of the agent affects the future. A detailed introduction of these learning models will be given in the next section. This dissertation mainly considers online MDP problems and presents algorithms with theoretically guaranteed bounds on their performance.

We introduce some real-world examples that motivate the online MDP setting. The common features of the examples to be presented are that: 1) the agent's action affects the future through a state, which evolves according to Markovian dynamics; 2) the cost incurred by the agent at each decision instant may change over time in an arbitrary fashion. First, consider the problem of intelligent patrol control. Traditional patrol system often suffers from passiveness, blindness, inefficiency and other disadvantages. Suppose we wish to design an intelligent patrol control system that makes the patrol more efficient and organized. The goal is to make good decisions to efficiently navigate the protecting area and try to find as many suspicious activities as possible. This problem can be modeled as an online MDP problem with arbitrary costs. The patrol team is driving around the protecting area, at each decision instant, the team is at some intersection, where the location is the state variable. The team can send "street view" images of the potential searching directions to the control center. The control center the will automatically rate the safety level of these possible searching areas based on their "street view" images and their experience

about past crimes. Then the team will choose which direction to go to based on this safety rating. Crime is often considered as unpredictable, so if we measure the distance between the patrol team and the crime scene as the cost, it may be arbitrarily changing. It fits the online MDP setting where we are doing online prediction with past observations in a dynamic and non-stationary environment. This type of intelligent patrol control based on online control technique will offer a better and more effective methodology for preventing crime, increasing the chance of successfully detect suspicious activities in a highly unpredictable environment. There is another example application of online MDPs. Consider the inventory control problem where the manager has to decide how much items to order from the supplier based on the amount of item he currently holds. This is an optimal control problem, where the state is the stock, and the action is the amount of item ordered. The goal is to maximize the revenue. Holding the stock has an inventory cost, and running out of supplies is a clear revenue loss. Moreover, the revenue depends on the price and demand of the products, which are assumed to be arbitrarily changing. This problem can be captured by the online MDP setting, where we use online control techniques to balance the trade-off between reducing inventory costs and increasing product revenues.

In recent years, developing algorithms for online MDPs has been an active research topic (McMahan (2003); Even-Dar et al. (2009); Yu et al. (2009); Neu et al. (2010); Arora et al. (2012); Guan et al. (2012); Abbasi-Yadkori et al. (2013); Dick et al. (2014)). However, most of the work in this area has been algorithmic: given a problem, one would develop an algorithm almost from scratch and then prove its performance guarantee. To provide a unifying view of existing methods and a general procedure for constructing new ones, we derive a general framework for developing online MDP algorithms in Chapter 2.

We can model a non-stationary environment as a source that generates a sequence

of observations in an unpredictable and arbitrary way. Sometimes it is reasonable to assume that the environment's behavior is unstable and seemingly arbitrary, due to the fact that the agent may interact with other agents that are irrational, oblivious, or have a varying and unclear objective. The non-stationary assumption also makes sense in the presence of many agents with conflicting interests. For example, in strategic and adversarial settings, such as stock markets or bidding auctions, the collective behavior of other agents is likely to change in unpredictable ways. A major advantage of considering non-stationary environments is that we end up with a solution of sequential decision making problems with basically no assumptions that would invalidate our result. A possible pitfall is that, by removing distributional assumptions on the data sequences, we have to consider the worst case scenario, which may render our solution to be rather conservative. In this dissertation, we address this downside by developing algorithms in Chapter 3 that take advantage of possible regularity (e.g. predictability) of observed sequences, yet retain worst-case performance guarantees in case the regularity assumption does not hold.

The presence of the state and the assumption of an arbitrarily varying environment complicate both the theoretical analysis and the development of computationally efficient methods. Therefore, most of the existing methods have computational issues for large-scale tasks. Moreover, although the optimal regret bound in terms of the time horizon is already achieved (Even-Dar et al. (2009); Dick et al. (2014)), additional assumptions on the state transition dynamics have to be imposed to ensure the optimal-order bounds hold. It is desirable to make those assumptions milder. These issues are investigated in Chapter 5 through a construction of a computationally efficient algorithm for a new class of MDPs. This new class of MDPs grants the agent direct control over the the state transitions, and the one-step cost not only measures the desirability of each state, but also penalizes the deviation of the transition probabilities specified by the chosen action from some fixed known dynamics.

We show that optimal policies are much easier to derive for this new class of MDPs in online setting and the construction of our algorithm requires milder assumptions on the state transition dynamics.

The online MDP problem can also be viewed as a dynamic game between the agent and the environment. In this case, a non-stationary environment can be modeled as an adaptive adversary from a game-theoretic perspective. In Chapter 6 of this dissertation, we explore this direction by studying a certain class of two-player stochastic games (Shapley (1953); Sorin (2002)). We provide computationally efficient near-optimal strategies for both players and prove that the actual payoff of these strategies converges to the value of the game at fast rates.

In the rest of this chapter, we first describe the basic models we use in order to fix ideas. We then briefly list the contributions of the dissertation. Specifically, in Section 1.1, we review the MDP and online learning frameworks and discuss the notion of regret. In Section 1.2, we summarize the contributions of this dissertation. In Section 1.3, we introduce some frequently used notations throughout this dissertation in more detail. We leave motivation, literature review and precise statement of contributions to the introduction section of each chapter.

The rest of the dissertation is organized as follows: Chapter 2 presents a general framework for deriving online MDP algorithms. Chapter 3 develops an algorithm that takes into account the "regularity" of the observed sequences. Chapter 4 considers traditional MDP problems and extends the LP approach to MDPs by adding relative entropy regularization. Chapter 5 gives an explicit description of a computationally efficient online strategy for a new class of MDPs. Chapter 6 considers single controller stochastic games, which can be thought of as a game-theoretic model that generalizes online MDPs, and quantifies the sub-optimality of stationary strategies for finite-horizon stochastic games. We close by summarizing our contributions and outlining some directions for future work in Chapter 7.

1.1 Review of existing learning models

MDPs comprise a popular framework for sequential decision-making in a random dynamic environment. At each time step, an agent observes the state of the system of interest and chooses an action. The system then transitions to its next state, with the transition probability determined by the current state and the action taken. There is a (possibly time-varying) cost associated with each admissible state-action pair, and a policy (feedback law) for mapping states to actions is selected to minimize average cost. In the basic MDP framework, it is assumed that the cost functions and the transition probabilities are known in advance. In this case, there are two ways of designing policies "offline"– via dynamic programming (Bertsekas (2001)) (where the construction of an optimal policy revolves around the computation of a relative value function), or via the linear programming (LP) approach (Manne (1960); Borkar (2002)), which reformulates the MDP problem as a "static" linear optimization problem over the so-called state-action polytope (Puterman (1994)). Here the optimality criterion is forward-looking, taking into account the effect of past actions on future costs. In many practical problems, however, this degree of advance knowledge is unavailable. When neither the transition probability nor the cost functions are known in advance, various RL methods, such as the celebrated Q-learning algorithm (Watkins and Dayan (1992); Tsitsiklis (1994)) and its variants, can be used to learn an optimal policy in an online regime. However, the key assumptions underlying RL are that the agent is operating in a stochastically stable environment, and that the state-action costs (or at least their expected values with respect to any environmental randomness) do not vary with time. These assumptions are needed to ensure that the agent is eventually able to learn an optimal stationary control policy.

Another framework for sequential decision-making, dating back to the seminal work of Robbins (1951) and Hannan (1957) and now widely used in the machine

learning community (Cesa-Bianchi and Lugosi (2006)), deals with nonstochastic, unpredictable environments. In this *online learning* (or *sequential prediction*) framework, the effects of the environment are modeled by an arbitrarily varying sequence of cost functions, where the cost function at each time step is revealed to the agent only *after* an action has been taken. There is no state, and the goal of the agent is to minimize *regret*, i.e., the difference between the total cost incurred using causally available information and the total cost of the best single action that could have been chosen in hindsight. In contrast with MDPs, the regret-based optimality criterion is necessarily myopic and backward-looking, since the cost incurred at each time step depends only on the action taken at that time step, so past actions have no effect on future costs. There is also a more stringent model of online learning, in which the agent observes not the entire cost function for each time step, but only the value of this cost at the currently taken action (Auer et al. (2002)). This model is inspired by the celebrated *multiarmed bandit* problem first introduced by Robbins (1952), and is referred to as the *nonstochastic bandit problem*. One widely used way of constructing regret-minimizing strategies for such bandit problems is to randomize the agent's actions (*exploration*) so that the random cost value revealed to the agent can be used to construct an unbiased estimate of the full cost function, which is then fed into a suitable strategy that minimizes regret under the assumption of full information (*exploitation*). We will not consider nonstochastic bandit problems in this dissertation. Instead, we refer the reader to a recent survey by Bubeck and Cesa-Bianchi (2012) that discusses both stochastic and nonstochastic bandit problems.

Recent work by Even-Dar et al. (2009) and Yu et al. (2009) combines the MDP and the online learning frameworks into what is called *online MDPs* with finite state and action spaces. Like in the traditional MDP setting, the agent observes the current state and chooses an action, and the system transitions to the next state according to a fixed and known Markov law. However, like in the online framework,

the one-step cost functions form an arbitrarily varying sequence, and the cost function corresponding to each time step is revealed to the agent after the action has been taken. The objective of the agent is to minimize regret relative to the best stationary Markov policy that could have been selected with full knowledge of the cost function sequence over the horizon of interest. The assumption of arbitrary time-varying cost functions makes sense in highly uncertain and complex environments whose temporal evolution may be difficult or costly to model, and it also accounts for collective (and possibly irrational) behavior of any other agents that may be present. The regret minimization viewpoint then ensures that the agent's *online* policy is robust against these effects.

Online MDP problems can be viewed as *online control problems*. The online aspect is due to the fact that the cost functions are generated by a dynamic environment under no distributional assumptions, and the agent learns the current state-action cost only after selecting an action. The control aspect comes from the fact that the choice of an action at each time step influences future states and costs. This online MDP problem reduces to a full information online learning problem in the absence of state variables. Table 1.1 is a summary of advantages of the online MDP framework over existing methods. Taking into account the effect of past actions on future costs in a dynamic distribution-free setting makes online MDPs hard to solve. Consequently, compared to the various algorithms developed over the past two decades for the online learning setting, there are not many algorithms derived for online MDPs. There are two distinct lines of methods: the algorithms presented by Even-Dar et al. (2009); Yu et al. (2009) require the computation of relative value functions at each time step, while the algorithms in Zimin and Neu (2013); Dick et al. (2014) reduce the online MDP problem into an online linear optimization problem and solve it by online learning methods. These two lines of methods correspond to the above mentioned two different ways of designing polices for basic MDPs.

Table 1.1: Advantages of the online MDP setting over existing methods.

Methods	Future impacts of actions	Applicable to arbitrary costs	Applicable to unknown costs
Basic MDPs	Yes	No	No
Reinforcement learning	Yes	No	Yes
Online learning	No	Yes	Yes
Onine MDPs	Yes	Yes	Yes

Even-Dar et al. (2009) first consider regret minimization in online MDPs and prove a sublinear regret relative to the best stationary policy in hindsight. They use an online learning algorithm in each state fed with the value function of their aggregated policy used in the last round. Yu et al. (2009) show that regret minimization is possible when the environment is oblivious (non-adaptive). Their proposed algorithm is a Follow-the-Perturbed-Leader-type algorithm with improved computational complexity but a sub-optimal regret bound. Zimin and Neu (2013); Dick et al. (2014) treat the online MDP problem as an online linear optimization problem and provide algorithms that achieve sublinear regret bounds. Neu et al. (2010); Arora et al. (2012) consider a similar online MDP setup by assuming only bandit-type feedback and prove sublinear regret bounds. Yu and Mannor (2009) consider a different setting where the state transition dynamics is also changing over time. The regret bound of their algorithm may potentially grow linearly with time. Abbasi-Yadkori et al. (2013) consider the same setting and prove sublinear regret bounds at the price of imposing stronger assumptions on the structure of the MDPs.

1.2 Contributions of the dissertation

This dissertation mainly considers online MDPs with finite state and action spaces. In Chapter 2 [1] , we give a general framework for deriving algorithms in the online

[1] This work is submitted to the journal *Machine Learning* (Guan et al. (2015)), and parts of this work are published in *Proceedings of American Control Conference* (Guan et al. (2014a)).

MDP setting. This framework brings existing methods under a single theoretical interpretation and provides a general procedure for constructing new ones. The general procedure greatly simplifies the development of algorithms for online MDPs and derivation of performance guarantees. One well-known algorithm is recovered as a special case by our framework and two new methods are presented.

In Chapter 3, we study the MDP problem with changing cost functions and look into the situation where the sequence being encountered may exhibit some "regularity". We present an algorithm that guarantees sublinear steady-state regret with respect to the best stationary policy. The regret scales with the error of our prediction model for the cost sequences and does not exceed the usual worst-case regret bounds if the sequence is actually not predictable by our model. This method addresses the downside of existing online MDP algorithms that are too conservative when only considering the worse-case scenario.

In Chapter 2 and Chapter 3, we sometimes need to reduce an online MDP problem to an online linear optimization problem, which requires a projection onto a certain convex polytope. In Chapter 4, we study this projection in-depth by relating it to an entropy-regularized approach for traditional MDP problems. We show that the ergodic occupation measure that minimizes the entropy-regularized ergodic control cost is a member of an exponential family of probability distributions, which has allowed us to exploit known results about exponential families to provide structural results for the optimizing measure. Specifically, we have derived various properties of the optimizing Gibbs measure, the corresponding ergodic cost and the relative value function. These results open the door to further in-depth analysis of the online MDP problem, such as optimal adaptation of the learning rate to observed data or efficient approximate implementation of online policy selection algorithms.

Existing methods can attain optimal regret bounds in terms of the time horizon

but run into computational issues. In Chapter 5 [2]we propose a computationally efficient online strategy with small regret under mild regularity conditions for a new class of MDPs proposed by Todorov (2007, 2008, 2009), which we call *MDPs with Kullback-Leibler (KL) control cost*. We also demonstrate the performance of the proposed strategy on a simulated target tracking problem. Apart from rigorously proving the performance guarantees, we also obtain a number of new results on MDPs with KL control cost. Sharp bounds on the sensitivity of optimal control laws to misspecification of state costs are provided, and our quantitative sensitivity estimates for optimal controllers may be used not only in the online setting, but also in the context of approximation algorithms for policy design – indeed, they can be used to establish error bounds for schemes like value or policy iteration.

Finally, in Chapter 6 [3], we transform online MDPs to single-controller stochastic games and derive near-optimal strategies. We use stationary policies that arise from regret minimization strategies and quantify the sub-optimality of such policies for finite-horizon stochastic games. We show that the actual payoff converges to the average minimax payoff at fast rates. Therefore, these stationary policies can be seen as efficient dynamics for zero-sum stochastic games to converge to minimax equilibrium. We also provide a computationally efficient way of approximately solving infinite-horizon stochastic games.

1.3 Notation

We will denote the underlying finite state space and action space by X and U, respectively. The set of all probability distributions on X will be denoted by $\mathcal{P}(\mathsf{X})$, and the same goes for U and $\mathcal{P}(\mathsf{U})$. A matrix $P = [P(u|x)]_{x\in\mathsf{X},u\in\mathsf{U}}$ with nonnegative

[2] This work is published in the journal *IEEE Transactions on Automatic Control* (Guan et al. (2014b)), and parts of this work are published in *Proceedings of American Control Conference* (Guan et al. (2012)).

[3] This work will be submitted to *Proceedings of American Control Conference*.

entries, and with rows and columns indexed by the elements of X and U respectively, is called *Markov* (or *stochastic*) if its rows sum to one: $\sum_{u\in\mathsf{U}} P(u|x) = 1, \forall x \in \mathsf{X}$. We will denote the set of all such Markov matrices (or randomized state feedback laws) by $\mathcal{M}(\mathsf{U}|\mathsf{X})$. Markov matrices in $\mathcal{M}(\mathsf{U}|\mathsf{X})$ transform probability distributions on X into probability distributions on U: for any $\mu \in \mathcal{P}(\mathsf{X})$ and any $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$, we have

$$\mu P(u) \triangleq \sum_{x\in\mathsf{X}} \mu(x)P(u|x), \qquad \forall u \in \mathsf{U}.$$

The same applies to Markov matrices on X and to their action on the elements of $\mathcal{P}(\mathsf{X})$.

The fixed and known stochastic transition kernel of the MDP will be denoted throughout by K – that is, $K(y|x,u)$ is the probability that the next state is y if the current state is x and the action u is taken. For any Markov matrix (randomized state feedback law) $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$, we will denote by $K(y|x,P)$ the Markov kernel

$$K(y|x,P) \triangleq \sum_{u\in\mathsf{U}} K(y|x,u)P(u|x).$$

Similarly, for any $\nu \in \mathcal{P}(\mathsf{U})$,

$$K(y|x,\nu) \triangleq \sum_{u\in\mathsf{U}} K(y|x,u)\nu(u)$$

(this can be viewed as a special case of the previous definition if we interpret ν as a state feedback law that ignores the state and draws a random action according to ν). For any $\mu \in \mathcal{P}(\mathsf{X})$ and $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$, $\mu \otimes P$ denotes the induced joint state-action distribution on $\mathsf{X} \times \mathsf{U}$:

$$\mu \otimes P(x,u) = \mu(x)P(u|x), \qquad \forall (x,u) \in \mathsf{X} \times \mathsf{U}.$$

We say that P is *unichain* (Hernández-Lerma and Lasserre (2003)) if the corresponding Markov chain with transition kernel $K(\cdot|\cdot,P)$ has a single recurrent class

of states (plus a possibly empty transient class). This is equivalent to the induced kernel $K(\cdot|P)$ having a unique invariant distribution π_P (Seneta (2006)).

The total variation (or L_1) distance between $\nu_1, \nu_2 \in \mathcal{P}(\mathsf{U})$ is

$$\|\nu_1 - \nu_2\|_1 \triangleq \sum_{u \in \mathsf{U}} |\nu_1(u) - \nu_2(u)|.$$

It admits the following variational representation:

$$\|\nu_1 - \nu_2\|_1 = \sup_{f:\, \|f\|_\infty \leqslant 1} \left| \langle \nu_1, f \rangle - \langle \nu_2, f \rangle \right|,$$

where the supremum is over all functions $f : \mathsf{U} \to \mathbb{R}$ with absolute value bounded by 1, and we define the *sup norm* $\|f\|_\infty \triangleq \max_{x \in \mathsf{X}} |f(x)|$. We are using the linear functional notation for expectations:

$$\langle \nu, f \rangle = \mathbb{E}_\nu[f] = \sum_{u \in \mathsf{U}} \nu(u) f(u).$$

The *Kullback–Leibler divergence* (or *relative entropy*) between ν_1 and ν_2 (Cover and Thomas (2006)) is

$$D(\nu_1 \| \nu_2) \triangleq \begin{cases} \displaystyle\sum_{u \in \mathsf{U}} \nu_1(u) \log \frac{\nu_1(u)}{\nu_2(u)} & \text{if } \operatorname{supp}(\nu_1) \subseteq \operatorname{supp}(\nu_2) \\ +\infty & \text{otherwise} \end{cases}$$

Here and in the sequel, we work with natural logarithms. The same applies, *mutatis mutandis*, to probability distributions on X. Here $\operatorname{supp}(\nu) \triangleq \{u \in \mathsf{U} : \nu(u) > 0\}$ is the *support* of ν.

Denote the set of all functions $f : \mathsf{X} \to \mathbb{R}$ by $\mathcal{C}(\mathsf{X})$. The *span seminorm* (also called the *oscillation*) of $f : \mathsf{X} \to \mathbb{R}$ is defined as

$$\|f\|_s \triangleq \max_{x \in \mathsf{X}} f(x) - \min_{x \in \mathsf{X}} f(x).$$

Note that $\|f\|_s = 0$ if and only if $f(x) = c$ for some constant $c \in \mathbb{R}$ and all $x \in \mathsf{X}$; $\|f\|_s = \|f + c\|_s$ for any $f \in \mathcal{C}(\mathsf{X})$ and $c \in \mathbb{R}$. Note that $\|f\|_s \leqslant 2\|f\|_\infty$.

We will also be dealing with binary trees that arise in symmetrization arguments, as in Rakhlin et al. (2012): Let ε be a vector $(\varepsilon_1, \ldots, \varepsilon_T)$ of i.i.d. Rademacher random variables, i.e., $\Pr(\varepsilon_i = \pm 1) = 1/2$. Let $\mathcal{H}$ be an arbitrary set. An $\mathcal{H}$-valued tree $\mathbf{h}$ of depth d is defined as a sequence $(\mathbf{h}_1, \ldots, \mathbf{h}_d)$ of mappings $\mathbf{h}_t : \{\pm 1\}^{t-1} \to \mathcal{H}$ for $t = 1, 2, \ldots, d$. Given a tuple $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d) \in \{\pm 1\}^d$, we will often write $\mathbf{h}_t(\varepsilon)$ instead of $\mathbf{h}_t(\varepsilon_{1:t-1})$.

2

From minimax value to low-regret algorithms

Online learning algorithms are designed to perform in non-stationary environments, but generally there is no notion of a dynamic *state* to model constraints on current and future actions as a function of past actions. State-based models are common in stochastic control settings, but commonly used frameworks such as MDPs assume a known stationary environment. In recent years, there has been a growing interest in combining the above two frameworks and considering an MDP setting in which the cost function is allowed to change arbitrarily after each time step. However, most of the work in this area has been algorithmic: given a problem, one would develop an algorithm almost from scratch. Moreover, the presence of the state and the assumption of an arbitrarily varying environment complicate both the theoretical analysis and the development of computationally efficient methods. This chapter describes a broad extension of the ideas proposed by Rakhlin et al. (2012) to give a general framework for deriving algorithms in an MDP setting with arbitrarily changing costs. This framework leads to a unifying view of existing methods and provides a general procedure for constructing new ones. Several new methods are presented, and one of them is shown to have important advantages over a similar method developed from

scratch via an online version of approximate dynamic programming.

2.1 Introduction

Online MDP problems can be viewed as *online control problems*. The online aspect is due to the fact that the cost functions are generated by a dynamic environment under no distributional assumptions, and the agent learns the current state-action cost only after committing to an action. The control aspect comes from the fact that the choice of an action at each time step influences future states and costs. Taking into account the effect of past actions on future costs in a dynamic distribution-free setting makes online MDPs hard to solve. To the best of our knowledge, only a few methods have been developed in this area over the past decade (McMahan (2003); Even-Dar et al. (2009); Yu et al. (2009); Neu et al. (2010); Arora et al. (2012); Guan et al. (2012); Abbasi-Yadkori et al. (2013); Zimin and Neu (2013); Dick et al. (2014)). Most research in this area has been algorithmic: given a problem, one would devise a method and prove a guarantee (i.e., a regret bound) on its performance. There are two distinct lines of methods: the algorithms presented by Even-Dar et al. (2009); Yu et al. (2009); Neu et al. (2010) require the computation of relative value functions at each time step, while the algorithms in Zimin and Neu (2013); Dick et al. (2014) reduce the online MDP problem to an online linear optimization problem and solve it by online learning methods. These two lines of methods correspond to the two different ways of designing polices for MDPs mentioned in Section 1.1. From a theoretical and conceptual standpoint, it is desirable to provide a unifying view of existing methods and a general procedure for constructing new ones. In this chapter, we present such a general framework for online MDP problems that subsumes the above two approaches. This general framework not only enables us to recover known algorithms, but it also gives us a generic toolbox for deriving new algorithms from a more principled perspective rather than from scratch.

Our general approach is motivated by recent work of Rakhlin et al. (2012), which gives a principled way of deriving online learning algorithms (and bounding their regret) from a minimax analysis. Of course, many online learning algorithms have been developed in various settings over the past few decades, but a comprehensive and systematic treatment was still lacking prior to Rakhlin et al. (2012). Starting from a general formulation of online learning as a (stateless) repeated game between a learner and an adversary, Rakhlin et al. (2012) analyze the minimax regret (value) of this online learning game, which is the regret (relative to a fixed competing strategy) that would be achieved if both the learner and the adversary play optimally. Rakhlin et al. (2010) point out before that one could derive sublinear upper bounds on the minimax value in a nonconstructive manner. However, algorithm design was done on a case-by-case basis, and custom analysis techniques were needed in each case to derive performance guarantees matching these upper bounds. The work of Rakhlin et al. (2012) bridges this gap between minimax value analysis and algorithm design: They have shown that, by choosing appropriate relaxations of a certain recursive decomposition of the minimax value, one can recover many known online learning algorithms and give a general recipe for developing new ones. In short, the framework proposed by Rakhlin et al. (2012) can be used to convert an upper bound on the value of the game into an algorithm.

Our main contribution is an extension of the framework of Rakhlin et al. (2012) to online MDPs. Since online learning problems are studied in a state-free setting, it is not straightforward to generalize the ideas of Rakhlin et al. (2012) to the case when the system has a state, and the technical nature of the arguments involved in online MDPs is significantly heavier than their state-free counterpart. We formulate the online MDP problem as a two-player repeated game with state variables and study its minimax value. We introduce the notion of an online MDP *relaxation* and show how it can be used to recover existing methods and to construct new

algorithms. More specifically, we present two distinct approaches of moving from the original dynamic setting, where the state evolves according to a controlled Markov chain, to simpler static settings and constructing corresponding relaxations. The first approach uses Poisson inequalities for MDPs (Meyn and Tweedie (2009)) to reformulate the original dynamic setting as a static setting, where each possible state is associated with a separate online learning algorithm. We show that the algorithm proposed by Even-Dar et al. (2009) arises from a particular relaxation, and we also derive a new algorithm in the spirit of Yu et al. (2009) which exhibits improved regret bounds. The second approach moves from the dynamic setting to a static setting by reducing the online MDP problem to an online linear optimization problem. After the reduction, we can directly capitalize on the framework of Rakhlin et al. (2012). We then derive a novel Online Mirror Descent (OMD) algorithm in the spirit of Zimin and Neu (2013) and Dick et al. (2014) under a carefully designed relaxation over a certain convex set. In short, while the existing methods fall into two major categories, they both can be captured by the above two approaches, and these two approaches arise from the same general idea: move from the original dynamic setting to a static setting, derive the corresponding relaxation, and convert the relaxation into an algorithm.

The remainder of the chapter is organized as follows. We close this section with a brief summary of our results. Section 2.2 contains precise formulation of the online MDP problem and points out the general idea and major challenges. Section 2.3 describes our proposed framework and contains the main result. The general framework includes two different methods of recovering and deriving algorithms. Section 2.4.1 uses the first method and shows the power of our framework by recovering an existing method proposed in Even-Dar et al. (2009) and further derives a new algorithm. Section 2.4.2 uses the second approach to derive a novel online MDP algorithm. Section 2.5 contains discussion about future research.

2.1.1 A summary of results

We start by recasting an MDP with arbitrary costs as a one-sided *stochastic game*, where an agent who wishes to minimize his long-term average cost is facing a Markovian environment, which is also affected by arbitrary actions of an opponent. A stochastic game (Shapley (1953); Sorin (2002)) is a repeated two-player game, where the state changes at every time step according to a transition law depending on the current state and the moves of both players. Here we are considering a special type of a stochastic game, where the agent controls the state transition alone and the opponent chooses the cost functions. By "one-sided", we mean that the utility of the opponent is left unspecified. In other words, we do not need to study the strategy and objectives of the opponent, and only assume that the changes in the environment in response to the opponent's moves occur arbitrarily. As a result, we simply model the opponent as the environment.

A popular and common objective in such settings is regret minimization. Regret is defined as the difference between the cost the agent actually incurred, and what could have been incurred if the agent knew the observed sequence of cost functions in advance. We will give the precise definition of this regret notion in Section 2.2. We start by studying the minimax regret, i.e., the regret the agent will suffer when both the agent and the environment play optimally. By applying the theory of dynamic programming for stochastic games (Sorin (2002)), we can give the strategy for the agent that achieves minimax regret (called the minimax strategy). It can be interpreted as choosing the best action that takes into account the current cost and the worst case future. Unfortunately, this minimax strategy in general is not computationally feasible due to the fact that the number of possible futures grows exponential with time. The idea is to find a way to approximate the term that represents the "future" and derive near-optimal strategy that is easy to compute

using the approximation.

Our main contribution is a construction of a general procedure for deriving algorithms in the online MDP setting. More specifically:

1. Just as in the state-free setting considered by Rakhlin et al. (2012), we argue that algorithms can be constructed systematically by first deriving a sequence of upper bounds (relaxations) on the minimax value of the game, and then choosing actions which minimize these upper bounds.

2. Once a relaxation and an algorithm are derived in this way, we give a general regret bound of that algorithm as follows:

   $$\text{Expected regret} \leqslant \text{Relaxation} + \text{Stationarization error}.$$

   The first term on the right-hand side of the above inequality is the expected relaxation, while the second term is an approximation error that results from approximating the Markovian evolution of the underlying process by a simpler stationary process using a procedure we refer to as *stationarization*. The first term can be analyzed using essentially the same techniques as the ones employed by Rakhlin et al. (2012), with some modifications; by contrast, the second term can be handled using only a novel combination of Markov chain methods. This approach significantly alleviates the technical burden of proving a regret bound as in the literature before our work.

3. Using the above procedure, we recover an existing method proposed in Even-Dar et al. (2009), which achieves $O(\sqrt{T})$ expected regret against the best stationary policy. We show that our derived relaxation gives us the same exponentially weighted average forecaster as in Even-Dar et al. (2009) and leads to the same regret bound. We also derive a new algorithm using our proposed framework and argue that, while this new algorithm is similar in nature to

the work of Yu et al. (2009), it has several advantages — in particular, better scaling of the regret with the horizon T. Both of these algorithms are based on introducing a sequence of appropriately defined relative value functions, and thus can be viewed as instantiations of the first approach to online MDPs — namely, the one rooted in dynamic programming.

4. We also present a different technique for deriving relaxations that implements the second approach to online MDPs — the one rooted in the LP method. This approach allows us to reduce the online MDP problem to an online linear optimization problem over the state-action polytope. This reduction enables us to use the framework of Rakhlin et al. (2012), and the resulting relaxation leads to a novel OMD algorithm that is similar in spirit to the work of Dick et al. (2014).

2.2 Problem formulation

We consider an online MDP with finite state and action spaces X and U and transition kernel $K(y|x,u)$. Let $\mathcal{F}$ be a fixed class of functions $f : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$, and let $x \in \mathsf{X}$ be a fixed initial state. Consider an agent performing a controlled random walk on X in response to signals coming from the environment. The agent is using mixed strategies to choose actions, where a mixed strategy is a probability distribution over the action space. The interaction between the agent and the environment proceeds as follows:

$X_1 = x$
for $t = 1, 2, \ldots, T$
  The agent observes the state X_t, selects a mixed strategy $P_t \in \mathcal{P}(\mathsf{U})$, and then draws an action U_t from P_t
  The environment simultaneously selects $f_t \in \mathcal{F}$ and announces it to the agent
  The agent incurs one-step cost $f_t(X_t, U_t)$
  The system transitions to the next state $X_{t+1} \sim K(\cdot|X_t, U_t)$
end for

Here, T is a fixed finite horizon.

Assumption 1 (Oblivious environment)**.** *The environment* E *is* oblivious *(or* non-adversarial, open-loop*), i.e., for every t, f_t depends only on f^{t-1}, but not on X^t and U^t.*

Assumption 1 is standard in the literature on sequential prediction (Cesa-Bianchi and Lugosi (2006)) (in particular, it is also imposed by Yu et al. (2009)). We view the above process as a two-player repeated game between the agent and the environment. At each $t \geqslant 1$, the process is at state $X_t = x_t$. The agent observes the current state x_t and selects the mixed strategy P_t, where $P_t(u|x_t) = \Pr\{U_t = u | X_t = x_t\}$, based on his knowledge of all the previous states and current state $x^t = (x_1, \ldots, x_t)$ and the previous moves of the environment $f^{t-1} = (f_1, \ldots, f_{t-1})$. After drawing the action U_t from P_t, the agent incurs the one-step cost $f_t(X_t, U_t)$. Adopting game-theoretic terminology (Başar and Olsder (1999)), we define the agent's closed-loop *behavioral strategy* as a tuple $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$, where $\gamma_t : \mathsf{X}^t \times \mathcal{F}^{t-1} \to \mathcal{P}(\mathsf{U})$. Similarly, the environment's open-loop behavioral strategy is a tuple $\boldsymbol{f} = (f_1, \ldots, f_t)$. Once the initial state $X_1 = x$ and the strategy pair $(\boldsymbol{\gamma}, \boldsymbol{f})$ are specified, the joint distribution of the state-action process (X^T, U^T) is well-defined.

Let $\mathcal{M}_0 = \mathcal{M}_0(\mathsf{U}|\mathsf{X}) \subseteq \mathcal{M}(\mathsf{U}|\mathsf{X})$ denote the subset of all Markov policies P, for which the induced state transition kernel $K(\cdot|\cdot, P)$ has a unique invariant distribution $\pi_P \in \mathcal{P}(\mathsf{X})$. The goal of the agent is to minimize the expected *steady-state regret*

$$R_x^{\boldsymbol{\gamma},\boldsymbol{f}} \triangleq \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left\{ \sum_{t=1}^{T} f_t(X_t, U_t) - \inf_{P \in \mathcal{M}_0} \mathbb{E}\left[\sum_{t=1}^{T} f_t(X, U) \right] \right\}, \tag{2.1}$$

where the outer expectation $\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}}$ is taken with respect to both the Markov chain induced by the agent's behavioral strategy $\boldsymbol{\gamma}$ (including randomization of the agent's actions), the environment's behavior strategy $\boldsymbol{f}$, and the initial state $X_1 = x$. The inner expectation (after the infimum) is with respect to the state-action distribution

$\pi_P \otimes P(x,u) = \pi_P(x)P(u|x)$, where π_P denotes the unique invariant distribution of $K(\cdot|\cdot,P)$. The regret $R_x^{\boldsymbol{\gamma},\boldsymbol{f}}$ can be interpreted as the gap between the expected cumulative cost of the agent using strategy $\boldsymbol{\gamma}$ and the best steady-state cost the agent could have achieved in hindsight by using the best stationary policy $P \in \mathcal{M}_0$ (with full knowledge of $\boldsymbol{f} = f^T$). This gap arises through the agent's lack of prior knowledge on the sequence of cost functions.

Here we consider the steady-state regret, so that the expectation with respect to the state evolution in the comparator term $\mathbb{E}\left[\sum_{t=1}^T f_t(X,U)\right]$ is taken over the invariant distribution π_P instead of the Markov transition law $K(\cdot|\cdot,P)$ induced by P. Under the additional assumptions that the cost functions f_t are uniformly bounded and the induced Markov chains $K(\cdot|\cdot,P)$ are uniformly exponentially mixing for all $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ (this assumption will be discussed in more detail in Section 2.3.1), the difference we introduce here by considering the steady state is bounded by a constant independent of T (Even-Dar et al. (2009); Yu et al. (2009)), and so is negligible in the long run. In our main results, we only consider baseline policies in $\mathcal{M}_0$ that are uniformly exponentially mixing, so we restrict our attention to the steady-state regret without any loss of generality.

While it is certainly true that some non-stationary policy with complete prior knowledge of the cost sequence may (and will) outperform any stationary policy, we limit our consideration to stationary reference policies because it is needed to have a fair comparison: indeed, no truly online strategy could compete with the best (i.e., omniscient) non-stationary policy.

2.2.1 Minimax value

We start our analysis by studying the value of the game (the minimax regret), which we first write down in *strategic form* as

$$V(x) \triangleq \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{f}} R_x^{\boldsymbol{\gamma},\boldsymbol{f}} = \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{f}} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right], \tag{2.2}$$

where we have introduced the shorthand Ψ for the comparator term:

$$\Psi(\boldsymbol{f}) \triangleq \inf_{P \in \mathcal{M}_0} \mathbb{E} \left[\sum_{t=1}^{T} f_t(X, U) \right].$$

In operational terms, $V(x)$ gives the best value of the regret the agent can secure by any closed-loop behavioral strategy against the worst-case choice of an open-loop behavioral strategy of the environment. However, the strategic form of the value hides the *timing protocol* of the game, which encodes the information available to the agent at each time step. To that end, we give the following equivalent expression of $V(x)$ in *extensive form*:

Proposition 1. *The minimax value* (2.2) *is given by*

$$V(x) = \inf_{P_1} \sup_{f_1} \ldots \inf_{P_T} \sup_{f_T} \mathbb{E} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right]. \tag{2.3}$$

Proof. The agent's closed-loop behavioral strategy $\boldsymbol{\gamma}$ is a tuple of mappings $\gamma_t : \mathcal{F}^{t-1} \to \mathcal{P}(\mathsf{U}), 1 \leqslant t \leqslant T$; the environment's open-loop behavior strategy $\boldsymbol{f}$ is a tuple of functions $(f_1, \ldots, f_T)$ in $\mathcal{F}$. Thus,

$$\begin{aligned} V(x) &= \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{f}} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right] \\ &= \inf_{\gamma_1} \ldots \inf_{\gamma_T} \sup_{f_1} \ldots \sup_{f_T} \mathbb{E}_x^{\gamma_1,\ldots,\gamma_T,f_1,\ldots,f_T} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right]. \end{aligned}$$

We start from the final step T and proceed by backward induction. Assuming $\gamma_1, \ldots, \gamma_{T-1}$ were already chosen, and using G_T as a shorthand for $f_T(X_T, U_T) - \Psi(f^T)$, we have

$$\begin{aligned}
&\inf_{\gamma_T} \sup_{f_1,\ldots,f_T} \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} \left\{ \sum_{t=1}^{T-1} [f_t(X_t, U_t)] + G_T \right\} \\
&= \inf_{\gamma_T} \sup_{f_1,\ldots,f_{T-1}} \sup_{f_T} \left\{ \mathbb{E}_x^{\gamma^{T-1},f^{T-1}} \left(\sum_{t=1}^{T-1} [f_t(X_t, U_t)] \right) + \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} [G_T] \right\} \\
&= \inf_{\gamma_T} \sup_{f_1,\ldots,f_{T-1}} \left\{ \mathbb{E}_x^{\gamma^{T-1},f^{T-1}} \left(\sum_{t=1}^{T-1} [f_t(X_t, U_t)] \right) + \sup_{f_T} \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} [G_T] \right\} \\
&= \sup_{f_1,\ldots,f_{T-1}} \left\{ \mathbb{E}_x^{\gamma^{T-1},f^{T-1}} \left(\sum_{t=1}^{T-1} [f_t(X_t, U_t)] \right) + \inf_{P_T(U_T|X_T)} \sup_{f_T} \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} [G_T] \right\}.
\end{aligned}$$

The last step is due to the easily proved fact that, for any two sets A, B and bounded functions $g_1 : A \to \mathbb{R}$, $g_2 : A \times B \to \mathbb{R}$,

$$\inf_{\gamma:A\to B} \sup_a \{g_1(a) + g_2(a, \gamma(a))\} = \sup_a \Big[g_1(a) + \inf_{b\in B} g_2(a, b) \Big]$$

(see, e.g., Lemma 1.6.1 in Bertsekas (2001)). Proceeding inductively in this way, we get (2.3). □

From this minimax formulation, we can immediately get an optimal algorithm that attains the minimax value. To see this, we give an equivalent recursive form for the value of the game. For any $t \in \{0, 1, \ldots, T-1\}$, any given prefix $f^t = (f_1, \ldots, f_t)$ (where we let f^0 be the empty tuple e), and any state $X_{t+1} = x$, define the conditional value for $t = T-1, \ldots, 0$

$$V_t(x, f^t) \triangleq \inf_{\nu\in\mathcal{P}(\mathsf{U})} \sup_f \left\{ \sum_{u\in\mathsf{U}} f(x, u)\nu(u) + \mathbb{E}\Big[V_{t+1}(Y, f_1, \ldots, f_t, f) \Big| x, \nu \Big] \right\}, \qquad (2.4a)$$

and $V_T(x, f^T) \triangleq -\Psi(\boldsymbol{f})$.

Remark 1. Recursive decompositions of this sort arise frequently in problems involving decision-making in the presence of uncertainty. For instance, we may view (2.4) as a dynamic program for a finite-horizon minimax control problem (Bertsekas and Rhodes (1973)). Alternatively, we can think of (2.4) as applying the *Shapley operator* (Sorin (2002)) to the conditional value in a two-player stochastic game, where one player controls only the state transitions, while the other player specifies the cost function. A promising direction for future work is to derive some characteristics of the conditional value from analytical properties of the Shapley operator.

From Proposition 1, we see that $V(x) = V_0(x, \mathsf{e})$. Moreover, we can immediately write down the minimax-optimal behavioral strategy for the agent at $t = 0, \ldots, T-1$:

$$\gamma_{t+1}(x, f^t) = \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathsf{U})} \sup_{f \in \mathcal{F}} \left\{ \sum_{u \in \mathsf{U}} f(x,u)\nu(u) + \mathbb{E}\Big[V_{t+1}(Y, f_1, \ldots, f_t, f) \Big| x, \nu \Big] \right\}.$$

Note that the expression being minimized is a supremum of affine functions of ν, so it is a lower-semicontinuous function of ν. Any lower-semicontinuous function achieves its infimum on a compact set. Since the probability simplex $\mathcal{P}(\mathsf{U})$ is compact, we are assured that a minimizing ν always exists. Using the above strategy at each time step, we can secure the minimax value in the worst-case scenario. Note also that this strategy is very intuitive: it balances the tendency to minimize the present cost against the risk of incurring high future costs. However, with all the future infimum and supremum pairs involved, computing this conditional value is intractable. As a result, the minimax optimal strategy is not computationally feasible. The idea is to give tight bounds of the conditional value, which can be minimized to form a near-optimal strategy. We address this challenge by developing computable bounds for the conditional value functions, choosing a strategy based on these bounds. In general, tighter bounds yield lower regret and looser bounds are easier to compute, and various online MDP methods occupy different points in this domain.

In the spirit of Rakhlin et al. (2012), we come up with approximations of the conditional value $V_t(x, f^t)$ in (2.4). We say that a sequence of functions $\widehat{V}_t : \mathsf{X} \times \mathcal{F}^t \to \mathbb{R}$ is an *admissible relaxation* if for $t = T-1, \dots, 0$:

$$\widehat{V}_t(x, f^t) \geqslant \inf_{\nu \in \mathcal{P}(\mathsf{U})} \sup_{f} \left\{ \sum_{u \in \mathsf{U}} f(x,u)\nu(u) + \mathbb{E}[\widehat{V}_{t+1}(Y, f_1, \dots, f_t, f) | x, \nu] \right\}, \tag{2.5a}$$

and $\widehat{V}_T(x, f^T) \geqslant -\Psi(\boldsymbol{f})$. We can associate a behavioral strategy $\widehat{\boldsymbol{\gamma}}$ to any admissible relaxation as follows:

$$\widehat{\gamma}_t(x, f^{t-1}) = \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathsf{U})} \sup_{f \in \mathcal{F}} \left\{ \sum_{u \in \mathsf{U}} f(x,u)\nu(u) + \mathbb{E}\Big[\widehat{V}_t(Y, f_1, \dots, f_{t-1}, f) \Big| x, \nu \Big] \right\}.$$

Proposition 2. *Given an admissible relaxation $\{\widehat{V}_t\}_{t=0}^T$ and the associated behavioral strategy $\widehat{\boldsymbol{\gamma}}$, for any open-loop strategy of the environment we have the regret bound*

$$R_x^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}} = \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right] \leqslant \widehat{V}_0(x).$$

Proof. The proof is by backward induction. Starting at time T and using the admis-

sibility condition (2.5), we write

$$
\begin{aligned}
&\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t)-\Psi(\boldsymbol{f})\right]\\
&\leqslant \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t)+\widehat{V}_T(X_{T+1},f^T)\right]\\
&= \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T-1} f_t(X_t,U_t)\right]+\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\Big[f_T(X_T,U_T)+\widehat{V}_T(X_{T+1},f^T)\Big]\\
&= \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T-1} f_t(X_t,U_t)\right]+\sum_{x_T}\mu_T(x_T)\Bigg\{\sum_{u\in\mathsf{U}} f_T(x_T,u)\Big[\widehat{\gamma}_T\big(x_T,f^{T-1}\big)\Big](u)\\
&\qquad\qquad +\mathbb{E}\Big[\widehat{V}_T(X_{T+1},f^T)\Big|x_T,\widehat{\gamma}_T\big(x_T,f^{T-1}\big)\Big]\Bigg\}\\
&\leqslant \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T-1} f_t(X_t,U_t)+\widehat{V}_{T-1}(X_T,f^{T-1})\right],
\end{aligned}
$$

where $\mu_T \in \mathcal{P}(\mathsf{X})$ denotes the probability distribution of X_T. The last inequality is due to the fact that $\widehat{\boldsymbol{\gamma}}$ is the behavioral strategy associated to the admissible relaxation $\{\widehat{V}_t\}_{t=0}^T$. Continuing in this manner, we complete the proof. □

Based on the above sequential decompositions, it suffices to restrict attention only to Markov strategies for the agent, i.e., sequences of mappings $\gamma_t : \mathsf{X}\times\mathcal{F}^{t-1}\to\mathcal{P}(\mathsf{U})$ for all t, so that U_t is conditionally independent of X^{t-1},U^{t-1} given X_t, f^{t-1}. From now on, we will just say "behavioral strategy" and really mean "Markov behavioral strategy." In other words, given X_t, f^{t-1}, the history of past states and actions is *irrelevant*, as far as the value of the game is concerned.

Remark 2. What happens if the environment is nonoblivious? Yu et al. (2009) gave a simple counterexample of an aperiodic and recurrent MDP to show that the regret is linear in T regardless of the agent's policy when the opponent can adapt to the agent's

state trajectory. We can gain additional insight into the challenges associated with an adaptive environment from the perspective of the minimax value. In particular, an adaptive environment's *closed-loop* behavioral strategy is $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_T)$ with $\delta_t : \mathsf{X}^t \times \mathsf{U}^{t-1} \to \mathcal{P}(\mathcal{F})$, and the corresponding regret will be given by

$$\begin{aligned}
&\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f})\right] \\
&\leqslant \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} f_t(X_t, U_t) + \widehat{V}_T(X_{T+1}, f^T)\right] \\
&= \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T-1} f_t(X_t, U_t)\right] + \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[f_T(X_T, U_T) + \widehat{V}_T(X_{T+1}, f^T)\right].
\end{aligned}$$

Let's analyze the last two terms:

$$\begin{aligned}
&\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[f_T(X_T, U_T) + V_T(X_{T+1}, f^T)\right] \\
&= \int_{\mathsf{X}^T, \mathcal{F}^T} \mathbb{P}(\mathrm{d}x^T, \mathrm{d}\boldsymbol{f}) \int_{\mathsf{U}} P(\mathrm{d}u_T | x_T, f^{T-1}) \left\{ f_T(x_T, u_T) + \mathbb{E}\Big[\widehat{V}_T\big(X_{T+1}, f^T\big)\Big| x^T, f^T\Big]\right\}.
\end{aligned}$$

In the above conditional expectation, $\boldsymbol{f}$ may depend on the entire x^T, so we cannot replace this conditional expectation by $\mathbb{E}[\cdot|x_T, \gamma_T(x_T)]$. This implies we cannot get similar results as in Proposition 2 in a fully adaptive environment.

2.2.2 Major challenges

From Proposition 2, we can see that we can bound the expected steady-state regret in terms of the chosen relaxation. Ideally, if we construct an admissible relaxation by deriving certain upper bounds on the conditional value and implement the associated behavioral strategy, we will obtain an algorithm that achieves the regret bound corresponding to the relaxation. In principle, this gives us a general framework to develop low-regret algorithms for online MDPs. However, with an additional state variable involved, it is difficult to derive admissible relaxations $\widehat{V}_t(x, f^t)$ to bound

the conditional value. The difficulty stems from the fact that now the current cost depends not only on the current action, but also on past actions. Our plan is to reduce this setting to a simpler setting where there is no Markov dynamics involved. In that setting, we will be able to capitalize on the ideas of Rakhlin et al. (2010, 2012) in two different ways. More specifically, using Rademacher complexity tools introduced by Rakhlin et al. (2010, 2012), we can derive algorithms in the simpler static setting and then transfer them to the original problem. In the same vein, we will also prove a general regret bound for the derived algorithms. Thus we will have a general recipe for developing algorithms and showing performance guarantees for online MDPs.

2.3 The general framework for constructing algorithms in online MDPs

As mentioned in the above section, the main challenge to overcome is the dependence of the conditional value in (2.5) on the state variable. Our plan is to reduce the original online MDP problem to a simpler one, where there is no Markov dynamics.

We proceed with our plan in several steps. First, we introduce a stationarization technique that will allow us to reduce the online MDP setting to a simpler setting without Markov dynamics. This effectively decouples current costs from past actions. Note that this reduction is fundamentally different from just naively applying stateless online learning methods in an online MDP setting, which would amount to a very poor stationarization strategy with larger errors and consequently large regret bounds. In contrast, our proposed stationarization performs the decoupling with minimal loss in accuracy by exploiting the transition kernel, yielding lower regret bounds. Using the stationarization idea, we present two different approaches to construct relaxations, aiming to recover and derive two distinct lines of existing methods. We call the first approach the *value-function approach*. Making use of

Poisson inequalities for MDPs (Meyn and Tweedie (2009)), we state a new admissibility condition for relaxations that differs from the admissibility condition in (2.5) in that there is no conditioning on the state variable. The advantage of working with this new type of relaxation is that the corresponding admissibility conditions are much easier to verify. The second approach is called the *convex-analytic approach*. By treating the online MDP problem as an online linear optimization problem, we are able to adopt the idea of Rakhlin et al. (2012) in a more straightforward way, and use the admissibility condition in Rakhlin et al. (2012) to construct relaxations and derive corresponding algorithms. These two approaches can recover different categories of existing methods (it should be pointed out, however, that there is a natural equivalence between these two approaches: the relative-value function arises as a Lagrange multiplier associated with the invariance constraint that defines the state-action polytope; cf. (Meyn, 2007, Sec. 9.2) for details). The main result of this section is that we can apply any algorithm derived in the simpler static setting to the original dynamic setting and automatically bound its regret.

2.3.1 Stationarization

As before, we let K denote the fixed and known transition law of the MDP. Following Even-Dar et al. (2009) and Yu et al. (2009), we assume the following "uniform mixing condition": There exists a finite constant $\tau > 0$ such that for all Markov policies $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ and all distributions $\mu_1, \mu_2 \in \mathcal{P}(\mathsf{X})$,

$$\|\mu_1 K(\cdot|P) - \mu_2 K(\cdot|P)\|_1 \leqslant e^{-1/\tau}\|\mu_1 - \mu_2\|_1, \tag{2.6}$$

where $K(\cdot|P) \in \mathcal{M}(\mathsf{X}|\mathsf{X})$ is the Markov matrix on the state space induced by P. In other words, the collection of all state transition laws induced by all Markov policies P is *uniformly mixing*. Here we assume, without loss of generality, that $\tau \geqslant 1$. As pointed out in Neu et al. (2014), this uniform mixing condition is actually stronger than the unichain assumption: $K(\cdot|\cdot, P)$ is unichain for any choice of $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$

— see Section 1.3 for definitions. The uniform mixing condition implies that the transition kernel of every policy is a scrambling matrix (A matrix $K \in \mathcal{M}(\mathsf{X}|\mathsf{X})$ is scrambling if and only if for any pair $x, x' \in \mathsf{X}$ there exists at least one $y \in \mathsf{X}$, such that y can be reached from both x and x' in one step with strictly positive probability using K as transition matrix).

Remark 3. The uniform mixing condition is rather strong, since it places significant simultaneous restrictions on an exponentially large family of Markov chains on the state space (each chain corresponds to a particular choice of a deterministic state feedback law, and there are $|\mathsf{U}|^{|\mathsf{X}|}$ such laws). It is also difficult to verify, since the problem of determining whether an MDP is uniform mixing requires verifying mixing for all the deterministic policies. Arora et al. (2012) relax this assumption by considering deterministic state transition dynamics and weakly communicating structure, under which it is possible to move from any state to any other state under *some* policy. Although it is not clear yet if we can derive positive results with stochastic state transition dynamics and weakly communicating structure, putting weaker assumptions on state connectivity is another interesting avenue for future research.

Consider now a behavioral strategy $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$ for the agent. For a given choice $\boldsymbol{f} = (f_1, \ldots, f_T)$ of costs, the following objects are well-defined:

- $P_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ — the Markov matrix that governs the conditional distribution of U_t given X_t, i.e.,
$$P_t^{\boldsymbol{\gamma},\boldsymbol{f}}(u|x) = \left[\gamma_t(x, f^{t-1})\right](u);$$
- $\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathcal{P}(\mathsf{X})$ — the distribution of X_t;
- $K_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathcal{M}(\mathsf{X}|\mathsf{X})$ — the Markov matrix that describes the state transition from

X_t to X_{t+1}, i.e.,

$$K_t^{\boldsymbol{\gamma},\boldsymbol{f}}(y|x) = K(y|x, P_t^{\boldsymbol{\gamma},\boldsymbol{f}}) \equiv \sum_u K(y|x,u) P_t^{\boldsymbol{\gamma},\boldsymbol{f}}(u|x);$$

- $\pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathcal{P}(\mathsf{X})$ — the unique stationary distribution of $K_t^{\boldsymbol{\gamma},\boldsymbol{f}}$, satisfying $\pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} = \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} K_t^{\boldsymbol{\gamma},\boldsymbol{f}}$, where existence and uniqueness are guaranteed by virtue of the unichain assumption;
- $\eta_t^{\boldsymbol{\gamma},\boldsymbol{f}} = \langle \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle$ — the steady-state cost at time t.

Moreover, for any other state feedback law $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$, we will denote by $\eta_t^{P,\boldsymbol{f}}$ the steady-state cost $\langle \pi_P \otimes P, f_t \rangle$, where π_P is the unique invariant distribution of $K(\cdot|\cdot, P)$.

It will be convenient to introduce the regret w.r.t. a fixed $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ with initial state $X_1 = x$:

$$\begin{aligned} R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P) &\triangleq \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left[\sum_{t=1}^T f_t(X_t, U_t) - \sum_{t=1}^T \eta_t^{P,\boldsymbol{f}} \right] \\ &= \sum_{t=1}^T \left[\langle \mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle - \langle \pi_P \otimes P, f_t \rangle \right], \end{aligned}$$

as well as the *stationarized regret*

$$\begin{aligned} \bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P) &\triangleq \sum_{t=1}^T \left(\eta_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \eta_t^{P,\boldsymbol{f}} \right) \\ &= \sum_{t=1}^T \left[\langle \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle - \langle \pi_P \otimes P, f_t \rangle \right]. \end{aligned}$$

Using the fact $\|\nu_1 - \nu_2\|_1 = \sup_{f:\, \|f\|_\infty \leqslant 1} |\langle \nu_1, f \rangle - \langle \nu_2, f \rangle|$, we get the bound

$$R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P) \leqslant \bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P) + \sum_{t=1}^T \|f_t\|_\infty \|\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}}\|_1. \tag{2.7}$$

The key observation here is that the task of analyzing the regret $R_x^{\gamma,\boldsymbol{f}}(P)$ splits into separately upper-bounding the two terms on the right-hand side of (2.7): the stationarized regret $\bar{R}^{\gamma,\boldsymbol{f}}(P)$ and the stationarization error $\sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1$. The latter can be handled using Markov chain techniques. We now present two distinct approaches to tackle the former: the value-function approach and the convex-analytic approach.

2.3.2 The value-function approach

The value-function approach relies on the availability of a so-called *reverse Poisson inequality*, which can be thought of as a generalization of the Poisson equation from the theory of MDPs (Meyn and Tweedie (2009)). Fix a Markov matrix $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ and let $\pi_P \in \mathcal{P}(\mathsf{X})$ be the (unique) invariant distribution of $K(\cdot|\cdot, P)$. Then we say that $\widehat{Q} : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ satisfies the reverse Poisson inequality with *forcing function* $g : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ if

$$\mathbb{E}\Big[\widehat{Q}(Y, P)\Big|x, u\Big] - \widehat{Q}(x, u) \geqslant -g(x, u) + \langle \pi_P \otimes P, g\rangle, \qquad \forall (x, u) \in \mathsf{X} \times \mathsf{U} \tag{2.8}$$

where

$$\widehat{Q}(y, P) \triangleq \sum_{u \in \mathsf{U}} P(u|y)\widehat{Q}(y, u)$$

and $\mathbb{E}[\cdot|x, u]$ is with respect to the transition law $K(y|x, u)$. We should think of this as a relaxation of the Poisson equation (Meyn and Tweedie (2009)), i.e., when (2.8) holds with equality. The Poisson equation arises naturally in the theory of Markov chains and Markov decision processes, where it provides a way to evaluate the long-term average cost along the trajectory of a Markov process. We are using the term "reverse Poisson inequality" to distinguish (2.8) from the Poisson inequality, which also arises in the theory of Markov chains and is obtained by replacing $\geqslant$ with $\leqslant$ in (2.8) (Meyn and Tweedie (2009)). Here we impose the following assumption that we use throughout the rest of the chapter:

Assumption 2. *For any $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ and any $f \in \mathcal{F}$, there exists some $\widehat{Q}_{P,f} : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ that solves the reverse Poisson inequality for P with forcing function f. Moreover,*

$$L(\mathsf{X}, \mathsf{U}, \mathcal{F}) \triangleq \sup_{P \in \mathcal{M}(\mathsf{U}|\mathsf{X})} \sup_{f \in \mathcal{F}} \|\widehat{Q}_{P,f}\|_\infty < \infty.$$

Remark 4. In Section 2.4, we will show this assumption is automatically satisfied when K is a unichain model (or, more generally, when all stationary Markov policies are uniformly mixing, as in Eq. (2.6)).

The main consequence of the reverse Poisson inequality is the following:

Lemma 3 (Comparison principle)**.** *Suppose that $\widehat{Q}$ satisfies the reverse Poisson inequality* (2.8) *with forcing function g. Then for any other Markov matrix P' we have*

$$\langle \pi_P \otimes P, g\rangle - \langle \pi_{P'} \otimes P', g\rangle \leqslant \sum_x \pi_{P'}(x) \sum_u \Big[P(u|x)\widehat{Q}(x,u) - P'(u|x)\widehat{Q}(x,u)\Big]$$

Proof. Let us take expectations of both sides of (2.8) w.r.t. $\pi_{P'} \otimes P'$:

$$\begin{aligned}
\langle \pi_P \otimes P, g\rangle - \langle \pi_{P'} \otimes P', g\rangle &\leqslant \mathbb{E}_{\pi_{P'} \otimes P'}\Big\{\mathbb{E}[\widehat{Q}(Y,P)|X,U] - \widehat{Q}(X,U)\Big\} \\
&= \sum_{x,u} \pi_{P'}(x)P'(u|x)\Big\{\mathbb{E}[\widehat{Q}(Y,P)|x,u] - \widehat{Q}(x,u)\Big\} \\
&= \sum_{x,u} \pi_{P'}(x)P'(u|x)\mathbb{E}[\widehat{Q}(Y,P)|x,u] - \sum_{x,u}\left(\sum_y \pi_{P'}(y)K(x|y,P')\right)P'(u|x)\widehat{Q}(x,u)
\end{aligned}$$

where in the third step we have used the fact that $\pi_{P'}$ is invariant w.r.t. $K(\cdot|\cdot, P')$.

Then we have

$$\sum_{x,u} \pi_{P'}(x) P'(u|x) \mathbb{E}[\widehat{Q}(Y,P)|x,u] - \sum_{x,u} \left(\sum_y \pi_{P'}(y) K(x|y,P') \right) P'(u|x) \widehat{Q}(x,u)$$

$$= \sum_x \pi_{P'}(x) \left\{ \sum_u P'(u|x) \mathbb{E}[\widehat{Q}(Y,P)|x,u] - \sum_{u,y} K(y|x,P') P'(u|y) \widehat{Q}(y,u) \right\}$$

$$= \sum_x \pi_{P'}(x) \left\{ \sum_u P'(u|x) \mathbb{E}[\widehat{Q}(Y,P)|x,u] - \sum_y K(y|x,P') \widehat{Q}(y,P') \right\},$$

where the second step is by definition of $\widehat{Q}(y,P')$. Then we can write

$$\sum_x \pi_{P'}(x) \left\{ \sum_u P'(u|x) \mathbb{E}[\widehat{Q}(Y,P)|x,u] - \sum_y K(y|x,P') \widehat{Q}(y,P') \right\}$$

$$\stackrel{\text{(a)}}{=} \sum_x \pi_{P'}(x) \left\{ \sum_u P'(u|x) \left(\mathbb{E}[\widehat{Q}(Y,P)|x,u] - \sum_y K(y|x,u) \widehat{Q}(y,P') \right) \right\}$$

$$= \sum_{x,u} \pi_{P'}(x) P'(u|x) \Big\{ \mathbb{E}[\widehat{Q}(Y,P)|x,u] - \mathbb{E}[\widehat{Q}(Y,P')|x,u] \Big\}$$

$$\stackrel{\text{(b)}}{=} \sum_{x,u,y} \pi_{P'}(x) P'(u|x) K(y|x,u) \left\{ \sum_{u'} P(u'|y) \widehat{Q}(y,u') - \sum_{u'} P'(u'|y) \widehat{Q}(y,u') \right\}$$

$$\stackrel{\text{(c)}}{=} \sum_{x,y} \pi_{P'}(x) K(y|x,P') \left\{ \sum_{u'} P(u'|y) \widehat{Q}(y,u') - \sum_{u'} P'(u'|y) \widehat{Q}(y,u') \right\}$$

$$\stackrel{\text{(d)}}{=} \sum_x \pi_{P'}(x) \sum_u \Big[P(u|x) \widehat{Q}(x,u) - P'(u|x) \widehat{Q}(x,u) \Big],$$

where (a) and (c) are by definition of $K(\cdot|\cdot,P')$; (b) is by definition of $\widehat{Q}(y,P')$; and in (d) we use the fact that $\pi_{P'}$ is invariant w.r.t. $K(\cdot|\cdot,P')$. □

Armed with this lemma, we can now analyze the stationarized regret $\bar{R}^{\gamma,\boldsymbol{f}}(P)$: suppose that, for each t, $\widehat{Q}_t^{\gamma,\boldsymbol{f}}$ satisfies reverse Poisson inequality for $P_t^{\gamma,\boldsymbol{f}}$ with forcing

function f_t. Then we apply the comparison principle to get

$$\eta_t^{\gamma,\boldsymbol{f}} - \eta_t^{P,\boldsymbol{f}} \leqslant \sum_x \pi_P(x) \left(\sum_u P_t^{\gamma,\boldsymbol{f}}(u|x) \widehat{Q}_t^{\gamma,\boldsymbol{f}}(x,u) - P(u|x) \widehat{Q}_t^{\gamma,\boldsymbol{f}}(x,u) \right).$$

This in turn yields

$$R_x^{\gamma,\boldsymbol{f}}(P) \leqslant \sum_x \pi_P(x) \sum_{t=1}^{T} \left(\sum_u P_t^{\gamma,\boldsymbol{f}}(u|x) \widehat{Q}_t^{\gamma,\boldsymbol{f}}(x,u) - P(u|x) \widehat{Q}_t^{\gamma,\boldsymbol{f}}(x,u) \right)$$
$$+ \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1.$$

Note that $\widehat{Q}_t^{\gamma,\boldsymbol{f}}$ depends functionally on $P_t^{\gamma,\boldsymbol{f}}$ and on f_t, which in turn depend functionally on f^t but not on $f_{t+1}, \ldots, f_T$. This ensures that any algorithm using $\widehat{Q}_t^{\gamma,\boldsymbol{f}}$ respects the causality constraint that any decision made at time t depends only on information available by time t.

Focusing on stationarized regret and upper-bounding it in terms of the $\widehat{Q}$-functions is one of the key steps that let us consider a simpler setting without Markov dynamics. The next step is to define a new type of relaxation with an accompanying new admissibility condition for this simpler setting. That is, we will find a relaxation and admissibility condition for the stationarized regret rather than for the expected steady-state regret directly. A new admissibility condition is needed because we have decoupled current costs from past actions, which makes the previous admissibility condition (2.5) inapplicable. The new admissibility condition is similar to the one in Rakhlin et al. (2012), which was derived in a stateless setting. The difference is that we are still in a *state-dependent* setting in the sense that the new type of relaxation is indexed by the state variable. Now instead of having Markov dynamics that depend on the state, we consider all the states in parallel and have a separate algorithm running on each state. The interaction between different states is generated by providing these algorithms with common information that comes from an

actual dynamical process. Thus, starting from this new admissibility condition, we further construct algorithms using relaxations and then use Lemma 3 to bound the regret of these algorithms.

Now we are in a position to pass to a simpler setting without Markov dynamics. Instead, we associate each state with a separate game. Within each game, the agent chooses an action and observes a signal from the environment, and the current cost in each state is independent from the past actions taken in that state. The signal generated by the environment is the $\widehat{Q}$-function mentioned above in Assumption 2. Although here we don't use the one-step cost functions f_t as the signal, we know that the $\widehat{Q}$-functions actually contain payoff-relevant information on f_t. From this perspective, the environment choosing one-step cost functions f_t is equivalent to letting the environment choose the corresponding $\widehat{Q}$-functions.

To proceed, we need to introduce a new type of relaxation with a new admissibility condition. The reason is that the relaxation $\widehat{V}$ defined in (2.5) is a sequence of functions with a state variable, and the state is changing at every time step according to the state transition dynamics and the agent's action. If we view the interaction between the agent and the environment as a stochastic game, then the relaxation is indeed a sequence of upper bounds on the conditional values of this game. However, after stationarization, we end up with a *family* of relaxations indexed by the state variable — for each state, we have a separate online learning game, and we have a separate relaxation for each of these online learning games. Consequently, there is no Markov dynamics involved in each of the new relaxations. The new relaxation at each state $x \in \mathsf{X}$, which we will denote by $\{\widehat{W}_{x,t}\}_{t=1}^{T}$, is a sequence of upper bounds on the conditional value of the corresponding online learning game. We define such a relaxation as follows.

For each $x \in \mathsf{X}$, let $\mathcal{H}_x$ denote the class of all functions $h_x : \mathsf{U} \to \mathbb{R}$ for which

there exist some $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ and $f \in \mathcal{F}$, such that

$$h_x(u) = \widehat{Q}_{P,f}(x,u), \qquad \forall u \in \mathsf{U}.$$

We say that a sequence of functions $\widehat{W}_{x,t} : \mathcal{H}_x^t \to \mathbb{R}, t = 0, \ldots, T$, is an admissible relaxation at state x if the following condition holds for any $h_{x,1}, \ldots, h_{x,T} \in \mathcal{H}_x$:

$$\widehat{W}_{x,T}(h_x^T) \geqslant -\inf_{\nu \in \mathcal{P}(\mathsf{U})} \mathbb{E}_{U\sim\nu}\left[\sum_{t=1}^{T} h_{x,t}(U)\right], \tag{2.9a}$$

$$\widehat{W}_{x,t}(h_x^t) \geqslant \inf_{\nu \in \mathcal{P}(\mathsf{U})} \sup_{h_x \in \mathcal{H}_x} \left\{\mathbb{E}_{U\sim\nu}[h_x(U)] + \widehat{W}_{x,t+1}(h_x^t, h_x)\right\}, \qquad t = T-1, \ldots, 0. \tag{2.9b}$$

Given such an admissible relaxation, we can associate to it a behavioral strategy

$$\widehat{\gamma}_t(x, f^{t-1}) = P_t^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}}(\cdot|x) = \arg\min_{\nu \in \mathcal{P}(\mathsf{U})} \sup_{h_x \in \mathcal{H}_x} \left\{\mathbb{E}_{U\sim\nu}[h_x(U)] + \widehat{W}_{x,t}(h_x^{t-1}, h_x)\right\}$$

$$h_{y,t} = \widehat{Q}_t^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}}(y, \cdot), \quad \forall y \in \mathsf{X}.$$

(Even though the above notation suggests the dependence of $h_{y,t}$ on the T-tuples $\boldsymbol{\gamma}$ and $\boldsymbol{f}$, this dependence at time t is only with respect to γ^t and f^t, so the resulting strategy is still causal.)

The relaxation $\{\widehat{W}_{x,t}\}_{t=1}^T$ at state x is a sequence of upper bounds on the conditional value of the online learning game associated with that state. In this game, at time step t, the agent chooses actions $u_t \in \mathsf{U}$ and the environment chooses function $h_{x,t} \in \mathcal{H}_x$. Although this relaxation is still state-dependent, there is no Markov dynamics involved here, which means that now the state-free techniques of Rakhlin et al. (2012) can be brought to bear on the problem of constructing algorithms and bounding their regret. Specifically, we derive a separate relaxation $\{\widehat{W}_{x,t}\}_{t=1}^T$ and the associated behavioral strategy for each state $x \in \mathsf{X}$. Then we assemble these into an overall algorithm for the MDP as follows: if at time t the state $X_t = x$, the agent will choose actions according to the corresponding behavioral strategy $\widehat{\gamma}_t(x, \cdot)$. Note

that although the agent's behavioral strategy switches between different relaxations depending on the current state, the agent still needs to update all the h-functions simultaneously for all the states. This is because the computation of the h-functions (in terms of the $\widehat{Q}$ functions) requires the knowledge of the behavioral strategy at other states. In other words, the algorithm has to keep updating all the relaxations in parallel for all states.

Under the constructed relaxation, the value-function approach amounts to the following:

Theorem 4. *Suppose that the MDP is unichain, the environment is oblivious, and Assumption 2 holds. Then, for any family of admissible relaxations given by* (2.9) *and the corresponding behavioral strategy* $\widehat{\boldsymbol{\gamma}}$*, we have*

$$R_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} = \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\widehat{W}_{x,0} + C_{\mathcal{F}} \sum_{t=1}^{T} \|\mu_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1 \tag{2.10}$$

where $C_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \|f\|_\infty$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \\
&\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})} \sum_{t=1}^{T}\Big[\langle \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\otimes P_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}, f_t\rangle - \langle \pi_P\otimes P, f_t\rangle\Big] + \sum_{t=1}^{T}\|f_t\|_\infty \|\mu_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1 \\
&\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x) \sum_{t=1}^{T}\left(\sum_u P_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u) - P(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u)\right) \\
&\quad + \sum_{t=1}^{T}\|f_t\|_\infty \|\mu_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1,
\end{aligned}$$

where in the first equality we have used (2.7), while the second inequality is by

Lemma 3. Then we write the last term out and get

$$\begin{aligned}
&\sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})}\sum_x \pi_P(x)\Bigg[\sum_{t=1}^{T-1}\left(\sum_u P_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u)\right)+\sum_u P_T^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(u|x)\widehat{Q}_T^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u)\\
&\quad-\sum_{t=1}^{T} P(u|x)\widehat{Q}_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u)\Bigg]+\sum_{t=1}^{T}\|f_t\|_\infty\|\mu_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}-\pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1\\
&\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})}\sum_x \pi_P(x)\Bigg[\sum_{t=1}^{T-1}\left(\sum_u P_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u)\right)+\sum_u P_T^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(u|x)\widehat{Q}_T^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u)\\
&\quad+\widehat{W}_{x,T}(h_x^T)\Bigg]+\sum_{t=1}^{T}\|f_t\|_\infty\|\mu_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}-\pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1\\
&\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})}\sum_x \pi_P(x)\left[\sum_{t=1}^{T-1}\left(\sum_u P_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}(x,u)\right)+\widehat{W}_{x,T-1}(h_x^{T-1})\right]\\
&\quad+\sum_{t=1}^{T}\|f_t\|_\infty\|\mu_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}-\pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1,
\end{aligned}$$

where the two inequalities are by the admissibility condition (2.9). Continuing this induction backward, and noting that $\sum_{t=1}^{T}\|f_t\|_\infty\|\mu_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}-\pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1 \leqslant C_{\mathcal{F}}\sum_{t=1}^{T}\|\mu_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}-\pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1$, we arrive at (2.10). □

This general framework gives us a recipe for deriving algorithms for online MDPs. First, we use stationarization to pass to a simpler setting without Markov dynamics. Here we need to find the $\widehat{Q}_t$ functions satisfying (2.8) with forcing function f_t at each time t. In this simpler setting, we associate each state with a separate online learning game. Next, we derive appropriate relaxations (upper bounds on the conditional values) for each of these online learning games. Then we plug the relaxation into the admissibility condition (2.9) to derive the associated algorithm. This algorithm in turn gives us a behavioral strategy for the original online MDP problem, and Theorem 4 automatically gives us a regret bound for this strategy. We emphasize that,

in general, multiple different relaxations are possible for a given problem, allowing for a flexible tradeoff between computational costs and regret.

We have reduced the original problem to a collection of standard online learning problems, each of which is associated with a particular state. We proceed by constructing a separate relaxation for each of these problems. Because we have removed the Markov dynamics, we may now use available techniques for constructing these relaxations. In particular, as shown by Rakhlin et al. (2010), a particularly versatile method for constructing relaxations relies on the notion of *sequential Rademacher complexity* (SRC).

2.3.3 The convex-analytic approach

In the preceding section, we have developed a procedure for recovering and deriving policies for online MDPs using relative-value functions that arise from revere Poisson inequalities. Now we show a complementary procedure that allows us to use an admissible relaxation with no conditioning on state variables. Specifically, we reduce the online MDP problem to an online linear optimization problem through stationarization, and then directly use the framework of Rakhlin et al. (2012) to derive a relaxation and an algorithm, which is similar in spirit to the algorithm proposed recently by Dick et al. (2014). The idea behind this convex-analytic method is closely related to the well-known fact that the dynamic optimization problem for an MDP can be reformulated as a "static" linear optimization problem a certain polytope, and therefore can be solved using LP methods (Manne (1960); Borkar (2002)). Under this reformulation, we are in a state-free setting in the sense that the relaxation is no longer indexed by the state variable, and the policy for the agent is computed from a certain joint distribution on states and actions via Bayes' rule.

As before, we start with the stationarization step. Recall that we decompose the regret $R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P)$ into two parts: the stationarized regret $\bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P)$ and the stationar-

ization error $\sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1$, that is:

$$\begin{aligned} R_x^{\gamma,\boldsymbol{f}}(P) &\leqslant \bar{R}^{\gamma,\boldsymbol{f}}(P) + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1 \\ &= \sum_{t=1}^{T} \left(\eta_t^{\gamma,\boldsymbol{f}} - \eta_t^{P,\boldsymbol{f}}\right) + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1 \\ &= \sum_{t=1}^{T} \left[\langle \pi_t^{\gamma,\boldsymbol{f}} \otimes P_t^{\gamma,\boldsymbol{f}}, f_t\rangle - \langle \pi_P \otimes P, f_t\rangle\right] + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1. \quad (2.11) \end{aligned}$$

Now, let $\mathcal{G} \subset \mathcal{P}(\mathsf{X} \times \mathsf{U})$ denote the set of all *ergodic occupation measures*

$$\mathcal{G} \triangleq \left\{\nu \in \mathcal{P}(\mathsf{X} \times \mathsf{U}) : \sum_{x,u} K(y|x,u)\nu(x,u) = \sum_{u} \nu(y,u), \forall y \in \mathsf{X}\right\}. \quad (2.12)$$

The set $\mathcal{G}$ is convex, and is defined by a finite collection of linear equality constraints. Hence, it is a convex polytope in $\mathbb{R}^{|\mathsf{X}\times\mathsf{U}|}$ (in fact, it is often referred to as the *state-action polytope* of the MDP (Puterman (1994))). Every element in $\mathcal{G}$ can be decomposed in the form

$$\nu(x,u) = \pi_P(x) \otimes P(u|x), \quad x \in \mathsf{X}, u \in \mathsf{U}$$

for some randomized Markov policy $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$, where π_P is the invariant distribution of the Markov kernel

$$K_P(x'|x) \triangleq \sum_{u\in\mathsf{U}} K(x'|x,u)P(u|x), \qquad \forall x, x' \in \mathsf{X}$$

induced by P. For this reason, the linear equality and inequality constraints that define $\mathcal{G}$ are also called the *invariance constraints*. Conversely, any element $\nu \in \mathcal{G}$ induces a Markov policy

$$P_\nu(u|x) \triangleq \frac{\nu(x,u)}{\sum_{v\in\mathsf{U}} \nu(x,v)}, \qquad \forall u \in \mathsf{U}(x) \quad (2.13)$$

where $\mathsf{U}(x)$ is the set of all states for which the denominator of (2.13) is nonzero.

With the definition of the set $\mathcal{G}$ at hand, now it is easy to see that the first term of (2.11) is the regret of an online *linear* optimization problem, where, at each time step t, the agent is choosing an occupation measure $\nu_t = \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}$ from the set $\mathcal{G}$ (here we omit the dependence of ν_t on $\boldsymbol{\gamma}, \boldsymbol{f}$ for simplicity), and the environment is choosing the one-step cost function f_t. The one-step linear cost function incurred by the agent is $\langle \nu_t, f_t \rangle$. Since we can recover a policy from an occupation measure, we just need to find a slowly changing sequence of occupation measures, to ensure simultaneously that the first term of (2.11) and the stationarization error are both small. Now we have mapped an online MDP problem to an online linear optimization problem. As we mentioned earlier, the idea behind this mapping is simply the fact that average-cost optimal control problem can be cast as a LP over the state-action polytope (Manne (1960); Borkar (2002)).

For reasons that will become apparent later, it is convenient to consider regret with respect to policies induced by elements of a given subset $\mathcal{G}'$ of $\mathcal{G}$. With that in mind, let us denote by $\mathcal{M}(\mathcal{G}')$ the set of all policies $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ that have the form P_ν for some $\nu \in \mathcal{G}'$. For the resulting online linear optimization problem, we can immediately apply the framework of Rakhlin et al. (2012) to derive novel relaxations and online MDP algorithms. For any $t \in \{0, 1, \ldots, T-1\}$, any given prefix $f^t = (f_1, \ldots, f_t)$, define the conditional value recursively via

$$V_T(\mathcal{G}'|f_1, \ldots, f_t) = \inf_{\nu \in \mathcal{G}'} \sup_{f \in \mathcal{F}} \left\{ \langle \nu, f \rangle + V_T(\mathcal{G}'|f_1, \ldots, f_t, f) \right\}, \tag{2.14}$$

where $V_T(\mathcal{G}'|f_1, \ldots, f_T) = -\inf_{\nu \in \mathcal{G}'} \sum_{t=1}^T \langle \nu, f_t \rangle$, and $V_T(\mathcal{G}') \equiv V_T(\mathcal{G}'|\mathsf{e})$ is the minimax regret of the game. Note that we are explicitly indicating the fact that the optimization takes place over the ergodic occupation measures in $\mathcal{G}'$. The minimax optimal algorithm specifying the mixed strategy of the player can be written as

$$\nu_t = \arg\min_{\nu \in \mathcal{G}'} \sup_{f \in \mathcal{F}} \left\{ \langle \nu, f \rangle + V_T(\mathcal{G}'|f_1, \ldots, f_{t-1}, f) \right\} \tag{2.15}$$

Following the formulation of Rakhlin et al. (2012), we say that a sequence of functions $\widehat{V}_T(\mathcal{G}'|f_1,\dots,f_t)$ is an *admissible relaxation* if for any $f_1,\dots,f_t \in \mathcal{F}$,

$$\widehat{V}_T(\mathcal{G}'|f_1,\dots,f_t) \geqslant \inf_{\nu\in\mathcal{G}'} \sup_f \left\{ \langle \nu, f\rangle + \widehat{V}_T(\mathcal{G}'|f_1,\dots,f_t,f) \right\}, \qquad t = T-1,\dots,0 \tag{2.16a}$$

$$\widehat{V}_T(\mathcal{G}'|f^T) \geqslant -\inf_{\nu\in\mathcal{G}'} \sum_{t=1}^{T} \langle \nu, f_t\rangle. \tag{2.16b}$$

We can associate a behavioral strategy to any admissible relaxation as follows:

$$\widehat{\gamma}_t(x, f^{t-1}) = \nu_t = \arg\min_{\nu\in\mathcal{G}'} \sup_{f\in\mathcal{F}} \left\{ \langle \nu, f\rangle + \widehat{V}_T(\mathcal{G}'|f_1,\dots,f_{t-1},f) \right\}.$$

In fact, as pointed out by Rakhlin et al. (2012), exact minimization is unnecessary: any choice $\nu_t = \widehat{\gamma}_t(x, f^{t-1})$ that satisfies

$$\widehat{V}_T(\mathcal{G}'|f_1,\dots,f_{t-1}) \geqslant \sup_{f\in\mathcal{F}} \left\{ \langle \nu_t, f\rangle + \widehat{V}_T(\mathcal{G}'|f_1,\dots,f_{t-1},f) \right\},$$

is admissible. The above admissibility condition of Rakhlin et al. (2012) is different from (2.5) in the sense that there is no conditioning on the state variable. It is also different from (2.9) because it is not indexed by the state variable. The following theorem provides the main regret bound for the convex-analytic approach:

Theorem 5. *Suppose that the MDP is unichain and the environment is oblivious. Then, for any family of admissible relaxations given by* (2.16) *and the corresponding behavioral strategy* $\widehat{\boldsymbol{\gamma}}$*, we have*

$$\begin{aligned} R_x^{\widehat{\gamma},\boldsymbol{f}}(\mathcal{G}') &\triangleq \mathbb{E}_x^{\widehat{\gamma},\boldsymbol{f}} \left\{ \sum_{t=1}^{T} f_t(X_t,U_t) - \inf_{P\in\mathcal{M}(\mathcal{G}')} \mathbb{E}\left[\sum_{t=1}^{T} f_t(X,U) \right] \right\} \\ &\leqslant \widehat{V}_T(\mathcal{G}'|\mathsf{e}) + C_{\mathcal{F}} \sum_{t=1}^{T} \| \mu_t^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}} \|_1 \end{aligned} \tag{2.17}$$

where $C_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \|f\|_\infty$.

Proof. Applying the same backward induction used in the proof of Proposition 2 (also see (Rakhlin et al., 2012, Prop. 1)), it is easy to show that

$$\sum_{t=1}^{T}\langle \nu_t, f_t\rangle - \inf_{\nu\in\mathcal{G}'}\sum_{t=1}^{T}\langle \nu, f_t\rangle \leqslant \widehat{V}_T(\mathcal{G}'|\mathsf{e}).$$

Then it is straightforward to see that

$$\begin{aligned}
&\mathbb{E}_x^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\left\{\sum_{t=1}^{T} f_t(X_t,U_t) - \inf_{P\in\mathcal{M}(\mathcal{G}')}\mathbb{E}\left[\sum_{t=1}^{T} f_t(X,U)\right]\right\}\\
&\leqslant \sum_{t=1}^{T}\left[\langle \pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\otimes P_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}, f_t\rangle - \inf_{P\in\mathcal{M}(\mathcal{G}')}\langle \pi_P\otimes P, f_t\rangle\right] + \sum_{t=1}^{T}\|f_t\|_\infty\|\mu_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1\\
&\leqslant \sum_{t=1}^{T}\langle \nu_t, f_t\rangle - \inf_{\nu\in\mathcal{G}'}\sum_{t=1}^{T}\langle \nu, f_t\rangle + \sum_{t=1}^{T}\|f_t\|_\infty\|\mu_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1,
\end{aligned}$$

where in the first equality we have used (2.7). □

2.4 Example derivations of explicit algorithms

In the preceding section, we have described two different approaches to construct relaxations and algorithms for online MDPs. Specifically, the value-function approach make use of Poisson inequalities for MDPs (cf. Meyn and Tweedie (2009)) to reduce the online MDP problem to a collection of standard online learning problems, each of which is associated with a particular state. We need to construct a separate relaxation for each of these problems. The convex-analytic approach reduces the online MDP problem to an online linear optimization problem, and uses a single relaxation (no longer indexed by the state) to derive algorithms. The common property of these two approaches is that we can apply any algorithm derived in the simpler static setting to the original dynamic setting and automatically bound its regret.

2.4.1 The value-function approach

In this section, we apply the value-function approach to recover and derive a class of online MDP algorithms (Even-Dar et al. (2009); Yu et al. (2009)). The common thread running through this class of algorithms is that value functions have to be computed in order to get the policy for each time step. The strategies derived in this section using our general framework also belong to a class of algorithms for *online prediction with expert advice* (Cesa-Bianchi and Lugosi (2006)). In this setting, the agent combines the recommendations of several individual "experts" into an overall strategy for choosing actions in real time in response to causally revealed information. Every expert is assigned a "weight" indicating how much the agent trusts that expert, based on the previous performance of the experts. One of the more popular algorithms for prediction with expert advice is the Randomized Weighted Majority (RWM) algorithm, which updates the expert weights multiplicatively (Littlestone and Warmuth (1994)). It has an alternative interpretation as a Follow the Regularized Leader (FRL) scheme (Shalev-Shwartz and Singer (2007)): The weights chosen by an RWM algorithm minimize a combination of empirical cost and an entropic regularization term. The entropy term (equal to the divergence between the current weight distribution and the uniform distribution over the experts) penalizes "spiky" weight vectors, thus guaranteeing that every expert has a nonzero probability of being selected at every time step, which in turn provides the algorithm with a degree of stability. The common feature of the strategies we consider in this section is that RWM algorithms are applied in parallel for each state.

We start by recovering an expert-based algorithm for online MDPs. Similar to our set-up, Even-Dar et al. (2009) consider an MDP with arbitrarily varying cost functions. The main idea of their work is to efficiently incorporate existing expert-based algorithms (Littlestone and Warmuth (1994); Cesa-Bianchi and Lugosi (2006))

into the MDP setting. For an MDP with state space X and action space U, there are $|\mathsf{U}|^{|\mathsf{X}|}$ deterministic Markov policies (state feedback laws), which renders the obvious approach of associating an expert with each possible deterministic policy computationally infeasible. Instead, they propose an alternative efficient scheme that works by associating a separate expert algorithm to each *state*, where experts correspond to *actions* and the feedback to provided each expert algorithm depends on the aggregate policy determined by the action choices of all the individual algorithms. Under a unichain assumption similar to the one we have made above, they show that the expected regret of their algorithm is sublinear in T and independent of the size of the state space. Their algorithm can be summarized as follows:

Put in every state x an expert algorithm $\mathcal{A}_x$
for $t = 1, 2, \ldots$ do
  Let $P_t(\cdot|x_t)$ be the distribution over actions of $\mathcal{A}_{x_t}$
  Use policy P_t and obtain f_t from the environment
  For every $x \in \mathsf{X}$
    Feed $\mathcal{A}_x$ with loss function $\widehat{Q}_{P_t,f_t}(x,\cdot) = \mathbb{E}\left[\sum_{i=0}^{\infty}\left(f_t(X_i,U_i) - \eta_t^{P_t,\boldsymbol{f}}\right)\right]$,
    where $\mathbb{E}$ is taken w.r.t. the Markov chain induced by P_t
    from the initial state x, and $\eta_t^{P_t,\boldsymbol{f}}$ is the steady-state cost $\langle \pi_{P_t} \otimes P_t, f_t\rangle$
  end for
end for

As we show next, the algorithm proposed by Even-Dar et al. (2009) arises from a particular relaxation of the kind that was introduced in the preceding section. For every possible state value $x \in \mathsf{X}$, we want to construct an admissible relaxation that satisfies (2.9). Here we show that the relaxation can be obtained as an upper bound of a quantity called *conditional sequential Rademacher complexity*, which is defined by Rakhlin et al. (2012) as follows. Let ε be a vector $(\varepsilon_1, \ldots, \varepsilon_T)$ of i.i.d. Rademacher random variables, i.e., $\Pr(\varepsilon_i = \pm 1) = 1/2$. For a given $x \in \mathsf{X}$, an $\mathcal{H}_x$-valued tree $\mathbf{h}$ of depth d is defined as a sequence $(\mathbf{h}_1, \ldots, \mathbf{h}_d)$ of mappings $\mathbf{h}_t : \{\pm 1\}^{t-1} \to \mathcal{H}_x$, where $\mathcal{H}_x$ is the function class defined in Section 2.3.2. Then the conditional sequential

Rademacher complexity at state x is defined as

$$\mathcal{R}_{x,t}(h_x^t) = \sup_{\mathbf{h}} \mathbb{E}_{\varepsilon_{t+1:T}} \max_{u \in \mathsf{U}} \left[2 \sum_{s=t+1}^{T} \varepsilon_s \left[\mathbf{h}_{s-t}(\varepsilon_{t+1:s-1}) \right](u) - \sum_{s=1}^{t} h_{x,s}(u) \right], \quad \forall h_x^t \in \mathcal{H}_x^t.$$

Here the supremum is taken over all $\mathcal{H}_x$-valued binary trees of depth $T-t$. The term containing the tree $\mathbf{h}$ can be seen as "future", while the term being subtracted off can be seen as "past". This quantity is conditioned on the already observed h_x^t, while for the future we consider the worst possible binary tree. As shown by Rakhlin et al. (2012), this Rademacher complexity is itself an admissible relaxation for standard (state-free) online optimization problems; moreover, one can obtain other relaxations by further upper-bounding the Rademacher complexity. As we will now show, because the action space U is finite and the functions in $\mathcal{H}_x$ are uniformly bounded (Assumption 2), the following upper bound on $\mathcal{R}_{x,t}(\cdot)$ is an admissible relaxation, i.e., it satisfies condition (2.9):

$$\widehat{W}_{x,t}(h_x^t) = \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u) \right) \right) + \frac{2}{\beta}(T-t)L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2, \tag{2.18}$$

where the learning rate $\beta > 0$ can be tuned to optimize the resulting regret bound. This relaxation leads to an algorithm that turns out to be exactly the scheme proposed by Even-Dar et al. (2009):

Proposition 6. *The relaxation* (2.18) *is admissible and it leads to a recursive exponential weights algorithm, specified recursively as follows: for all* $x \in \mathsf{X}$, $u \in \mathsf{U}$, *at* $t = 0, \ldots, T-1$,

$$P_{t+1}(u|x) = \frac{P_t(u|x) \exp\left(-\frac{1}{\beta} h_{x,t}(u)\right)}{\left\langle P_t(\cdot|x), \exp\left(-\frac{1}{\beta} h_{x,t}\right) \right\rangle} = \frac{\nu_1(u) \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u)\right)}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}\right) \right\rangle}, \tag{2.19}$$

where ν_1 *is the uniform distribution on* U.

Proof. First we show that the relaxation (2.18) arises as an upper bound on the conditional SRC. The proof of this is similar to the one given by Rakhlin et al. (2010), except that they also optimize over the choice of the learning rate β. For any $\beta > 0$,

$$\begin{aligned}
&\mathbb{E}_\varepsilon \left[\max_{u \in \mathsf{U}} \left\{ 2 \sum_{i=1}^{T-t} \varepsilon_i \left[\mathbf{h}_i(\varepsilon) \right](u) - \sum_{s=1}^{t} h_{x,s}(u) \right\} \right] \\
&\leqslant \beta \log \left(\mathbb{E}_\varepsilon \left[\max_{u \in \mathsf{U}} \exp \left(\frac{2}{\beta} \sum_{i=1}^{T-t} \varepsilon_i \left[\mathbf{h}_i(\varepsilon) \right](u) - \frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u) \right) \right] \right) \\
&\leqslant \beta \log \left(\mathbb{E}_\varepsilon \left[\sum_{u \in \mathsf{U}} \exp \left(\frac{2}{\beta} \sum_{i=1}^{T-t} \varepsilon_i \left[\mathbf{h}_i(\varepsilon) \right](u) - \frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u) \right) \right] \right),
\end{aligned}$$

where the first inequality is by Jensen's inequality, while the second inequality is due to the non-negativity of exponential function. Then we pull out the second term inside the expectation $\mathbb{E}_\varepsilon$ and get

$$\begin{aligned}
&\beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u) \right) \mathbb{E}_\varepsilon \left[\prod_{i=1}^{T-t} \exp \left(\frac{2}{\beta} \varepsilon_i \left[\mathbf{h}_i(\varepsilon) \right](u) \right) \right] \right) \\
&\leqslant \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u) \right) \times \exp \left(\frac{2}{\beta^2} \max_{\varepsilon_1, \ldots, \varepsilon_{T-t} \in \{\pm 1\}} \sum_{i=1}^{T-t} \left(\left[\mathbf{h}_i(\varepsilon) \right](u) \right)^2 \right) \right) \\
&\leqslant \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u) \right) \max_{u} \exp \left(\frac{2}{\beta^2} \max_{\varepsilon_1, \ldots, \varepsilon_{T-t} \in \{\pm 1\}} \sum_{i=1}^{T-t} \left(\left[\mathbf{h}_i(\varepsilon) \right](u) \right)^2 \right) \right) \\
&\leqslant \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{s=1}^{t} h_{x,s}(u) \right) \right) + \frac{2}{\beta} \sup_{\mathbf{h}} \max_{u \in \mathsf{U}} \max_{\varepsilon_1, \ldots, \varepsilon_{T-t} \in \{\pm 1\}} \sum_{i=1}^{T-t} \left([\mathbf{h}_i(\varepsilon)](u) \right)^2,
\end{aligned}$$

where the first inequality is due to Hoeffding's lemma (Hoeffding (1963))(also see, e.g., Lemma A.1 in Cesa-Bianchi and Lugosi (2006)) applied to the expectation w.r.t. ε. The last term, representing the worst-case future, is upper bounded by $\frac{2}{\beta}(T-t)L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2$. We thus obtain our exponential weight relaxation from (2.18).

Next we prove that the relaxation (2.18) is admissible and leads to the recursive algorithm (2.19). To keep the notation simple, we drop the subscript x in the following. In particular, we use h_t for $h_{x,t}$, $\widehat{W}_t$ for $\widehat{W}_{x,t}$, ν_t for $P_t(\cdot|x)$, etc. The admissibility condition to be proved is

$$\sup_{h_t \in \mathcal{H}_x} \left\{ \mathbb{E}_{U \sim \nu_t} \left[h_t(U) \right] + \widehat{W}_t(h^t) \right\} \leqslant \widehat{W}_{t-1}(h^{t-1}).$$

Note that

$$\begin{aligned}
\left\langle \nu_t, \exp\left(-\frac{1}{\beta} h_t\right) \right\rangle &= \sum_{u \in \mathsf{U}} \frac{\nu_1(u) \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t-1} h_s(u)\right)}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t-1} h_s\right) \right\rangle} \exp\left(-\frac{1}{\beta} h_t(u)\right) \\
&= \frac{\left\langle \nu_1, \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t} h_s\right) \right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t-1} h_s\right) \right\rangle}.
\end{aligned}$$

We have

$$\begin{aligned}
&\beta \log \left(\sum_{u \in \mathsf{U}} \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t} h_s(u) \right) \right) \\
&= \beta \log \left\langle \nu_1, \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t} h_s \right) \right\rangle + \beta \log |\mathsf{U}| \\
&= \beta \log \left\langle \nu_t, \exp\left(-\frac{1}{\beta} h_t \right) \right\rangle + \beta \log \left\langle \nu_1, \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t-1} h_s \right) \right\rangle + \beta \log |\mathsf{U}| \\
&\leqslant -\mathbb{E}_{U \sim \nu_t} h_t(U) + \frac{L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2}{2\beta} + \beta \log \left(\sum_{u \in \mathsf{U}} \exp\left(-\frac{1}{\beta} \sum_{s=1}^{t-1} h_s(u) \right) \right),
\end{aligned}$$

where the first equality is due to the fact that ν_1 is the uniform distribution on U, while the inequality is due to Hoeffding's lemma. Plugging the resulting bound into

the admissibility condition, we get

$$\begin{aligned}
&\sup_{h_t \in \mathcal{H}_x} \left\{ \mathbb{E}_{U \sim \nu_t} [h_t(U)] + \widehat{W}_{x,t}(h^t) \right\} \\
&\leqslant \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{s=1}^{t-1} h_s(u) \right) \right) + 2\frac{1}{\beta}(T-t+1)L(\mathsf{X},\mathsf{U},\mathcal{F})^2 \\
&= \widehat{W}_{x,t-1}(h^{t-1}).
\end{aligned}$$

Thus, the recursive algorithm (2.19) is admissible for the relaxation (2.18). □

The above algorithm works with any collection of $\widehat{Q}$ functions satisfying the reverse Poisson inequalities determined by the f_t's (recall Assumption 2). Here is one particular example of such a function — the usual Q-function that arises in reinforcement learning and that was used by Even-Dar et al. (2009). Recall our assumption that every randomized state feedback law $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ has a unique stationary distribution π_P. For given choices of $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ and $f \in \mathcal{F}$, consider the function

$$\widehat{Q}_{P,f}(x,u) = \lim_{T \to \infty} \mathbb{E}_P \left[\sum_{t=1}^{T} f(X_t, U_t) - \langle \pi_P \otimes P, f \rangle \middle| X_1 = x, U_1 = u \right],$$

where X_t and U_t are the state and action at time step t after starting from the initial state $X_1 = x$, applying the immediate action $U_1 = u$, and following P onwards. It is easy to check that $\widehat{Q}_{P,f}(x,u)$ satisfies the reverse Poisson inequality for P with forcing function f. In fact, it satisfies (2.8) with equality. We can also derive a bound on the Q-function in terms of the mixing time τ. Let us first bound $\widehat{Q}_{P,f}(x,P)$ where P is used on the first step instead of u. For all t, let $\mu_{x,t}^{P,f}$ be the state distribution

at time t starting from x and following P onwards. So we have

$$\begin{aligned}\widehat{Q}_{P,f}(x,P) = \lim_{T\to\infty}\sum_{t=1}^{T}\Big[\langle \mu_{x,t}^{P,f}\otimes P, f\rangle - \langle \pi_P \otimes P, f\rangle\Big] &\leqslant \|f\|_\infty \sum_{t=1}^{T}\|\delta_x P^t - \pi_P P^t\|_1 \\ &\leqslant 2\|f\|_\infty \sum_{t=1}^{T} e^{-t/\tau} \\ &\leqslant 2\tau\|f\|_\infty,\end{aligned}$$

where $\delta_x \in \mathcal{P}(\mathsf{X})$ is the Dirac distribution centered at x, and the first inequality results from repeated application of the uniform mixing bound (2.6). Due to the fact that the one-step cost is bounded by $C_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\|f\|_\infty$, we have

$$\widehat{Q}_{P,f}(x,u) \leqslant \widehat{Q}_{P,f}(x,P) + f(x,u) - \langle \mu_{x,1}^{P,f}\otimes P, f\rangle \leqslant 2\tau C_{\mathcal{F}} + C_{\mathcal{F}} \leqslant 3\tau C_{\mathcal{F}}.$$

We can now establish the following regret bound for the exponential weights strategy (2.19):

Theorem 7. *Let $L \triangleq L(\mathsf{X},\mathsf{U},\mathcal{F})$. Assume the uniform mixing condition is satisfied by the controlled transition kernel K. Then for the relaxation* (2.18) *and the corresponding behavioral strategy $\widehat{\boldsymbol{\gamma}}$ given by* (2.19) *with $\beta = \sqrt{\frac{2TL^2}{\log|\mathsf{U}|}}$, we have*

$$\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \leqslant 2L\sqrt{2T\log|\mathsf{U}|} + C_{\mathcal{F}}(\tau+1)^2\sqrt{\frac{\log|\mathsf{U}|T}{2}} + (2\tau+2)C_{\mathcal{F}}.$$

Proof. Again, we drop the subscript x and write ν_t for $P_t(\cdot|x)$, etc. We have

$$\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})}\sum_x \pi_P(x)\widehat{W}_{x,0} + C_{\mathcal{F}}\sum_{t=1}^{T}\|\mu_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\|_1. \tag{2.20}$$

From the relaxation (2.18), it is easy to see $\widehat{W}_{x,0} \leqslant 2L\sqrt{2T\log|\mathsf{U}|}$ for all states x (in fact, the bound is met with equality with the optimal choice of $\beta = \sqrt{\frac{2TL^2}{\log|\mathsf{U}|}}$). Since

we have bounded the first term, now we focus on bounding the second term of the regret bound.

The relative entropy between ν_t and ν_{t-1} is given by

$$D(\nu_t \| \nu_{t-1}) = \left\langle \nu_t, \log \frac{\exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-1} h_s\right)}{\exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-2} h_s\right)} \right\rangle + \log \frac{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-1} h_s\right)\right\rangle}$$

$$= -\frac{1}{\beta}\langle \nu_t, h_{t-1}\rangle + \log \frac{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-1} h_s\right)\right\rangle}, \tag{2.21}$$

where

$$\frac{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-1} h_s\right)\right\rangle} = \frac{\sum_{u\in\mathsf{U}} \nu_1(u) \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-1} h_s(u)\right) \exp\left(\frac{1}{\beta} h_{t-1}(u)\right)}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-1} h_s\right)\right\rangle}$$

$$= \left\langle \nu_t, \exp\left(\frac{1}{\beta} h_{t-1}\right)\right\rangle.$$

Using Hoeffding's lemma, we can write

$$\log \frac{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta}\sum_{s=1}^{t-1} h_s\right)\right\rangle} \leqslant \frac{1}{\beta}\langle \nu_t, h_{t-1}\rangle + \frac{L^2}{2\beta^2}$$

Substituting this bound into (2.21), we see that the terms involving the expectation of h_{t-1} w.r.t. ν_t cancel, and we are left with

$$D(\nu_t \| \nu_{t-1}) \leqslant \frac{L^2}{2\beta^2}.$$

Plugging in the optimal value of β and using Pinsker's inequality (Cover and Thomas (2006)), we find

$$\|\nu_t - \nu_{t-1}\|_1 \leqslant \sqrt{\frac{\log|\mathsf{U}|}{2T}}.$$

So far, we have been working with a fixed state $x \in \mathsf{X}$, so we had $\nu_t = P_t^{\widehat{\gamma},\boldsymbol{f}}(\cdot|x)$, where $\widehat{\boldsymbol{\gamma}}$ is the agent's behavioral strategy induced by the relaxation (2.18). Since x was arbitrary, we get the uniform bound

$$\max_{x\in\mathsf{X}} \left\| P_t^{\widehat{\gamma},\boldsymbol{f}}(\cdot|x) - P_{t-1}^{\widehat{\gamma},\boldsymbol{f}}(\cdot|x) \right\|_1 \leqslant \sqrt{\frac{\log \mathsf{U}|}{2T}}. \tag{2.22}$$

Armed with this estimate, we now bound the total variation distance between the actual state distribution at time t and the unique invariant distribution of $K_t^{\widehat{\gamma},\boldsymbol{f}}$. For any time $k \leqslant t$, we have

$$\begin{aligned}
\left\| \mu_k^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 &= \left\| \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_{k-1}^{\widehat{\gamma},\boldsymbol{f}} - \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_t^{\widehat{\gamma},\boldsymbol{f}} + \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_t^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 \\
&\overset{\text{(a)}}{\leqslant} \left\| \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_t^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 + \left\| \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_{k-1}^{\widehat{\gamma},\boldsymbol{f}} - \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 \\
&\overset{\text{(b)}}{=} \left\| \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_t^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}} K_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 + \left\| \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_{k-1}^{\widehat{\gamma},\boldsymbol{f}} - \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} K_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 \\
&\overset{\text{(c)}}{\leqslant} e^{-1/\tau} \left\| \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 + \max_{x\in\mathsf{X}} \left\| P_{k-1}^{\widehat{\gamma},\boldsymbol{f}}(\cdot|x) - P_t^{\widehat{\gamma}}(\cdot|x) \right\|_1 \\
&\overset{\text{(d)}}{\leqslant} e^{-1/\tau} \left\| \mu_{k-1}^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}} \right\|_1 + \sum_{i=k-1}^{t-1} \sqrt{\frac{\log |\mathsf{U}|}{2T}},
\end{aligned} \tag{2.23}$$

where (a) is by triangle inequality; (b) is by invariance of $\pi_t^{\widehat{\gamma},\boldsymbol{f}}$ w.r.t. $K_t^{\widehat{\gamma},\boldsymbol{f}}$; (c) is by the uniform mixing bound (2.6); and (d) follows from repeatedly using (2.22) together with triangle inequality and the easily proved fact that, for any state distribution $\mu \in \mathcal{P}(\mathsf{X})$ and any two Markov kernels $P, P' \in \mathcal{M}(\mathsf{U}|\mathsf{X})$,

$$\|\mu K(\cdot|P) - \mu K(\cdot|P')\|_1 \leqslant \max_{x\in\mathsf{X}} \|P(\cdot|x) - P'(\cdot|x)\|_1 .$$

Letting now the initial state distribution be μ_1, we can apply the bound (2.23)

recursively to obtain

$$\begin{aligned}
\left\|\mu_t^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}}\right\|_1 &\leqslant e^{-(t-1)/\tau}\left\|\mu_1 - \pi_t^{\widehat{\gamma},\boldsymbol{f}}\right\|_1 + \sum_{k=2}^{t} e^{-\frac{t-k}{\tau}} \sum_{i=k-1}^{t} \sqrt{\frac{\log|\mathsf{U}|}{2T}} \\
&\leqslant 2e^{-(t-1)/\tau} + \sum_{k=2}^{t} e^{-\frac{t-k}{\tau}}(t-k+1)\sqrt{\frac{\log|\mathsf{U}|}{2T}} \\
&\leqslant 2e^{-(t-1)/\tau} + \sqrt{\frac{\log|\mathsf{U}|}{2T}} \sum_{k=0}^{\infty}(k+1)e^{-\frac{k}{\tau}} \\
&\leqslant 2e^{-(t-1)/\tau} + (\tau+1)^2\sqrt{\frac{\log|\mathsf{U}|}{2T}}.
\end{aligned}$$

So, the second term on the right-hand side of (2.20) can bounded by

$$C_{\mathcal{F}} \sum_{t=1}^{T} \|\mu_t^{\widehat{\gamma},\boldsymbol{f}} - \pi_t^{\widehat{\gamma},\boldsymbol{f}}\|_1 \leqslant C_{\mathcal{F}}(\tau+1)^2\sqrt{\frac{\log|\mathsf{U}|T}{2}} + (2\tau+2)C_{\mathcal{F}},$$

which completes the proof. □

As we can see, this regret bound is consistent with the bound derived in Even-Dar et al. (2009). Therefore, we have shown that our framework, with a specific choice of relaxation, can recover their algorithm. The advantage of our general framework is that we can analyze the part of the corresponding regret bound simply by instantiating our analysis on specific relaxations, without the need of ad-hoc proof techniques applied in Even-Dar et al. (2009).

The above policy relies on exponential weight updates. We now present a "lazy" version of that policy, wherein time is divided into phases of increasing length, and during each phase the agent applies a fixed state feedback law. The main advantage of lazy strategies is their computational efficiency, which is the result of a looser relaxation and hence suboptimal scaling of the regret with the time horizon.

We partition the set of time indices $1, 2, \ldots$ into nonoverlapping contiguous phases of (possibly) increasing duration. The phases are indexed by $m \in \mathbb{N}$, where we denote

the mth phase by $\mathcal{T}_m$ and its duration by τ_m. We also define $\mathcal{T}_{1:m} \triangleq \mathcal{T}_1 \cup \ldots \cup \mathcal{T}_m$ (the union of phases 1 through m) and denote its duration by $\tau_{1:m}$. Let $M \leqslant T$ denote the number of complete phases concluded before time T. Here we need to describe a generic algorithm that works in phases:

```
Initialize at t = 0 and phases T_1, ..., T_M s.t. τ_{1:M} = T
For t ∈ T_1, choose u_t uniformly at random over U
for m = 2, 3, ...
  for t ∈ T_m do
    if the process is at state x, choose action u_t randomly according to P_m(u|x)
    where P_m(u|x) is the state feedback law only using information from
    phase 1 to m − 1
  end for
end for
```

Because now we work in phases instead of time steps, we need to provide an alternative definition of relaxations and admissibility condition. For every state $x \in \mathsf{X}$, we denote by h_x^m the τ_m-tuple $(h_{x,s} : s \in \mathcal{T}_m)$, and by $h_{x,1:m}$ the $\tau_{1:m}$-tuple $(h_{x,1}, h_{x,2}, \ldots, h_{x,\tau_{1:m}})$. For each $x \in \mathsf{X}$, we will say that a sequence of functions $\widehat{W}_{x,m} : \mathcal{H}_x^{\tau_{1:m}} \to \mathbb{R}, m = 1, \ldots, M$, is an admissible relaxation if

$$\widehat{W}_{x,M}(h_{x,1:M}) \geqslant -\inf_{\nu \in \mathcal{P}(\mathsf{U})} \mathbb{E}_{U\sim\nu}\left[\sum_{t=1}^{T} h_{x,t}(U)\right]$$

$$\widehat{W}_{x,m}(h_{x,1:m}) \geqslant \inf_{\nu \in \mathcal{P}(\mathsf{U})} \sup_{h_x^m \in \mathcal{H}_x^{\tau_m}} \left\{\mathbb{E}_{U\sim\nu}\left[\sum_{s\in\mathcal{T}_m} h_{x,s}(U)\right] + \widehat{W}_{x,m+1}(h_{x,1:m}, h_x^{m+1})\right\}.$$

For a given state x, we also define the conditional SRC in terms of phases:

$$\mathcal{R}_{x,m}(h_{x,1:m}) = \sup_{\mathbf{h}} \mathbb{E}_{\varepsilon_{m+1:M}} \max_{u\in\mathsf{U}} \left[2\sum_{j=m+1}^{M} \varepsilon_j \sum_{t\in\mathcal{T}_j} \left[\mathbf{h}_{x,t}(\varepsilon)\right](u) - \sum_{i=1}^{m}\sum_{s\in\mathcal{T}_i} h_{x,s}(u)\right].$$

Here the supremum is taken over all $\mathcal{H}_x$-valued binary trees of depth $M - m$. When recovering the method in Even-Dar et al. (2009), we replaced the actual future induced by the infimum and supremum pairs in the conditional value by the "worst

future" binary tree, which involves expectation over a sequence of coin flips in every time step. By contrast, in the above quantity we replace the real future by the "worst future" binary tree that branches only once per phase. Now we can construct the following relaxation:

$$\widehat{W}_{x,m}(h_{x,1:m}) = \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{i=1}^{m} \sum_{s \in \mathcal{T}_i} h_{x,s}(u) \right) \right) + \frac{2L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2}{\beta} \sum_{j=m+1}^{M} \tau_j^2. \tag{2.24}$$

The corresponding algorithm, specified in (2.25) below, uses a fixed state feedback law throughout each phase:

Proposition 8. *The relaxation* (2.24) *is admissible and it leads to the following Markov policy for phase* m*:*

$$P_m(u|x) = \frac{\nu_1(u) \exp\left(-\frac{1}{\beta} \sum_{i=1}^{m-1} \sum_{s \in \mathcal{T}_i} h_{x,s}(u) \right)}{\left\langle \nu_1, \exp\left(-\frac{1}{\beta} \sum_{i=1}^{m-1} \sum_{s \in \mathcal{T}_i} h_{x,s} \right) \right\rangle}, \tag{2.25}$$

where ν_1 *is the uniform distribution on* U*.*

Proof. First we show that the relaxation (2.24) arises as an upper bound on the conditional SRC. Once again, we omit the subscript x from $h_{x,t}$ etc. to keep the notation light. Following the same steps as in the proof for Proposition 6, we have, for any $\beta > 0$,

$$\begin{aligned}
&\mathbb{E}_\varepsilon \left[\max_{u \in \mathsf{U}} \left\{ 2 \sum_{j=m+1}^{M} \varepsilon_j \sum_{t \in \mathcal{T}_j} \left[\mathbf{h}_t(\varepsilon) \right](u) - \sum_{i=1}^{m} \sum_{s \in \mathcal{T}_i} h_s(u) \right\} \right] \\
&\leqslant \beta \log \left(\mathbb{E}_\varepsilon \left[\max_{u \in \mathsf{U}} \exp \left(\frac{2}{\beta} \sum_{j=m+1}^{M} \varepsilon_j \sum_{t \in \mathcal{T}_j} \left[\mathbf{h}_t(\varepsilon) \right](u) - \frac{1}{\beta} \sum_{i=1}^{m} \sum_{s \in \mathcal{T}_i} h_s(u) \right) \right] \right) \\
&\leqslant \beta \log \left(\mathbb{E}_\varepsilon \left[\sum_{u \in \mathsf{U}} \exp \left(\frac{2}{\beta} \sum_{j=m+1}^{M} \varepsilon_j \sum_{t \in \mathcal{T}_j} \left[\mathbf{h}_t(\varepsilon) \right](u) - \frac{1}{\beta} \sum_{i=1}^{m} \sum_{s \in \mathcal{T}_i} h_s(u) \right) \right] \right).
\end{aligned}$$

Denote $g_m = \sum_{u\in\mathsf{U}} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{m}\sum_{s\in\mathcal{T}_i} h_s(u)\right)$. In the same vein,

$$\begin{aligned}
&\beta \log\left(g_m \mathbb{E}_\varepsilon\left[\prod_{j=m+1}^{M} \exp\left(\frac{2}{\beta}\varepsilon_j \sum_{t\in\mathcal{T}_j}\left[\mathbf{h}_t(\varepsilon)\right](u)\right)\right]\right)\\
&\leqslant \beta \log\left(g_m \times \exp\left(\frac{2}{\beta^2}\max_{\varepsilon_{m+1},\ldots,\varepsilon_M\in\{\pm 1\}} \sum_{j=m+1}^{M}\left(\tau_j\left[\mathbf{h}(\varepsilon)\right](u)\right)^2\right)\right)\\
&\leqslant \beta \log\left(g_m \max_u \exp\left(\frac{2}{\beta^2}\max_{\varepsilon_{m+1},\ldots,\varepsilon_M\in\{\pm 1\}} \sum_{j=m+1}^{M}\left(\tau_j\left[\mathbf{h}(\varepsilon\right](u)\right)^2\right)\right)\\
&\leqslant \beta \log\left(g_m\right) + \frac{2}{\beta}\sup_{\mathbf{h}}\max_{u\in\mathsf{U}}\max_{\varepsilon_{m+1},\ldots,\varepsilon_M\in\{\pm 1\}} \sum_{j=m+1}^{M}\left(\tau_j\left[\mathbf{h}(\varepsilon)\right](u)\right)^2\\
&\leqslant \beta \log\left(\sum_{u\in\mathsf{U}} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{m}\sum_{s\in\mathcal{T}_i} h_s(u)\right)\right) + \frac{2}{\beta}\sum_{j=m+1}^{M}\tau_j^2 L(\mathsf{X},\mathsf{U},\mathcal{F})^2,
\end{aligned}$$

where the first inequality is due to Hoeffding's lemma, while the last inequality is by Assumption 2. We thus derive the relaxation in (2.24).

Now we prove that this relaxation is admissible, and leads to the lazy algorithm (2.25) The admissibility condition to be proved is

$$\sup_{h_m\in\mathcal{H}_x^{\tau_m}}\left\{\mathbb{E}_{U\sim\nu_m}\left[\sum_{s\in\mathcal{T}_m} h_s(U)\right] + \widehat{W}_{x,m}(h_{1:m})\right\} \leqslant \widehat{W}_{x,m-1}(h_{1:m-1}),$$

where $\nu_m = P_m(\cdot|x)$ is the Markov policy used in phase m. We have

$$\begin{aligned}
&\beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{i=1}^{m} \sum_{s \in \mathcal{T}_i} h_s(u) \right) \right) \\
&= \beta \log \left\langle \nu_1, \exp \left(-\frac{1}{\beta} \sum_{i=1}^{m} \sum_{s \in \mathcal{T}_i} h_s \right) \right\rangle + \beta \log |\mathsf{U}| \\
&= \beta \log \left\langle \nu_m, \exp \left(-\frac{1}{\beta} \sum_{s \in \mathcal{T}_m} h_s \right) \right\rangle + \beta \log \left\langle \nu_1, \exp \left(-\frac{1}{\beta} \sum_{i=1}^{m-1} \sum_{s \in \mathcal{T}_i} h_s \right) \right\rangle + \beta \log |\mathsf{U}| \\
&\leqslant -\mathbb{E}_{U \sim \nu_m} \left[\sum_{s \in \mathcal{T}_m} h_s(U) \right] + \frac{\tau_m^2 L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2}{2\beta} + \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{i=1}^{m-1} \sum_{s \in \mathcal{T}_i} h_s(u) \right) \right),
\end{aligned}$$

Plugging this into the admissibility condition, we have

$$\begin{aligned}
&\sup_{h_m \in \mathcal{H}_x^{\tau_m}} \left\{ \mathbb{E}_{U \sim \nu_m} \left[\sum_{s \in \mathcal{T}_m} h_s(U) \right] + \widehat{W}_{x,m}(h_{1:m}) \right\} \\
&\leqslant \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{i=1}^{m-1} \sum_{s \in \mathcal{T}_i} h_s(u) \right) \right) + \frac{2}{\beta} \sum_{j=m+1}^{M} \tau_j^2 L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2 + \frac{\tau_m^2 L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2}{2\beta} \\
&\leqslant \beta \log \left(\sum_{u \in \mathsf{U}} \exp \left(-\frac{1}{\beta} \sum_{i=1}^{m-1} \sum_{s \in \mathcal{T}_i} h_s(u) \right) \right) + \frac{2}{\beta} \sum_{j=m}^{M} \tau_j^2 L(\mathsf{X}, \mathsf{U}, \mathcal{F})^2 \\
&= \widehat{W}_{x,m-1}(h^{m-1}).
\end{aligned}$$

So the lazy algorithm (2.25) is an admissible strategy for the relaxation (2.24). □

Now we derive the regret bound for (2.25):

Theorem 9. *Let $L \triangleq L(\mathsf{X}, \mathsf{U}, \mathcal{F})$. Under the same assumptions as before, the behavioral strategy $\widehat{\boldsymbol{\gamma}}$ corresponding to* (2.25) *enjoys the following regret bound when $\beta = \sqrt{\frac{2\sum_{i=1}^{M} \tau_i^2 L^2}{\log |\mathsf{U}|}}$:*

$$\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right] \leqslant 2L \sqrt{2 \log |\mathsf{U}| \sum_{i=1}^{M} \tau_i^2} + \frac{2 C_{\mathcal{F}} M}{1 - e^{-1/\tau}}. \tag{2.26}$$

Proof. The state feedback law $P_t^{\hat{\gamma},\boldsymbol{f}}(\cdot|x)$ that the agent applies within phase m is the same for all $t \in \mathcal{T}_m$, and we denote it by $P_m^{\hat{\gamma},\boldsymbol{f}}(\cdot|x)$. Let $K_m^{\hat{\gamma},\boldsymbol{f}}$ denote the Markov matrix that describes the state transition from X_t to X_{t+1} if $t \in \mathcal{T}_m$. Thus, we can write

$$K_m^{\hat{\gamma},\boldsymbol{f}}(y|x) = \sum_u K(y|x,u) P_t^{\hat{\gamma},\boldsymbol{f}}(u|x), \qquad \forall x, y \in \mathsf{X}.$$

First, we show that

$$\mathbb{E}_x^{\hat{\gamma},\boldsymbol{f}}\left[\sum_{t=1}^T f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\widehat{W}_{x,0} + C_{\mathcal{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\hat{\gamma},\boldsymbol{f}} - \pi_m^{\hat{\gamma},\boldsymbol{f}}\|_1, \tag{2.27}$$

where $\pi_m^{\hat{\gamma},\boldsymbol{f}}$ is the invariant distribution of $K_m^{\hat{\gamma},\boldsymbol{f}}$.

To prove (2.27), we write

$$\begin{aligned}
&\mathbb{E}_x^{\hat{\gamma},\boldsymbol{f}}\left[\sum_{t=1}^T f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \\
&\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})} \sum_{t=1}^T \left[\langle \pi_t^{\hat{\gamma},\boldsymbol{f}} \otimes P_t^{\hat{\gamma},\boldsymbol{f}}, f_t\rangle - \langle \pi_P \otimes P, f_t\rangle\right] + \sum_{t=1}^T \|f_t\|_\infty \|\mu_t^{\hat{\gamma},\boldsymbol{f}} - \pi_t^{\hat{\gamma},\boldsymbol{f}}\|_1 \\
&\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \left[\langle \pi_t^{\hat{\gamma},\boldsymbol{f}} \otimes P_t^{\hat{\gamma},\boldsymbol{f}}, f_t\rangle - \langle \pi_P \otimes P, f_t\rangle\right] + C_{\mathcal{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\hat{\gamma},\boldsymbol{f}} - \pi_m^{\hat{\gamma},\boldsymbol{f}}\|_1 \\
&\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x) \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \left(\sum_u P_m^{\hat{\gamma},\boldsymbol{f}}(u|x) \widehat{Q}_t^{\hat{\gamma},\boldsymbol{f}}(x,u) - P(u|x) \widehat{Q}_t^{\hat{\gamma},\boldsymbol{f}}(x,u)\right) \\
&\quad + C_{\mathcal{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\hat{\gamma},\boldsymbol{f}} - \pi_m^{\hat{\gamma},\boldsymbol{f}}\|_1,
\end{aligned}$$

where the last inequality is by Lemma 3. By writing out the first term in the right

hand side, we get

$$\sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})}\sum_x \pi_P(x)\Bigg[\sum_{m=1}^{M-1}\sum_{t\in\mathcal{T}_m}\left(\sum_u P_m^{\widehat{\gamma},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\gamma},\boldsymbol{f}}(x,u)\right)+\sum_u \nu_M(u|x)\sum_{t\in\mathcal{T}_M}\widehat{Q}_t^{\widehat{\gamma},\boldsymbol{f}}(x,u)$$

$$-\sum_{t=1}^{T}P(u|x)\widehat{Q}_t^{\widehat{\gamma},\boldsymbol{f}}(x,u)\Bigg]+C_{\mathcal{F}}\sum_{m=1}^{M}\sum_{t\in\mathcal{T}_m}\|\mu_t^{\widehat{\gamma},\boldsymbol{f}}-\pi_m^{\widehat{\gamma},\boldsymbol{f}}\|_1$$

$$\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})}\sum_x \pi_P(x)\Bigg[\sum_{m=1}^{M-1}\sum_{t\in\mathcal{T}_m}\left(\sum_u P_m^{\widehat{\gamma},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\gamma},\boldsymbol{f}}(x,u)\right)$$

$$+\sum_u \nu_M(u|x)\sum_{t\in\mathcal{T}_M}\widehat{Q}_t^{\widehat{\gamma},\boldsymbol{f}}(x,u)+\widehat{W}_{x,M}(h^M)\Bigg]+C_{\mathcal{F}}\sum_{m=1}^{M}\sum_{t\in\mathcal{T}_m}\|\mu_t^{\widehat{\gamma},\boldsymbol{f}}-\pi_m^{\widehat{\gamma},\boldsymbol{f}}\|_1$$

$$\leqslant \sup_{P\in\mathcal{M}(\mathsf{U}|\mathsf{X})}\sum_x \pi_P(x)\left[\sum_{m=1}^{M-1}\sum_{t\in\mathcal{T}_m}\left(\sum_u P_m^{\widehat{\gamma},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\gamma},\boldsymbol{f}}(x,u)\right)+\widehat{W}_{x,M-1}(h^{M-1})\right]$$

$$+\sum_{t=1}^{T}\|f_t\|_\infty\left\|\mu_t^{\widehat{\gamma},\boldsymbol{f}}-\pi_t^{\widehat{\gamma},\boldsymbol{f}}\right\|_1.$$

The last inequality is due to the fact that $\widehat{\boldsymbol{\gamma}}$ is the behavioral strategy associated to the admissible relaxation $\{\widehat{W}_{x,m}\}_{m=1}^M$. Continuing this induction backwards, we arrive at (2.27).

Next, we bound the two terms on the right-hand side of (2.27). From the form of the relaxation (2.24), it is easy to see $\widehat{W}_{x,0}\leqslant 2L\sqrt{2\log|\mathsf{U}|\sum_{i=1}^M\tau_i^2}$ for all states x; in fact, this bound is attained with equality if we use the optimal choice $\beta=\sqrt{\frac{2\sum_{i=1}^M\tau_i^2L^2}{\log|\mathsf{U}|}}$. Since we have bounded the first term, now we focus on bounding the second term of (2.27).

From the contraction inequality (2.6) it follows that, for every $k\in\{0,1,\ldots,\tau_m-$

1}, we have

$$\begin{aligned}\left\|\mu^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{\tau_{1:m-1}+k+1}-\pi^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{m}\right\|_1 &= \left\|\mu^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{\tau_{1:m-1}+1}(K^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{m})^{k}-\pi^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{m}(K^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{m})^{k}\right\|_1\\ &\leqslant e^{-k/\tau}\left\|\mu^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{\tau_{1:m-1}+1}-\pi^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{m}\right\|_1\\ &\leqslant 2e^{-k/\tau}.\end{aligned}$$

Hence,

$$\sum_{t\in\mathcal{T}_m}\left\|\mu^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{t}-\pi^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{m}\right\|_1 \leqslant 2\sum_{k=0}^{\tau_m-1}e^{-k/\tau}\leqslant\frac{2}{1-e^{-1/\tau}}.$$

Plugging it in (2.27), we have shown that

$$\mathbb{E}^{\hat{\boldsymbol{\gamma}},\boldsymbol{f}}_{x}\left[\sum_{t=1}^{T}f_t(X_t,U_t)-\Psi(\boldsymbol{f})\right]\leqslant 2L\sqrt{2\log|\mathsf{U}|\sum_{i=1}^{M}\tau_i^2}+\frac{2C_{\mathcal{F}}M}{1-e^{-1/\tau}}.$$

□

Our behavioral strategy (2.19) is a novel RWM algorithm for online MDPs. Yu et al. (2009) also consider a similar model, where the decision-maker has full knowledge of the transition kernel, and the costs are chosen by an oblivious (open-loop) adversary. They propose an algorithm that computes and changes the policy periodically according to a perturbed version of the empirically observed cost functions, and then follows the computed stationary policy for increasingly long time intervals. As a result, their algorithm achieves sublinear regret and has diminishing computational effort per time step; in particular, it is computationally more efficient than that of Even-Dar et al. (2009).

Although our new algorithm is similar in nature to the algorithm of Yu et al. (2009), it has several advantages. First, in the algorithm of Yu et al. (2009), the policy computation at the beginning of each phase requires solving a linear program and then adding a carefully tuned random perturbation to the solution. As a result, the

performance analysis in Yu et al. (2009) is rather lengthy and technical (in particular, it invokes several advanced results from perturbation theory for linear programs). By contrast, our strategy is automatically randomized, and the performance analysis is a lot simpler. Second, the regret bound of Theorem 9 shows that we can control the scaling of the regret with T by choosing the duration of each phase, whereas the algorithm of Yu et al. (2009) relies on a specific choice of phase durations in order to guarantee that the regret is sublinear in T and scales as $O(T^{3/4})$. We show that if the horizon T is known in advance, then it is possible to choose the phase durations to secure $O(T^{2/3})$ regret, which is better than the $O(T^{3/4})$ bound derived by Yu et al. (2009).

Corollary 10. *Consider the setting of Theorem 9. For a given horizon T, the optimal choice of phase lengths is $T^{1/3}$, which gives the regret of $O(T^{2/3})$.*

Proof. Let us inspect the right-hand side of (2.26). We see that both $\sqrt{\sum_{j=1}^{M} \tau_j^2}$ and M have to be sublinear in T. Since $\sum_{i=1}^{M} \tau_i = T$ and $\sqrt{\sum_{i=1}^{M} \tau_i^2} < \sqrt{(\sum_{i=1}^{M} \tau_i)^2}$, at least the first of these terms can be made sublinear, e.g., by having $\tau_j = 1$ for all j. Of course, this means that $M = T$, so we need longer phases. For example, if we follow Yu et al. (2009) and let $\tau_m = \lceil m^{1/3-\varepsilon} \rceil$ for some $\varepsilon \in (0, 1/3)$, then a straightforward if tedious algebraic calculation shows that $M = O(T^{3/4})$ and $\sqrt{\sum_{j=1}^{M} \tau_j^2} = O(T^{5/8})$, which yields the regret of $O(T^{3/4})$.

However, if T is known in advance, then we can do better: ignoring the rounding issues, for any constants $A_1, A_2 > 0$,

$$\min_{1 \leqslant M \leqslant T} \min \left\{ A_1 \sqrt{\sum_{j=1}^{M} \tau_j^2} + A_2 M : \sum_{j=1}^{M} \tau_j = T \right\} = O(T^{2/3}), \tag{2.28}$$

To see this, let us first fix M and optimize the choice of the τ_j's:

$$\min \sum_{j=1}^{M} \tau_j^2 \quad \text{subject to } \sum_{j=1}^{M} \tau_j = T.$$

By the Cauchy–Schwarz inequality, we have

$$\sum_{j=1}^{M} \tau_j \leqslant \sqrt{M \sum_{j=1}^{M} \tau_j^2}.$$

Thus, $\sum_{j=1}^{M} \tau_j^2$ achieves its minimum when the above bound is met with equality. This will happen only if all the τ_j's are equal, i.e., $\tau_j = \frac{T}{M}$ for every j (for simplicity, we assume that M divides T — otherwise, the remainder term will be strictly smaller than M, and the bound in (2.28) will still hold, but with a larger multiplicative constant). Therefore,

$$\min_{1 \leqslant M \leqslant T} \min \left\{ A_1 \sqrt{\sum_{j=1}^{M} \tau_j^2} + A_2 M : \sum_{j=1}^{M} \tau_j = T \right\} = \min_{1 \leqslant M \leqslant T} \left(\frac{A_1 T}{\sqrt{M}} + A_2 M \right) = O(T^{2/3}),$$

where the minimum on the right-hand side (again, ignoring rounding issues) is achieved by $M = T^{2/3}$ and $\tau_j = T^{1/3}$ for all j. This shows that, for a given horizon T, the optimal choice of phase lengths is $T^{1/3}$, which gives the regret of $O(T^{2/3})$, better than the $O(T^{3/4})$ bound derived by Yu et al. (2009). □

2.4.2 The convex-analytic approach

In this section, we use the convex-analytic approach to derive an algorithm that relies on the reduction of the online MDP problem to an online linear optimization problem over the state-action polytope [recall the definition in Eq. (2.12)]. Structurally, this algorithm is similar to the Online Mirror Descent scheme proposed and analyzed recently by Dick et al. (2014); however, its key ingredients and the resulting performance guarantee on the regret are more closely related to interior-point methods of

Abernethy et al. (2012). Moreover, we will show that this algorithm arises from an admissible relaxation with respect to (2.16).

We start by introducing the definition of a *self-concordant barrier*, which is basic to the theory of interior point methods (Nesterov and Nemirovski (1994); Nemirovski and Todd (2008)): Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a closed convex set with nonempty interior. A function $F : \operatorname{int}(\mathcal{K}) \to \mathbb{R}$ is a *barrier* on $\mathcal{K}$ if $F(x_i) \to +\infty$ along any sequence $\{x_i\}_{i=1}^{\infty} \subset \mathcal{K}$ that converges to a boundary point of $\mathcal{K}$. Moreover, F is *self-concordant* if it is a convex C^3 function, such that the inequality

$$\nabla^3 F(v)[h,h,h] \leqslant 2\left(\nabla^2 F(v)[h,h]\right)^{3/2}$$

holds for all $v \in \operatorname{int}(\mathcal{K})$ and $h \in \mathbb{R}^n$. Here, $\nabla^2 F(v)$ and $\nabla^3 F(v)$ are the Hessian and the third-derivative tensor of F at v, respectively. We also need some geometric quantities induced by F. The first is the *Bregman divergence* $D_F : \operatorname{int}(\mathcal{K}) \times \operatorname{int}(\mathcal{K}) \to \mathbb{R}^+$, defined by

$$D_F(v,w) \triangleq F(v) - F(w) - \langle \nabla F(w), v - w \rangle, \qquad v, w \in \operatorname{int}(\mathcal{K}). \tag{2.29}$$

The second is the *local norm* of $h \in \mathbb{R}^n$ around a point $v \in \operatorname{int}(\mathcal{K})$ (assuming $\nabla^2 F(v)$ is nondegenerate):

$$\|h\|_v \triangleq \sqrt{\nabla^2 F(v)[h,h]}.$$

Finally, if F is self-concodrant, then so is its *Legendre–Fenchel dual* $F^*(h) \triangleq \sup_{v \in \operatorname{int}(\mathcal{K})} \{\langle h, v \rangle - F(v)\}$. Thus, the definitions of the Bregman divergence and the local norm carry over to F^*. Specifically,

$$\|f\|_h^* \triangleq \sqrt{\langle f, \nabla^2 F^*(h) f \rangle} \equiv \sqrt{\nabla^2 F^*(h)[f,f]}$$

is the local norm of f at h induced by F^* (by the following assumption that F^* is strictly convex, this local norm is well-defined everywhere).

Both our algorithm and the relaxation that induces it revolve around a self-concordant barrier for the set $\mathcal{K} = \mathcal{G}$, the state-action polytope of our MDP. This

set is a compact convex subset of $\mathbb{R}^{|\mathsf{X}|\times|\mathsf{U}|}$ with nonempty interior. We make the following assumption:

Assumption 3. *The state-action polytope $\mathcal{G}$ associated to the MDP with controlled transition law K admits a self-concordant barrier $F : \mathrm{int}(\mathcal{G}) \to \mathbb{R}$ with the following properties:*

1. *F is strictly convex on $\mathrm{int}(\mathcal{G})$, and its dual F^* is strictly convex on $\mathbb{R}^{|\mathsf{X}|\times|\mathsf{U}|}$.*
2. *The gradient map $\nu \mapsto \nabla F(\nu)$ is a bijection between $\mathrm{int}(\mathcal{G})$ and $\mathbb{R}^{|\mathsf{X}|\times|\mathsf{U}|}$, and admits the map $h \mapsto \nabla F^*(h)$ as inverse.*
3. *The minimum value of F on $\mathrm{int}(\mathcal{G})$ is equal to 0.*

This assumption is not difficult to meet in practice. For example, the *universal entropic barrier* of Bubeck and Eldan (2015) (which can be constructed for any compact convex polytope) satisfies these requirements.

We are now ready to present our algorithm and the associated relaxation. We start by describing the former:

For $t = 1, 2, \ldots$ do
  If $t = 1$, choose $\nu_t = \nu^* \equiv \arg\min_{\nu\in\mathrm{int}(\mathcal{G})} F(\nu)$;
  else choose $\nu_t = \nabla F^* \left(\nabla F(\nu_{t-1}) - \beta f_{t-1}\right)$
  Construct the policy $P_t = P_{\nu_t}$ according to Eq. (2.13)
  Observe the state X_t
  Draw the action $U_t \sim P_t(\cdot|X_t)$ and obtain f_t from the environment
end for

Here, $\beta > 0$ is the tunable learning rate. Note also that, by virtue of Assumption 3, the sequence of measures $\{\nu_t\}$ lies in $\mathrm{int}(\mathcal{G})$. Next, we describe the relaxation. For reasons that will be spelled out shortly, we focus on the regret with respect to policies induced by elements of a given subset $\mathcal{G}'$ of $\mathrm{int}(\mathcal{G})$. For $t = 0, \ldots, T$, we let

$$\widehat{V}_T(\mathcal{G}'|f_1, \ldots, f_t) = \sup_{\mu\in\mathcal{G}'} \left\{ \sum_{s=1}^{t} \langle \mu, -f_s \rangle + \frac{1}{\beta} D_F(\mu, \nu_{t+1}) \right\} + 2\beta(T - t). \tag{2.30}$$

Note that ν_{t+1} is a deterministic function of $f_1, \ldots, f_t$, and therefore the relaxation is well-defined.

Proposition 11. *Suppose that the learning rate β is such that $\beta\|f_t\|^*_{\nabla F(\nu_t)} \leqslant 1/2$ for all $t = 1, \ldots, T$. Assume that $\|f_t\|^*_{\nabla F(\nu_t)} \leqslant 1$, for all $t = 1, \ldots, T$. The relaxation* (2.30) *is admissible, and the algorithm that generates the sequence $\{\nu_t\}$ is also admissible:*

$$\widehat{V}_T(\mathcal{G}'|f_1, \ldots, f_{t-1}) \geqslant \sup_{f \in \mathcal{F}} \left\{ \langle \nu_t, f \rangle + \widehat{V}_T(\mathcal{G}'|f_1, \ldots, f_{t-1}, f) \right\}$$

Proof. First, we check the admissibility condition at time $t = T$. Since the Bregman divergence is nonnegative, we have

$$\begin{aligned} \widehat{V}_T(\mathcal{G}'|f_1, \ldots, f_T) &= \sup_{\mu \in \mathcal{G}'} \left\{ \sum_{s=1}^{T} \langle \mu, -f_s \rangle + \frac{1}{\beta} D_F(\mu, \nu_{T+1}) \right\} \\ &\geqslant - \inf_{\mu \in \mathcal{G}'} \left\langle \mu, \sum_{s=1}^{T} f_s \right\rangle. \end{aligned}$$

Now let us consider an arbitrary t. From the construction of our relaxation, we have

$$\begin{aligned} &\sup_{f_t \in \mathcal{F}} \left\{ \langle \nu_t, f_t \rangle + \widehat{V}_T(\mathcal{G}'|f_1, \ldots, f_t) \right\} \\ &= \sup_{f_t \in \mathcal{F}} \sup_{\mu \in \mathcal{G}'} \left\{ \sum_{s=1}^{t-1} \langle \mu, -f_s \rangle + \langle \nu_t - \mu, f_t \rangle + \frac{1}{\beta} D_F(\mu, \nu_{t+1}) + 2\beta(T-t) \right\} \end{aligned}$$

for all $t = 1, \ldots, T$. From the definition (2.29) of the Bregman divergence, the following equality holds for any three $\mu, \nu, \lambda \in \mathcal{G}$:

$$D_F(\mu, \nu) + D_F(\nu, \lambda) = D_F(\mu, \lambda) + \langle \nabla F(\lambda) - \nabla F(\nu), \mu - \nu \rangle. \tag{2.31}$$

Since ∇F and ∇F^* are inverses of each other, we have $-\beta f_t = \nabla F(\nu_{t+1}) - \nabla F(\nu_t)$. Using this fact together with (2.31), for any $\mu \in \mathcal{G}'$ we can write

$$\begin{aligned} \langle \nu_t - \mu, \beta f_t \rangle &= \langle \nabla F(\nu_{t+1}) - \nabla F(\nu_t), \mu - \nu_t \rangle \\ &= D_F(\mu, \nu_t) - D_F(\mu, \nu_{t+1}) + D_F(\nu_t, \nu_{t+1}). \end{aligned} \tag{2.32}$$

Moreover, once again using the fact that ∇F and ∇F^* are inverses of one another, we have

$$\begin{aligned} D_F(\nu_t, \nu_{t+1}) &= D_{F^*}(\nabla F(\nu_{t+1}), \nabla F(\nu_t)) \\ &= F^*(\nabla F(\nu_{t+1})) - F^*(\nabla F(\nu_t)) - \langle \nabla F^*(\nabla F(\nu_t)), \nabla F(\nu_{t+1}) - \nabla F(\nu_t) \rangle \\ &\leqslant \Lambda\left(\beta \|f_t\|^*_{\nabla F(\nu_t)}\right), \end{aligned} \tag{2.33}$$

where

$$\Lambda(r) \triangleq -\log(1-r) - r = \frac{r^2}{2} + \frac{r^3}{3} + \frac{r^4}{4} + \dots.$$

Note that, because of the definition of Λ, the learning rate β must be chosen in such a way that $\beta \|f_t\|^*_{\nabla F(\mu_t)} < 1$ for all $t = 1, \dots, T$. By hypothesis, we have $\beta \|f_t\|^*_{\nabla F(\nu_t)} \leqslant 1/2$ for all t. The first line of Eq. (2.33) is by Prop. 11.1 in Cesa-Bianchi and Lugosi (2006), the second is by definition of the Bregman divergence, and the third follows from from a local second-order Taylor formula for a self-concordant function (Nemirovski and Todd, 2008, Eq. (2.5)) (which is applicable because, by hypothesis, $\beta \|f_t\|^*_{\nabla F(\nu_t)} \leqslant 1/2 < 1$ for all t) and the fact that the Legendre–Fenchel dual of a self-concodrant function is also self-concordant.

Using the inequality $\log r \leqslant r - 1$, we can upper-bound

$$\Lambda(r) = -\log(1-r) - r \leqslant \frac{1}{1-r} - 1 - r = \frac{1 - (1-r)(1+r)}{1-r} = \frac{r^2}{1-r}.$$

Moreover, since $\beta \|f_t\|^*_{\nabla F(\nu_t)} \leqslant 1/2$ and $\|f_t\|^*_{\nabla F(\nu_t)} \leqslant 1$ for all $t \in \{1, \dots, T\}$ by hypothesis, we can further bound

$$\Lambda\left(\beta \|f_t\|^*_{\nabla F(\nu_t)}\right) \leqslant 2\beta^2 \|f_t\|^{*2}_{\nabla F(\nu_t)} \leqslant 2\beta^2.$$

Applying Eqs. (2.32) and (2.33), we arrive at

$$\begin{aligned}
&\sup_{f_t\in\mathcal{F}}\left\{\langle\nu_t,f_t\rangle+\widehat{V}_T(\mathcal{G}'|f_1,\ldots,f_t)\right\}\\
&=\sup_{f_t\in\mathcal{F}}\sup_{\mu\in\mathcal{G}'}\left\{\sum_{s=1}^{t-1}\langle\mu,-f_s\rangle+\langle\nu_t-\mu,f_t\rangle+\frac{1}{\beta}D_F(\mu,\nu_{t+1})+2\beta(T-t)\right\}\\
&\leqslant\sup_{f_t\in\mathcal{F}}\sup_{\mu\in\mathcal{G}'}\left\{\sum_{s=1}^{t-1}\langle\mu,-f_s\rangle+\frac{1}{\beta}D_F(\mu,\nu_t)+\frac{1}{\beta}\Lambda\left(\beta\|f_t\|^*_{\nabla F(\nu_t)}\right)+2\beta(T-t)\right\}\\
&\leqslant\sup_{f_t\in\mathcal{F}}\sup_{\mu\in\mathcal{G}'}\left\{\sum_{s=1}^{t-1}\langle\mu,-f_s\rangle+\frac{1}{\beta}D_F(\mu,\nu_t)+2\beta(T-t+1)\right\}\\
&=\widehat{V}_T(\mathcal{G}'|f_1,\ldots,f_{t-1}).
\end{aligned}$$

This shows that the proposed algorithm (behavoiral strategy) is admissible, and the proof is complete. □

Here we impose the assumption that $\beta\|f_t\|^*_{\nabla F(\nu_t)}\leqslant 1/2$ for all $t=1,\ldots,T$. A restriction of this kind is necessary when using interior-point methods to construct online optimization schemes — see, for example, the condition of Theorem 4.1 and 4.2 in Abernethy et al. (2012). The boundedness of the dual local norm $\|f_t\|^*_{\nabla F(\nu_t)}$ is a reasonable assumption as well. In particular, Abernethy et al. (2012) point out that if a large number of the points ν_t are close to the boundary of $\mathcal{G}'$, then the eigenvalues of the Hessian of F at those points will be large due to the large curvature of the barrier near the boundary of $\mathcal{G}'$. This will imply, in turn, that the dual local norm $\|f_t\|^*_{\nabla F(\nu_t)}$ is expected to be small.

Now we are ready to present the online-learning (i.e., steady-state) part of the regret bound for the above algorithm:

Theorem 12. *Let $D_F(\mathcal{G}')\triangleq\sup_{\nu\in\mathcal{G}'}D_F(\nu,\nu_1)$. Suppose that the learning rate β is such that $\beta\|f_t\|^*_{\nabla F(\nu_t)}\leqslant 1/2$ for all $t=1,\ldots,T$. Assume that $\|f_t\|^*_{\nabla F(\nu_t)}\leqslant 1$, for all*

$t = 1, \ldots, T$. *Then for the relaxation* (2.30) *and the corresponding algorithm, we can bound the online learning part of the regret as*

$$\sum_{t=1}^{T} \langle \nu_t, f_t \rangle - \inf_{\nu \in \mathcal{G}'} \sum_{t=1}^{T} \langle \nu, f_t \rangle \leqslant \frac{D_F(\mathcal{G}')}{\beta} + 2\beta T. \tag{2.34}$$

Proof. Since the relaxation (2.30) is admissible by Proposition 11, we have

$$\begin{aligned} \sum_{t=1}^{T} \langle \nu_t, f_t \rangle - \inf_{\nu \in \mathcal{G}'} \sum_{t=1}^{T} \langle \nu, f_t \rangle &\leqslant \widehat{V}_T(\mathcal{G}' | \mathsf{e}) \\ &= \sup_{\mu \in \mathcal{G}'} \left\{ \frac{D_F(\mu, \nu_1)}{\beta} + 2\beta T \right\} \\ &= \frac{D_F(\mathcal{G}')}{\beta} + 2\beta T. \end{aligned}$$

□

Remark 5. Since F is a barrier, $D_F(\mathcal{G}')$ will be finite only if all the elements of $\mathcal{G}'$ are not too close to the boundary of $\mathcal{G}$. This motivates our restriction of the comparator term to a proper subset $\mathcal{G}' \subset \operatorname{int}(\mathcal{G})$.

Finally, we present the total regret bound for the above algorithm, including the stationarization error:

Theorem 13. *Suppose that all of our earlier assumptions are in place, and also that the uniform mixing condition is satisfied. Then for the relaxation* (2.30) *and the corresponding algorithm, we have*

$$\begin{aligned} &\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \inf_{P \in \mathcal{M}(\mathcal{G}')} \mathbb{E} \left[\sum_{t=1}^{T} f_t(X, U) \right] \right] \\ &\qquad \leqslant \frac{D_F(\mathcal{G}')}{\beta} + 2\beta T + C_{\mathcal{F}} (\tau + 1)^2 T \Delta_T + (2\tau + 2) C_{\mathcal{F}}, \end{aligned} \tag{2.35}$$

where $\Delta_T \triangleq \max_{1 \leqslant t \leqslant T} \max_{x \in \mathsf{X}} \| P_{t-1}(\cdot | x) - P_t(\cdot | x) \|_1$.

Proof. Let us denote by $P_t = P_{\nu_t}$ the policy extracted from ν_t, and the induced marginal distribution of X_t by $\mu_t \in \mathcal{P}(\mathsf{X})$. We also denote by K_t the Markov matrix that describes the state transition from X_t to X_{t+1}, and by $\pi_t \in \mathcal{P}(\mathsf{X})$ its the unique invariant distribution. Finally, we denote by $\widehat{\boldsymbol{\gamma}}$ the behavioral strategy corresponding to our algorithm. Then, for any $\boldsymbol{f} \in \mathcal{F}^T$, we can upper-bound the regret by

$$\begin{aligned}
R_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(\mathcal{G}') &= \sum_{t=1}^{T} \langle \mu_t \otimes P_t, f_t \rangle - \inf_{\nu \in \mathcal{G}'} \sum_{t=1}^{T} \langle \nu, f_t \rangle \\
&\leqslant \sum_{t=1}^{T} \left[\langle \pi_t \otimes P_t, f_t \rangle - \inf_{P \in \mathcal{M}(\mathcal{G}')} \langle \pi_P \otimes P, f_t \rangle \right] + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t - \pi_t\|_1 \\
&= \sum_{t=1}^{T} \langle \nu_t, f_t \rangle - \inf_{\nu \in \mathcal{G}'} \sum_{t=1}^{T} \langle \nu, f_t \rangle + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t - \pi_t\|_1 \\
&\leqslant \frac{D_F(\mathcal{G}')}{\beta} + 2\beta T + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t - \pi_t\|_1. \qquad (2.36)
\end{aligned}$$

Now we focus on bounding the third term of the regret bound. For any time $k \leqslant t$, we have

$$\begin{aligned}
\|\mu_k - \pi_t\|_1 &= \|\mu_{k-1}K_{k-1} - \mu_{k-1}K_t + \mu_{k-1}K_t - \pi_t\|_1 \\
&\overset{(a)}{\leqslant} \|\mu_{k-1}K_t - \pi_t\|_1 + \|\mu_{k-1}K_{k-1} - \mu_{k-1}K_t\|_1 \\
&\overset{(b)}{=} \|\mu_{k-1}K_t - \pi_t K_t\|_1 + \|\mu_{k-1}K_{k-1} - \mu_{k-1}K_t\|_1 \\
&\overset{(c)}{\leqslant} e^{-1/\tau} \|\mu_{k-1} - \pi_t\|_1 + \max_{x \in \mathsf{X}} \|P_{k-1}(\cdot|x) - P_t(\cdot|x)\|_1 , \\
&\overset{(d)}{\leqslant} e^{-1/\tau} \|\mu_{k-1} - \pi_t\|_1 + \sum_{j=k-1}^{t-1} \max_{x \in \mathsf{X}} \|P_j(\cdot|x) - P_{j+1}(\cdot|x)\|_1, \qquad (2.37)
\end{aligned}$$

where (a) is by triangle inequality; (b) is by invariance of π_t w.r.t. $K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}$; and (c) follows from the uniform mixing bound (2.6), and (d) follows from the triangle inequality and the easily proved fact that, for any state distribution $\mu \in \mathcal{P}(\mathsf{X})$ and any

two Markov kernels $P, P' \in \mathcal{M}(\mathsf{U}|\mathsf{X})$,

$$\|\mu K(\cdot|P) - \mu K(\cdot|P')\|_1 \leqslant \max_{x \in \mathsf{X}} \|P(\cdot|x) - P'(\cdot|x)\|_1 .$$

Letting now the initial state distribution be μ_1, we can apply the bound (2.37) recursively to obtain

$$\begin{aligned}
\|\mu_t - \pi_t\|_1 &\leqslant e^{-(t-1)/\tau} \|\mu_1 - \pi_t\|_1 + \sum_{k=2}^{t} e^{-\frac{t-k}{\tau}} \sum_{j=k-1}^{t-1} \max_{x \in \mathsf{X}} \|P_j(\cdot|x) - P_{j+1}(\cdot|x)\|_1 \\
&\leqslant 2e^{-(t-1)/\tau} + \sum_{k=2}^{t} e^{-\frac{t-k}{\tau}} (t-k+1)\Delta_T \\
&\leqslant 2e^{-(t-1)/\tau} + B \sum_{k=0}^{\infty} (k+1) e^{-\frac{k}{\tau}} \\
&\leqslant 2e^{-(t-1)/\tau} + (\tau+1)^2 \Delta_T.
\end{aligned}$$

So, the second term on the right-hand side of (2.36) can bounded by

$$C_{\mathcal{F}} \sum_{t=1}^{T} \|\mu_t^{\widehat{\gamma}, \boldsymbol{f}} - \pi_t^{\widehat{\gamma}, \boldsymbol{f}}\|_1 \leqslant C_{\mathcal{F}} (\tau+1)^2 T \Delta_T + (2\tau+2) C_{\mathcal{F}},$$

which completes the proof. □

The third term on the right-hand side of (2.35) quantifies the drift of the policies generated by the algorithm. A similar term appears in all of the regret bounds of Dick et al. (see, e.g., the bound of Lemma 1 in Dick et al. (2014)). Moreover, just like the Mirror Descent scheme of Dick et al. (2014), our algorithm may run into implementation issues, since in general it may be difficult to compute the gradient mappings ∇F and ∇F^* associated to the self-concordant barrier F. We refer the reader to the discussion in the paper by Bubeck and Eldan (2015) pertaining to computational feasibility of their universal entropic barrier.

2.5 Conclusions

We have provided a unified viewpoint on the design and the analysis of online MDPs algorithms, which is an extension of a general relaxation-based approach of Rakhlin et al. (2010) to a certain class of stochastic game models. We have unified two distinct categories of existing methods (those based on the relative-function approach and those based on the convex-analytic approach) under a general framework. We have shown that an algorithm previously proposed by Even-Dar et al. (2009) naturally arises from our framework via a specific relaxation. Moreover, we have shown that one can obtain lazy strategies (where time is split into phases, and a different stationary policy is followed in each phase) by means of relaxations as well. In particular, we have obtained a new strategy, which is similar in spirit to the one previously proposed by Yu et al. (2009), but with several advantages, including better scaling of the regret. The above two algorithms are based on the relative-function approach via reverse Poisson inequalities. Finally, using a different type of a relaxation, we have derived another algorithm for online MDPs, which relies on interior-point methods and belongs to the class of algorithms derived using the convex-analytic approach. The takeaway point is that our general technique of constructing relaxations after a stationarization step brings all of the existing methods under the same umbrella and paves the way toward constructing new algorithms for online MDPs.

3

Online MDPs with predictable sequences

3.1 Introduction

In the preceding chapters, we have presented no-regret algorithms with performance guarantees for all possible cost sequences, including the worst-case scenario. Therefore, the regret bounds are often conservative. More optimistic results are desirable when the cost sequence may have some level of "regularity". This chapter addresses this concern by introducing a different model of non-stationary environments, where algorithms with more favorable performance guarantees can be designed. The goal is to develop algorithms that enjoy tighter regret bounds for "more predictable" cost sequences while still providing the same guarantees as before in the worst case. For example, in managing a company's inventories, we often have some prior knowledge of the future demands due to past experiences and records, and we would like to form this prior knowledge into a predictive model and incorporate the model into the decision loop.

We present methods for online MDP problems that take advantage of the so-called "predictable sequences". This notion is proposed by Rakhlin and Sridharan

(2013a,b), where they present methods for online linear optimization that take advantage of benign (as opposed to worst-case) sequences. Our approach can be seen as an extension of their work to the setting with state variable. We assume the agent has a predictive model that returns an estimate of the current cost function given all the previously revealed cost functions. We incorporate this estimate along with all the previously revealed cost functions into choosing the current action. The predictive model can be seen as our guess of how the actual cost sequence is generated. We present an algorithm that incurs smaller regret when our guesses become more accurate over time.

The remainder of the chapter is organized as follows. Section 3.2 contains the precise formulation of the online MDP problem with predictable sequences. Section 3.3 describes our proposed algorithm and contains the main result. We close this chapter in Section 3.4.

3.2 Problem setup

We consider an online MDP with finite state and action spaces X and U and a fixed and known transition kernel K. Let $\mathcal{F}$ be a fixed and known class of functions $f : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$, and let $x \in \mathsf{X}$ be a fixed initial state.

Here we follow the prediction model in Rakhlin and Sridharan (2013b) to define predictable sequences. For each $t \in \{1, \dots, T\}$, fix a sequence of functions $M_t : \mathcal{F}^{t-1} \to \mathcal{F}$, and let $\widehat{f}_t = M_t(f_1, \dots, f_{t-1})$ be our estimate (or prediction) of f_t based on the revealed realization of $f_1, \dots, f_{t-1}$. We call $\{M_t\}_{t=0}^{T}$ the prediction model of the cost functions, and it is supposed to be available to the agent as prior knowledge. Ideally, the prediction model should capture the predictable part of the evolution of the cost sequence, with the underlying assumption that f_t can be approximated as a function of the previous revealed cost functions $\{f_1, \dots, f_{t-1}\}$ plus some unpredictable

noise. If $f_t = M_t(f_1, \ldots, f_{t-1})$, it means we have a perfect prediction model for the cost sequence. Our goal is to incur smaller regret if the estimates are close to the true cost functions.

Consider an agent (A) performing a controlled random walk on X in response to signals chosen by the environment (E). The agent is using mixed strategies to choose actions, where a mixed strategy is a probability distribution over the action space. The interaction between the agent (A) and the environment (E) proceeds as follows:

$X_1 = x$
for $t = 1, 2, \ldots, T$
  A knows f_1^{t-1} and prediction model $\widehat{f}_t = M_t(f_1, \ldots, f_{t-1})$;
A observes state X_t, selects $P_t \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ based on X_1^t, f_1^{t-1} and M_t
  E selects $f_t \in \mathcal{F}$ and announces it to A
  A draws $U_t \sim P_t(\cdot|X_t)$ and incurs cost $f_t(X_t, U_t)$
  The system transitions to $X_{t+1} \sim K(\cdot|X_t, U_t)$
end for

This setup is very similar to what we had in Section 2.2. The difference lies in the way the agent uses his information to form his decision. At each $t \geqslant 1$, the agent observes the current state x_t and selects the mixed strategy P_t, where $P_t(u|x_t) = \Pr\{U_t = u|X_t = x_t\}$, using his knowledge of all the previous states and current state x^t, the previous moves of E, $f^{t-1} = (f_1, \ldots, f_{t-1})$, and the estimate of f_t, i.e., $\widehat{f}_t = M_t(f_1, \ldots, f_{t-1})$. As before, we define A's closed-loop *behavioral strategy* as a tuple $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$, where $\gamma_t : \mathsf{X}^t \times \mathcal{F}^{t-1} \to \mathcal{P}(\mathsf{U})$. Similarly, E's *open-loop* behavioral strategy is a tuple $\boldsymbol{f} = (f_1, \ldots, f_t)$. We impose the same assumptions as in Section 2.2: We assume throughout that E is *oblivious* (or *open-loop*) in the sense that we assume that the evolution of the sequence $\{f_t\}$ is not affected by the state and action sequences $\{X_t\}$ and $\{U_t\}$. We also assume the "uniform mixing condition" (see Section 2.3.1 for details). After drawing the action U_t from P_t, A incurs the one-step cost $f_t(X_t, U_t)$. Once the initial state $X_1 = x$ and the strategy pair $(\boldsymbol{\gamma}, \boldsymbol{f})$ are specified, the joint distribution of the state-action process (X^T, U^T)

is well-defined.

The goal of the agent is to minimize the expected *steady-state regret*

$$R_x^{\boldsymbol{\gamma},\boldsymbol{f}} \triangleq \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left\{ \sum_{t=1}^{T} f_t(X_t, U_t) - \inf_{P \in \mathcal{M}_0} \mathbb{E}\left[\sum_{t=1}^{T} f_t(X, U) \right] \right\}, \tag{3.1}$$

where the outer expectation $\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}}$ is taken w.r.t. the distribution induced by A's behavioral strategy $\boldsymbol{\gamma}$, E's behavioral strategy $\boldsymbol{f}$, and the initial state $X_1 = x$. As in Section 2.2, we consider the steady-state regret, so that the expectation in the comparator term $\mathbb{E}\left[\sum_{t=1}^{T} f_t(X, U)\right]$ is taken over $\pi_P \otimes P(x, u) = \pi_P(x)P(u|x)$, where π_P is the unique invariant distribution of $K(\cdot|\cdot, P)$. The regret $R_x^{\boldsymbol{\gamma},\boldsymbol{f}}$ can be interpreted as the gap between the expected cumulative cost of A using strategy $\boldsymbol{\gamma}$ and the best steady-state cost A could have achieved in hindsight by using the best stationary policy $P \in \mathcal{M}_0$ (with full knowledge of $\boldsymbol{f} = f^T$).

3.3 The proposed strategy

In this section, we will present an algorithm that takes advantage of the prediction model in the sense that the regret bound scales with the error of the prediction model.

Here we need to reapply the stationarization idea to reduce the online MDP problem to an online linear optimization problem, decoupling the current state and the past actions. The reader can see Section 2.3.3 for details. Here we briefly review the reduction to fix ideas. Recall that we define the regret w.r.t. a fixed $P \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ with initial state $X_1 = x$:

$$\begin{aligned} R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P) &\triangleq \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \sum_{t=1}^{T} \eta_t^{P,\boldsymbol{f}} \right] \\ &= \sum_{t=1}^{T} \left[\langle \mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle - \langle \pi_P \otimes P, f_t \rangle \right], \end{aligned}$$

and decompose the regret $R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P)$ into two parts: the stationarized regret $\bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P)$ and the stationarization error $\sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}}\|_1$, that is

$$\begin{aligned} R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P) &\leqslant \bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P) + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}}\|_1 \\ &= \sum_{t=1}^{T} \left[\langle \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle - \langle \pi_P \otimes P, f_t \rangle \right] + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}}\|_1. \end{aligned} \tag{3.2}$$

Now we end up with an online linear optimization problem where at each time step t, the agent is choosing an occupation measure $\nu_t = \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}$ from the set $\mathcal{G}$ (here we omit the dependence of ν_t on $\boldsymbol{\gamma}, \boldsymbol{f}$ for simplicity), and the environment is choosing the one-step cost function f_t. The agent then incurs the cost $\langle \nu_t, f_t \rangle$. We will derive an algorithm for the agent to choose occupation measures from $\mathcal{G}$, so that the total cost incurred by the agent is not much worse than the total cost incurred by the best fixed occupation measure in $\mathcal{G}$. Since we can recover a policy from an occupation measure, we just need to find a slowly changing sequence of occupation measures such that the online learning part of the regret is small, and the slowly changing aspect will ensure the stationarization error is also small. We then utilize the prediction model to bound the stationarized regret.

3.3.1 Algorithm description

The algorithm we use can be seen as a modified version of the FRL algorithm. In particular, our algorithm selects the occupancy measure $\nu_t \in \mathcal{G}$ that solves the optimization problem

$$\mu^{P_t} = \nu_t = \arg\min_{\nu \in \mathcal{G}} \left\{ \left\langle \nu, \frac{1}{\beta} \sum_{s=1}^{t-1} f_s + \frac{1}{\beta} \widehat{f}_t \right\rangle + \Phi(\nu) \right\}, \tag{3.3}$$

where $\Phi(\nu)$ is the relative entropy regularization term $D(\nu \| \nu_1)$, where $\nu_1 \in \mathcal{P}(\mathsf{X} \times \mathsf{U})$ is a given reference measure that assigns positive mass to every state-action pair

(x, u). Note that ν_1 does not have to be a member of $\mathcal{G}$. Here we choose it to be the uniform measure over all state-action pairs. The entropic regularization term penalizes "spiky" occupation measures, thus guaranteeing that every action has a nonzero probability of being selected at every time step, which in turn provides the algorithm with a degree of stability. This algorithm is very similar to the Optimistic Follow the Regularized Leader (OFRL) algorithm in Rakhlin and Sridharan (2013b). Here we incorporate the prediction model of the cost function into the decision loop. If the prediction turns out to be accurate (which means that $\widehat{f_t}$ is close to f_t), the method should yield tighter regret bounds. This method is "optimistic", because it incorporates M_t into the calculation of the next decision as if it were true.

3.3.2 Main result

Before proving the regret bound of the above algorithm, we present the following lemma:

Lemma 14. *Let* $\nu_g^* = \arg\min_{\nu \in \mathcal{G}} \{\langle g, \nu \rangle + \beta\Phi(\nu)\}$ *and* $\nu_{g'}^* = \arg\min_{\nu \in \mathcal{G}} \{\langle g', \nu \rangle + \beta\Phi(\nu)\}$. *Then we have*

$$\beta\|\nu_g^* - \nu_{g'}^*\|_1 \leqslant \|g - g'\|_\infty.$$

Proof. For each function $g \in \mathcal{F}$, define $F_g : \mathcal{P}(\mathsf{X} \times \mathsf{U}) \to \mathbb{R}$ by

$$F_g(\nu) = \langle g, \nu \rangle + \beta\Phi(\nu). \tag{3.4}$$

We say the function F_g is β-strongly convex with respect to ℓ_1 norm if for all $\nu, \nu' \in \mathcal{P}(\mathsf{X} \times \mathsf{U})$,

$$F_g(\nu) \geqslant F_g(\nu') + \langle \nabla F_g(\nu'), \nu - \nu' \rangle + \frac{\beta}{2}\|\nu - \nu'\|_1^2. \tag{3.5}$$

It is well-known that $\Phi(\nu)$ is 1-strongly convex with respect to the ℓ_1 norm over $\mathcal{P}(\mathsf{X} \times \mathsf{U})$ (see the proof of Example 2.5 in Shalev-Shwartz (2011). By Lemma 2.9 in Shalev-Shwartz (2011), we know that if Φ is 1-strongly convex with respect to some

norm, then $\beta\Phi$ is β-strongly convex with respect to the same norm. It is easy to see that the linear term $\langle g, \nu \rangle$ does not affect the strong convexity. Therefore, we know the function $F_g(\nu)$ is β-strongly convex over $\mathcal{P}(\mathsf{X} \times \mathsf{U})$. The set $\mathcal{G}$ of occupation measures is a convex subset of $\mathcal{P}(\mathsf{X} \times \mathsf{U})$. Thus, $F_g(\nu)$ is β-strongly convex on $\mathcal{G}$ as well.

Let $\nu_g^* = \arg\min_{\nu \in \mathcal{G}} F_g(\nu)$ and $\nu_{g'}^* = \arg\min_{\nu \in \mathcal{G}} F_{g'}(\nu)$. By the strong convexity of F_g, for any $\nu \in \mathcal{G}$, we have

$$F_g(\nu) - F_g(\nu_g^*) \geqslant \frac{\beta}{2} \|\nu - \nu_g^*\|_1^2. \tag{3.6}$$

and

$$F_{g'}(\nu) - F_{g'}(\nu_{g'}^*) \geqslant \frac{\beta}{2} \|\nu - \nu_{g'}^*\|_1^2. \tag{3.7}$$

These holds due to (3.5) and the fact that the gradient terms $\nabla F_g(\nu_g^*)$ and $\nabla F_{g'}(\nu_{g'}^*)$ are zero. Replacing ν in (3.6) with $\nu_{g'}^*$, and ν in (3.7) with ν_g^*, we have

$$F_g(\nu_{g'}^*) - F_g(\nu_g^*) \geqslant \frac{\beta}{2} \|\nu_{g'}^* - \nu_g^*\|_1^2.$$

and

$$F_{g'}(\nu_g^*) - F_{g'}(\nu_{g'}^*) \geqslant \frac{\beta}{2} \|\nu_g^* - \nu_{g'}^*\|_1^2.$$

Adding the above two inequalities together, we have

$$(F_g - F_{g'})(\nu_{g'}^*) + (F_{g'} - F_g)(\nu_g^*) \geqslant \beta \|\nu_g^* - \nu_{g'}^*\|_1^2.$$

From the definition of F_g, we have

$$\langle g - g', \nu_{g'}^* \rangle + \langle g' - g, \nu_g^* \rangle \geqslant \beta \|\nu_g^* - \nu_{g'}^*\|_1^2.$$

Rearranging, we obtain

$$\langle g - g', \nu_{g'}^* - \nu_g^* \rangle \geqslant \beta \|\nu_g^* - \nu_{g'}^*\|_1^2$$

Then by Hölder's inequality,

$$\|g - g'\|_\infty \|\nu_g^* - \nu_{g'}^*\|_1 \geqslant \beta \|\nu_g^* - \nu_{g'}^*\|_1^2,$$

and if the term $\|\nu_g^* - \nu_{g'}^*\|_1$ is nonzero, i.e. if ν_g^* and $\nu_{g'}^*$ are different, by cancelling out $\|\nu_g^* - \nu_{g'}^*\|_1$ on both sides, we end up with

$$\|g - g'\|_\infty \geqslant \beta \|\nu_g^* - \nu_{g'}^*\|_1.$$

If $\|\nu_g^* - \nu_{g'}^*\|_1 = 0$, then the desired bound holds a fortiori. □

Theorem 15. *The proposed algorithm* (3.3) *enjoys the following regret bound:*

$$R_x^{\gamma,\boldsymbol{f}} \leqslant \sum_{t=1}^T \frac{1}{\beta} \|f_t - \widehat{f}_t\|_\infty^2 + \beta \log |\mathsf{X} \times \mathsf{U}| + C_{\mathcal{F}} \sum_{t=1}^T \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1.$$

Proof. First, let $\nu^* = \arg\min_{\nu \in \mathcal{G}} \langle \nu, \sum_{t=1}^T f_t \rangle$ and $\nu_t = \pi_t^{\gamma,\boldsymbol{f}} \otimes P_t^{\gamma,\boldsymbol{f}}$. From (3.2), we get

$$\begin{aligned} R_x^{\gamma,\boldsymbol{f}} &= \mathbb{E}_x^{\gamma,\boldsymbol{f}} \left\{ \sum_{t=1}^T f_t(X_t, U_t) - \inf_{P \in \mathcal{M}_0} \mathbb{E} \left[\sum_{t=1}^T f_t(X, U) \right] \right\} \\ &\leqslant \sum_{t=1}^T \langle \nu_t - \nu^*, f_t \rangle + C_{\mathcal{F}} \sum_{t=1}^T \|\mu_t^{\gamma,\boldsymbol{f}} - \pi_t^{\gamma,\boldsymbol{f}}\|_1. \end{aligned}$$

where $C_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \|f\|_\infty$.

Define $\mu_{t+1} = \arg\min_{\nu \in \mathcal{G}} \left\{ \frac{1}{\beta} \langle \nu, \sum_{s=1}^t f_s \rangle + \Phi(\nu) \right\}$ to be the unmodified Follow the Regularized Leader. We have

$$\sum_{t=1}^T \langle \nu_t - \nu^*, f_t \rangle = \sum_{t=1}^T \langle \nu_t - \mu_{t+1}, f_t - \widehat{f}_t \rangle + \sum_{t=1}^T \langle \nu_t - \mu_{t+1}, \widehat{f}_t \rangle + \sum_{t=1}^T \langle \mu_{t+1} - \nu^*, f_t \rangle$$

Using induction, we will now show that

$$\sum_{t=1}^\tau \langle \nu_t - \mu_{t+1}, \widehat{f}_t \rangle + \sum_{t=1}^\tau \langle \mu_{t+1}, f_t \rangle \leqslant \sum_{t=1}^\tau \langle \nu^*, f_t \rangle + \beta \Phi(\nu^*) \tag{3.8}$$

The base case $\tau = 1$ is immediate because we can simply assume $\widehat{f}_1 = 0$. Suppose that the above inequality holds for $\tau = T - 1$. Using $\nu^* = \nu_T$ and adding $\langle \nu_T - \mu_{T+1}, \widehat{f}_T \rangle + \langle \mu_{T+1}, f_T \rangle$ to both sides, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \nu_t - \mu_{t+1}, \widehat{f}_t \rangle + \sum_{t=1}^{T} \langle \mu_{t+1}, f_t \rangle \\
&\leqslant \sum_{t=1}^{T-1} \langle \nu_T, f_t \rangle + \beta\Phi(\nu_T) + \langle \nu_T - \mu_{T+1}, f_T \rangle + \langle \mu_{T+1}, f_T \rangle \\
&\leqslant \left\langle \nu_T, \sum_{t=1}^{T-1} f_t + \widehat{f}_T \right\rangle + \beta\Phi(\nu_T) - \langle \mu_{T+1}, \widehat{f}_T \rangle + \langle \mu_{T+1}, f_T \rangle \\
&\leqslant \left\langle \mu_{T+1}, \sum_{t=1}^{T-1} f_t + \widehat{f}_T \right\rangle + \beta\Phi(\mu_{T+1}) - \langle \mu_{T+1}, \widehat{f}_T \rangle + \langle \mu_{T+1}, f_T \rangle \\
&\leqslant \left\langle \nu^*, \sum_{t=1}^{T} f_T \right\rangle + \beta\Phi(\nu^*)
\end{aligned}$$

by the optimality of ν_T and μ_{T+1}. This concludes the inductive argument, and we have

$$\sum_{t=1}^{T} \langle \nu_t - \nu^*, f_t \rangle \leqslant \sum_{t=1}^{T} \langle \nu_t - \mu_{t+1}, f_t - \widehat{f}_t \rangle + \beta\Phi(\nu^*) \tag{3.9}$$

Armed with Lemma 14, it is easy to see that

$$\begin{aligned}
\sum_{t=1}^{T} \langle \nu_t - \nu^*, f_t \rangle &\leqslant \sum_{t=1}^{T} \langle \nu_t - \mu_{t+1}, f_t - \widehat{f}_t \rangle + \beta\Phi(\nu^*) \\
&\leqslant \sum_{t=1}^{T} \|\nu_t - \mu_{t+1}\|_1 \|f_t - \widehat{f}_t\|_\infty + \beta\Phi(\nu^*) \\
&\leqslant \sum_{t=1}^{T} \frac{1}{\beta} \|f_t - \widehat{f}_t\|_\infty^2 + \beta\Phi(\nu^*) \\
&\leqslant \sum_{t=1}^{T} \frac{1}{\beta} \|f_t - \widehat{f}_t\|_\infty^2 + \beta \log |\mathsf{X} \times \mathsf{U}|.
\end{aligned}$$

□

Note that if we have prior knowledge about the quantity $\sum_{t=1}^{T} \|f_t - \widehat{f}_t\|_\infty$, we can optimize the online learning part of the regret by choosing $\beta = \sqrt{\sum_{t=1}^{T} \frac{\|f_t - \widehat{f}_t\|_\infty^2}{\log |\mathsf{X} \times \mathsf{U}|}}$, and then we have:

$$R_x^{\gamma, \boldsymbol{f}} \leqslant 2\sqrt{\log |\mathsf{X} \times \mathsf{U}| \sum_{t=1}^{T} \|f_t - \widehat{f}_t\|_\infty^2} + C_{\mathcal{F}} \sum_{t=1}^{T} \|\mu_t^{\gamma, \boldsymbol{f}} - \pi_t^{\gamma, \boldsymbol{f}}\|_1.$$

The second term is the stationarization error term, and it can be treated the same way as before. The key point here is that the regret will become smaller as the prediction model becomes more accurate. Contrary to the online learning setting, it is not possible to achieve zero regret even if we have a perfect prediction model. It is inevitable to have at least some stationarizaton error no matter how accurate the prediction model is. This result also shows how the presence of the state in the online MDP problem fundamentally complicates both the theoretical analysis and the development of no-regret methods. In general, the stationarization error is difficult to analyze, and it is closely related to the quantitative study of merging and stability of time inhomogeneous Markov chains on a finite state space. The readers can see Saloff-Coste and Zuniga (2009) for details.

3.4 Conclusion

We have presented an algorithm that seeks to improve performance by exploiting observed deviations of the arbitrary player from a worst-case strategy. We have extended the work of Rakhlin and Sridharan (2013a,b) to the online MDP setting and shown that the regret of our proposed strategy decreases when the prediction model becomes more accurate.

4

Information projection onto the state-action polytope

The convex-analytic approach in Chapter 2 and the algorithm developed in Chapter 3 all require a projection onto the state-action polytope $\mathcal{G}$, by which we reduce the online MDP problem to an online linear optimization problem. In this chapter, we study this projection in the traditional MDP setting by relating it to a certain class of optimization problems. More specifically, the average-cost optimal control problem for an MDP can be formulated as a linear program (LP) over the convex polytope of *ergodic occupation measures* that characterize steady-state frequencies of states and actions along the trajectory of the controlled chain, which in fact is the state-action polytope. We extend this convex-analytic method by adding a relative entropy regularization to the objective function, which allows us to minimize the long-term average cost while staying close to some reference measure. We further relate the resulting optimization problem to the theory of exponential families of probability distributions and exploit this relation to derive structural results for the optimizing measure, its associated policy, and the corresponding ergodic cost. These properties

are then used to establish sensitivity and stability results, which can provide novel insights into online MDP algorithms beyond establishing sublinear regret bounds.

4.1 Introduction

Consider the standard set-up for an MDP with finite state and action space X and U (Puterman (1994); Arapostathis et al. (1993)). The *average cost* of a stationary Markov policy $w : \mathsf{X} \to \mathsf{U}$ with the initial condition $X_1 = x_1$ is given by

$$J(w, x_1) \triangleq \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^w_{x_1}\left[\sum_{t=1}^{T} f(X_t, w(X_t))\right], \tag{4.1}$$

where $f : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ is the one-step state-action cost, and $\mathbb{E}^w_{x_1}[\cdot]$ is the expectation w.r.t. the law of the Markov chain $\boldsymbol{X} = \{X_t\}$ with the deterministic initial condition $X_1 = x_1$, and with controlled transition probabilities

$$\Pr\{X_{t+1} = y | X_t = x\} = K(y|x, w(x)).$$

Here, $K(y|x, u)$ is the controlled *transition kernel*. The construction of an optimal policy that minimizes (4.1) for every initial condition x_1 is a well-known dynamic optimization problem whose solution revolves around the average cost optimality equation (ACOE) (Arapostathis et al. (1993)). Alternatively, it can be reformulated as a "static" optimization problem over a convex subset $\mathcal{G}$ of the simplex of all probability distributions on $\mathsf{X}\times\mathsf{U}$ (the so-called *state-action polytope*, see Section 2.3.3 for details). As a result, minimizing the long-term expected average cost can be solved using a LP. The idea behind this convex-analytic method dates back to the work of Manne in the early 1960s (Manne (1960)) (see Borkar (2002) for a detailed introduction).

While the LP approach to MDPs is widely studied and used, there are some situations where a direct application of this approach may not be suitable. We may have some extra knowledge or side information that could be used to further constrain

our optimization problem. For example, we could directly get a nominal state-action distribution from an expert in the field and want to use it as a reference during the optimization. Alternatively, there are situations where we only have access to certain sample state-action trajectories, which could be obtained from observing an expert performing a task; the goal is to build a controller that mimics the expert while minimizing costs. In this case, we could use the observed trajectories to form an estimation of the state-action distribution and then use it as a reference. In those settings, if we apply the traditional LP approach without any regularization, we will be ignoring important data that could possibly improve our performance.

To address the gap between the traditional LP approach and the above-mentioned situations, we add a relative entropy term to the original LP optimization objective. Suppose we have computed a reference measure as a prior from observed data or an external advice; the relative entropy between any candidate measure in $\mathcal{G}$ and the reference measure (which may or may not be a member of $\mathcal{G}$) will act as a regularizer. By relaxing the LP approach with an entropy regularization, we find a policy that makes the expected ergodic cost as small as possible while staying close to the reference measure. Moreover, we also have a regularization parameter that controls how much we should trust the side information; in the online setting, this parameter can be interpreted as the learning rate. We will motivate our formulation in more detail within the context of online MDP policy learning problem in the next section.

We relate this entropy-regularized approach to the theory of exponential families of probability distributions (Wainwright and Jordan (2008)), and extract various useful properties of the optimizing measure, its associated policy, and the corresponding ergodic cost, including their dependence on the state-action costs and the regularization parameter. To the best of our knowledge, this connection is not addressed in existing works. Understanding the behavior of the optimizing measure and the re-

sulting policy gives an insight into several interesting problems, such as online MDP policy learning. From previous chapters and recent work (Dick et al. (2014); Zimin and Neu (2013)), it is known that we can solve an online MDP policy learning problem by using online linear optimization methods, where at each time step we deal with an optimization problem that exactly falls in our formulation. However, existing works focus mainly on the regret analysis, so that the stability and sensitivity of the derived policy are unaddressed. Our results on the sensitivity of the optimizing measure can be used to study those kinds of properties of the resulting policies. We also identify the role of the relative value function within our framework by relating it to the canonical parameters of the exponential family, and derive its dependence on the regularization parameter.

4.2 Entropy-regularized MDPs

In this section, we start by briefly reviewing the LP approach to MDPs, and then give a formal description of our entropy-regularized approach.

Let X and U be finite state and action spaces, and $K(y|x,u)$ be the state transition kernel of a controlled Markov chain. Let $\mathcal{G} \subset \mathcal{P}(\mathsf{X} \times \mathsf{U})$ be *state-action polytope*: the set of all *ergodic occupation measures* (see Section 2.3.3 for details).

It is well-known that the average cost optimal control problem can be cast as an LP over the state-action polytope $\mathcal{G}$ (Manne (1960); Borkar (2002)). Specifically, under some mild assumptions (Manne (1960); Meyn (2008)), the optimal average cost is independent of the initial state and can be expressed as the value of

$$\min_{\nu} \langle \nu, f \rangle \quad \text{subject to } \nu \in \mathcal{G}. \tag{4.2}$$

The basic idea is that we can express the expected long-term average cost of any Markov policy as an expectation of the state-action cost w.r.t. the empirical distribution of the states and the actions. Via limiting arguments relying on Birkhoff's

Individual Ergodic Theorem (Birkhoff (1931)), it can be shown that this empirical distribution converges to an element of $\mathcal{G}$. Thus, instead of solving a dynamic problem involving the long-term average cost, we can solve a static problem of minimizing a linear functional of the joint distribution of the state and the action, subject to the invariance constraints. In fact, the minimizing ν^* must be an extreme point of $\mathcal{G}$, and it is well-known that the extreme points of $\mathcal{G}$ correspond precisely to *deterministic* control policies (Borkar (2002); Meyn (2008)). Therefore, the resulting policy P_{ν^*} computed according to (2.13) is a Dirac measure for each $x \in \mathsf{X}$.

The ACOE can be seen to arise from this framework as well. Indeed, we can account for the invariance constraint $\nu \in \mathcal{G}$ by introducing the Lagrangian

$$\mathcal{L}_f(\nu, h) = \langle \nu, f + Kh - h \otimes 1 \rangle, \tag{4.3}$$

where the Lagrange multiplier h takes values in $\mathcal{C}(\mathsf{X})$, Kh is the element of $\mathcal{C}(\mathsf{X} \times \mathsf{U})$ given by

$$Kh(x, u) \triangleq \sum_{y \in \mathsf{X}} K(y|x, u)h(y), \qquad \forall (x, u) \in \mathsf{X} \times \mathsf{U} \tag{4.4}$$

and $h \otimes 1$ is the function $(x, u) \mapsto h(x)$. It can then be shown using the convex duality (see Proposition 9.2.11 in Meyn (2008)) that the maximizer of $\mathcal{L}_f(\nu, h)$ over $\mathcal{C}(\mathsf{X})$ is precisely the *relative value function* $h_f : \mathsf{X} \to \mathbb{R}$ that solves the ACOE

$$h_f(x) + \eta = \min_{u \in \mathsf{U}(x)} \{ f(x, u) + Kh_f(x, u) \}, \ x \in \mathsf{X} \tag{4.5}$$

where $\mathsf{U}(x) \subseteq \mathsf{U}$ is the set of allowable actions in state x and η is the value of the linear program (4.2).

4.3 Problem formulation

We now consider a relaxation of (4.2) that adds a relative entropy regularization term. Fix a state-action cost function $f : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ and consider the following

problem:

$$\min_{\nu \in \mathcal{G}} \{\beta \langle \nu, f \rangle + D(\nu \| \nu_0)\}, \tag{4.6}$$

where $\beta \geqslant 0$ is the regularization parameter, and $\nu_0 \in \mathcal{P}(\mathsf{X} \times \mathsf{U})$ is a given reference measure. As we mentioned above, the motivation for this is that the Markov policy induced by the optimizing occupation measure will minimize the expected ergodic cost while staying close to the reference measure ν_0. Note that ν_0 does not have to be a member of $\mathcal{G}$. For instance, if ν_0 is an empirical state-action distribution computed from a finite-time trajectory of some MDP, it might not necessarily satisfy the invariance constraints. We also point out that (4.6) can be thought of as a scalarized version of the problem

$$\begin{aligned} &\text{minimize } D(\nu \| \nu_0) \\ &\text{subject to } \langle \nu, f \rangle = c, \quad \nu \in \mathcal{G} \end{aligned}$$

for an appropriately chosen $c \in \mathbb{R}$. We recognize this as the so-called *information projection* (or I-projection) (Csiszár (1975)) of ν_0 onto the set of all $\nu \in \mathcal{G}$ satisfying a given linear equality constraint. For this reason, we may refer to the problem of finding the minimizer in (4.6) as an *information projection onto the state-action polytope.*

We now describe the basic structure of the optimizing measure in $\mathcal{G}$. Let $N = |\mathsf{X}|$, and let $\mathsf{X} = \{x^{(1)}, \ldots, x^{(N)}\}$ be an arbitrary enumeration of the states. Consider N functions $r_i : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ defined for each $i = 1, \ldots, N$ by

$$r_i(x, u) \triangleq K(x^{(i)}|x, u) - \mathbf{1}\{x = x^{(i)}\}. \tag{4.7}$$

The constraint $\nu \in \mathcal{G}$ is equivalent to N linear constraints $\langle \nu, r_i \rangle = 0, i = 1, \ldots, N$. From (4.7), we see that

$$\sum_{i=1}^{N} r_i(x, u) = 0, \qquad \forall (x, u) \in \mathsf{X} \times \mathsf{U}.$$

That is, the functions $r_1, \ldots, r_N$ are linearly dependent. Fortunately, this is easy to fix: if we remove one of the invariance constraints (say, the Nth one), the remaining $N-1$ constraints $\langle \nu, r_i \rangle = 0$, $1 \leqslant i \leqslant N-1$, still ensure that $\nu \in \mathcal{G}$. Then we can rewrite the problem in (4.6) as

$$\min_{\nu \in \mathcal{P}(\mathsf{X} \times \mathsf{U})} \{\beta \langle \nu, f \rangle + D(\nu \| \nu_0)\}$$
$$\text{subject to } \langle \nu, r_i \rangle = 0, i = 1, \ldots, N-1. \tag{4.8}$$

This is a convex program, since the relative entropy $D(\cdot\|\cdot)$ is convex in each of its arguments. Introducing Lagrange multipliers for the invariance constraints and using the nonnegativity of the relative entropy, it can be shown that the solution is given by the *Gibbs distribution*

$$\nu_f^\beta = \frac{\nu_0 \exp\left\{-\left(\beta f + \sum_{i=1}^{N-1} \lambda_i^\beta r_i\right)\right\}}{\left\langle \nu_0, \exp\left\{-\left(\beta f + \sum_{i=1}^{N-1} \lambda_i^\beta r_i\right)\right\}\right\rangle}, \tag{4.9}$$

where the parameters $\lambda_1^\beta, \ldots, \lambda_{N-1}^\beta$ are chosen to satisfy the invariance constraints. We can now extract the desired control policy:

$$\begin{aligned} P_f^\beta(u|x) &= P_{\nu_f^\beta}(u|x) \\ &= \frac{\nu_0(x,u) \exp\left\{-\beta f(x,u) - \sum_{i=1}^{N-1} \lambda_i^\beta r_i(x,u)\right\}}{\sum_{v \in \mathsf{U}} \nu_0(x,v) \exp\left\{-\beta f(x,v) - \sum_{i=1}^{N-1} \lambda_i^\beta r_i(x,v)\right\}}. \end{aligned} \tag{4.10}$$

Note that this policy is randomized, in contrast to the deterministic nature of the optimizing policy in the usual average-cost setting without entropy regularization.

Remark 6. By eliminating the Nth invariance constraint, we have resolved a nonuniqueness issue involving λ^β. If we had kept all N invariance constraints, then the solution to (4.6) would have been of the same Gibbs form (4.9), except with N

parameters $\lambda_1^\beta, \ldots, \lambda_N^\beta$. However, in that case adding an arbitrary $\alpha \neq 0$ to each λ_i^β, $1 \leqslant i \leqslant N$, would have resulted in the same distribution ν_f^β. Removing the Nth constraint is thus equivalent to setting $\lambda_N^\beta \equiv 0$.

4.3.1 Motivation

Now we take a look at a situation where an optimization problem like (4.6) could arise. Let us consider the online MDP setting with arbitrarily varying state-action costs. In Section 2.3.3 we reduced the MDP problem to an online linear optimization over $\mathcal{G}$: the agent must select a sequence $\{\nu_t\}_{t=1}^\infty$ in $\mathcal{G}$ to minimize the regret

$$\sum_{t=1}^T \langle \nu_t, f_t \rangle - \inf_{\nu \in \mathcal{G}} \sum_{t=1}^T \langle \nu, f_t \rangle$$

subject to the causality requirement that ν_t may depend only on $f_1, \ldots, f_{t-1}$. The Markov policy P_t at time t is then given by P_{ν_t}, according to (2.13). Dick et al. (2014) propose an online policy based on the (inexact) MD algorithm (Cesa-Bianchi and Lugosi, 2006, Ch. 11) over the polytope $\mathcal{G}$, and we derived an OMD algorithm in Section 2.3.3. Alternatively, one may consider the Follow the Regularized Leader (FRL) algorithm (Cesa-Bianchi and Lugosi (2006)): at each time step t, the algorithm selects

$$\nu_t = \arg\min_{\nu \in \mathcal{G}} \left\{ \beta \left\langle \nu, \sum_{s=1}^{t-1} f_s \right\rangle + D(\nu \| \nu_0) \right\}, \tag{4.11}$$

where the reference measure $\nu_0 \in \mathcal{P}(\mathsf{X} \times \mathsf{U})$ assigns positive probability to each state-action pair $(x, u) \in \mathsf{X} \times \mathsf{U}$. The intuition behind (4.11) is to balance exploitation (where we minimize the expected cost given by the average of all previously revealed cost functions) against exploration (where we may want to visit each state-action pair infinitely often with probability one).

Both the FRL and the MD algorithms require a projection onto the the convex polytope $\mathcal{G}$. Existing works (Dick et al. (2014); Zimin and Neu (2013)) have so

far focused mainly on deriving sublinear regret bounds. Much less is known about the behavior of the optimizing measure and its associated policy. Here are some questions that can naturally arise in this setting from a control-theoretic point of view:

- How does a no-regret policy behave if we change the learning rate β or state-action costs?
- What plays the role of the relative value function in this entropy-regularized setting, and how does it depend on state-action costs and learning rates?
- How do we efficiently implement this projection step?

Answers to these questions are crucial for understanding stability and sensitivity properties of the resulting policies, but most of them have not been addressed in existing works. Peters et al. (2010) consider a similar setup and propose a reinforcement learning method which identifies the role of the relative value function in the presence of a relative entropy regularization. However, the sensitivity of their solution to the step size and costs are not mentioned. Dick et al. (2014) deal with the projection by an approximating algorithm, and Zimin and Neu (2013) only study the regret bound and present the optimization problem without providing the solution. This chapter aims to address these questions or at least shed some light on them. Indeed, the optimization problem (4.11) has the exact same form as our entropy-regularized formulation (4.6). Consequently, the optimizing occupation measure selected at each time step t is of the Gibbs form (4.9). In next section we study this Gibbs measure, and relate the projection to the relative value function.

4.4 Main results on the Gibbs policy

We now turn to our analysis of the Gibbs distribution (4.9) and the induced policy (4.10). In particular, we address the following questions:

1. How do the invariance parameters λ_i^β, $i \in \{1, \dots, N\}$, depend on the state-action cost f and on the regularization parameter β?

2. How do the Gibbs distributions ν_f^β behave as we vary β and/or f? Note that increasing β corresponds to relaxing the relative entropy constraint in favor of keeping the ergodic cost small.

3. What plays the role of the relative value function h_f in the presence of entropy regularization? What is the corresponding analogue of the ACOE?

To that end, we exploit the fact that the optimizing distribution ν_f^β is a member of an *exponential family* (Wainwright and Jordan (2008)), with some additional structure arising from the invariance constraints.

4.4.1 Embedding into an exponential family

Let us fix a state-action cost function $f \in \mathcal{C}(\mathsf{X} \times \mathsf{U})$. We impose the following assumption:

Assumption 4. *The N functions $f, r_1, \dots, r_{N-1}$ are* affinely independent*, i.e., the functions $r_1 - f, \dots, r_{N-1} - f$ are linearly independent.*

Next, we define the *log-partition function*[1]

$$\Phi_f(\lambda) \triangleq \log \left\langle \nu_0, \exp \left\{ -\lambda_0 f - \sum_{i=1}^{N-1} \lambda_i r_i \right\} \right\rangle$$

[1] This terminology comes from statistical physics.

of a vector $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_{N-1})^T$ of N real parameters. Since f and all r_i's are bounded, $\Phi_f(\lambda) < \infty$ for all $\lambda \in \mathbb{R}^N$. We denote by $\mathcal{E}_f$ the family of all Gibbs distributions

$$\nu_{f,\lambda} \triangleq \nu_0 \exp\left\{-\lambda_0 f - \sum_{i=1}^{N-1} \lambda_i r_i - \Phi_f(\lambda)\right\}, \ \lambda \in \mathbb{R}^N. \tag{4.12}$$

Since Φ_f is everywhere finite, and with Assumption 4 in force, this is a *regular* and *minimal exponential family* of distributions (Wainwright and Jordan (2008)), where the functions $f, r_1, \ldots, r_{N-1}$ are called the *sufficient statistics* and $\lambda \in \mathbb{R}^N$ is the vector of *canonical parameters*. The following standard result on regular and minimal exponential families will be useful:

Lemma 16. *The function Φ_λ has the following properties:*

1. *It is strictly convex.*

2. *It is infinitely differentiable, and its derivatives are the cumulants of the sufficient statistics. In particular:*

$$\nabla\Phi_f(\lambda) = -\mathbb{E}_{f,\lambda}[\varphi_f], \ \nabla^2\Phi_f(\lambda) = \mathrm{Cov}_{f,\lambda}\left[\varphi_f\right], \tag{4.13}$$

where $\mathbb{E}_{f,\lambda}[\cdot]$ and $\mathrm{Cov}_{f,\lambda}(\cdot)$ denote, respectively, the expectation and covariance w.r.t. $\nu_{f,\lambda}$, and $\varphi_f : \mathsf{X} \times \mathsf{U} \to \mathbb{R}^N$ is the vector of sufficient statistics:

$$\varphi_f(x,u) \triangleq \big(f(x,u), r_1(x,u), \ldots, r_{N-1}(x,u)\big).$$

4.4.2 Basic properties of the Gibbs policy

The Gibbs distribution ν_f^β that solves the optimization problem (4.6) has the form (4.12) with $\lambda_0 = \beta$ and $\lambda_i = \lambda_i^\beta$ for $i = 1, \ldots, N-1$. Moreover, it is also an element of the state-action polytope $\mathcal{G}$. Thus, it is contained in the set

$$\mathcal{G}_f \triangleq \mathcal{E}_f \cap \mathcal{G}.$$

Unlike $\mathcal{G}$, this set is not convex (in fact, neither is $\mathcal{E}_f$). Let us introduce some notation: We let $\lambda^\beta \in \mathbb{R}^N$ denote the vector $(\beta, \lambda_1^\beta, \ldots, \lambda_{N-1}^\beta)^T$, and we let

$$E_f(\beta) \triangleq \left\langle \nu_f^\beta, f \right\rangle$$

$$D_f(\beta) \triangleq D\big(\nu_f^\beta \| \nu_0\big)$$

$$A_f(\beta) \triangleq -\beta^{-1}\Phi_f(\lambda^\beta).$$

With this, we have the following result:

Theorem 17. *The ergodic occupation measure ν_f^β that solves* (4.6) *is an element of $\mathcal{G}_f$, and has the following properties:*

1. *It satisfies the* Helmholtz formula $A_f(\beta) = E_f(\beta) + \beta^{-1}D_f(\beta)$.

2. *It has the* equalizer property: *Any ergodic occupation measure $\mu \in \mathcal{G}$ with the same ergodic cost as ν_f^β, i.e., $\langle \mu, f \rangle = E_f(\beta)$, satisfies*

$$\left\langle \mu, \log \frac{\nu_f^\beta}{\nu_0} \right\rangle = D_f(\beta). \tag{4.14}$$

3. *It has the* I-projection property (Csiszár (1975)): *for any $\mu \in \mathcal{G}$ with $\langle \mu, f \rangle = E_f(\beta)$,*

$$D(\mu\|\nu_0) \geqslant D(\mu\|\nu_f^\beta) + D(\nu_f^\beta\|\nu_0), \tag{4.15}$$

   with equality if $D(\mu\|\nu_0) < \infty$.

4. *The vector λ^β solves the equation*

$$\nabla\Phi_f(\lambda^\beta) = -E_f(\beta)e_0, \tag{4.16}$$

   where $e_0 \triangleq (1, 0, \ldots, 0)^T \in \mathbb{R}^N$.

Remark 7. The term "Helmholtz formula" comes from statistical physics, where the functional $A_f(\mu) \triangleq \langle \mu, f \rangle + \beta^{-1} D(\mu \| \nu_0)$ is known as the Helmholtz free energy of μ at the inverse temperature β (Ellis (1985); Streater (2009)). Then the fact that

$$A_f(\beta) = A_f(\nu_f^\beta) = \min_{\mu \in \mathcal{P}(\mathsf{X} \times \mathsf{U})} A_f(\mu) \tag{4.17}$$

is referred to as the *thermodynamic stability* of ν_f^β.

Proof. By the definition of the relative entropy,

$$\begin{aligned} D_f(\beta) &= \left\langle \nu_f^\beta, \log \frac{\nu_f^\beta}{\nu_0} \right\rangle \\ &= -\left\langle \nu_f^\beta, \beta f + \sum_{i=1}^{N-1} \lambda_i^\beta r_i \right\rangle - \Phi_f(\lambda^\beta) \\ &= -\beta \langle \nu_f^\beta, f \rangle - \Phi_f(\lambda^\beta) \\ &= \beta \left(A_f(\beta) - E_f(\beta) \right), \end{aligned}$$

where we have used the fact that $\nu_f^\beta \in \mathcal{G}$, so that $\langle \nu_f^\beta, r_i \rangle = 0$ for all i. Next, we write

$$\begin{aligned} \left\langle \mu, \log \frac{\nu_f^\beta}{\nu_0} \right\rangle &= -\left\langle \mu, \beta f + \sum_{i=1}^{N-1} \lambda_i^\beta r_i \right\rangle - \Phi_f(\lambda^\beta) \\ &= -\beta \langle \mu, f \rangle - \Phi_f(\lambda^\beta) \\ &= -\beta E_f(\beta) + \beta A_f(\beta) \\ &= D_f(\beta), \end{aligned}$$

where we have used the assumptions that $\mu \in \mathcal{G}$ and $\langle \mu, f \rangle = E_f(\beta)$. Now, if $D(\mu \| \nu_0) = \infty$, then (4.15) holds trivially. Thus, we assume $D(\mu \| \nu_0) < \infty$, i.e., $\mathrm{supp}(\mu) \subseteq \mathrm{supp}(\nu_0)$. But then, since $\mathrm{supp}(\nu_f^\beta) = \mathrm{supp}(\nu_0)$, we have $D(\mu \| \nu_f^\beta) < \infty$, and therefore

$$\begin{aligned} D(\mu \| \nu_0) &= \left\langle \mu, \log \frac{\mu}{\nu_0} \right\rangle = \left\langle \mu, \log \frac{\mu}{\nu_f^\beta} \right\rangle + \left\langle \mu, \log \frac{\nu_f^\beta}{\nu_0} \right\rangle \\ &= D(\mu \| \nu_f^\beta) + D(\nu_f^\beta \| \nu_0), \end{aligned}$$

where in the last step we have used the equalizer property (4.14). Finally, we note that

$$\langle \nu_f^\beta, f \rangle = E_f(\beta) \quad \text{and} \quad \langle \nu_f^\beta, r_i \rangle = 0, \, \forall i,$$

and then use (4.13). □

4.4.3 Sensitivity to perturbations in f and β

Next, we investigate the effect of varying the state-action cost f or the entropy regularization parameter β. When f stays fixed, we are still in the same exponential family $\mathcal{E}_f$. When we change f to f', we must consider another exponential family $\mathcal{E}_{f'}$, where we assume that f' still satisfies Assumption 4. Consequently, we now use the notation λ_f^β, to indicate the dependence of the canonical parameters on the underlying state-action cost explicitly. We start by analyzing the effect of varying the cost function f:

Theorem 18 (Sensitivity to perturbations in f)**.** *For a fixed $\beta \geqslant 0$ and for any two $f, f' \in \mathcal{C}(\mathsf{X} \times \mathsf{U})$,*

$$D(\nu_f^\beta \| \nu_{f'}^\beta) \leqslant \frac{\beta^2}{2} \|f - f'\|_\infty^2. \tag{4.18}$$

Moreover, the ergodic cost is Lipschitz-continuous on bounded subsets of $\mathcal{C}(\mathsf{X} \times \mathsf{U})$:

$$|E_f(\beta) - E_{f'}(\beta)| \leqslant (M\beta + 1) \, \|f - f'\|_\infty, \tag{4.19}$$

where $M \triangleq \|f\|_\infty \vee \|f'\|_\infty$.

Proof. The bound (4.18) is proved as follows:

$$
\begin{aligned}
D(\nu_f^\beta \| \nu_{f'}^\beta) &= \left\langle \nu_f^\beta, \log \frac{\nu_f^\beta}{\nu_{f'}^\beta} \right\rangle \\
&= \beta \langle \nu_f^\beta, f' - f \rangle + \beta \left(A_f(\beta) - A_{f'}(\beta) \right) \\
&= \beta \langle \nu_f^\beta, f' - f \rangle + \log \frac{\left\langle \nu_0, \exp\left\{ -\beta f' - \sum_{i=1}^{N-1} \lambda_i^\beta r_i \right\} \right\rangle}{\left\langle \nu_0, \exp\left\{ -\beta f - \sum_{i=1}^{N-1} \lambda_i^\beta r_i \right\} \right\rangle} \\
&= \beta \langle \nu_f^\beta, f' - f \rangle + \log \left\langle \nu_f^\beta, e^{\beta(f - f')} \right\rangle \\
&\leqslant \frac{\beta^2 \| f - f' \|_\infty^2}{2},
\end{aligned}
$$

where we have used Hoeffding's lemma (Hoeffding (1963)) (also see (Cesa-Bianchi and Lugosi, 2006, Lemma A.1)). This proves (4.18). Now, we prove (4.19):

$$
\begin{aligned}
|E_f(\beta) - E_{f'}(\beta)| &= \left| \langle \nu_f^\beta, f \rangle - \langle \nu_{f'}^\beta, f' \rangle \right| \\
&\leqslant \left| \langle \nu_f^\beta - \nu_{f'}^\beta, f \rangle \right| + \left| \langle \nu_{f'}^\beta, f - f' \rangle \right| \\
&\leqslant M \| \nu_f^\beta - \nu_{f'}^\beta \|_{\mathrm{TV}} + \| f - f' \|_\infty \\
&\leqslant M \sqrt{2 D(\nu_f^\beta \| \nu_{f'}^\beta)} + \| f - f' \|_\infty \\
&\leqslant (M\beta + 1) \| f - f' \|_\infty,
\end{aligned}
$$

where in the last two lines we have used the Pinsker's inequality (Cover and Thomas (2006)), as well as (4.18). □

Now we consider the effect of varying β while keeping f fixed. It will be convenient to introduce the following shorthand notation:

$$D_f(\beta | \beta') \triangleq D(\nu_f^\beta \| \nu_f^{\beta'}).$$

Theorem 19 (Sensitivity to perturbations in β)**.** *For any $f \in \mathcal{C}(\mathsf{X} \times \mathsf{U})$ and any pair $\beta, \beta' \geqslant 0$,*

$$D_f(\beta | \beta') + D_f(\beta' | \beta) = \left(E_f(\beta) - E_f(\beta') \right) (\beta' - \beta). \tag{4.20}$$

The corresponding ergodic costs satisfy

$$|E_f(\beta) - E_f(\beta')| \leqslant \|f\|_\infty^2 |\beta - \beta'|, \tag{4.21}$$

i.e., the function $\beta \mapsto E_f(\beta)$ is Lipschitz-continuous.

Proof. A simple calculation using the definitions gives

$$D_f(\beta|\beta') = \Phi_f(\lambda^{\beta'}) - \Phi_f(\lambda^\beta) + E_f(\beta)(\beta' - \beta).$$

Interchanging the roles of β and β' and adding the two equations, we see that the terms involving Φ_f cancel, and we are left with (4.20). To prove (4.21), we write

$$\begin{aligned}
|E_f(\beta) - E_f(\beta')| &= \left| \left\langle \nu_f^\beta - \nu_f^{\beta'}, f \right\rangle \right| \\
&\leqslant \|f\|_\infty \|\nu_f^\beta - \nu_f^{\beta'}\|_{\text{TV}} \\
&\leqslant \frac{1}{2}\|f\|_\infty \left(\sqrt{2D_f(\beta|\beta')} + \sqrt{2D_f(\beta'|\beta)} \right) \\
&\leqslant \|f\|_\infty \sqrt{D_f(\beta|\beta') + D_f(\beta'|\beta)} \\
&= \|f\|_\infty \sqrt{(E_f(\beta) - E_f(\beta'))\,(\beta' - \beta)},
\end{aligned}$$

where the third line is by the Pinsker's inequality, the fourth line is by concavity of the square root, and the last line is by (4.20). Squaring both sides and rearranging, we get (4.21). □

4.4.4 Canonical parameters and the relative value function

The vector of canonical parameters $\lambda^\beta \in \mathbb{R}^N$ that identifies the Gibbs distribution (4.9) as a member of the exponential family $\mathcal{E}_f$ satisfies the relation (4.16). This relation couples the coordinates $\lambda_0^\beta \equiv \beta$ and $\lambda_1^\beta, \ldots, \lambda_{N-1}^\beta$ in a nontrivial way. We now show that the parameters λ_i^β, $1 \leqslant i \leqslant N-1$, can be used to define a function $h_f^\beta : \mathsf{X} \to \mathbb{R}$ that plays the same role in the presence of the entropy regularization as

the relative value function h in the entropy-unconstrained optimal control problem (4.2).

Specifically, we define h_f^β by setting

$$h_f^\beta(x) = \begin{cases} \beta^{-1}\lambda_{f,i}^\beta, & \text{if } x = x^{(i)},\ i = 1, \dots, N-1 \\ 0, & \text{if } x = x^{(N)}. \end{cases} \tag{4.22}$$

With this, we can write

$$\begin{aligned} \sum_{i=1}^{N-1} \lambda_{f,i}^\beta r_i(x,u) &= \beta \left(\sum_{y \in \mathsf{X}} K(y|x,u) h_f^\beta(y) - h_f^\beta(x) \right) \\ &\equiv \beta \left(K h_f^\beta(x,u) - h_f^\beta(x) \right). \end{aligned}$$

Now the ergodic occupation measure ν_f^β and the induced policy P_f^β can be expressed as

$$\begin{aligned} &\nu_f^\beta(x,u) \\ &\quad = \frac{\nu_0(x,u) \exp\left\{ -\beta \left(f(x,u) + K h_f^\beta(x,u) - h_f^\beta(x) \right) \right\}}{\left\langle \nu_0, \exp\left\{ -\beta \left(f + Kh - h \otimes 1 \right) \right\} \right\rangle} \end{aligned}$$

and

$$P_f^\beta(u|x) = \frac{\nu_0(x,u) \exp\left\{ -\beta \left(f(x,u) + K h_f^\beta(x,u) \right) \right\}}{\sum_{v \in \mathsf{U}} \nu_0(x,v) \exp\left\{ -\beta \left(f(x,v) + K h_f^\beta(x,v) \right) \right\}}$$

respectively. We can now recognize the policy P_f^β as a "softmax" relaxation of the deterministic optimal policy

$$w^*(x) = \arg\min_{u \in \mathsf{U}(x)} \left\{ f(x,u) + Kh(x,u) \right\},$$

where $h = h_f$ solves the ACOE (4.5), except now we have $\mathsf{U}(x) = \{u \in \mathsf{U} : \nu_0(x,u) \neq 0\}$ and, instead of choosing an action u that minimizes the sum

of the immediate cost $f(x,u)$ and the forecast future value $Kh(x,u)$, the controller samples each action u with a probability that decays exponentially with $f(x,u) + Kh_f^\beta(x,u)$.

Moreover, the gradient relation (4.16) that determines λ_f^β can be derived alternatively by considering the dual of (4.6), in the same way that the ACOE (4.5) arises from the dual of (4.2) (Meyn (2008)). Specifically, let us introduce the Lagrangian

$$\mathcal{L}_f^\beta(\nu, h) \triangleq \beta\langle \nu, f + Kh - h \otimes 1\rangle + D(\nu\|\nu_0), \tag{4.23}$$

where the Lagrange multiplier h takes values in $\mathcal{C}(\mathsf{X} \times \mathsf{U})$. Since the constraints in (4.6) are given by linear (in)equalities, strong duality holds by Slater's condition:

$$\inf_{\nu \in \mathcal{G}} \{\beta\langle \nu, f\rangle + D(\nu\|\nu_0)\} = \sup_{h \in \mathcal{C}(\mathsf{X})} \inf_{\nu \in \mathcal{P}(\mathsf{X}\times\mathsf{U})} \mathcal{L}_f^\beta(\nu, h).$$

Using the nonnegativity of the relative entropy, it is easy to show that the infimum of $\mathcal{L}_f^\beta(\nu, h)$ over $\nu \in \mathcal{P}(\mathsf{X} \times \mathsf{U})$ is achieved by the Gibbs distribution

$$\nu_{f,h}^\beta \triangleq \frac{\nu_0 \exp\{-\beta\,(f + Kh - h \otimes 1)\}}{\langle \nu_0, \exp\{-\beta\,(f + Kh - h \otimes 1)\}\rangle},$$

and that

$$\begin{aligned}&\inf_{\nu \in \mathcal{P}(\mathsf{X}\times\mathsf{U})} \mathcal{L}_f^\beta(\nu, h)\\ &\qquad = -\log\langle \nu_0, \exp\{-\beta\,(f + Kh - h \otimes 1)\}\rangle.\end{aligned}$$

This is a concave and differentiable functional of h, and writing out the corresponding first-order optimality condition gives precisely (4.16) once we relate the maximizer h_f^β to the canonical parameters $\lambda_1^\beta, \ldots, \lambda_{N-1}^\beta$ through (4.22). We also observe that $\beta^{-1}\mathcal{L}_f^\beta(\nu, h)$ converges, as $\beta \to +\infty$, to the Lagrangian (4.3) for the average-cost optimal control problem without the entropy regularization. We can also write down the analogue of the ACOE (4.5) in the presence of the entropy regularization:

$$A_f(\beta) = -\frac{1}{\beta}\log\left\langle \nu_0, \exp\left\{-\beta\left(f + Kh_f^\beta - h_f^\beta \otimes 1\right)\right\}\right\rangle,$$

and the usual ACOE is recovered in the limit as $\beta \to \infty$.

There is still a question of how the canonical parameters $\lambda_i^{(\beta)}$, $1 \leqslant i \leqslant N-1$, or, equivalently, the relative value function h_f^β, change as we vary β. The following theorem provides a partial answer, provided the log-partition function Φ_f is strongly convex:

Theorem 20. *Fix a cost function $f : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$. Suppose that there exists a constant $c > 0$, such that*

$$\Phi_f(\lambda) \geqslant \Phi_f(\lambda') + \nabla\Phi_f(\lambda')^T(\lambda - \lambda') + \frac{c}{2}\|\lambda - \lambda'\|^2 \tag{4.24}$$

for all $\lambda, \lambda' \in \mathbb{R}^N$, where $\|\cdot\|$ denotes the Euclidean (ℓ_2) norm. Then

$$\|\lambda_f^\beta - \lambda_f^{\beta'}\| \leqslant \frac{\|f\|_\infty^2}{c}|\beta - \beta'|$$

for any $\beta, \beta' \geqslant 0$.

Proof. From the properties of the log-partition function Φ_f (see Lemma 16) it can be shown (Wainwright and Jordan (2008)) that the gradient mapping $\lambda \mapsto \nabla\Phi_f(\lambda)$ is invertible, and the inverse mapping is given by $\xi \mapsto \nabla\Phi_f^*(\xi)$, where

$$\Phi_f^*(\xi) \triangleq \sup_\lambda \{\langle \xi, \lambda\rangle - \Phi_f(\lambda)\}$$

is the Legendre–Fenchel conjugate of Φ_f (Hiriart-Urruty and Lemarechal (2013)). Moreover, it follows from the strong convexity property (4.24) of Φ_f that $\nabla\Phi_f^*$ is Lipschitz-continuous with constant $1/c$:

$$\|\nabla\Phi_f^*(\xi) - \nabla\Phi_f^*(\xi')\| \leqslant \frac{1}{c}\|\xi - \xi'\| \tag{4.25}$$

(see (Hiriart-Urruty and Lemarechal, 2013, Theorem 4.2.1)). From this, we obtain

$$\begin{aligned}\|\lambda_f^\beta - \lambda_f^{\beta'}\| &= \left\|\nabla\Phi_f^*\left(E_f(\beta)e_0\right) - \nabla\Phi_f^*\left(E_f(\beta')e_0\right)\right\| \\ &\leqslant \frac{1}{c}\left\|\left(E_f(\beta) - E_f(\beta')\right)e_0\right\| \\ &= \frac{1}{c}\left|E_f(\beta) - E_f(\beta')\right| \\ &\leqslant \frac{\|f\|_\infty^2}{c}|\beta - \beta'|,\end{aligned}$$

where the first line is by (4.16), the second line is by (4.25), and the last line is by (4.21). □

In general, it may not be straightforward to verify the strong convexity condition (4.24), which can be thought of as a form of ergodicity.

4.5 Conclusion

We have extended the LP approach to MDPs by adding relative entropy regularization. We have shown that the ergodic occupation measure that minimizes the entropy-regularized ergodic control cost is a member of an exponential family of probability distributions, which has allowed us to exploit known results about exponential families to provide structural results for the optimizing measure. These results open the door to further in-depth analysis of the online MDP problem, such as optimal adaptation of the learning rate to observed data or efficient approximate implementation of online policy selection algorithms. We plan to explore these directions in future work.

5

Online MDPs with Kullback-Leibler control cost

5.1 Introduction

The set-up considered in preceding chapters is motivated by problems in machine learning and artificial intelligence, where the actions are the main object of interest, and the state merely represents memory effects present in the system. In this chapter, we take a more control-oriented view: the emphasis is on steering the system along a desirable state trajectory through actions selected according to a state feedback law. Following the formulation proposed by Todorov (2007, 2008, 2009), we allow the agent to modulate the state transitions directly, so that actions (resp., state feedback laws) correspond to probability distributions (resp., Markov kernels) on the underlying state space. As in Todorov (2007, 2008, 2009), the one-step cost is a sum of two terms: the state cost, which measures how "desirable" each state is, and the control cost, which measures the deviation of the transition probabilities specified by the chosen action from some fixed *default* or *passive dynamics*. (We also refer the reader to a recent paper by Kappen et al. (2012), which interprets Todorov's set-up as an inference problem for probabilistic graphical models.) An

explicit construction of a computationally efficient strategy with small regret (i.e., expected difference between its actual total cost and the smallest cost attainable using noncausal knowledge of the state costs) under mild regularity conditions is presented, along with a demonstration of the performance of the proposed strategy on a simulated target tracking problem.

More precisely, we consider a new class of MDP with a finite state space X, where the action space U is the simplex $\mathcal{P}(\mathsf{X})$ of probability distributions over X. A fixed Markov matrix (transition kernel) $P^* = [P^*(x,y)]_{x,y\in\mathsf{X}}$ is given. A stationary Markov policy (state feedback law) is a mapping $w : \mathsf{X} \to \mathcal{P}(\mathsf{X})$, so if the system is in state $x \in \mathsf{X}$, then the transition to the next state is stochastic, as determined by the probability distribution $u(\cdot) = w(x) \in \mathcal{P}(\mathsf{X})$. In other words, if we denote the next state by X^+, then the state transitions induced by the action u are governed by the conditional probability law

$$\Pr\{X^+ = x^+ | X = x\} = P(x, x^+) = u(x^+) = [w(x)](x^+).$$

The one-step state-action cost $c(x,u)$ consists of two terms, the *state cost* $f(x)$, where $f : \mathsf{X} \mapsto \mathbb{R}_+$ is a given function, and the *control cost*, which penalizes any deviation of the next-state distribution $u(\cdot) = w(x)$ from the one prescribed by $P^*(x,\cdot)$, the row of P^* corresponding to x. To motivate the introduction of such control costs, we can imagine the situation, in which implementing the state transitions according to P^* can be done "for free". However, it may very well be the case that following P^* will be in conflict with the goal of keeping the state cost low. From this perspective, it may actually be desirable to deviate from P^*. Any such deviation may be viewed as an *active perturbation* of the *passive dynamics* prescribed by P^*, and the agent should attempt to balance the tendency to keep the state costs low against allowing too strong of a perturbation of P^*. Our choice of control cost is inspired by the work of Todorov (2007, 2009), and is given by the KL divergence (Cover and Thomas

(2006)) $D(u\|P^*(x,\cdot))$ between the proposed next-state distribution $u(\cdot)$ and the next-state distribution prescribed by the passive dynamics P^*. One useful property of this control cost is that it automatically forbids all those state transitions that are already forbidden by P^*. Indeed, if for a given $x \in \mathsf{X}$ there exists some $y \in \mathsf{X}$ such that $u(y) = [w(x)](y) > 0$, while $P^*(x,y) = 0$, then $D(u\|P^*(x,\cdot)) = +\infty$. Thus, the overall one-step state-action cost is given by

$$c(x,u) = f(x) + D(u\|P^*(x,\cdot)), \quad \forall x \in \mathsf{X}, u \in \mathcal{P}(\mathsf{X}). \tag{5.1}$$

In the online version of this problem (detailed in Section 5.2.1), the state costs form an arbitrarily varying sequence $\{f_t\}_{t=1}^{\infty}$, and the agent learns the state cost for each time step only after having selected the transition law to determine the next state. For any given value of the horizon, the regret is computed with respect to the best stationary Markov policy (state feedback law) that could have been chosen in hindsight. The precise definition of regret is given in Section 5.2.2.

Since this is a nonstandard set-up, we take a moment to situate it in the context of usual models of MDPs. In a standard MDP with finite state and action spaces, we have a finite collection of Markov matrices P_u on X indexed by the actions $u \in \mathsf{U}$. State feedback laws are functions $w : \mathsf{X} \to \mathsf{U}$, and the set of all such functions is finite with cardinality $|\mathsf{U}|^{|\mathsf{X}|}$. Therefore, in each state $x \in \mathsf{X}$ the agent has at most $|\mathsf{U}|^{|\mathsf{X}|}$ choices for the distribution of the next state X^+, and we may equivalently represent each state feedback law w as a mapping from X into $\mathcal{P}(\mathsf{X})$ with $x \mapsto P_{w(x)}(x,\cdot)$. Since the state space U is finite, the range of this mapping is a finite subset of the probability simplex $\mathcal{P}(\mathsf{X})$. The criterion for selecting this next-state distribution pertains to minimization of the expectation of the immediate state-action cost plus a suitable value function that accounts for the effect of the current action on future costs. In many cases, the one-step state-action cost $c(x,u)$ decomposes into a sum of state cost $f(x)$ and control cost $g(x,u)$, where $f(x)$ quantifies the (un)desirability

of the state x, while $g(x, u)$ represents the effort required to apply action u in state x.

In the set-up of Todorov (2007, 2008, 2009), the collection of all possible next-state distributions is unrestricted. As a consequence, any mapping $w : \mathsf{X} \to \mathcal{P}(\mathsf{X})$ is a feasible stationary Markov policy. Since any Markov matrix P on X can be equivalently represented as a mapping from X into $\mathcal{P}(\mathsf{X})$ with $x \mapsto P(x, \cdot)$, there is thus a one-to-one correspondence between state feedback laws and Markov matrices on X. In contrast to the case when the agent may choose among a finite set of actions, the probability simplex $\mathcal{P}(\mathsf{X})$ is an uncountable set, so the agent has considerably greater freedom to choose the next state distribution. As before, we introduce a state-action cost $c(x, u) = f(x) + g(x, u)$, where $f(x)$ measures the (un)desirability of state x, while $g(x, u)$ quantifies the difficulty of executing action u in state x. Since actions u now correspond to probability distributions, and we choose the Kullback–Leibler control cost $g(x, u) = D(u\|P^*(x, \cdot))$, where P^* is a fixed Markov matrix on the state space X that may represent, e.g., the "free" dynamics of the system in the absence of external controls.

The KL divergence is widely used in stochastic control and inference. First of all, it has many desirable properties, such as nonnegativity and convexity (Cover and Thomas (2006)). Secondly, if we adopt the viewpoint that the purpose of a control policy is to shape the joint distribution of all relevant variables describing the closed-loop behavior of the system, then using the relative entropy to compare the distribution induced by any control law to some reference model leads to functional equations for the optimal policy that are often easier to solve than the corresponding dynamical programming recursion (Kárný (1996); Šindelář et al. (2008)) (e.g., see Kárný (1996) for an alternative derivation of the optimal controller in an LQG problem using relative entropy instead of dynamic programming); similar ideas are fruitful in the context of robust control, where the relative entropy is used to quan-

tify the radius of uncertainty around some nominal system (Petersen et al. (2000); Charalambous and Rezaei (2007); Hansen and Sargent (2008)). Moreover, the relative entropy is a canonical regularization functional for stochastic nonlinear filtering problems (Mitter and Newton (2003)): an optimal Bayesian filter is the solution of a variational problem that entails minimization of the sum of expected negative log-likelihood (which can be interpreted as state cost) and a relative entropy with respect to the prior measure on the state space.

To further motivate our interest in problems of this sort, let us consider two examples. One is *target tracking* with an arbitrarily moving target (or multiple targets). In this example, the state space X is the vertex set of an undirected graph, and the passive dynamics P^* specifies some default *random walk* on this graph. The tracker's discrete-time motion is constrained by the topology of the graph, while the targets' motions are not. At each time t, the state cost f_t is the tracking error, which quantifies how far the tracker is from the targets. For instance, it may be given by the graph distance (length of shortest path) between the tracker's current location and the location of the closest target. Other possibilities can also be considered, including some based on noisy information on the location of the targets. The control cost penalizes the tracker's deviation from P^* as it attempts to track the targets. The passive dynamics P^* can be seen as the tracker's prior model for the targets' motion. Moreover, if P^* is sufficiently rapidly mixing, then any tracker that follows P^* will visit every vertex of the graph infinitely often with probability one; however, there is no guarantee that the tracker's prior model is correct (i.e., that the tracker will be anywhere near the targets). Hence, the state-action cost will trade off the tendency of the tracker to "cover" the graph as much as possible (exploration) against the tendency to follow a potentially faulty model of the targets (exploitation).

Another example setting is real-time control of a *brain-machine interface*. There, the state space X may be the set of possible positions or modes of a neural prosthetic

device, and the passive dynamics P^* may encode the "natural" (free) dynamics of the device in the absence of user control; we may assume, for instance, that the state transitions prescribed by P^* correspond to "minimum-energy" operating mode of the device. If the user wishes to make the device execute some trajectory, the state cost f_t at time t may represent the deviation of the current point on the trajectory from the one intended by the user. Since the user is a human operator with conscious intent, we may not want to ascribe an a priori model to her intended trajectory, and instead treat it as an individual sequence modulating the state costs $\{f_t\}_{t=1}^{\infty}$. In this setting, the Kullback–Leibler control cost penalizes significant deviations from the free dynamics P^*, since these will typically be associated with energy expenditures.

The common thread running through these two examples (and it is certainly possible to construct many others) is that they model real-time interaction of a particular system with some well-defined "reference" or "nominal" dynamics P^* with a potentially unpredictable environment (which may include hard-to-model adversaries or rational agents, etc.), and we must balance the tendency to respond to immediate changes in the environment against the need to operate the system near the nominal mode. Since no offline policy design is possible in such circumstances, the regret minimization framework offers a meaningful alternative.

5.1.1 Our contributions and comparison with relevant literature

In this chapter, we give an explicit construction of a strategy for the agent, such that the regret relative to any uniformly ergodic class of stationary Markov policies grows *sublinearly* as a function of the horizon. The only regularity conditions needed for this result to hold are (a) uniform boundedness of the state costs (the agent need not know the bound, only that it exists); and (b) ergodicity of the passive dynamics. Moreover, our strategy is computationally efficient: the time is divided into phases of increasing length, and during each phase the agent applies a stationary Markov

policy optimized for the average of the state cost functions revealed during all of the preceding phases. Thus, our strategy belongs to the class of so-called "lazy" strategies for online decision-making problems (Vovk (1990); Merhav et al. (2002); Kalai and Vempala (2005)); a similar approach was also taken by Yu et al. (2009) in their paper on online MDPs with finite state and action spaces. The main advantage of lazy strategies is their computational efficiency, which, however, comes at the price of suboptimal scaling of the regret with the time horizon. We comment on this issue further in the sequel.

Our main contribution is an extension of the theory of online MDPs to a wide class of control problems that lie outside the scope of existing approaches (Even-Dar et al. (2009); Yu et al. (2009)). More specifically:

1. While in Even-Dar et al. (2009); Yu et al. (2009) both the state and the action spaces are finite, we only assume this for the state space. Our action space is the simplex of probability distributions on the state space, which is a compact subset of a Euclidean space. Hence, the techniques used in the existing literature are no longer directly applicable. (It is also possible to extend our approach to continuous state spaces, but additional regularity conditions will be needed. This extension will be the focus of our future work.)

2. Yu et al. (2009) assume that the underlying MDP is unichain (Puterman, 1994, Sec. 8.3) and satisfies a certain uniform ergodicity condition (a similar assumption is also needed by Even-Dar et al. (2009)). Their assumption is rather strong, since it places significant simultaneous restrictions on an exponentially large family of Markov chains on the state space (each chain corresponds to a particular choice of state feedback law, and there are $|\mathsf{U}|^{|\mathsf{X}|}$ such laws). It is also difficult to verify, since the problem of determining whether an MDP is unichain is NP-hard (Tsitisklis (2007)). By contrast, our ergodicity assumption

pertains to only *one* Markov chain (the passive dynamics P^*), it can be efficiently verified in polynomial time, and we prove that it automatically implies uniform ergodicity of all stationary control laws that could possibly be invoked by our strategy.

3. Because these stationary control laws correspond to solutions of certain average-cost optimality equations (ACOEs) in the set-up of Todorov (2007, 2008, 2009), we establish and subsequently exploit several useful and previously unknown results concerning the continuity and uniform ergodicity of optimal policies for Todorov's problem. These results, as well as the techniques used to prove them, play a very important role in our overall contribution. Indeed, in the online setting, the state cost functions are revealed to the agent in real time. Hence, any policy used by the agent must rely on estimates (or forecasts) of future state costs based on currently available information. Our new results on Todorov's optimal control laws provide sharp bounds on the sensitivity of these laws to misspecification of state costs, and may be of independent interest.

4. In Yu et al. (2009), the policy computation at the beginning of each phase requires solving a linear program and then adding a carefully tuned random perturbation to the solution. As a result, the performance analysis in Yu et al. (2009) is rather lengthy and technical (in particular, it invokes several advanced results from perturbation theory for linear programs). By contrast, even though we are working with a continuous action space, all policy computations in our case reduce to solving finite-dimensional eigenvalue problems, without any need for additional randomization. Moreover, even though the overall scheme of our analysis is similar to the one in Yu et al. (2009) (which, in turn, is inspired by existing work on lazy strategies (Vovk (1990); Kalai and Vempala (2005); Merhav et al. (2002))), the proof is self-contained and much less technical,

relying on our new results pertaining to Todorov-type optimal control laws.

5.1.2 Organization of the chapter

The remainder of the chapter is organized as follows. We close this section with a brief summary of frequently used notation. Section 5.2 contains precise formulation of the new online MDP problem and presents our main result, Theorem 21. In preparation for the proof of the theorem, Section 5.3 contains preliminaries on MDPs with KL control cost (Todorov (2007, 2008, 2009)), including a number of new results pertaining to optimal policies. Section 5.4 then describes our proposed strategy, whose performance is then analyzed in Section 5.5 in order to prove Theorem 21. Some simulation results are presented in Section 5.6. We close by summarizing our contributions and outlining some directions for future work.

5.1.3 Notation

The notation used in this chapter is slightly different from previous chapters due to the fact that we are working on a different action space. Here we introduce the notation to avoid the possible confusion. We will denote the underlying finite state space by X. A matrix $P = [P(x,y)]_{x,y\in\mathsf{X}}$ with nonnegative entries, and with the rows and the columns indexed by the elements of X, is called *stochastic* (or *Markov*) if its rows sum to one: $\sum_{y\in\mathsf{X}} P(x,y) = 1, \forall x \in \mathsf{X}$.

We will denote the set of all such stochastic matrices by $\mathcal{M}(\mathsf{X})$, the set of all probability distributions over X by $\mathcal{P}(\mathsf{X})$, the set of all functions $f : \mathsf{X} \to \mathbb{R}$ by $\mathcal{C}(\mathsf{X})$, and the cone of all nonnegative functions $f : \mathsf{X} \to \mathbb{R}_+$ by $\mathcal{C}_+(\mathsf{X})$. We will represent the elements of $\mathcal{P}(\mathsf{X})$ by row vectors and denote them by π, μ, ν, etc., and the elements of $\mathcal{C}(\mathsf{X})$ by column vectors and denote them by f, g, h, etc.

Any Markov matrix $P \in \mathcal{M}(\mathsf{X})$ acts on probability distributions from the right

and on functions from the left:

$$\mu P(y) = \sum_{x \in \mathsf{X}} \mu(x) P(x, y), \qquad Pf(x) = \sum_{y \in \mathsf{X}} P(x, y) f(y).$$

We will denote the set of all unchain Markov matrices over X by $\mathcal{M}_1(\mathsf{X})$. Given $\rho \in [0, 1]$, we say that P is *ρ-contractive* if

$$\|\mu P - \nu P\|_1 \leqslant \rho \|\mu - \nu\|_1, \qquad \forall \mu, \nu \in \mathcal{P}(\mathsf{X})$$

(in fact, every $P \in \mathcal{M}(\mathsf{X})$ is 1-contractive). We will denote the set of ρ-contractive Markov matrices by $\mathcal{M}_1^\rho(\mathsf{X})$. It is easy to show that, for every $0 \leqslant \rho < 1$, $\mathcal{M}_1^\rho(\mathsf{X}) \subset \mathcal{M}_1(\mathsf{X})$. The *Dobrushin ergodicity coefficient* (Seneta (2006); Cappé et al. (2005)) of $P \in \mathcal{M}(\mathsf{X})$ is given by

$$\alpha(P) \triangleq \frac{1}{2} \max_{x, x' \in \mathsf{X}} \|P(x, \cdot) - P(x', \cdot)\|_1,$$

and it can be shown that any $P \in \mathcal{M}(\mathsf{X})$ is $\alpha(P)$-contractive (Seneta (2006); Cappé et al. (2005)). Finally, for any $P, P' \in \mathcal{M}(\mathsf{X})$ we define the supremum distance

$$\|P - P'\|_\infty \triangleq \max_{x \in \mathsf{X}} \|P(x, \cdot) - P'(x, \cdot)\|_1.$$

5.2 Problem formulation and the main result

5.2.1 The model

Denote the set of all functions $f : \mathsf{X} \to \mathbb{R}$ by $\mathcal{C}(\mathsf{X})$, and the cone of all nonnegative functions $f : \mathsf{X} \to \mathbb{R}_+$ by $\mathcal{C}_+(\mathsf{X})$. Given the finite state space X, let $\mathcal{F}$ be a fixed subset of $\mathcal{C}_+(\mathsf{X})$, and let $x_1 \in \mathsf{X}$ be a fixed initial state. Consider an agent (A) performing a controlled random walk on X in response to a dynamic environment (E). The interaction between A and E proceeds as follows:

$X_1 = x_1$
for $t = 1, 2, \ldots$
  A selects $P_t \in \mathcal{M}(\mathsf{X})$ and draws $X_{t+1} \sim P_t(X_t, \cdot)$
  E selects $f_t \in \mathcal{F}$ and announces it to A
end for

At each $t \geqslant 1$, A selects the transition probabilities $P_t(x, y) = \Pr\{X_{t+1} = y | X_t = x\}$ based on his knowledge $f^{t-1} = (f_1, \ldots, f_{t-1})$, and incurs the *state cost* $f_t(X_t)$ and the *control cost* $D(P_t(X_t, \cdot) \| P^*(X_t, \cdot))$. The total cost incurred by the agent A at time t is given by

$$c_t(X_t, P_t) = f_t(X_t) + D(P_t(X_t, \cdot) \| P^*(X_t, \cdot)).$$

and the objective is to minimize a suitable notion of regret.

5.2.2 Strategies and regret

A *strategy* for the agent A is a collection of mappings $\gamma = \{\gamma_t\}_{t=1}^{\infty}$ where $\gamma_t : \mathcal{F}^{t-1} \to \mathcal{M}(\mathsf{X})$, so that $P_t = \gamma_t(f^{t-1})$. This means our strategy is based on the complete knowledge of all the past cost functions. The cumulative cost of γ after T steps is

$$C_T = \sum_{t=1}^{T} c_t(X_t, P_t) = \sum_{t=1}^{T} c_t(X_t, \gamma_t(f^{t-1})).$$

To define the regret after T steps, we will consider the gap between C_T and the expected cumulative cost that A could have achieved in hindsight by using a stationary unichain random walk on X (with full knowledge of f^T). This gap arises through the agent's lack of prior knowledge on the sequence of state cost functions. Formally, we define the regret of γ after T steps w.r.t. a particular $P \in \mathcal{M}_1(\mathsf{X})$ by[1]

$$R_T(P) \triangleq C_T - \mathbb{E}_{x_1}^{P}\left[\sum_{t=1}^{T} c_t(X_t, P)\right],$$

where the expectation is taken over the Markov chain induced the by *comparison* transition kernel P with initial state $X_1 = x_1$. As before, we assume throughout that the environment is *oblivious* (or *open-loop*), in the sense that the evolution of the sequence $\{f_t\}$ is not affected by the state sequences $\{X_t\}$. In our case, it implies

[1] To keep the notation clean, we will suppress the dependence of the cumulative cost C_T and the regret R_T on the strategy γ and on the state costs $f_1, \ldots, f_T$.

that, for a fixed sequence $f_1, f_2, \ldots$ of state costs chosen by E, the state process $\boldsymbol{X} = \{X_t\}_{t=1}^{\infty}$ induced by A's choices $P_1, P_2, \ldots$ is a (time-inhomogeneous) Markov chain. Now consider some set $\mathcal{N} \subset \mathcal{M}_1(\mathsf{X})$. Adopting standard terminology (Cesa-Bianchi and Lugosi (2006)), we will say that γ is *Hannan-consistent* w.r.t. $\mathcal{N}$ if

$$\limsup_{T\to\infty} \sup_{P\in\mathcal{N}} \sup_{f_1,\ldots,f_T\in\mathcal{F}} \frac{\mathbb{E}R_T(P)}{T} \leqslant 0, \tag{5.2}$$

where the expectation is w.r.t. the law of the process $\boldsymbol{X}$ starting at $X_1 = x_1$. In other words, a strategy is Hannan-consistent if its worst case (over $\mathcal{F}$) expected per-round regret converges to zero uniformly over $\mathcal{N}$. We can alternatively interpret the Hannan consistency condition (5.2) as follows: as the horizon T increases, the smallest average cost achievable by a strategy which is Hannan-consistent w.r.t. $\mathcal{N}$ will converge to the smallest long-term average cost achievable by any stationary Markov strategy in $\mathcal{N}$ on an MDP with the state cost given by the *empirical average* $(1/T)\sum_{t=1}^{T} f_t$ of the state costs revealed up to time T.

5.2.3 The main result

Our main result (Theorem 21 below) guarantees the existence of a Hannan-consistent strategy against any uniformly ergodic collection of stationary unichain policies under the following two assumptions on the passive dynamics P^*:

Assumption 5 (Irreducibility and aperiodicity)**.** *The passive dynamics P^* is* irreducible *and* aperiodic*, where the former means that, for every $x, y \in \mathsf{X}$, there exists some $n \in \mathbb{N}$ such that $(P^*)^n(x, y) > 0$, while the latter means that, for every $x \in \mathsf{X}$, the greatest common divisor of the set $\{n \in \mathbb{N} : (P^*)^n(x, x) > 0\}$ is equal to 1.*

Assumption 6 (Ergodicity)**.** *The Dobrushin ergodicity coefficient $\alpha(P^*)$ is strictly less than 1.*

Assumption 5 ensures that P^* has a unique everywhere positive invariant distribution π^* (Seneta (2006)) and, for a finite X, it is equivalent to the existence of some $\bar{n} \in \mathbb{N}$,

such that

$$\theta \triangleq \min_{x,y \in \mathsf{X}} (P^*)^{\bar{n}}(x,y) > 0$$

(see, e.g., Theorem 1.4 in Seneta (2006)). Assumption 6, which is frequently used in the study of MDPs with average cost criterion (Hernández-Lerma (1989); Di Masi and Stettner (1999); Arapostathis et al. (1993)), guarantees that the convergence to π^* is exponentially fast (so that P^* is geometrically ergodic), and it also imposes a stronger type of ergodicity, since a Markov matrix $P \in \mathcal{M}(\mathsf{X})$ has $\alpha(P) < 1$ if and only if for any pair $x, x' \in \mathsf{X}$ there exists at least one $y \in \mathsf{X}$, such that y can be reached from both x and x' in one step with strictly positive probability. For example, if P^* satisfies the *Doeblin minorization condition* (Cappé et al. (2005); Meyn and Tweedie (2009)), i.e., if there exist some $\delta \in (0,1]$ and some $\mu \in \mathcal{P}(\mathsf{X})$, such that $P^*(x,y) \geqslant \delta\mu(y)$ for all $x, y \in \mathsf{X}$, then we will have $\alpha(P^*) \leqslant 1 - \delta < 1$ (see, e.g., Lemma 4.3.13 in Cappé et al. (2005)).

Remark 8. As we pointed out in Section 5.1.1, our assumption is actually much milder than the assumptions made in related literature (Even-Dar et al. (2009); Yu et al. (2009); Neu et al. (2014)). Recent work by Neu et al. (2014) shows that the assumptions made in Even-Dar et al. (2009); Yu et al. (2009); Neu et al. (2014) are valid only if the Dobrushin ergodicity coefficient of the state transition kernel induced by every policy is strictly smaller than one. By contrast, we only assume this for the passive dynamics P^*.

With these assumptions in place, we are now ready to state our main result:

Theorem 21. *Let $\mathcal{F}$ consist of all $f \in \mathcal{C}_+(\mathsf{X})$ with $\|f\|_\infty \leqslant 1$. Fix an arbitrary $\epsilon \in (0, 1/3)$. Under Assumptions 1–6, there exists a strategy γ, such that for any $\rho \in [0,1)$,*

$$\sup_{P \in \mathcal{M}_1^\rho(\mathsf{X})} \sup_{f_1, \ldots, f_T \in \mathcal{F}} \frac{\mathbb{E} R_T(P)}{T} = O(T^{-1/4+\epsilon}). \tag{5.3}$$

As a consequence, the strategy γ is Hannan-consistent w.r.t. $\mathcal{M}_1^\rho(\mathsf{X})$.

Remark 9. The constant hidden in the $O(\cdot)$ notation depends only on the passive dynamics P^* and on the contraction rate ρ of the baseline policies in $\mathcal{M}_1^\rho(\mathsf{X})$; cf. Eq. (5.44), and the discussion preceding it, for details.

5.3 Preliminaries

Our construction of a Hannan-consistent strategy in Theorem 21 relies on Todorov's theory of MDPs with KL control cost (Todorov (2007, 2008, 2009)). In this section, we give an overview of this theory and present several new results that will be used later on.

First, let us recall the general set-up for MDPs with finite state space X and compact action space U under the average cost criterion (see e.g., Hernández-Lerma and Lasserre (1996) or Arapostathis et al. (1993)). It involves a family of Markov matrices $P_u \in \mathcal{M}(\mathsf{X})$ indexed by actions $u \in \mathsf{U}$. The (long-term) *average cost* of a stationary Markov policy (state feedback law) $w : \mathsf{X} \to \mathsf{U}$ with initial state $X_1 = x_1$ is given by

$$J(w, x_1) \triangleq \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{x_1}^w \left[\sum_{t=1}^{T} c(X_t, w(X_t)) \right], \tag{5.4}$$

where the expectation $\mathbb{E}_{x_1}^w[\cdot]$ is w.r.t. the law of the Markov chain $\boldsymbol{X} = \{X_t\}$ with controlled transition probabilities

$$\Pr\{X_{t+1} = y | X_t = x\} = P_{w(x)}(x, y), \qquad X_1 = x_1$$

and $c : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ is the one step state-action cost. The construction of an optimal policy to minimize (5.4) for every x_1 revolves around the ACOE

$$h(x) + \lambda = \min_{u \in \mathsf{U}(x)} \{c(x, u) + P_u h(x)\}, \ x \in \mathsf{X} \tag{5.5}$$

where $\mathsf{U}(x) \subseteq \mathsf{U}$ is the set of allowable actions in state x. If a solution pair $(\lambda, h) \in \mathbb{R} \times \mathcal{C}(\mathsf{X})$ exists with $\|h\|_s < +\infty$, then it can be shown (Hernández-Lerma and Lasserre (1996); Arapostathis et al. (1993)) that the stationary policy

$$w_*(x) = \arg\min_{u \in \mathsf{U}(x)} \{c(x,u) + P_u h(x)\}$$

is optimal, and has average cost λ for every x. The function h is called the *relative value function.*

5.3.1 Linearly solvable MDPs

In a series of papers (Todorov (2007, 2008, 2009)), Todorov has introduced a class of Markov decision processes, for which solving the ACOE reduces to solving an eigenvalue problem. In this set-up, which we have described informally before, the action space U is the probability simplex $\mathcal{P}(\mathsf{X})$, which is compact in the Euclidean topology, and for each $u \in \mathcal{P}(\mathsf{X})$ we have $P_u(x,y) \triangleq u(y), \forall (x,y) \in \mathsf{X} \times \mathsf{X}$. Thus, any state feedback law (Markov policy) $w : \mathsf{X} \to \mathcal{P}(\mathsf{X})$ induces the state transitions directly via

$$\Pr\{X_{t+1} = y | X_t = x\} = P_{w(x)}(x,y) = [w(x)](y), \quad t \geqslant 1.$$

In other words, if $X_t = x$, then $u(\cdot) = w(x)$ is the probability distribution of the next state X_{t+1}. Hence, there is a one-to-one correspondence between Markov policies w and Markov matrices $P \in \mathcal{M}(\mathsf{X})$, given by $w(x) = P(x, \cdot)$.

To specify an MDP, we fix a state cost function $f \in \mathcal{C}_+(\mathsf{X})$ and a Markov matrix P^* as the passive dynamics, which specifies the state transition probabilities in the absence of control. The one-step state-action cost function $c(x,u)$ is given by (5.1). If we use the shorthand $c(x,P)$ for $c(x, P(x,\cdot))$, then the average cost of a policy $P \in \mathcal{M}(\mathsf{X})$ starting at $X_1 = x_1$ can be written as

$$J(P, x_1) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^P_{x_1} \left[\sum_{t=1}^{T} c(X_t, P) \right].$$

Intuitively, if P has a small average cost, then the induced Markov chain $\boldsymbol{X} = \{X_t\}$ has a small average state cost, and its one-step transitions stay close to those prescribed by P^*.

The ACOE for this problem takes the form

$$h(x) + \lambda = f(x) + \min_{u \in \mathcal{P}(\mathsf{X})} \{D(u \| P^*(x, \cdot)) + \mathbb{E}_u h\}. \tag{5.6}$$

For a given $h \in \mathcal{C}(\mathsf{X})$, the minimization of the right-hand side of (5.6) can be done in closed form. To see this, let us define, for every $\varphi \in \mathcal{C}(\mathsf{X})$, the *twisted kernel* (Balaji and Meyn (2000))

$$\check{P}_\varphi(x, \cdot) \triangleq \frac{P^*(x, \cdot) e^{-\varphi(\cdot)}}{P^* e^{-\varphi}(x)}, \qquad x \in \mathsf{X} \tag{5.7}$$

which is obviously an element of $\mathcal{M}(\mathsf{X})$. Then we have

$$\begin{aligned}
&\min_{u \in \mathcal{P}(\mathsf{X})} \{D(u \| P^*(x, \cdot)) + \mathbb{E}_u h\} \\
&= \min_{u \in \mathcal{P}(\mathsf{X})} \left\{ \mathbb{E}_u \left[\log \frac{u(Y)}{P^*(x, Y)} + h(Y) \right] \right\} \\
&= \min_{u \in \mathcal{P}(\mathsf{X})} \left\{ \mathbb{E}_u \left[\log \frac{u(Y)}{\check{P}_h(x, Y)} \right] - \log P^* e^{-h}(x) \right\}
\end{aligned} \tag{5.8}$$

If we further define $\Lambda_h(x) \triangleq P^* e^{-h}(x)$, then the quantity in braces in (5.8) can be written as $D(u \| \check{P}_h(x, \cdot)) - \log \Lambda_h(x)$. Using the fact that the divergence $D(\mu \| \nu)$ between any two $\mu, \nu \in \mathcal{P}(\mathsf{X})$ is nonnegative and equal to zero if and only if $\mu = \nu$ (Cover and Thomas (2006)), we see that the minimum value in (5.8) is uniquely achieved by $u_*(x) = \check{P}_h(x, \cdot)$ and is equal to $-\log \Lambda_h(x)$. Thus, we can rewrite the ACOE (5.6) as

$$h(x) + \lambda = f(x) - \log \Lambda_h(x), \qquad \forall x \in \mathsf{X}. \tag{5.9}$$

If we now consider the *exponentiated* relative value function $V \triangleq e^{-h}$, then (5.9) can be also written as $e^{-f} P^* V(x) = e^{-\lambda} V(x)$. Expressing this in vector form, we obtain

the so-called *multiplicative Poisson equation* (MPE) (Balaji and Meyn (2000)):

$$e^{-f} P^* V = e^{-\lambda} V \tag{5.10}$$

To construct the optimal policy for our MDP, we first solve the MPE (5.10) for λ and V, obtain h, and then compute the twisted kernel $\check{P}_h(x, \cdot)$ for every $x \in \mathsf{X}$. The MPE is an instance of a so-called *Frobenius–Perron eigenvalue* (FPE) problem (Seneta (2006)); there exist efficient methods for solving such problems, e.g., a recent algorithm due to Chanchana (2007). We also should point out that, for each $x \in \mathsf{X}$, the twisted kernel (5.7) is a *Boltzmann–Gibbs distribution* on the state space X with energy function h and base measure $P^*(x, \cdot)$. Boltzmann–Gibbs distributions arise in various contexts, e.g., in statistical physics and in the theory of large deviations (Ellis (1985); Streater (2009)), as solutions of variational problems over the space of probability measures that involve minimization of a Gibbs-type free energy functional, consisting of an affine "energy" term and a convex "entropy" term (given by the divergence relative to the base measure). Indeed, the functional being minimized on the right-hand side of (5.6) is precisely of this form.

In the sequel, we will often need to consider simultaneously several MDPs with different state costs f. Thus, whenever need arises, we will indicate the dependence on f using appropriate subscripts, as in c_f, λ_f, h_f, V_f, etc. For instance, the MPE (5.10) for a given state cost f is

$$P_f^* V_f = e^{-\lambda_f} V_f, \tag{5.11}$$

where $P_f^* \triangleq e^{-f} P^*$, i.e., $P_f^*(x, y) = e^{-f(x)} P^*(x, y)$ for all $x, y \in \mathsf{X}$.

5.3.2 Some properties of Todorov's optimal policy

We now investigate the properties of Todorov's optimal policy under the assumptions on the passive dynamics P^* that are listed in Section 5.2.3. Most of the results of this section are new (with some exceptions, which we point out explicitly).

We start with the following basic existence and uniqueness result, which is implicit in Todorov (2007):

Proposition 22. *Under Assumption 5, for any state cost $f \in \mathcal{C}_+(\mathsf{X})$ the MPE* (5.11) *has a strictly positive solution $V_f \in \mathcal{C}_+(\mathsf{X})$ with the associated strictly positive eigenvalue $e^{-\lambda_f}$, and the only nonnegative solutions of* (5.11) *are positive multiples of V_f. Moreover, the corresponding twisted kernel $\check{P}_{h_f}$ is also irreducible and aperiodic, and has a unique invariant distribution $\check{\pi}_f = \check{\pi}_f \check{P}_f \in \mathcal{P}(\mathsf{X})$.*

Proof. Consider the matrix $P_f^* \triangleq e^{-f} P^*$ with entries $P_f^*(x,y) = e^{-f(x)} P^*(x,y)$. For any $n \in \mathbb{N}$ and any $x, y \in \mathsf{X}$, $(P_f^*)^n(x,y) \geqslant e^{-n\|f\|_\infty} (P^*)^n(x,y)$. Since P^* is irreducible (Assumption 5), for any pair $x, y \in \mathsf{X}$ of states there exists some $n \in \mathbb{N}$, such that $(P^*)^n(x,y) > 0$. But then $(P_f^*)^n(x,y) > 0$ as well, which means that P_f^* is also irreducible. Therefore, by the Frobenius–Perron theorem (Seneta (2006)), P_f^* has a strictly positive right eigenvector V_f with a positive eigenvalue r (the Frobenius-Perron eigenvalue): $P_f^* V_f = r V_f$. Thus, $e^{-\lambda_f} = r$. Moreover, the FP eigenvalue is simple, and P_f^* has no nonnegative right eigenvectors other than the positive multiples of V_f (Seneta (2006)). This proves the existence and uniqueness part.

Now, using the fact that $V_f = e^{-h_f}$ solves the MPE (5.11), we can show that

$$\check{P}_{h_f}(x,y) = e^{\lambda_f} \frac{V_f(y)}{V_f(x)} P_f^*(x,y),$$

whence it follows that

$$(\check{P}_{h_f})^n(x,y) = e^{n\lambda_f^*} \frac{V_f(y)}{V_f(x)} (P_f^*)^n(x,y),$$

As was just proved, P_f^* is irreducible, and V_f is strictly positive. Hence, for any pair $(x,y) \in \mathsf{X} \times \mathsf{X}$ there exists some $n \in \mathbb{N}$, such that $(\check{P}_{h_f})^n(x,y) > 0$ as well. This proves the irreducibility of $\check{P}_{h_f}$. Now, since P_f^* is irreducible, the Frobenius–Perron theorem

says that there exists a unique strictly positive $\mu \in \mathcal{P}(\mathsf{X})$, such that $\mu P_f^* = e^{-\lambda_f}\mu$ (Seneta (2006)). Now define $\check{\pi}_f \in \mathcal{P}(\mathsf{X})$ through

$$\check{\pi}_f(x) \triangleq \frac{\mu(x)V_f(x)}{\sum_{y\in\mathsf{X}}\mu(y)V_f(y)} \equiv \frac{\mu(x)V_f(x)}{\mathbb{E}_\mu V_f}, \qquad x \in \mathsf{X}.$$

A straightforward calculation shows that $\check{\pi}_f$ is an invariant distribution of $\check{P}_{h_f}$:

$$\begin{aligned}
\check{\pi}_f\check{P}_{h_f}(y) &= \sum_{x\in\mathsf{X}} \check{\pi}_f(x)\check{P}_{h_f}(x,y) \\
&= e^{\lambda_f}V_f(y)\sum_{x\in\mathsf{X}} \frac{\check{\pi}_f(x)P_f^*(x,y)}{V_f(x)} \\
&= \frac{e^{\lambda_f}V_f(y)}{\mathbb{E}_\mu V_f}\sum_{x\in\mathsf{X}} \mu(x)P_f^*(x,y) \\
&= \frac{e^{\lambda_f}V_f(y)}{\mathbb{E}_\mu V_f}\mu P_f^*(y) \\
&= \frac{e^{\lambda_f}V_f(y)}{\mathbb{E}_\mu V_f}e^{-\lambda_f}\mu(y) \\
&= \frac{\mu(y)V_f(y)}{\mathbb{E}_\mu V_f} \\
&= \check{\pi}_f(y).
\end{aligned}$$

The uniqueness of $\check{\pi}_f$ follows from the irreducibility of $\check{P}_{h_f}$. □

Since $V_f = e^{-h_f}$, the fact that any positive multiple of V_f is a solution of the MPE is equivalent to the well-known fact that the relative value function h_f as a solution of the ACOE (5.6) is unique up to additive constants. That is, if a particular h_f solves (5.6), then so does any $h_f + c$ for any additive constant $c \in \mathbb{R}$. For this reason, we can fix an arbitrary $x^\circ \in \mathsf{X}$ and assume that $h_f(x^\circ) = 0$ for any f. This ensures that the mapping

$$f \longmapsto h_f, \qquad h_f(x^\circ) = 0 \tag{5.12}$$

is well-defined. The following results are new:

Proposition 23. *Under Assumption 5, the mapping* (5.12) *is bounded on compact subsets of the cone* $\mathcal{C}_+(\mathsf{X})$*: for any* $f \in \mathcal{C}_+(\mathsf{X})$*,*

$$\|h_f\|_s \leqslant \log \theta^{-1} + \bar{n}\|f\|_\infty, \tag{5.13}$$

where $\bar{n}$ *and* θ *are defined in Section 5.2.3. Hence,*

$$\sup_{f \in \mathcal{C}_+(\mathsf{X});\, \|f\|_\infty \leqslant C} \|h_f\|_s \leqslant \log \theta^{-1} + \bar{n}C. \tag{5.14}$$

Proof. We essentially follow the proof of Theorem 3.2 in Fleming and Hernández-Hernández (1997), with some simplifications. For each $T \in \mathbb{N}$, define the function $W_T : \mathsf{X} \to \mathbb{R}$ via

$$e^{-W_T(x)} \triangleq \mathbb{E}_x \left[\exp\left(-\sum_{t=1}^{T} f(X_t) - h_f(X_{T+1}) \right) \right],$$

where $\mathbb{E}_x[\cdot]$ denotes the expectation w.r.t. the Markov chain $\boldsymbol{X} = (X_1, X_2, \ldots)$ with initial state $X_1 = x$ and transition matrix P^*. Then a simple inductive argument shows that

$$e^{-W_T(x)} = e^{-T\lambda_f - h_f(x)}. \tag{5.15}$$

Indeed, for each t let $\Psi_t \triangleq \prod_{s=1}^{t} \frac{V_f(X_s)}{P^* V_f(X_s)}$. Then, since $e^{-f(x)} = \frac{e^{-\lambda_f} V_f(x)}{P^* V_f(x)}$ by (5.11), we can write

$$\begin{aligned} e^{-W_T(x)} &= e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_T V_f(X_{T+1}) \right] && (5.16)\\ &= e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_T \mathbb{E}[V_f(X_{T+1}) | X_T] \right] && (5.17)\\ &= e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_T P^* V_f(X_T) \right] && (5.18)\\ &= e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_{T-1} V_f(X_T) \right], && (5.19) \end{aligned}$$

where (5.16) follows from definitions, (5.17) and (5.18) use the Markov property, and (5.19) again follows from definitions. Proceeding backwards, we get

$$\begin{aligned} \mathbb{E}_x[\Psi_T V_f(X_{T+1})] &= \mathbb{E}_x[\Psi_1 V_f(X_2)] = \mathbb{E}_x[\Psi_1 P^* V_f(X_1)] \\ &= V_f(x) = e^{-h_f(x)}. \end{aligned}$$

Substituting this into (5.16), we get (5.15), which in turn implies that $h_f(x) = W_T(x) - T\lambda_f$ for all $x \in \mathsf{X}, T \in \mathbb{N}$. Since $h_f(x^\circ) = 0$, we can write $h_f(x) = W_T(x) - W_T(x^\circ), \forall x \in \mathsf{X}, T \in \mathbb{N}$. Let ν (respectively, ν°) be the distribution of $X_{\bar{n}+1}$ in the Markov chain with transition matrix P^* and initial state $X_1 = x$ (respectively, $X_1 = x^\circ$). Then

$$\frac{\nu(y)}{\nu^\circ(y)} = \frac{(P^*)^{\bar{n}}(x,y)}{\nu^\circ(y)} \geqslant \theta > 0 \tag{5.20}$$

for every $y \in \mathsf{X}$. Consequently, for any $T > \bar{n}$ we have

$$\begin{aligned}
& e^{-W_T(x)} \\
& = \mathbb{E}_x \left[e^{-\sum_{t=1}^{\bar{n}} f(X_t)} e^{-\sum_{t=\bar{n}+1}^{T} f(X_t) - h_f(X_{T+1})} \right] && (5.21) \\
& \geqslant e^{-\bar{n}\|f\|_\infty} \mathbb{E}_x \left[e^{-\sum_{t=\bar{n}+1}^{T} f(X_t) - h_f(X_{T+1})} \right] && (5.22) \\
& = e^{-\bar{n}\|f\|_\infty} \mathbb{E}_{x^\circ} \left[e^{-\sum_{t=\bar{n}+1}^{T} f(X_t) - h_f(X_{T+1})} \frac{\nu(X_{\bar{n}+1})}{\nu^\circ(X_{\bar{n}+1})} \right] && (5.23) \\
& \geqslant \theta e^{-\bar{n}\|f\|_\infty} \mathbb{E}_{x^\circ} \left[e^{-\sum_{t=1}^{T} f(X_t) - h_f(X_{T+1})} \right] && (5.24) \\
& = \theta e^{-\bar{n}\|f\|_\infty} e^{-W_T(x^\circ)}, && (5.25)
\end{aligned}$$

where (5.21) is by definition, (5.23) follows from the Markov property and a change of measure, (5.24) follows from (5.20) and from the fact that $f \geqslant 0$, and (5.25) is again by definition. Taking logarithms, we get $W_T(x) - W_T(x^\circ) \leqslant \log \theta^{-1} + \bar{n}\|f\|_\infty, \forall T > \bar{n}$. Interchanging the roles of x and x°, we get $|h_f(x)| \leqslant \log \theta^{-1} + \bar{n}\|f\|_\infty$. This proves (5.13); (5.14) follows immediately. □

Moreover, the dependence of the relative value function h_f on the state cost f is *continuous*:

Proposition 24. *Under Assumptions 5 and 6, the mapping* (5.12) *is Lipschitz-continuous on compact subsets of* $\mathcal{C}_+(\mathsf{X})$*: for every* $C > 0$ *there exists a constant*

$K = K(C) > 0$, *such that for any two* $f, g \in \mathcal{C}_+(\mathsf{X})$ *with* $\|f\|_\infty, \|g\|_\infty \leqslant C$ *we have*

$$\|h_f - h_g\|_s \leqslant K\|f - g\|_\infty. \tag{5.26}$$

Proof. The basic idea is as follows. For a given $f \in \mathcal{C}_+(\mathsf{X})$, let us introduce the dynamic programming operator $\mathbb{T}_f$ that maps any $\varphi \in \mathcal{C}(\mathsf{X})$ to $\mathbb{T}_f\varphi \in \mathcal{C}(\mathsf{X})$, where $\forall \varphi \in \mathcal{C}(\mathsf{X}), x \in \mathsf{X}$,

$$\mathbb{T}_f\varphi(x) \triangleq f(x) + \inf_{\mu \in \mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu \varphi + D(\mu \| P^*(x, \cdot))\}.$$

Then we can express the ACOE (5.6) as $h_f + \lambda_f = \mathbb{T}_f h_f$. Hence, for any $f, g \in \mathcal{C}_+(\mathsf{X})$,

$$\begin{aligned} \|h_f - h_g\|_s &= \|(\mathbb{T}_f h_f - \lambda_f) - (\mathbb{T}_g h_g - \lambda_g)\|_s \\ &= \|\mathbb{T}_f h_f - \mathbb{T}_g h_g\|_s \qquad (5.27) \\ &\leqslant \|\mathbb{T}_f h_f - \mathbb{T}_g h_f\|_s + \|\mathbb{T}_g h_f - \mathbb{T}_g h_g\|_s, \qquad (5.28) \end{aligned}$$

where (5.27) uses the fact that the span seminorm is unchanged after adding a constant, and (5.28) is by the triangle inequality. We will then show the following:

1. For any $\varphi \in \mathcal{C}(\mathsf{X})$ and any $f, g \in \mathcal{C}_+(\mathsf{X})$,
$$\|\mathbb{T}_f\varphi - \mathbb{T}_g\varphi\|_s \leqslant 2\|f - g\|_\infty. \tag{5.29}$$

2. For a fixed $f \in \mathcal{C}_+(\mathsf{X})$, the dynamic programming operator $\mathbb{T}_f : \mathcal{C}(\mathsf{X}) \to \mathcal{C}(\mathsf{X})$ is a contraction in the span seminorm: for every $M > 0$, there exists a constant $K' = K'(M) \in (0, 1)$, such that for any two $\varphi, \varphi' \in \mathcal{C}(\mathsf{X})$ with $\|\varphi\|_s, \|\varphi'\|_s \leqslant M$ we have
$$\|\mathbb{T}_f\varphi - \mathbb{T}_f\varphi'\|_s \leqslant K'\|\varphi - \varphi'\|_s. \tag{5.30}$$

Assuming that items 1) and 2) above are proved, we proceed as follows. First of all, the first term in (5.28) is bounded by $2\|f - g\|_\infty$ by (5.29). Next, since $\|f\|_\infty, \|g\|_\infty \leqslant C$, Proposition 23 guarantees that there exists some $M = M(C) < \infty$, such that $\|h_f\|_s, \|h_g\|_s \leqslant M$. Therefore, there exists a constant $K' = K'(M) < 1$, such that

the second term in (5.28) is bounded by $K'\|h_f - h_g\|_s$. Therefore, $\|h_f - h_g\|_s \leqslant \frac{2}{1-K'}\|f-g\|_\infty$, which gives (5.26) with $K = 2/(1-K')$.

We now prove 1) and 2). For any function $\varphi \in \mathcal{C}(\mathsf{X})$ and any two $f, g \in \mathcal{C}_+(\mathsf{X})$, we have

$$\begin{aligned}
&\max_{x\in\mathsf{X}} \{\mathbb{T}_f\varphi(x) - \mathbb{T}_g\varphi(x)\} \\
&= \max_{x\in\mathsf{X}} \left\{ \left[f(x) + \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x,\cdot))\} \right] \right. \\
&\qquad \left. - \left[g(x) + \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x,\cdot))\} \right] \right\} \\
&= \max_{x\in\mathsf{X}} [f(x) - g(x)] .
\end{aligned}$$

Similarly, we get $\min_{x\in\mathsf{X}} \{\mathbb{T}_f\varphi(x) - \mathbb{T}_g\varphi(x)\} = \min_{x\in\mathsf{X}} [f(x) - g(x)]$. Thus, $\|\mathbb{T}_f\varphi - \mathbb{T}_g\varphi\|_s = \|f-g\|_s \leqslant 2\|f-g\|_\infty$, so we have proved (5.29).

To establish (5.30), we follow the proof of Proposition 2.2 in Di Masi and Stettner (1999) with some simplifications. Pick any $x, x' \in \mathsf{X}$ and let

$$\begin{aligned}
\nu &= \arg\min_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi' + D(\mu\|P^*(x,\cdot))\} , \\
\nu' &= \arg\min_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x',\cdot))\} ,
\end{aligned}$$

where explicitly $\nu(\cdot) = \check{P}_{\varphi'}(x,\cdot)$ and $\nu'(\cdot) = \check{P}_{\varphi}(x',\cdot)$. Then

$$
\begin{aligned}
&[\mathbb{T}_f\varphi(x) - \mathbb{T}_f\varphi'(x)] - [\mathbb{T}_f\varphi(x') - \mathbb{T}_f\varphi'(x')] \\
&= \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x,\cdot))\} \\
&\qquad - \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi' + D(\mu\|P^*(x,\cdot))\} \\
&\qquad - \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x',\cdot))\} \\
&\qquad + \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi' + D(\mu\|P^*(x',\cdot))\} \\
&\leqslant \mathbb{E}_\nu\varphi + D(\nu\|P^*(x,\cdot)) - \mathbb{E}_\nu\varphi' - D(\nu\|P^*(x,\cdot)) \\
&\qquad - \mathbb{E}_{\nu'}\varphi - D(\nu'\|P^*(x',\cdot)) + \mathbb{E}_{\nu'}\varphi' + D(\nu'\|P^*(x',\cdot)) \\
&= \int(\varphi-\varphi')\mathrm{d}(\nu-\nu').
\end{aligned}
$$

A standard argument shows that that $\int(\varphi-\varphi')\mathrm{d}(\nu-\nu') \leqslant \frac{1}{2}\|\varphi-\varphi'\|_s\|\nu-\nu'\|_1$. Consequently,

$$
\begin{aligned}
&\left\|\mathbb{T}_f\varphi - \mathbb{T}_f\varphi'\right\|_s \\
&\leqslant \frac{1}{2}\left\|\varphi-\varphi'\right\|_s \cdot \max_{x,x'\in\mathsf{X}} \left\|\check{P}_\varphi(x,\cdot) - \check{P}_{\varphi'}(x',\cdot)\right\|_1 .
\end{aligned}
$$

Then the proof of (5.30) will be complete if we can show that

$$
\begin{aligned}
K'(M) &\triangleq \frac{1}{2} \sup_{\varphi,\varphi';\|\varphi\|_s,\|\varphi'\|_s\leqslant M} \max_{x,x'\in\mathsf{X}} \left\|\check{P}_\varphi(x,\cdot) - \check{P}_{\varphi'}(x',\cdot)\right\|_1 \\
&< 1. \qquad (5.31)
\end{aligned}
$$

Suppose that (5.31) does not hold. Then there exist sequences $\{\varphi_n\}$, $\{\varphi'_n\}$ of functions with $\|\varphi_n\|_s, \|\varphi'_n\|_s \leqslant M, \forall n$, a set $B \subset \mathsf{X}$, and a pair of points $x, x' \in \mathsf{X}$, such that

$$
\lim_{n\to\infty} \left[\check{P}_{\varphi_n}(x,B) - \check{P}_{\varphi'_n}(x',B)\right] = 1,
$$

where for any $P \in \mathcal{M}(\mathsf{X})$ we denote $P(x,B) \triangleq \sum_{y\in B} P(x,y)$. This implies in turn

that

$$\lim_{n\to\infty} \check{P}_{\varphi_n}(x, \mathsf{X}\backslash B) = \lim_{n\to\infty} \check{P}_{\varphi'_n}(x', B) = 0. \tag{5.32}$$

Since $\check{P}_\varphi(x, B) \geqslant e^{-\|\varphi\|_s} P^*(x, B)$, (5.32) implies that $P^*(x, \mathsf{X}\backslash B) = P^*(x', B) = 0$. But this means that $P^*(x, B) - P^*(x', B) = 1$, which contradicts Assumption 6. Hence, (5.31) holds. □

More generally, the twisted kernel $\check{P}_\varphi$ depends smoothly on the "twisting function" φ:

Proposition 25. *Fix any two functions $\varphi, \varphi' \in \mathcal{C}(\mathsf{X})$. Then the twisted kernels* (5.7) *have the following properties: for any $x \in \mathsf{X}$,*

$$D(\check{P}_\varphi(x, \cdot) \| \check{P}_{\varphi'}(x, \cdot)) \leqslant \frac{1}{8} \|\varphi - \varphi'\|_s^2 \tag{5.33}$$

$$\|\check{P}_\varphi(x, \cdot) - \check{P}_{\varphi'}(x, \cdot)\|_1 \leqslant \frac{1}{2} \|\varphi - \varphi'\|_s. \tag{5.34}$$

Moreover, if Assumptions 5 and 6 hold, then there exists a mapping $\kappa : \mathbb{R}_+ \to [0, 1)$, such that

$$\|\varphi\|_s \leqslant C \quad \Longrightarrow \quad \alpha(\check{P}_\varphi) \leqslant \kappa(C). \tag{5.35}$$

Proof. We begin with (5.33). From definitions, we have

$$\begin{aligned} &D(\check{P}_\varphi(x, \cdot) \| \check{P}_{\varphi'}(x, \cdot)) \\ &\quad = \mathbb{E}_{\check{P}_\varphi(x,\cdot)}[\varphi'(Y) - \varphi(Y)] + \log \frac{\Lambda_{\varphi'}(x)}{\Lambda_\varphi(x)}. \end{aligned} \tag{5.36}$$

A simple change-of-measure calculation shows that

$$\begin{aligned} \frac{\Lambda_{\varphi'}(x)}{\Lambda_\varphi(x)} &= \frac{\sum_y P^*(x, y) e^{-\varphi'(y)}}{\Lambda_\varphi(x)} \\ &= \frac{\sum_y e^{\varphi(y) - \varphi'(y)} P^*(x, y) e^{-\varphi(y)}}{\Lambda_\varphi(x)} \\ &= \mathbb{E}_{\check{P}_\varphi(x,\cdot)}[e^{\varphi(Y) - \varphi'(Y)}]. \end{aligned} \tag{5.37}$$

To bound the right-hand side of (5.37), we recall the well-known *Hoeffding bound* (Hoeffding (1963)), which for our purposes can be stated as follows: For any $\mu \in \mathcal{P}(\mathsf{X})$ and any $\psi \in \mathcal{C}(\mathsf{X})$,

$$\log \mathbb{E}_\mu e^\psi \leqslant \mathbb{E}_\mu \psi + \frac{\|\psi\|_s^2}{8}.$$

Applying this bound gives

$$\log \frac{\Lambda_{\varphi'}(x)}{\Lambda_\varphi(x)} \leqslant \mathbb{E}_{\check{P}_\varphi(x,\cdot)}[\varphi(Y) - \varphi'(Y)] + \frac{\|\varphi - \varphi'\|_s^2}{8}.$$

Substituting this bound into (5.36), we see that the terms involving the expectation of the difference $\varphi - \varphi'$ cancel, and we are left with (5.33). To prove (5.34), we use Pinsker's inequality, $\|P_1 - P_2\|_1 \leqslant \sqrt{2D(P_1\|P_2)}$ (Cover and Thomas (2006)). To prove (5.35), we follow essentially the same strategy as in the proof of Proposition 24 to show that $\kappa(C) \triangleq \sup_{\varphi;\|\varphi\|_s \leqslant C} \alpha(\check{P}_\varphi) < 1$ for every $C > 0$.

□

We close with the following basic but important result on steady-state optimality:

Proposition 26. *For any $f \in \mathcal{C}_+(\mathsf{X})$ and any $P \in \mathcal{M}_1(\mathsf{X})$, define*

$$\bar{J}_f(P) \triangleq \mathbb{E}_{\pi_P}[c_f(X, P)] \equiv \mathbb{E}_{\pi_P}[J_f(P, X)].$$

Then

$$\bar{J}_f(\check{P}_{h_f}) = \inf_{P \in \mathcal{M}_1(\mathsf{X})} \bar{J}_f(P).$$

Proof. Fix some $P \in \mathcal{M}_1(\mathsf{X})$. If there exists some $x \in \mathsf{X}$ such that $\pi_P(x) > 0$ and $D(P(x,\cdot)\|P^*(x,\cdot)) = +\infty$, then Proposition 26 holds trivially. Thus, there is no loss

of generality if we assume that $D(P(x,\cdot)\|P^*(x,\cdot)) < +\infty, \forall x \in \mathsf{X}$. Then

$$\begin{aligned}
\bar{J}_f(P) &= \mathbb{E}_{\pi_P}[f(X) + D(P(X,\cdot)\|P^*(X,\cdot))] \\
&= \sum_{x\in\mathsf{X}} \pi_P(x) \Bigg[f(x) + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{P(x,y)}{\check{P}_{h_f}(x,y)} \\
&\qquad + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{\check{P}_{h_f}(x,y)}{P^*(x,y)} \Bigg] \\
&= \sum_{x\in\mathsf{X}} \pi_P(x) \Bigg[f(x) + \mathbb{E}_{\pi_P} D(P(X,\cdot)\|\check{P}_{h_f}(X,\cdot)) \\
&\qquad + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{e^{-h_f(y)}}{\Lambda_{h_f}(x)} \Bigg] \\
&\geqslant \sum_{x\in\mathsf{X}} \pi_P(x) \Bigg[f(x) + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{e^{-h_f(y)}}{\Lambda_{h_f}(x)} \Bigg] \\
&= \mathbb{E}_{\pi_P}[f(X) - Ph_f(X) - \log \Lambda_{h_f}(X)] \\
&= \mathbb{E}_{\pi_P}[f(X) - h_f(X) - \log \Lambda_{h_f}(X)],
\end{aligned}$$

where the inequality is due to the fact that the KL divergence is always nonnegative, and the last step is due to the fact that π_P is the invariant distribution of P. By the ACOE (5.9), we know that $f(x) - h_f(x) - \log \Lambda_{h_f}(x) = \lambda_f$ for every $x \in \mathsf{X}$. So we have $\bar{J}_f(P) \geqslant \lambda_f, \forall P \in \mathcal{M}_1(\mathsf{X})$. Note that if we take the expectation $\mathbb{E}_{\check{\pi}_f}[\cdot]$ of both sides of the ACOE (5.6), we get

$$\begin{aligned}
&\mathbb{E}_{\check{\pi}_f}[h_f(X) + \lambda_f] \\
&= \mathbb{E}_{\check{\pi}_f}[f(X) + D(\check{P}_{h_f}(X,\cdot)\|P^*(X,\cdot)) + \check{P}_{h_f} h_f(X)] \\
&= \mathbb{E}_{\check{\pi}_f}[f(X) + D(\check{P}_{h_f}(X,\cdot)\|P^*(X,\cdot))] + \mathbb{E}_{\check{\pi}_f}[h_f(X)],
\end{aligned}$$

where the last equality is due to the fact that $\check{\pi}_f$ is the invariant distribution of $\check{P}_{h_f}$. Therefore, $\mathbb{E}_{\check{\pi}_f}[f(X) + D(\check{P}_{h_f}(X,\cdot)\|P^*(X,\cdot))] = \bar{J}_f(\check{P}_{h_f}) = \lambda_f$. So we now have

$\bar{J}_f(P) \geqslant \bar{J}_f(\check{P}_{h_f})$ for any $P \in \mathcal{M}_1(\mathsf{X})$, which completes the proof of Proposition 26.

□

5.4 The proposed strategy

Our construction of a Hannan-consistent strategy is similar to the approach of Yu et al. (2009). The main idea behind it is as follows. We partition the set of time indices $1, 2, \ldots$ into nonoverlapping contiguous segments (phases) of increasing duration and, during each phase, use Todorov's optimal policy matched to the average of the state cost functions revealed during the preceding phases. As in Yu et al. (2009), the phases are sufficiently long to ensure convergence to the steady state within each phase, and yet are sufficiently short, so that the policies used during successive phases are reasonably close to one another.

The phases are indexed by $m \in \mathbb{N}$, where we denote the mth phase by $\mathcal{T}_m$ and its duration by τ_m. Given $\epsilon \in (0, 1/3)$, we let $\tau_m = \lceil m^{1/3-\epsilon} \rceil$. The parameter ϵ is needed to control the growth of the total length of each fixed number of phases relative to the length of the most recent phase (we comment upon this in more detail in the next section). We also define $\mathcal{T}_{1:m} \triangleq \mathcal{T}_1 \cup \ldots \cup \mathcal{T}_m$ (the union of phases 1 through m) and denote its duration by $\tau_{1:m}$. Given a sequence $\{f_t\}$ of state cost functions, we define for each m the average state costs

$$\widehat{f}^{(m)} \triangleq \frac{1}{\tau_m} \sum_{t \in \mathcal{T}_m} f_t, \qquad \widehat{f}^{(1:m)} \triangleq \frac{1}{\tau_{1:m}} \sum_{t \in \mathcal{T}_{1:m}} f_t$$

and let $\widehat{f}^{(0)} = \widehat{f}^{(1:0)} = 0$. Our strategy takes the following form:

```
for m = 1, 2, ...
  solve the MPE e^{-f̂^(1:m-1)} P* e^{-h^(m)} = e^{-λ^(m)} e^{-h^(m)}
  let P^(m) = P̌_{h^(m)}
  for t ∈ T_m
    draw X_{t+1} ~ P^(m)(X_t, ·)
  end for
end for
```

Since we use the same policy throughout each phase, the evolution of the state induced by the above algorithm is described by the following inhomogeneous Markov chain:

$$X_1 \xrightarrow{P^{(1)}} X_2 \xrightarrow{P^{(1)}} \dots \xrightarrow{P^{(1)}} X_{\tau_1} \xrightarrow{P^{(2)}} X_{\tau_1+1} \xrightarrow{P^{(2)}} \dots$$

The implementation of this strategy reduces to solving a finite-dimensional FPE problem (Seneta (2006)) at the beginning of each phase to obtain a Todorov-type relative value function. The corresponding twisted kernel then determines the stationary policy to be followed throughout that phase. An efficient method for solving FPE problems was developed by Chanchana (2007). This method makes use of the well-known Collatz formula for the FPE(Seneta (2006)) and Elsner's inverse iteration algorithm for computing the spectral radius of a nonnegative irreducible matrix (Elsner (1976)). It is an iterative algorithm, which at each iteration performs an LU factorization of an $|\mathsf{X}| \times |\mathsf{X}|$ matrix. The time complexity of each iteration is $O(|\mathsf{X}|^3)$. Chanchana's algorithm outperforms the three best known algorithms for solving FPE problems, which all rely on Elsner's inverse iteration and have quadratic convergence. Numerical experimental results can be found in (Chanchana, 2007, Section 3.5).

5.5 Proof of Theorem 21

5.5.1 The main idea

Following the general outline in Yu et al. (2009), the proof of Theorem 21 can be divided into four major steps. The first step is to show that there is no loss of generality in considering a different notion of regret, i.e., the *steady-state regret*, which is the difference between the cumulative cost of the proposed strategy and the steady-state cost of a fixed stationary policy. The second step is to bound the difference between the expected total cost of our strategy and the sum of expected steady-state costs within each phase. That is, for each m, the steady-state expectation of the cost

incurred in phase m is taken w.r.t. the unique invariant distribution of $P^{(m)}$. After this step, we may only concentrate on expectations over invariant state distributions, which renders the problem much easier. For the third step, we show that the sum of steady-state expected costs is not much worse than what we would get if, at the start of each phase m, we also knew all the state cost functions to be revealed during phase m, i.e., if we used the "clairvoyant" policy $P^{(m+1)}$ in phase m. In the fourth step, we consider the sum of expected costs in each phase that could be attained if we knew all the state cost functions in advance and used the optimal policy w.r.t. the average of all the state cost functions throughout all the phases. We show that this expected cost is actually greater than the sum of expected costs of each phase when we only know the state cost functions one phase ahead. We then assemble the bounds obtained in these four steps to obtain the final bound on the regret of our strategy.

5.5.2 Preliminary lemmas

Before proceeding to the proof of Theorem 21, we present two lemmas that will be used throughout. The proofs of the lemmas rely heavily on the results of Section 5.3.2.

Lemma 27 (Uniform bounds)**.** *There exists constants $K_0 \geqslant 0$, $K_1 \geqslant 0$ and $0 \leqslant \alpha < 1$, such that, for every $f \in \mathcal{F}$ and every $m \in \mathbb{N}$,*

$$\|c_f(\cdot, P^{(m)})\|_\infty \leqslant K_0, \quad \|h^{(m)}\|_s \leqslant K_1, \quad \alpha(P^{(m)}) \leqslant \alpha$$

Moreover, the bound $\|c_f(\cdot, P)\|_\infty \leqslant K_0$ holds for all $P \in \mathcal{M}_1(\mathsf{X})$, such that $D(P(x,\cdot)\|P^*(x,\cdot)) < \infty$ for all $x \in \mathsf{X}$.

Proof. For every state $x \in \mathsf{X}$, let G_x denote the set of states that can be reached from x in one step by the passive dynamics P^*, i.e., $G_x \triangleq \{y : P^*(x,y) > 0\}$. Let us also define $p^*_x = \min_{y \in G_x} P^*(x,y)$ and $p^* = \min_{x \in \mathsf{X}} p^*_x$. Since $P^{(m)} = \check{P}_{h^{(m)}}$, and $h^{(m)}$ is bounded

by Proposition 23, we have $\mathrm{supp}(P^{(m)}(x,\cdot)) \subseteq \mathrm{supp}(P^*(x,\cdot)) \equiv G_x$. Therefore,

$$D(P^{(m)}(x,\cdot)\|P^*(x,\cdot)) = \sum_{y\in G_x} P^{(m)}(x,y)\log\frac{P^{(m)}(x,y)}{P^*(x,y)}$$

$$\leqslant \log\frac{1}{p^*_x}, \qquad \forall x\in\mathsf{X}, m\in\mathbb{N}$$

and for any $f\in\mathcal{F}$

$$\|c_f(\cdot,P^{(m)})\|_\infty$$

$$\leqslant \|f\|_\infty + \max_{x\in\mathsf{X}} D(P^{(m)}(x,\cdot)\|P^*(x,\cdot)) \leqslant 1+\log\frac{1}{p^*}.$$

Thus, the first bound of Lemma 27 holds with $K_0 = 1+\log\frac{1}{p^*}$. The same argument works for any $P\in\mathcal{M}_1(\mathsf{X})$ that satisfies $D(P(x,\cdot)\|P^*(x,\cdot)) < \infty, \forall x\in\mathsf{X}$. The second bound holds by Proposition 23, where $K_1 = \log\theta^{-1} + \bar{n}$. The third bound follows from the second bound, $\|h^{(m)}\|_s \leqslant K_1$, and by Proposition 25 with $\alpha = \kappa(K_1)$. □

Lemma 28 (Policy continuity)**.** *There exists a constant $K_2 \geqslant 0$, such that, for every $m\in\mathbb{N}$,*

$$\|P^{(m+1)}(x,\cdot) - P^{(m)}(x,\cdot)\|_1 \leqslant \frac{K_2\tau_m}{\tau_{1:m}}, \tag{5.38}$$

and

$$\|\pi^{(m+1)} - \pi^{(m)}\|_1 \leqslant \frac{K_2\tau_m}{(1-\alpha)\tau_{1:m}} \tag{5.39}$$

where $\pi^{(m)}$ is the unique invariant distribution of $P^{(m)}$. Moreover, there exists a constant $K_3 \geqslant 0$, such that for $D^{(m)}(x) \triangleq D(P^{(m)}(x,\cdot)\|P^*(x,\cdot)), \forall x\in\mathsf{X}$, we have

$$\|D^{(m)}(x) - D^{(m+1)}(x)\|_1 \leqslant \frac{K_3\tau_m}{\tau_{1:m}}. \tag{5.40}$$

Proof. Let us recall that each $P^{(m)}$ is given by the twisted kernel $\check{P}_{h^{(m)}}$, where the relative value function $h^{(m)}$ arises from the solution of the MPE (5.11) with state

cost $\widehat{f}^{(1:m-1)}$. Then

$$\|P^{(m+1)}(x,\cdot) - P^{(m)}(x,\cdot)\|_1 \leqslant \frac{1}{2}\|h^{(m+1)} - h^{(m)}\|_s$$
$$\leqslant \frac{K_2}{2}\|\hat{f}^{(1:m)} - \hat{f}^{(1:m-1)}\|_\infty \leqslant \frac{K_2\tau_m}{\tau_{1:m}} \tag{5.41}$$

where the first step is by Proposition 25, the second by Proposition 24 with $K_2 = K(1)$, and the third by Lemma 4.3 in Yu et al. (2009). This proves (5.38). Moreover, Proposition 22 guarantees that $P^{(m)} = \check{P}_{h^{(m)}}$ has a unique invariant distribution $\pi^{(m)}$. Therefore,

$$\begin{aligned}
&\|\pi^{(m)} - \pi^{(m+1)}\|_1 \\
&= \|\pi^{(m)}P^{(m)} - \pi^{(m+1)}P^{(m+1)}\|_1 \\
&\leqslant \|\pi^{(m)}P^{(m)} - \pi^{(m)}P^{(m+1)}\|_1 \\
&\qquad + \|\pi^{(m)}P^{(m+1)} - \pi^{(m+1)}P^{(m+1)}\|_1 \\
&\leqslant \|P^{(m)} - P^{(m+1)}\|_\infty + \alpha\|\pi^{(m)} - \pi^{(m+1)}\|_1 \\
&\leqslant \frac{K_2\tau_m}{\tau_{1:m}} + \alpha\|\pi^{(m)} - \pi^{(m+1)}\|_1,
\end{aligned}$$

where the third inequality follows from (5.41). Rearranging, we get (5.39).

Next, from the form of $P^{(m)}$ and $P^{(m+1)}$ we have

$$\begin{aligned}
&D^{(m)}(x) - D^{(m+1)}(x) \\
&= \mathbb{E}_x^{(m+1)}[h^{(m+1)}] - \mathbb{E}_x^{(m)}[h^{(m)}] + \log\frac{\Lambda_{h^{(m+1)}}(x)}{\Lambda_{h^{(m)}}(x)},
\end{aligned} \tag{5.42}$$

where $\mathbb{E}_x^{(m)}[\cdot]$ denotes expectation w.r.t. $P^{(m)}(x,\cdot)$, and we can follow the same steps we have used in (5.37) to show that

$$\frac{\Lambda_{h^{(m+1)}}(x)}{\Lambda_{h^{(m)}}(x)} = \mathbb{E}_x^{(m)}\left[e^{h^{(m)} - h^{(m+1)}}\right].$$

Using the Hoeffding bound (Hoeffding (1963)), we can write

$$\log \frac{\Lambda_{h^{(m+1)}}(x)}{\Lambda_{h^{(m)}}(x)} \leqslant \mathbb{E}_x^{(m)}[h^{(m)} - h^{(m+1)}] + \frac{\|h^{(m)} - h^{(m+1)}\|_s^2}{8} \tag{5.43}$$

Substituting (5.43) into (5.42) and simplifying, we get

$$\begin{aligned}
&D^{(m)}(x) - D^{(m+1)}(x) \\
&\leqslant \mathbb{E}_x^{(m+1)}[h^{(m+1)}] - \mathbb{E}_x^{(m)}[h^{(m+1)}] + \frac{1}{8}\|h^{(m)} - h^{(m+1)}\|_s^2 \\
&\leqslant \|h^{(m+1)}\|_s \cdot \|P^{(m)}(x,\cdot) - P^{(m+1)}(x,\cdot)\|_1 \\
&\qquad + \frac{1}{8}\|h^{(m)} - h^{(m+1)}\|_s^2 \\
&\leqslant \frac{K_1 K_2 \tau_m}{\tau_{1:m}} + \frac{1}{8}\|h^{(m)} - h^{(m+1)}\|_s^2 \\
&\leqslant \frac{K_1 K_2 \tau_m}{\tau_{1:m}} + \frac{K_2^2 \tau_m^2}{2\tau_{1:m}^2} \\
&\leqslant \left(K_1 K_2 + \frac{K_2^2}{2}\right) \frac{\tau_m}{\tau_{1:m}}.
\end{aligned}$$

Here, the third step uses the fact that $\|h^{(m)}\|_s \leqslant K_1$ (Lemma 28) and (5.41), the fourth also uses (5.41), and the last is due to the fact that $\frac{\tau_m}{\tau_{1:m}} < 1$. Letting $K_3 = K_1 K_2 + \frac{K_2^2}{2}$, we get (5.40). □

Remark 10. As will be evident from the proof below, we can specify the precise form of the regret bound in (5.3) using the constants from the above lemmas:

$$\begin{aligned}
&\sup_{P \in \mathcal{M}_1^\rho(\mathsf{X})} \sup_{f_1,\ldots,f_T \in \mathcal{F}} \frac{\mathbb{E}R_T(P)}{T} \\
&\leqslant \frac{4}{3}\left(\frac{K_0(K_2+2)}{1-\alpha} + K_0 + K_3\right) T^{-1/4+\epsilon} + \frac{2K_0}{(1-\rho)T}.
\end{aligned} \tag{5.44}$$

5.5.3 Details

We are now ready to present the detailed proof of Theorem 21.

Step 1: Reduction to the steady-state case. For any $P \in \mathcal{M}_1(\mathsf{X})$, let us define the *steady-state regret* of our strategy γ w.r.t. P by

$$R_T^{\mathrm{ss}}(P) \triangleq C_T - \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{T} c_t(X, P)\right],$$

which is the difference between the actual cumulative cost of γ and the steady-state cost of the stationary unichain policy P initialized with π_P. Now let us fix some $\rho \in [0, 1)$ and consider an arbitrary $P \in \mathcal{M}_1^{\rho}(\mathsf{X})$, where without loss of generality we can assume $D(P(x,\cdot)\|P^*(x,\cdot)) < \infty$ for all $x \in \mathsf{X}$. For each $t \geqslant 1$, let $\nu_t = \delta_{x_1} P^{t-1}$ be the distribution of X_t in the Markov chain induced by the transition matrix P and initial state $X_1 = x_1$. For any T, we have

$$\begin{aligned}
&|R_T^{\mathrm{ss}}(P) - R_T(P)| \\
&= \left| \mathbb{E}_{x_1}^{P}\left[\sum_{t=1}^{T} c_t(X_t, P)\right] - \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{T} c_t(X, P)\right] \right| \\
&= \left| \sum_{t=1}^{T} \{\mathbb{E}_{\nu_t}[c_t(X_t, P)] - \mathbb{E}_{\pi_P}[c_t(X, P)]\} \right| \\
&\leqslant \sum_{t=1}^{T} \|c_t(\cdot, P)\|_{\infty} \|\nu_t - \pi_P\|_1 \\
&\leqslant 2K_0 \sum_{t=1}^{T} \rho^{t-1} \leqslant \frac{2K_0}{1-\rho}, \qquad (5.45)
\end{aligned}$$

where the second inequality is by Lemma 27 and the fact that $P \in \mathcal{M}_1^{\rho}(\mathsf{X})$. Therefore, it suffices to show that the bound in (5.3) holds with $\mathbb{E}R_T^{\mathrm{ss}}(P)$ in place of $\mathbb{E}R_T(P)$.

Step 2: Steady-state approximation within phases. In this step, we approximate the cumulative cost within each phase by its steady-state value. Let M denote the number of complete phases up to time T, i.e. $\tau_{1:M} \leqslant T < \tau_{1:M+1}$ (simple algebra

gives $M \leqslant (4/3)T^{3/4+\epsilon}$). Then we can decompose the total cost as

$$\begin{aligned} C_T &= \sum_{t=1}^{\tau_{1:M}} c_t(X_t, P_t) + \sum_{t=\tau_{1:M}+1}^{T} c_t(X_t, P_t) \\ &\leqslant \sum_{t=1}^{\tau_{1:M}} c_t(X_t, P_t) + K_0 \tau_{M+1} = C_{\tau_{1:M}} + K_0 \tau_{M+1}, \end{aligned}$$

where the inequality is by Lemma 27. Since all state costs are nonnegative by hypothesis,

$$\mathbb{E}_{\pi_P}\left[\sum_{t=1}^{T} c_t(X, P)\right] \geqslant \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X, P)\right],$$

which implies that

$$R_T^{\mathrm{ss}}(P) \leqslant R_{\tau_{1:M}}^{\mathrm{ss}}(P) + K_0 \tau_{M+1}. \tag{5.46}$$

For every time step t, let μ_t be the state distribution induced by our strategy when starting from initial state distribution $\mu_1 = \delta_{x_1}$. Note that the transition matrix at time t is $P_t = P^{(m)}$ if $t \in \mathcal{T}_m$. We can decompose the expected cost in the first M phases as

$$\mathbb{E}C_{\tau_{1:M}} = \sum_{m=1}^{M} \sum_{t \in \mathcal{T}_m} \mathbb{E}_{\mu_t}\left[c_t(X, P^{(m)})\right], \tag{5.47}$$

and for every $t \in \mathcal{T}_m$ we have

$$\begin{aligned} &\mathbb{E}_{\mu_t}\left[c_t(X, P^{(m)})\right] \\ &\leqslant \mathbb{E}_{\pi^{(m)}}\left[c_t(X, P^{(m)})\right] + \|c_t(\cdot, P^{(m)})\|_\infty \|\mu_t - \pi^{(m)}\|_1 \\ &\leqslant \mathbb{E}_{\pi^{(m)}}\left[c_t(X, P^{(m)})\right] + K_0 \|\mu_t - \pi^{(m)}\|_1, \end{aligned}$$

where the last step is by Lemma 27. In addition, for every $k \in \{0, 1, \ldots, \tau_m - 1\}$, we have

$$\begin{aligned} &\|\mu_{\tau_{1:m-1}+k+1} - \pi^{(m)}\|_1 \\ &= \left\|\mu_{\tau_{1:m-1}+1}(P^{(m)})^k - \pi^{(m)}(P^{(m)})^k\right\|_1 \\ &\leqslant \alpha^k \|\mu_{\tau_{1:m-1}+1} - \pi^{(m)}\|_1 \leqslant 2\alpha^k, \end{aligned}$$

where the first inequality is due to Lemma 27. Hence,

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_m} \mathbb{E}_{\mu_t}\left[c_t(X, P^{(m)})\right] \\
&\leqslant \sum_{t\in\mathcal{T}_m} \mathbb{E}_{\pi^{(m)}}\left[c_t(X, P^{(m)})\right] + 2K_0 \sum_{k=0}^{\tau_m-1} \alpha^k \\
&\leqslant \sum_{t\in\mathcal{T}_m} \mathbb{E}_{\pi^{(m)}}\left[c_t(X, P^{(m)})\right] + \frac{2K_0}{1-\alpha}.
\end{aligned}$$

Substituting this into (5.47), we have

$$\mathbb{E}C_{\tau_{1:M}} \leqslant \sum_{m=1}^{M} \sum_{t\in\mathcal{T}_m} \mathbb{E}_{\pi^{(m)}}\left[c_t(X, P^{(m)})\right] + \frac{2K_0 M}{1-\alpha}.$$

Step 3: Looking one phase ahead. In this step, we show that the steady-state cost in each phase is not much worse than what we could get if we knew everything one phase ahead. For every $m \in \{1, \ldots, M\}$, we have

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_m} \mathbb{E}_{\pi^{(m)}}\left[c_t(X, P^{(m)})\right] \\
&\leqslant \sum_{t\in\mathcal{T}_m} \mathbb{E}_{\pi^{(m+1)}}\left[c_t(X, P^{(m)})\right] + K_0\tau_m \|\pi^{(m+1)} - \pi^{(m)}\|_1 \\
&\leqslant \tau_m \mathbb{E}_{\pi^{(m+1)}}\left[\widehat{f}^{(m)} + D^{(m)}\right] + \frac{K_0 K_2 \tau_m^2}{(1-\alpha)\tau_{1:m}} \\
&= \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) + \tau_m \mathbb{E}_{\pi^{(m+1)}}\left[D^{(m)} - D^{(m+1)}\right] \\
&\qquad + \frac{K_0 K_2 \tau_m^2}{(1-\alpha)\tau_{1:m}} \\
&\leqslant \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) + \left(\frac{K_0 K_2}{1-\alpha} + K_3\right) \frac{\tau_m^2}{\tau_{1:m}},
\end{aligned}$$

where the first inequality is by Lemma 27, the second inequality is by Lemma 28,

and the last inequality is due to (5.40) in Lemma 28. So we now have

$$\mathbb{E}C_{\tau_{1:M}} \leqslant \sum_{m=1}^{M} \tau_m \bar{J}_{\hat{f}^{(m)}}(P^{(m+1)}) + \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3 \right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha}. \tag{5.48}$$

Step 4: Looking M phases ahead. In this step, we consider the fictitious situation where we know everything M phases ahead, and show that the resulting steady-state value is actually greater than what we could get if we knew everything just one phase ahead. In other words, we claim that

$$\sum_{m=1}^{M} \tau_m \bar{J}_{\hat{f}^{(m)}}(P^{(m+1)}) \leqslant \sum_{m=1}^{M} \tau_m \bar{J}_{\hat{f}^{(m)}}(P^{(M+1)}). \tag{5.49}$$

To see that this claim is true, we apply backward induction:

$$\begin{aligned}
&\sum_{m=1}^{M} \tau_m \bar{J}_{\hat{f}^{(m)}}(P^{(M+1)}) \\
&= \sum_{m=1}^{M-1} \tau_m \bar{J}_{\hat{f}^{(m)}}(P^{(M+1)}) + \tau_M \bar{J}_{\hat{f}^{(M)}}(P^{(M+1)}) \\
&= \tau_{1:M-1} \bar{J}_{\hat{f}^{(1:M-1)}}(P^{(M+1)}) + \tau_M \bar{J}_{\hat{f}^{(M)}}(P^{(M+1)}) \\
&\geqslant \tau_{1:M-1} \bar{J}_{\hat{f}^{(1:M-1)}}(P^{(M)}) + \tau_M \bar{J}_{\hat{f}^{(M)}}(P^{(M+1)}) \\
&= \sum_{m=1}^{M-1} \tau_m \bar{J}_{\hat{f}^{(m)}}(P^{(M)}) + \tau_M \bar{J}_{\hat{f}^{(M)}}(P^{(M+1)}),
\end{aligned}$$

where the second equality is due to the fact that $\tau_{1:M-1}\widehat{f}^{(1:M-1)} = \sum_{t \in \mathcal{T}_{1:M-1}} f_t = \sum_{m=1}^{M} \tau_m \widehat{f}^{(m)}$, while the inequality is by Proposition 26 and the fact that $P^{(M)} = \check{P}_{h^{(M)}}$, where $h^{(M)}$ is the relative value function for state cost $\widehat{f}^{(1:M-1)}$. Repeating

this argument, we obtain (5.49). Moreover,

$$\begin{aligned}\sum_{m=1}^{M} \tau_m \bar{J}_{\hat{f}^{(m)}}(P^{(M+1)}) &= \tau_{1:M} \bar{J}_{\hat{f}^{(1:M)}}(P^{(M+1)}) \\ &= \tau_{1:M} \inf_{P \in \mathcal{M}_1(\mathsf{X})} \bar{J}_{\hat{f}^{(1:M)}}(P) \\ &= \inf_{P \in \mathcal{M}_1(\mathsf{X})} \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right]. \end{aligned} \tag{5.50}$$

After these four steps, we are finally in a position to bound the expected steady-state regret. Combining (5.48)–(5.50), we can write

$$\begin{aligned}\mathbb{E}C_{\tau_{1:M}} \leqslant \inf_{P \in \mathcal{M}_1(\mathsf{X})} \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right] \\ + \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3\right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}R^{\text{ss}}_{\tau_{1:M}}(P) &= \mathbb{E}C_{\tau_{1:M}} - \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right] \\ &\leqslant \mathbb{E}C_{\tau_{1:M}} - \inf_{P \in \mathcal{M}_1(\mathsf{X})} \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right] \\ &\leqslant \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3\right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha}. \end{aligned} \tag{5.51}$$

Next we show that the right-hand side of (5.51) can be bounded by a quantity that is sublinear in T. From (5.46), we have

$$\begin{aligned}\mathbb{E}R^{\text{ss}}_T(P) &\leqslant \mathbb{E}R^{\text{ss}}_{\tau_{1:M}}(P) + K_0 \tau_{M+1} \\ &\leqslant \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3\right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha} + K_0 \tau_{M+1}.\end{aligned}$$

Due to our construction of the phases, $M \leqslant (4/3)T^{3/4+\epsilon}$ and $\tau_{M+1} \leqslant M$ if $M > 1$. Moreover, it is a matter of routine but tedious algebraic calculations to show that the choice $\tau_m = \lceil m^{\frac{1}{3}-\epsilon} \rceil$ for $m = 1, \ldots, M$ for any $\epsilon \in (0, 1/3)$ is sufficient to guarantee that $\tau_m^2 \leqslant \sqrt{\tau_{1:m}}$. Thus, we obtain

$$\begin{aligned}
\mathbb{E}R_T^{\mathrm{ss}}(P) &\leqslant \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3 \right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha} + K_0 M \\
&\leqslant M \left(\frac{K_0(K_2+2)}{1-\alpha} + K_0 + K_3 \right) \\
&\leqslant \frac{4}{3} \left(\frac{K_0(K_2+2)}{1-\alpha} + K_0 + K_3 \right) T^{3/4+\epsilon}.
\end{aligned}$$

Therefore, recalling (5.45), we finally obtain

$$\begin{aligned}
&\frac{\mathbb{E}R_T(P)}{T} \\
&\leqslant \frac{\mathbb{E}R_T^{\mathrm{ss}}(P)}{T} + \frac{2K_0}{T(1-\rho)} \\
&\leqslant \frac{4}{3} \left(\frac{K_0(K_2+2)}{1-\alpha} + K_0 + K_3 \right) T^{-1/4+\epsilon} + \frac{2K_0}{T(1-\rho)},
\end{aligned}$$

which completes the proof of Theorem 21.

5.6 Simulations

In this section, we demonstrate the performance of our proposed strategy on a simulated problem involving online (real-time) tracking of a moving target on a large, connected, undirected graph G, which models a terrain with obstacles. The state space is the set of all vertices (nodes) of G. The target is executing a stationary random walk on G with a randomly sampled transition probability matrix, which is different from the one that governs the passive dynamics P^*. The motion of both the tracking agent and the target must conform to the topology of G, in the sense that

both can only move between neighboring vertices. The graph used in our simulation has 564 vertices.

To make sure that Assumptions 5 and 6 are satisfied, we construct the passive dynamics in the form $P^* = (1-\delta)P_1 + \delta P_0$ for some $\delta \in (0,1)$. Here, P_1 is a random walk that represents environmental constraints, allowing the agent to go from a given node either to any adjacent node (with equal probability) or to remain at the current location. To ensure that the agent is sufficiently mobile, the probability of not moving is chosen to be relatively small (in our case, 0.01) compared to the probability of transitioning to any of the neighboring nodes. Since the underlying graph is connected, the random walk P_1 is irreducible; it is also aperiodic since $P_1(x,x) > 0$ for all vertices x. We also add a perturbation random walk P_0, which has a fixed column of ones (we can think of the node indexing that column as a "home base" for the agent), and zeros elsewhere. The "size" of the perturbation is controlled by δ, which is set to be small (we have chosen $\delta = 0.01$), so the agent only has a slight chance of returning to "home base" from any given node within one step. This perturbation ensures that no two rows of P^* are orthogonal, and $\alpha(P^*) \leqslant 1 - \delta = 0.99$.

The simulation consists of a number of independent experiments. Each individual experiment runs for $T = 1000$ time steps. We first randomly sample a transition matrix for the target motion. After simulating the target's random walk for T steps, we record the target locations and use them to generate a sequence of state cost functions $\{f_t\}_{t=1}^T$. Then we feed these 1000 state cost functions sequentially to our online algorithm and compute the resulting cumulative cost C_T. At each time t, the tracking agent is in state (location) x_t, the target is at location s_t, and the agent's action is P_t. The cumulative cost after T time steps is

$$C_T = \sum_{t=1}^{T} \left[f_t(x_t) + D\left(P_t(x_t,\cdot) \| P^*(x_t,\cdot)\right) \right], \tag{5.52}$$

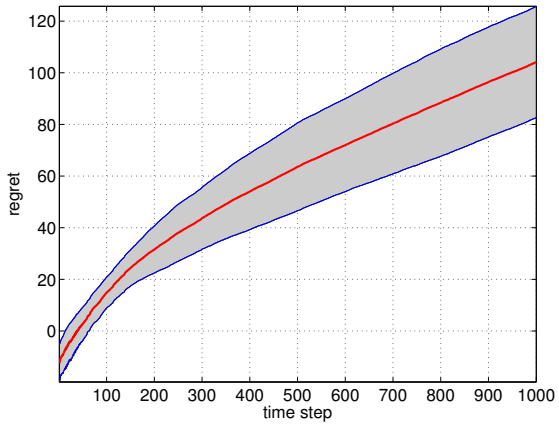

FIGURE 5.1: Regret versus time. The red curve shows the average of the regret (the difference between the total cost of our strategy up to each time t and the total cost of the best stationary policy up to that time) over 100 independent realizations of the simulation. At each time t, the height of the gray region corresponds to one sample standard deviation.

with state costs $f_t(x_t) = d_G(x_t, s_t)$, where $d_G(\cdot, \cdot)$ is the graph distance (number of edges in the shortest path) between the agent's current location and the location of the target, normalized by the diameter of G. Then we compute the best stationary policy P in hindsight for the average of all the state costs by solving the MPE

$$e^{-\widehat{f}} P^* e^{-h} = e^{-\lambda} e^{-h}$$

for the relative value function h, where $\widehat{f} = \frac{1}{T} \sum_{t=1}^{T} f_t$, and then setting

$$P(x, \cdot) = \frac{P^*(x, \cdot) e^{-h(\cdot)}}{P^* e^{-h}(x)}, \qquad x \in \mathsf{X}$$

The regret is then computed with respect to the steady-state cost of this best stationary policy:

$$R_T(P) = C_T - \mathbb{E}_{\pi_P} \left[\sum_{t=1}^{T} c_t(X, P) \right],$$

where

$$c_t(X, P) = f_t(X) + D\left(P(X, \cdot) \| P^*(X, \cdot)\right),$$

and π_P is the unique invariant distribution of P.

To plot the regret versus time with error bars, we implement the experiment 100 times and compute the empirical average of the regret across experiments. For each realization the agent was initialized with the same starting state. The evolution of the regret versus time is shown in Figure 5.1, where the regret at time t is defined as the total cost of our strategy up to time t minus the total cost of the best stationary policy up to time t. We can see that the regret is growing sublinearly, as stated in Theorem 21.

We also compare the total cost of our strategy to that of the best stationary baseline policy among a set $\tilde{\mathcal{N}}$ of 10^5 randomly sampled stationary policies. Once again, each experiment runs for $T = 1000$ time steps. The baseline policy P_{baseline} is the one that has the smallest total cost

$$C_T(P_{\text{baseline}}) = \min_{P \in \tilde{\mathcal{N}}} \sum_{t=1}^{T} \left[f_t(x_t) + D\left(P(x_t, \cdot) \| P^*(x_t, \cdot)\right) \right].$$

among the 10^5 randomly sampled policies. The regret of our adaptive strategy is thus given by $C_T - C_T(P_{\text{baseline}})$.

As before, there are 100 independent experiments, where in each experiment the agent using our strategy and the agent using the best sampled stationary policy were initialized with the same starting state. The evolution of the regret versus time is shown in Figure 5.2, where the regret at time t is defined as the total cost of our strategy up to time t minus the total cost of the best sampled stationary policy up to time t. We can see that the regret is negative, which implies that our strategy outperforms the best sampled stationary policy for each particular realization of the state cost sequence.

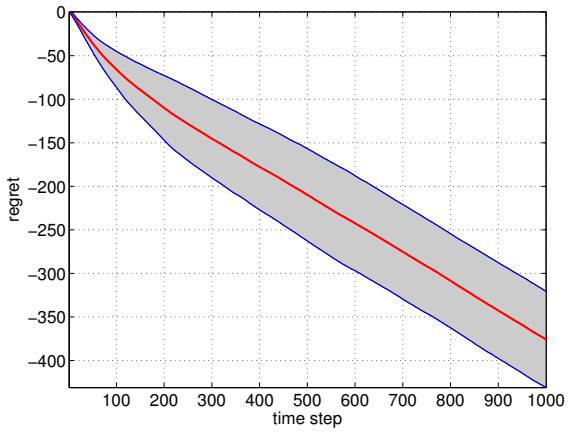

FIGURE 5.2: Comparison of our proposed strategy to the best stationary policy in a set of 10^5 randomly sampled policies. The red curve shows the average of the regret (the difference between the total cost of our strategy up to each time t and the total cost of the best stationary policy up to that time) over 100 independent realizations of the simulation. At each time t, the height of the gray area corresponds to one sample standard deviation.

5.7 Conclusion and future work

The problem studied in this chapter combines aspects of both stochastic control and online learning. In particular, our construction of a Hannan-consistent strategy (a concept from the theory of online learning (Cesa-Bianchi and Lugosi (2006))) uses several ideas and techniques from the theory of MDPs with average cost criterion, including some new results concerning optimal policies for MDPs with KL control costs (Todorov (2007, 2008, 2009)).

We have proved that, for any horizon T, our strategy achieves sublinear $O(T^{3/4})$ regret relative to any uniformly ergodic class of stationary policies, which is similar to the results of Yu et al. (2009) for online MDPs with finite state and action spaces. However, while our strategy (like that of Yu et al. (2009)) is computationally efficient, we believe that the $O(T^{3/4})$ scaling of regret with T is suboptimal. Indeed, in the case when both the state and the action spaces are finite, Even-Dar et al. (2009) present a strategy that achieves a much better $O(\sqrt{T})$ regret. Of course, the strategy of

Even-Dar et al. (2009) involves recomputing the policy at *every* time step (rather than in phases, as is done here and in Yu et al. (2009)), which results in a significant loss of efficiency. An interesting open question, which we plan to address in our future work, is whether it is possible to attain $O(\sqrt{T})$ regret for online MDPs with KL control costs.

6

No-regret algorithms for zero-sum stochastic games

We can view online MDPs as a special type of *stochastic games* (Shapley (1953); Sorin (2002)), where an agent who wishes to minimize his long-term average cost is controlling the state transition alone, while an oblivious environment chooses the cost functions. The underlying assumption is that the environment is acting arbitrarily and likely sub-optimally. This obliviousness is exploited by the rational agent. Earlier work (Even-Dar et al. (2009); Yu et al. (2009); Guan et al. (2012); Dick et al. (2014)) in the literature concentrated on developing algorithms that achieve sublinear regret bounds. The goal of this chapter is to study regret minimization in a richer setting. We remove the oblivious environment assumption and let the environment be a rational opponent who is adaptive to the agent's actions and also has a specific goal, which is to maximize his long-term average reward. In that case, the agent and the environment are playing a zero-sum stochastic game, where the loss of the agent is the reward of the environment and vice-versa. As before, we let the agent control the state transition alone. Now both players may adapt their strategies based on history, i.e., this is a subclass of zero-sum stochastic games where the the state transition is

determined by one player.

Von Neumann (1928) laid the foundations of the field of game theory and algorithms by proving the celebrated *minimax theorem*: for every two-person, zero-sum game with finitely many pure strategies, there exists an equilibrium mixed strategy, i.e., a mixed strategy for each player, such that no player can improve his payoff by unilaterally deviating. Recent work by Daskalakis et al. (2011); Rakhlin and Sridharan (2013b) has shown that online optimization can be used to develop efficient dynamics that lead the two players of the game to converge to an equilibrium. In other words, if both players of a game use an online learning algorithm to compete against their opponent, then the average payoffs of the players converge to the minimax value and their averaged strategies constitute an approximate minimax equilibrium. Motivated by this result, we investigate the following questions: *what is the role of regret minimization strategies for zero-sum single-controller stochastic games? Do they lead the game to converge to the minimax equilibrium?* We answer these questions by showing that, if the two players follow stationary policies derived from regret minimization algorithms, then their average payoff will converge to the average minimax payoff at fast rates.

Existing work on stochastic games has mainly focused on finding and computing equilibria for these games (Shapley (1953); Mertens and Neyman (1981); Parthasarathy and Raghavan (1981)). There is a well-developed theory for regret minimization algorithms in repeated games (Cesa-Bianchi and Lugosi (2006)). Even-Dar et al. (2009) first considered regret minimization in single-controller stochastic games (online MDPs) and proved a sublinear regret relative to the best stationary policy in hindsight. They also showed that, for standard stochastic games where the environment controls both the costs and the state transitions, it is NP-hard to approximate the best fixed strategy within any factor better than 0.875, which means it is NP-hard to derive an algorithm that has an expected average reward larger

than $0.875R^*$, where R^* is the optimal average reward obtained by a stationary policy. Yu et al. (2009) also proved a hardness result for achieving sublinear regret for standard stochastic games, and showed that regret minimization is possible for single-controller stochastic games, where the environment is oblivious (non-adaptive). Dick et al. (2014) treated the single-controller stochastic game as an online linear optimization problem and provided algorithms that achieve sublinear regret bounds. Guan et al. (2014a) provided a general framework for deriving regret minimization algorithms in single-controller stochastic games under the assumption that the environment is oblivious, and unified existing methods (Even-Dar et al. (2009); Yu et al. (2009); Dick et al. (2014)) under a single theoretical interpretation.

In this chapter, we look at regret minimization and stochastic games from a different and novel prospective. We are not interested in minimizing the regret against the best stationary strategy; instead, we shift our focus toward using regret minimization strategies to derive near-optimal strategies for single-controller stochastic games. To the best of our knowledge, this topic has not been the subject of previous work. We first reduce a single-controller stochastic game to an online linear optimization problem and use the resulting regret minimization strategies for both players to generate a pair of stationary policies. We let both players follow these stationary policies and show that the average payoff converges to the average minimax payoff at fast rates.

The range of applications of stochastic games is severely limited by their high computational cost. The computational burden is significant due to the curse of dimensionality, since the computational complexity scales quadratically with the number of states when dynamic programming is used. Our approach can be seen as an alternative way of approximately solving two-player stochastic games by using stationary policies derived from regret minimization strategies. Moreover, by providing the rate at which the payoff resulting from these policies converges to the minimax

value of the game, we are able to precisely quantify the degree of sub-optimality of these policies.

The rest of the chapter is organized as follows. Section 6.1 introduces the single-controller stochastic game setting and shows how to reduce it to an online linear optimization problem. Section 6.2 presents our main results. Section 6.3 summarizes our contributions.

6.1 Problem formulation

We consider a single-controller stochastic game, i.e., only one player controls the state transitions (Parthasarathy and Raghavan (1981)). The state space is denoted by X, and there are two action spaces for two players, U_1 and U_2. The cost function is $c : \mathsf{X} \times \mathsf{U}_1 \times \mathsf{U}_2 \to [0, 1]$. Player 1's closed-loop behavioral strategy is denoted by the tuple $\boldsymbol{\gamma} = (\gamma_1, \ldots)$, where $\gamma_t : \mathsf{X}^t \times \mathsf{U}_1^{t-1} \times \mathsf{U}_2^{t-1} \to \mathcal{P}(\mathsf{U}_1)$, and his mixed strategy at time t is denoted by $P_{1,t}$. Similarly, Player 2 also has a closed-loop behavioral strategy denoted by the tuple $\boldsymbol{\delta} = (\delta_1, \ldots)$ with $\delta_t : \mathsf{X}^t \times \mathsf{U}_1^{t-1} \times \mathsf{U}_2^{t-1} \to \mathcal{P}(\mathsf{U}_2)$, and his mixed strategy at time t is denoted by $P_{2,t}$. Since Player 1 controls the state transition alone, the controlled transition kernel is given by $K(y|x, u_1)$, which specifies the probability of moving to state y given the current state x and Player 1's action u_1. The game protocol is the following:

$X_1 = x$
for $t = 1, 2, \ldots$
  Player 1 knows $x^t, u_1^{t-1}, u_2^{t-1}$,
  Player 1 picks $u_{1,t} \sim P_{1,t}$;
  Simultaneously
  Player 2 knows $x^t, u_1^{t-1}, u_2^{t-1}$,
  Player 2 picks $u_{2,t} \sim P_{2,t}$;
  $u_{1,t}$ and $u_{2,t}$ are revealed to each other at the same time
  Player 1 incurs cost $c(x_t, u_{1,t}, u_{2,t})$, Player 2 incurs cost $-c(x_t, u_{1,t}, u_{2,t})$
  The next state is drawn by $X_{t+1} \sim K(\cdot|x_t, u_{1,t})$
end for

A basic result in the theory of stochastic games is that, for every initial state x, the game has a value given by

$$V(x) = \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{\delta}} \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}} \left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t}) \right]. \tag{6.1}$$

This corresponds to an infinite-horizon stochastic game with optimal stationary strategies for both players (Mertens and Neyman (1981); Parthasarathy and Raghavan (1981)). On the other hand, we can fix a time horizon T, and look at this quantity:

$$V_T(x) = \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{\delta}} \frac{1}{T} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}} \left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t}) \right]. \tag{6.2}$$

In this case, instead, the optimal strategies of both players are non-stationary. Our goal is to explore whether there exist regret-minimization strategies for the two players, such that, if both players honestly follow *stationary* policies derived from the prescribed strategies, their average payoffs will approach the average minimax payoff at some fast rate.

Our approach is motivated by Rakhlin and Sridharan (2013b), who solve a saddle-point optimization problem by playing two online convex optimization algorithms against each other. Hence, we first need to reduce the single-controller stochastic game to an online linear optimization problem. We start by introducing some necessary details about this reduction, and then give a formal description of our approach.

6.1.1 Preliminaries

As mentioned earlier, a single-controller stochastic game can be seen as an online MDP problem, where the cost functions are chosen by the environment and the state transition is controlled by the agent alone. In Section 2.3.3, we have reduced the online MDP problem to an online linear optimization problem. Here we adopt

the exact same idea to reduce a single-controller stochastic game to an online linear optimization problem.

By definition of the state-action polytope $\mathcal{G}$, it is easy to see that we can associate each policy with an occupation measure. Thus, we can let Player 1 choose an occupation measure from $\mathcal{G}$ at each time step, instead of choosing his action $u_1 \in \mathsf{U}_1$. Next, by allowing Player 2 to choose his action from a certain set, we can let the cost function itself reflect Player 2's action: for each $u_2 \in U_2$, we have a state-action cost function $c_{u_2}(x, u_1) = c(x, u_1, u_2)$. Let $\mathcal{F}$ be the closed convex hull of all such functions: $\mathcal{F} = \{\sum_{u_2 \in \mathsf{U}_2} \rho(u_2) c(\cdot, \cdot, u_2) : \rho \in \mathcal{P}(\mathsf{U}_2)\}$, where choosing a point within this set is equivalent to simply picking out a mixed strategy of Player 2. Here we slightly change the notation by viewing the cost function at each time step as the mixed strategy picked by Player 2. Namely, for each mixed strategy ρ_t of Player 2 at time t, there exists a corresponding cost function $f_t \in \mathcal{F}$, such that $\sum_{u_2 \in \mathsf{U}_2} \rho_t(u_2) c(\cdot, \cdot, u_2) = f_t(\cdot, \cdot)$. The key point is that the expected cost of following a policy can be then approximated by the inner product of the occupation measure associated with that policy and the new cost function chosen from $\mathcal{F}$. This is shown by Lemma 1 in Dick et al. (2014). Consequently, we are able to reduce a single-controller stochastic game to an online linear optimization problem, where at each time step t, Player 1 is choosing the occupation measure ν_t from the state-action polytope $\mathcal{G}$, and Player 2 is choosing the cost function $f_t \in \mathcal{F}$ (here we omit the dependence of ν_t, f_t on $\boldsymbol{\gamma}, \boldsymbol{\delta}$ for simplicity). The one step costs incurred by Player 1 and 2 are $\langle \nu_t, f_t \rangle$ and $-\langle \nu_t, f_t \rangle$, respectively. We would like to emaphsize that in this online linear optimization problem, at time step t, Player 1 can only observe the quantity $\sum_{u_2 \in \mathsf{U}_2} \rho_t(u_{2,t}) c(x_t, u_{1,t}, u_{2,t})$; Player 2 can only observe $\sum_{x_t, u_{1,t}} \nu_t(x_t, u_{1,t}) c(x_t, u_{1,t}, u_{2,t})$. They don't have access to their opponents' mixed strategies.

6.1.2 From online linear optimization to a single-stage game

Once the problem is reduced to an online linear optimization, we let both players adopt regret minimization strategies, and we look at their online learning regrets:

$$\frac{1}{T}\sum_{t=1}^{T}\langle \nu_t, f_t\rangle - \inf_{\nu\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\langle \nu, f_t\rangle \leqslant R^1(f_1,\ldots,f_T), \tag{6.3}$$

$$\frac{1}{T}\sum_{t=1}^{T}(-\langle \nu_t, f_t\rangle) - \inf_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}(-\langle \nu_t, f\rangle) \leqslant R^2(\nu_1,\ldots,\nu_T). \tag{6.4}$$

Here we assume that Player 1 produces a sequence of $\nu_1,\ldots,\nu_T$ by using a regret minimization algorithm, and his regret is denoted by $R^1(f_1,\ldots,f_T)$ given his observations, which are averages of Player 2's actions. Respectively, Player 2 produces a sequence of $f_1,\ldots,f_T$ by using a regret minimization algorithm, and his regret is denoted by $R^2(\nu_1,\ldots,\nu_T)$ given averages of Player 1's actions. Note that there is no Markov chain involved, and the strategies for both players depends only on the averages of previous moves of their opponents. This online linear optimization problem refers to the online learning (steady-state) component of the game. Since this is a standard online learning problem, we know that there exist numerous regret minimization strategies for both players such that R^1 and R^2 are sublinear (Cesa-Bianchi and Lugosi (2006)). We refer to this online learning game by G1, and it serves as a starting point of our analysis. Both players adopt regret minimization algorithms in G1 and get sublinear regret bounds $R^1(f_1,\ldots,f_T)$ and $R^2(\nu_1,\ldots,\nu_T)$.

We denote the averages of the sequences of two players' actions by $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^{T}\nu_t$ and $\bar{f}_T = \frac{1}{T}\sum_{t=1}^{T} f_t$. They can be seen as stationary policies (strategies) for the corresponding player. The following proposition will be exploited in the proofs of the next section.

Proposition 29. *Suppose both players adopt regret minimization strategies in G1. Let* $\{\nu_t\}_{t=1}^{T}$ *and* $\{f_t\}_{t=1}^{T}$ *denote the sequences of the players' choices. There exists a*

stationary policy for Player 1 to choose the occupation measure, given by $\bar{\nu}_T$, i.e., the average occupation measure of all T measures drawn from the regret minimization algorithm, such that, given any choice of Player 2, for a large enough horizon T, the one-shot game where Player 1 chooses the occupation measure and Player 2 chooses the cost function, converges to an equilibrium at fast rates:

$$\sup_{f\in\mathcal{F}}\langle\bar{\nu}_T, f\rangle - \inf_{\nu\in\mathcal{G}}\sup_{f\in\mathcal{F}}\langle\nu, f\rangle \leqslant R^1(f_1,\ldots,f_T) + R^2(\nu_1,\ldots,\nu_T). \tag{6.5}$$

Proof. By linearity, we have

$$\inf_{\nu\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\langle\nu, f_t\rangle = \inf_{\nu\in\mathcal{G}}\langle\nu, \bar{f}_T\rangle \leqslant \sup_{f\in\mathcal{F}}\inf_{\nu\in\mathcal{G}}\langle\nu, f\rangle,$$

and

$$\inf_{\nu\in\mathcal{G}}\sup_{f\in\mathcal{F}}\langle\nu, f\rangle \leqslant \sup_{f\in\mathcal{F}}\langle\bar{\nu}_T, f\rangle = \sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\langle\nu_t, f\rangle.$$

Because both $\mathcal{G}$ and $\mathcal{F}$ are convex, using von Neumann's minimax theorem (see Theorem 7.1 in Cesa-Bianchi and Lugosi (2006) for details), we have

$$\sup_{f\in\mathcal{F}}\inf_{\nu\in\mathcal{G}}\langle\nu, f\rangle = \inf_{\nu\in\mathcal{G}}\sup_{f\in\mathcal{F}}\langle\nu, f\rangle.$$

Therefore, we get $\inf_{\nu\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\langle\nu, f_t\rangle \leqslant \inf_{\nu\in\mathcal{G}}\sup_{f\in\mathcal{F}}\langle\nu, f\rangle$. By adding the regret of the two players, we have

$$\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\langle\nu_t, f\rangle - \inf_{\nu\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\langle\nu, f_t\rangle \leqslant R^1(f_1,\ldots,f_T) + R^2(\nu_1,\ldots,\nu_T). \tag{6.6}$$

Using $\sup_{f\in\mathcal{F}}\langle\bar{\nu}_T, f\rangle = \sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\langle\nu_t, f\rangle$ and $\inf_{\nu\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\langle\nu, f_t\rangle \leqslant \inf_{\nu\in\mathcal{G}}\sup_{f\in\mathcal{F}}\langle\nu, f\rangle$, we get (6.5). □

This naturally leads to another question: *how can we translate this result back to the original dynamic game that has multiple rounds?* We will answer this question by relating the quantity $\inf_{\nu\in\mathcal{G}}\sup_{f\in\mathcal{F}}\langle\nu, f\rangle$ to V_T and relating the quantity $\sup_{f\in\mathcal{F}}\langle\bar{\nu}_T, f\rangle$ to the payoff of a dynamic game.

6.2 Regret minimization in stochastic games

6.2.1 Stationary strategies for finite-horizon stochastic games

Consider a one-shot game, where Player 1 chooses the occupation measure from $\mathcal{G}$ and Player 2 chooses the cost function from $\mathcal{F}$. The following lemma relates this game to a finite-horizon two player zero-sum single-controller stochastic game, whose value is given by (6.2).

Lemma 30. *Suppose both players adopt regret minimization strategies in G1, and* $\{\nu_t\}_{t=1}^T$ *and* $\{f_t\}_{t=1}^T$ *are the sequences of the players' choices. If Player 1 uses the average occupation measure* $\bar{\nu}_T$ *for the above one-shot game, then regardless of Player 2's choice, Player 1's payoff approaches the minimax payoff of the* T*-step finite-horizon stochastic game starting at any initial state* x *at fast rates. Specifically, let* $R^1(f_1, \ldots, f_T)$ *and* $R^2(\nu_1, \ldots, \nu_T)$ *be the online learning regrets of the corresponding regret minimization algorithms used in G1. Then there exists some constant* C *such that*

$$\sup_{f \in \mathcal{F}} \langle \bar{\nu}_T, f \rangle \leqslant V_T(x) + \frac{C}{T} + R^1(f_1, \ldots, f_T) + R^2(\nu_1, \ldots, \nu_T). \tag{6.7}$$

Proof. To lower-bound $V_T(x)$, we restrict Player 2 only to those strategies that draw actions i.i.d. from some fixed distribution, ignoring the state. We denote the set of all such strategies by $\Delta_{iid}^{\boldsymbol{\delta}}$. For each such strategy of Player 2, Player 1 faces a T-step

MDP:

$$
\begin{aligned}
V_T(x) &= \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{\delta}} \frac{1}{T} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}} \left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t}) \right] \\
&\geqslant \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{\delta} \in \Delta_{iid}^{\boldsymbol{\delta}}} \frac{1}{T} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}} \left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t}) \right] \\
&= \inf_{\boldsymbol{\gamma}} \sup_{f \in \mathcal{F}} \frac{1}{T} \mathbb{E}_x^{\boldsymbol{\gamma},f} \left[\sum_{t=1}^{T} f(X_t, U_{1,t}) \right] \\
&\geqslant \sup_{f \in \mathcal{F}} \inf_{\boldsymbol{\gamma}} \mathbb{E}_x^{\boldsymbol{\gamma},f} \left[\sum_{t=1}^{T} f(X_t, U_{1,t}) \right].
\end{aligned}
$$

For a fixed choice of $f \in \mathcal{F}$, we denote the optimal payoff of that MDP starting from initial state $x \in \mathsf{X}$ by $v_T^f(x) = \inf_{\boldsymbol{\gamma}} \mathbb{E}_x^{\boldsymbol{\gamma},f} \left[\sum_{t=1}^{T} f(X_t, U_{1,t}) \right]$ and denote $v_\infty^f(x) \triangleq \lim_{T \to \infty} v_T^f(x)$.

By Proposition 5.21 in Sorin (2002), we know that $T|v_T^f(x) - v_\infty^f(x)|$ is uniformly bounded. Thus, for each $f \in \mathcal{F}$, there exists a constant $C(f)$, such that

$$
\max_{x \in \mathsf{X}} \left| v_T^f(x) - v_\infty^f(x) \right| \leqslant \frac{C(f)}{T}. \tag{6.8}
$$

Using the fact that $\mathcal{F}$ is a convex hull of finitely many functions, we know that there exists a constant $C = \max_{f \in \mathcal{F}} C(f)$, such that

$$
\max_{f \in \mathcal{F}} \max_{x \in \mathsf{X}} \left| v_T^f(x) - v_\infty^f(x) \right| \leqslant \frac{C}{T}.
$$

It is shown in Borkar (1988) that

$$
v_\infty^f(x) = \inf_{\nu \in \mathcal{G}} \langle \nu, f \rangle. \tag{6.9}
$$

Consequently,

$$\begin{aligned} V_T(x) &\geqslant \sup_{f\in\mathcal{F}} \inf_{\boldsymbol{\gamma}} \mathbb{E}_x^{\boldsymbol{\gamma},f}\left[\sum_{t=1}^{T} f(X_t, U_{1,t})\right] \\ &= \sup_{f\in\mathcal{F}} v_T^f(x) \\ &\geqslant \sup_{f\in\mathcal{F}}\{\inf_{\nu\in\mathcal{G}}\langle\nu, f\rangle - \frac{C}{T}\} \\ &= \inf_{\nu\in\mathcal{G}} \sup_{f\in\mathcal{F}}\langle\nu, f\rangle - \frac{C}{T}. \end{aligned}$$

Combining this with (6.5), we get

$$\sup_{f\in\mathcal{F}}\langle\bar{\nu}_T, f\rangle \leqslant V_T(x) + \frac{C}{T} + R^1(f_1, \ldots, f_T) + R^2(\nu_1, \ldots, \nu_T).$$

□

Next we relate the quantity $\sup_{f\in\mathcal{F}}\langle\bar{\nu}_T, f\rangle$ to the payoff of different dynamic games. We start with the online learning game G1. Recall that Player 1 uses a regret-minimization algorithm to produce a sequence of T occupation measures on the space of state-action pairs $\mathsf{X}\times\mathsf{U}_1$, and uses the average of these measures $\bar{\nu}_T$ to recover a stationary policy $\bar{P}_T : \mathsf{X}\to\mathcal{P}(\mathsf{U}_1)$. $\pi_{\bar{P}_T}$ is the invariant distribution induced by policy $\bar{P}_T$, and we have $\bar{\nu}_T = \pi_{\bar{P}_T}\otimes\bar{P}_T$. Player 2 also adopts a regret-minimization algorithm to generate a sequence of $\{f_t\}_{t=1}^T$ and computes the average $\bar{f}_T$. Now we consider a T-round stochastic game, throughout which Player 1 uses the stationary policy $\bar{P}_T$ and Player 2 uses $\bar{f}_T$.

Armed with Lemma 30, we get the following theorem:

Theorem 31. *Assume the uniform mixing condition is satisfied by the controlled transition kernel K. Let $C_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\|f\|_\infty$. When both players follow stationary policies $\boldsymbol{\gamma} = \bar{P}_T$ and $\boldsymbol{\delta} = \bar{f}_T$ derived from regret minimization strategies in G1, we*

have

$$\left|\frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right] - V_T(x)\right| \leqslant \max\{M_1, M_2\}, \tag{6.10}$$

where $M_1 = \frac{C}{T} + R^1(f_1, \ldots, f_T) + R^2(\nu_1, \ldots, \nu_T) + \frac{2(\tau+1)C_{\mathcal{F}}}{T}$ *and* $M_2 = R^1(f_1, \ldots, f_T) + R^2(\nu_1, \ldots, \nu_T) + \frac{4(\tau+1)C_{\mathcal{F}}}{T}$.

Proof. Now both $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are stationary policies. We let μ_t be the actual state distribution at time t. We have

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right] = \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma}}\left[\sum_{t=1}^{T} \bar{f}_T(X_t, U_{1,t})\right] \\
&\leqslant \frac{1}{T}\sum_{t=1}^{T}\left\langle \bar{\nu}_T, \bar{f}_T\right\rangle + \frac{1}{T}\sum_{t=1}^{T}\|\pi_{\bar{P}_T} - \mu_t\|_1\|f\|_\infty \\
&\leqslant \left\langle \bar{\nu}_T, \bar{f}_T\right\rangle + \frac{1}{T}\sum_{t=1}^{T} 2e^{-t/\tau}\|f\|_\infty \\
&\leqslant \sup_f \langle \bar{\nu}_T, f\rangle + \frac{2C_{\mathcal{F}}}{T(1-e^{-1/\tau})} \\
&\leqslant V_T(x) + \frac{C}{T} + R^1(f_1, \ldots, f_T) + R^2(\nu_1, \ldots, \nu_T) + \frac{2(\tau+1)C_{\mathcal{F}}}{T},
\end{aligned}$$

where the second inequality is by the uniform mixing condition, and the last inequal-

ity is by Lemma 30. On the other hand, we can upper-bound $V_T(x)$ by

$$\begin{aligned}
V_T(x) &\leqslant \sup_{\boldsymbol{\delta}} \frac{1}{T}\mathbb{E}_x^{\bar{P}_T,\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right] \\
&= \frac{1}{T}\sup_{f_1,\ldots,f_T} \mathbb{E}_x^{\bar{P}_T}\left[\sum_{t=1}^{T} f_t(X_t, U_{1,t})\right] \\
&\leqslant \frac{1}{T}\sup_{f_1,\ldots,f_T} \sum_{t=1}^{T}\left(\left\langle \pi_{\bar{P}_T}\otimes\bar{P}_T, f_t\right\rangle + \left\langle \mu_t\otimes\bar{P}_T - \pi_{\bar{P}_T}\otimes\bar{P}_T, f_t\right\rangle\right) \\
&\leqslant \frac{1}{T}\sup_{f_1,\ldots,f_T} \sum_{t=1}^{T}\left\langle \pi_{\bar{P}_T}\otimes\bar{P}_T, f_t\right\rangle + \frac{1}{T}\sum_{t=1}^{T}\|\pi_{\bar{P}_T}-\mu_t\|\|f\|_\infty \\
&\leqslant \sup_f \langle \bar{\nu}_T, f\rangle + \frac{2(\tau+1)C_{\mathcal{F}}}{T}.
\end{aligned}$$

We further upper bound it through:

$$\begin{aligned}
V_T(x) &\leqslant \inf_{\nu\in\mathcal{G}}\langle \nu, \bar{f}_T\rangle + R^1(f_1,\ldots,f_T) + R^2(\nu_1,\ldots,\nu_T) + \frac{2(\tau+1)C_{\mathcal{F}}}{T} \\
&\leqslant \left\langle \bar{\nu}_T, \bar{f}_T\right\rangle + R^1(f_1,\ldots,f_T) + R^2(\nu_1,\ldots,\nu_T) + \frac{2(\tau+1)C_{\mathcal{F}}}{T} \\
&\leqslant \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right] + \frac{1}{T}\sum_{t=1}^{T}\|\pi_{\bar{P}_T}-\mu_t\|\|f\|_\infty \\
&\quad + R^1(f_1,\ldots,f_T) + R^2(\nu_1,\ldots,\nu_T) + \frac{2(\tau+1)C_{\mathcal{F}}}{T} \\
&\leqslant \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right] + R^1(f_1,\ldots,f_T) + R^2(\nu_1,\ldots,\nu_T) + \frac{4(\tau+1)C_{\mathcal{F}}}{T},
\end{aligned}$$

where the first inequality is by (6.6), and the second-to-last inequality is by $\left|\left\langle \bar{\nu}_T, \bar{f}_T\right\rangle - \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right]\right| \leqslant \frac{1}{T}\sum_{t=1}^{T}\|\pi_{\bar{P}_T}-\mu_t\|_1\|f\|_\infty$. □

This result quantifies the sub-optimality of stationary polices $\bar{P}_T$ and $\bar{f}_T$, and shows that, when the two players use these near-optimal stationary strategies, the

actual payoff converges to the average minimax payoff $V_T(x)$ for the finite-horizon stochastic game at fast rates.

6.2.2 Stationary strategies for infinite-horizon stochastic games

We can also extend the above result to infinite-horizon stochastic games. Bewley and Kohlberg (1978) have shown that if both players follow optimal stationary strategies in the stochastic game with limit expected average criterion:

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right],$$

or the limit average criterion:

$$\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}} \limsup_{T\to\infty} \frac{1}{T}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right],$$

then these strategies are optimal in both games and the value of the game is

$$V_\infty(x) = \lim_{T\to\infty} V_T(x). \tag{6.11}$$

This result also applies to the weak criterion, where one replaces $\limsup_{T\to\infty}$ by $\liminf_{T\to\infty}$.

We also know that, in an undiscounted single-controller stochastic game, a pair of optimal stationary strategies for both players always exists (Parthasarathy and Raghavan (1981)). In this case, we have the following corollary:

Corollary 32. *When both players follow stationary policies* $\boldsymbol{\gamma} = \bar{P}_T$ *and* $\boldsymbol{\delta} = \bar{f}_T$ *derived from regret minimization strategies, the average payoff* $\frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} c(X_t, U_{1,t}, U_{2,t})\right]$ *converges to the value of the game* $V(x)$ *as* $T \to \infty$.

This corollary shows that we can compute near-optimal stationary strategies for both players by letting both players adopt an online learning algorithm to compete

against their opponent. Compared to traditional methods, where we use a linear program to find the value and optimal stationary strategies, this construction of near-optimal strategies may be more computationally efficient for large-scale tasks, in the sense that we can process the data in a stream rather than a batch.

6.3 Conclusion

We have derived stationary policies from regret minimization strategies for online learning games, and shown that they can be used as efficient dynamics for zero-sum stochastic games to converge to minimax equilibrium. In particular, if regret minimization algorithms achieve sublinear regret, then the convergence to the minimax payoff is also at sublinear rates. This result provides a novel method for approximately solving zero-sum stochastic games by computing near-optimal strategies that approximate the minimax equilibrium efficiently.

7

Conclusion

7.1 Summary of contributions

This dissertation has addressed the problem of online MDPs by providing advances in the following key areas:

- We have provided a unified viewpoint on the design and the analysis of online MDP algorithms by constructing a general framework for deriving algorithms. This framework significantly alleviates the burden of the algorithm design and proving performance guarantees. This has been accomplished by coming up with a sequence of upper bounds on a certain quantity and plugging those upper bounds into a recursively defined system of inequalities. Existing methods are proven to arise from this framework via specific upper bounds, and new algorithms are derived using this framework.

- We have moved away from the common assumption that the cost sequences are generated completely arbitrarily and developed an algorithm that takes advantage of possible "regularity" of the observed sequences. We have proved an optimal regret bound in terms of the time horizon, which only pays for

the unpredictable part of the cost sequence and retains the usual worst-case guarantees if the sequence is indeed arbitrary.

- We have also studied the traditional MDP problem and extended the LP approach to MDPs by adding relative entropy regularization. The goal is to provide additional insights into the convex-analytic methods applied in previous chapters. We have found a tight connection between ergodic occupation measures that minimize the entropy-regularized ergodic control cost and a particular exponential family of probability distributions, which has allowed us to exploit known results about exponential families to provide structural results for the optimizing measure. These properties are then used to establish sensitivity and stability results, which open the door to further in-depth analysis of the online MDP problem, such as optimal adaptation of the learning rate to observed data or efficient approximate implementation of online policy selection algorithms.

- We have proposed a computationally efficient online strategy with sub-optimal regret under mild regularity conditions for a new class of MDPs. A more control-oriented formulation is considered: the emphasis is on steering the system along a desirable state trajectory. The proposed algorithm is much easier to implement and compute than existing online MDP algorithms, and milder assumptions are needed to prove performance guarantees. Along the way, sharp bounds on the sensitivity of optimal control laws are provided. We have also demonstrated the performance of the proposed strategy on a simulated target tracking problem.

- We have transformed online MDPs to single controller stochastic games by making the environment able to respond to the agent's previous actions. We

have derived near-optimal stationary strategies for finite-horizon stochastic games, and the sub-optimality of such strategies is quantified. We have shown that they can be used as efficient dynamics for zero-sum stochastic games to converge to minimax equilibrium. We have also provided a computationally efficient method of approximately addressing the infinite-horizon case

7.2 Future work

This dissertation opens up possible directions for future research.

In Chapter 2, we have shown that, if we construct an admissible relaxation by deriving certain upper bounds on the conditional value and implement the associated behavioral strategy, we will obtain an algorithm that achieves the regret bound corresponding to the relaxation. In principle, this gives us a general framework to develop low-regret algorithms for online MDPs. However, with an additional state variable involved, it is difficult to derive admissible relaxations to bound the conditional value. We bypassed this problem by presenting two alternative frameworks where we decouple the current action from future costs. An interesting prospect for future research is to see if we can derive algorithms directly from the admissibility condition (2.4). This could be possible if we find a way to derive some characteristics of the conditional value from analytical properties of the Shapley operator. We may also apply the idea of empirical dynamic programming (Haskell et al. (2013)) to approximate the conditional value. If we can obtain a sub-linear and computationally feasible upper bound of the minimax value, it would be possible to explore the optimal strategy for the adversary.

In Chapter 4, we have derived various properties of the optimizing Gibbs measure, the corresponding ergodic cost and the relative value function. These results open the door to further in-depth analysis of the online MDP problem, such as optimal adaptation of the learning rate to observed data. It would be interesting to explore

different ways to efficient approximate implementation of online policy selection algorithms. Another possible extension of our work is to introduce some uncertainty about the state transition kernel, which will draw tight connection between our setting and the robust optimization problem.

Finally, in Chapter 5, we have proposed a computationally efficient algorithm following the Todorov set-up. An interesting line of research is to explore if we could improve the regret bound and extend our results to more general state spaces. This will require more sophisticated machinery, e.g., Foster–Lyapunov criteria and ergodicity with respect to weighted norms (Hernández-Lerma and Lasserre (2003); Meyn and Tweedie (2009)), as well as the spectral theory of the MPE for Markov chains with general state spaces (Kontoyiannis and Meyn (2003)). Another promising avenue for future research is related to the apparent duality between our set-up and the theory of *risk-sensitive control* of Markov processes (Hernández-Hernández and Marcus (1996); Fleming and Hernández-Hernández (1997); Moldovan and Abbeel (2012)).

Bibliography

Abbasi-Yadkori, Y., Bartlett, P. L., Kanade, V., and Seldin, Y. (2013), "Online learning in Markovian Decision Processes with adversarially chosen transition probability distributions," *Adv. Neural Inform. Processing Systems*.

Abernethy, J. D., Hazan, E., and Rakhlin, A. (2012), "Interior-point methods for full-information and bandit online learning," *IEEE Trans. Inform. Theory*, 58, 4164–4175.

Anthony, M. and Bartlett, P. L. (1999), *Neural Network Learning: Theoretical Foundations*, Cambridge University Press.

Arapostathis, A., Borkar, V. S., Fernández-Gaucherand, E., Ghosh, M. K., and Marcus, S. I. (1993), "Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM J. Control Optim.*, 31, 282–344.

Arora, R., Dekel, O., and Tewari, A. (2012), "Deterministic MDPs with adversarial rewards and bandit feedback," in *Proceedings of the 28th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 93–101, AUAI Press.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002), "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, 32, 48–77.

Başar, T. and Olsder, G. J. (1999), *Dynamic Noncooperative Game Theory*, SIAM, Philadelphia, PA, 2nd edn.

Balaji, S. and Meyn, S. P. (2000), "Multiplicative ergodicity and large deviations for an irreducible Markov chain," *Stochastic Process. Appl.*, 90, 123–144.

Bertsekas, D. P. (2001), *Dynamic Programming and Optimal Control*, Athena Scientific, Nashua, NH.

Bertsekas, D. P. and Rhodes, I. B. (1973), "Sufficiently informative functions and the minimax feedback control of uncertain dynamic systems," *IEEE Trans. Automat. Control*, 18, 117–124.

Bewley, T. and Kohlberg, E. (1978), "On stochastic games with stationary optimal strategies," *Math. Oper. Res.*, 3.

Birkhoff, G. D. (1931), "Proof of the ergodic theorem," *Proc. Natl. Acad. Sci. USA*, 17, 656–660.

Borkar, V. S. (1988), "A convex analytic approach to Markov decision processes," *Probab. Th. Rel. Fields*, 78, 583–602.

Borkar, V. S. (2002), "Convex analytic methods in Markov decision processes," in *Handbook of Markov decision processes*, Kluwer Academic Publishers.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004), "Introduction to Statistical Learning Theory," in *Advanced Lectures in Machine Learning*, Springer.

Bubeck, S. and Cesa-Bianchi, N. (2012), "Regret analysis of stochastic and nonstochastic multiarmed bandit problems," *Foundations and Trends in Machine Learning*, to appear.

Bubeck, S. and Eldan, R. (2015), "The entropic barrier: a simple and optimal universal self-concordant barrier," *In proceedings of the 28th Annual Conference on Learning Theory (COLT).*

Cappé, O., Moulines, E., and Rydén, T. (2005), *Inference in Hidden Markov Models*, Springer Series in Statistics., Springer, New York.

Cesa-Bianchi, N. and Lugosi, G. (2006), *Prediction, Learning and Games*, Cambridge Univ. Press.

Chanchana, P. (2007), "An algorithm for computing the Perron root of a nonnegative irreducible matrix," Ph.D. thesis, North Carolina State University, Raleigh, NC.

Charalambous, C. D. and Rezaei, F. (2007), "Stochastic uncertain systems subject to relative entropy constraints: induced norms and monotonicity properties of minimax games," *IEEE Trans. Automat. Control*, 52, 647–660.

Cover, T. M. and Thomas, J. A. (2006), *Elements of Information Theory*, Wiley, 2 edn.

Csiszár, I. (1975), "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, 3, 146–158.

Daskalakis, C., Deckelbaum, A., and Kim, A. (2011), "Near-optimal no-regret algorithms for zero-sum games," *In Proceedings of the 22nd Annual ACM-SIAM symposium on Discrete Algorithms*, pp. 235–254.

Di Masi, G. B. and Stettner, L. (1999), "Risk-sensitive control of discrete-time Markov processes with infinite horizon," *SIAM J. Control Optim.*, 38, 61–78.

Dick, T., György, A., and Szepesvári, C. (2014), "Online learning in Markov decision processes with changing cost sequences," *ICML*.

Ellis, R. S. (1985), *Entropy, Large Deviations, and Statistical Mechanics*, Springer.

Elsner, L. (1976), "Inverse iteration for calculating the spectral radius of a non-negative irreducible matrix," *Lin. Algebra Appl.*, pp. 235–242.

Even-Dar, E., Kakade, S. M., and Mansour, Y. (2009), "Online Markov decision processes," *Math. Oper. Res.*, 34, 726–736.

Fleming, W. H. and Hernández-Hernández, D. (1997), "Risk-sensitive control of finite state machines on an infinite horizon I," *SIAM J. Control Optim.*, 35, 1790–1810.

Guan, P., Raginsky, M., and Willett, R. (2012), "Online Markov decision processes with Kullback–Leibler control cost," In Proceedings of American Control Conference.

Guan, P., Raginsky, M., and Willett, R. (2014a), "From minimax value to low-regret algorithms for online Markov decision processes," *In Proceedings of American Control Conference*.

Guan, P., Raginsky, M., and Willett, R. (2014b), "Online Markov decision processes with Kullback-Leibler control cost," *IEEE Trans. Automat. Control*, 59, 1423–1438.

Guan, P., Raginsky, M., and Willett, R. (2015), "Relax but stay in control: from value to algorithms for online Markov decision processes," *Submitted to Machine Learning*.

Hannan, J. (1957), "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games*, vol. 3, pp. 97–139, Princeton Univ. Press.

Hansen, L. P. and Sargent, T. J. (2008), *Robustness*, Princeton University Press.

Haskell, W. B., Jain, R., and Kalathil, D. (2013), "Empirical dynamic programming," *http://arxiv.org/1311.5918*.

Hernández-Hernández, D. and Marcus, S. I. (1996), "Risk sensitive control of Markov processes in countable state space," *Systems Control Lett.*, 29, 147–155.

Hernández-Lerma, O. (1989), *Adaptive Markov Control Processes*, Springer.

Hernández-Lerma, O. and Lasserre, J. B. (1996), *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer.

Hernández-Lerma, O. and Lasserre, J. B. (2003), *Markov Chains and Invariant Probabilities*, Birkhäuser.

Hiriart-Urruty, J. B. and Lemarechal, C. (2013), *Convex analysis and minimization algorithms II: advanced theory and bundle methods*, Springer Science & Business Media.

Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, 58, 13–30.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996), "Reinforcement learning: a survey," *J. Artif. Intel. Res.*, pp. 237–285.

Kalai, A. and Vempala, S. (2005), "Efficient algorithms for online decision problems," *J. Comput. Sys. Sci.*, 71, 291–307.

Kappen, B., Gomez, V., and Opper, M. (2012), "Optimal control as a graphical model inference problem," *Machine Learning*, 87, 159–182.

Kárný, M. (1996), "Towards fully probabilistic control design," *Automatica*, 32, 1719–1722.

Kontoyiannis, I. and Meyn, S. P. (2003), "Spectral theory and limit theorems for geometrically ergodic Markov processes," *Ann. Appl. Probab.*, 13, 304–362.

Littlestone, N. and Warmuth, M. K. (1994), "The weighted majority algorithm," *Inform. Comput.*, 108, 212–261.

Manne, A. S. (1960), "Linear programming and sequential decisions," *Management Science*, 6, 259–267.

McMahan, H. (2003), "Planning in the presence of cost functions controlled by an adversary," *The 20th International Conference on Machine Learning*, pp. 536–543.

Merhav, N., Ordentlich, E., Seroussi, G., and Weinberger, M. J. (2002), "On sequential strategies for loss functions with memory," *IEEE Trans. Inform. Theory*, 48, 1947–1958.

Mertens, J. F. and Neyman, A. (1981), "Stochastic games," *International Journal of Game Theory*, 10, 53–66.

Meyn, S. (2008), *Control techniques for complex networks*, Cambrdige Univ. Press.

Meyn, S. and Tweedie, R. L. (2009), *Markov Chains and Stochastic Stability*, Cambrdige Univ. Press, 2nd edn.

Meyn, S. P. (2007), *Control Techniques for Complex Networks*, Cambridge Univ. Press.

Mitter, S. K. and Newton, N. J. (2003), "A variational approach to nonlinear estimation," *SIAM J. Control Optim.*, 42, 1813–1833.

Moldovan, T. M. and Abbeel, P. (2012), "Risk aversion in Markov decision processes via near-optimal Chernoff bounds," *Adv. Neural Inform. Processing Systems*.

Nemirovski, A. S. and Todd, M. J. (2008), "Interior-point methods for optimization," *Acta Numerica*, pp. 191–234.

Nesterov, Y. and Nemirovski, A. (1994), "Interior-point polynomial algorithms in convex programming," *SIAM*.

Neu, G., György, A., Szepesvári, C., and Antos, A. (2010), "Online Markov decision processes under bandit feedback," in *Advances in Neural Information Processing Systems 23*, eds. J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, pp. 1804–1812.

Neu, G., György, A., Szepesvári, C., and Antos, A. (2014), "Online Markov decision processes under bandit feedback," *IEEE Trans. Automat. Control*, 59, 676–691.

Neumann, J. V. (1928), "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen*, 100, 295–320.

Parthasarathy, T. and Raghavan, T. (1981), "An order field property for stochastic games when one player controls transition probabilities," *J. Opt. Theory. Appl*, pp. 375–392.

Peters, J., Ḿulling, K., and Alt́un, Y. (2010), *Relative entropy policy search*, In National Conference on Artifical Intelligence (AAAI).

Petersen, I. R., James, M. R., and Dupuis, P. (2000), "Minimax optimal control of stochastic systems with relative entropy constraints," *IEEE Trans. Automat. Control*, 45, 398–412.

Puterman, M. (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley.

Rakhlin, A. and Sridharan, K. (2013a), "Online learning with predictable sequences," *In proceedings of the 26th Annual Conference on Learning Theory (COLT)*.

Rakhlin, A. and Sridharan, K. (2013b), "Optimization, Learning, and Games with predictable sequences," *Adv. Neural Inform. Processing Systems*.

Rakhlin, A., Sridharan, K., and Tewari, A. (2010), "Online learning: random averages, combinatorial parameters, and learnability," *Adv. Neural Inform. Processing Systems*.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012), "Relax and randomize: from value to algorithms," *Adv. Neural Inform. Processing Systems*.

Robbins, H. (1951), "Asymptotically subminimax solutions of compound statistical decision problems," in *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability 1950*, pp. 131–148, University of California Press, Berkeley, CA.

Robbins, H. (1952), "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, pp. 527–535.

Saloff-Coste, L. and Zuniga, J. (2009), "Merging for time inhomogeneous finite Markov chains. I. Singular values and stability," *Electron. J. Probab*, 14, 1456–1494.

Seneta, E. (2006), *Nonnegative Matrices and Markov Chains*, Springer.

Shalev-Shwartz, S. (2011), "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, 4, 107–194.

Shalev-Shwartz, S. and Singer, Y. (2007), "A primal-dual perspective of online learning algorithms," *Machine Learning*, 69, 115–142.

Shapley, L. S. (1953), "Stochastic games," *Proceedings of the National Academy of Sciences*, 39, 1095–1100.

Sorin, S. (2002), *A first course on zero-sum repeated games*, Springer.

Streater, R. F. (2009), *Statistical Dynamics: A Stochastic Approach to Nonequilibrum Thermodynamics*, Imperial College Press, London, 2nd edn.

Todorov, E. (2007), "Linearly-solvable Markov decision problems," in *Advances in Neural Information Processing Systems 19*, eds. B. Schölkopf, J. Platt, and T. Hoffman, pp. 1369–1376, MIT Press, Cambridge, MA.

Todorov, E. (2008), "General duality between optimal control and estimation," in *Proc. 47th IEEE Conf. on Decision and Control*, pp. 4286–4292.

Todorov, E. (2009), "Efficient computation of optimal actions," *Proc. Nat. Acad. Sci.*, 106, 11478–11483.

Tsitisklis, J. N. (2007), "NP-hardness of checking the unichain condition in average cost MDPs," *Oper. Res. Lett.*, 35, 319–323.

Tsitsiklis, J. N. (1994), "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, 16, 185–202.

Vapnik, V. (1998), *Statistical Learning Theory*, John Wiley, New York.

Vovk, V. G. (1990), "Aggregating strategies," in *Proc. 3rd Annual Workshop on Computational Learning Theory*, pp. 372–383, San Mateo, CA.

Šindelář, J., Vajda, I., and Kárný, M. (2008), "Stochastic control optimal in the Kullback sense," *Kybernetika*, 44, 53–60.

Wainwright, M. J. and Jordan, M. I. (2008), "Graphic models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 1.

Watkins, C. J. C. H. and Dayan, P. (1992), "Q-learning," *Machine Learning*, 8, 279–292.

Yu, J. Y. and Mannor, S. (2009), "Online learning in Markov decision processes with arbitrarily changing rewards and transitions," *GameNets*.

Yu, J. Y., Mannor, S., and Shimkin, N. (2009), "Markov decision processes with arbitrary reward processes," *Math. Oper. Res.*, 34, 737–757.

Zimin, A. and Neu, G. (2013), "Online learning in episodic Markovian Decision Processes by relative entropy policy search," *Adv. Neural Inform. Processing Systems*.

Biography

Peng Guan was born in Beijing, China on June 10, 1984. He received the B.E. and M.Sc. degrees in Department of Automation from Tsinghua University, Beijing, China, in 2006 and 2009, respectively. In the fall of 2010, he began to study towards a Ph.D. degree in Electrical and Computer Engineering at Duke University, under the guidance of Prof. Rebecca Willett and Prof. Maxim Raginsky. His research interests include optimization, signal processing, stochastic control, online learning and reinforcement learning theory.