Online Learning in Markovian Decision Processes

By
Alexander Zimin

Submitted to
Central European University
Department of Mathematics and its Applications

In partial fulfilment of the requirements for the degree of
Master of Science

Supervisor: Dr. László Györfi

Budapest, Hungary
2013

Table of Contents

Introduction

The theory of Markovian Decision Processes (MDPs) provide a popular framework to model operation research and control problems. As an illustrative example, consider an inventory control problem (Puterman (1994); Szepesvári (2010)). We control an inventory of a fixed maximal size and every day we have to order the quantity for the next day and, afterwards, we observe stochastic demand. Our revenue on each day depends on the demand and on the prices at which products are sold. The goal is to maximize the expected total future income. There are several key features that distinguishes this problem. Among them there are the sequential fashion of decision making, the stochastic nature of the reward (revenue) we receive and the cumulative form of the objective we want to maximize.

In general, Markovian Decision Process models the interaction between an agent (inventory manager) and the environment (market). In each time step (day) the agent observes the current state of the environment (the current size of the inventory, past prices and demands) and then makes a decision about the next action (order) that it sends to the environment. The environment then samples the new state based on the previous one and the action received (this corresponds to the assumption that demand and prices change stochastically). The agent then observes the reward (or, alternatively, the loss) for this round. The goal of an agent is to maximize (minimize, in case of losses) the expected cumulative reward (loss) up to some finite horizon.

MDPs have been successfully applied to many real world problems (Sutton and Barto (1998)). However, there are some aspects of real life that Markovian models fail to address. For example, in the inventory control problem, prices can depend on a lot of unobservable events and, thus, change arbitrarily. To address this issue, Even-Dar et al. (2004, 2009) propose a way to relax the assumption that the states are completely Markovian (i.e. that the next state of the environment depends on the previous state and action). Instead, they assume that there are external factors that can not be modelled, however, they only influence the rewards that the agent receives, but not the transitions between the states. This naturally leads to a formulation of the problem that enjoy similar properties with the framework of online prediction with expert advice (or, in short, experts framework).

In the experts framework (Cesa-Bianchi and Lugosi (2006)) we face the sequential decision problem of choosing an decision-making experts from some finite set. Afterwards, we obtain the reward of the expert chosen and the goal is to perform not worse than the best fixed expert chosen in hindsight. If we allow arbitrary rewards in MDP, then it is natural to aim to perform not worse than some reference class of agents. A good reference class, as proposed in Even-Dar et al. (2004, 2009), is the class of so-called stochastic stationary policies, since there is always a member of this class that achieves the best possible performance (Puterman (1994)). The difference between our cumulative reward and the cumulative reward of the best member of the class is called a regret. Therefore,

following the ideas of the experts framework, we set the goal of minimizing the regret. The resulting problem is called an online learning problem in MDPs.

In the experts problems the performance guarantees for the algorithms are stated as bounds on the regret. These bounds depend on different parameters, but the one that we are interested the most is the dependence on the total number of the time steps T (which is fixed in advance). The desired property of the algorithms is Hannan-consistency, i.e. a sublinear growth of the regret, which guarantees vanishing average regret. However, the rate of growth is also a quantity of the interest.

Bringing some more ideas from the experts framework, we introduce two types of the main problem: the full-information and the bandit one. They differ in the amount of the information available to the learner (agent) after each round of the interaction. In the full-information case the agent observes the rewards for all state-action pairs that could possibly occur. It is obvious that this is a very strong requirement, since, for example, in the inventory control problem this corresponds to the knowledge of the demand and prices if we would have chosen a different amount to order. Hence, we also need to consider the so-called bandit version, when we observe only the reward for the actual state and action. However, even if quite unrealistic, the full-information case is an important problem, since it provides us with the useful insight to the problem and usually serves as an intermediate step towards the solution of the bandit case.

In their seminal paper Even-Dar et al. (2004, 2009) consider the full-information case and the provide an algorithm that obtains the optimal $\mathcal{O}(\sqrt{T})$ bound. Later, Yu et al. (2009) present a new algorithm and prove $\mathcal{O}(T^{3/4+\epsilon})$ bound for the full-information case. While the regret of their algorithm is higher, it has less computation complexity.

A related work is due to Yu and Mannor (2009a,b). Their problem is more general than ours, since they assume that the transition probabilities can also change arbitrarily (while we require them to be fixed). In addition, as pointed out in Neu et al. (2010a), their analysis seems to have gaps and, therefore, we can rely only on the Hannan-consistent property of their algorithm.

In the bandit setting the first Hannan-consistent algorithm is given by Yu et al. (2009). Subsequently, Neu et al. (2010b) extend the algorithm of Even-Dar et al. (2004, 2009) to the bandit case and prove suboptimal $\mathcal{O}(T^{2/3})$ bound.

In this work we also consider another modification of the original problem, which is called an online stochastic shortest path problem and can be thought of as an extension of the stochastic shortest path problem (Bertsekas and Tsitsiklis (1996)) to the online setting and as a stochastic version of the online shortest path problem (György et al. (2007)). In this problem the interaction between the agent and the environment is divided into episodes, and the rewards are allowed to change arbitrarily between the episodes. This is an important problem, since it naturally models some real life situation like the routing in virtual networks. The problem was first considered by Neu et al. (2010a). They introduce loop-free assumption on MDP and provide an algorithm that achieves $\mathcal{O}(L^2\sqrt{T})$ bound on the regret in the full-information case, where L is the number of layers. Assuming that all states are reachable with probability $\alpha > 0$ under all policies, they also show $\mathcal{O}(L^2\sqrt{T}/\alpha)$ bound in the bandit case. A related problem in the full-information case is presented by Neu et al. (2012) with the difference that they relax the assumption that the stochastic model of the environment is known. They give an algorithm with an $\mathcal{O}(L\sqrt{T})$ bound on the regret.

The motivation for this thesis is to study the application of the Proximal Point Algorithm to the problems described. Our work is inspired by the insightful article of Peters

et al. (2010), who present a new approach called REPS. Motivated by the need to constrain the information loss, they derive the algorithm and compare it numerically with the existing approaches. Our study shows that the idea the authors came up with is an instance of the Proximal Point Algorithm, a deeply investigated algorithm from the field of online linear optimization.

The primal idea of the algorithm is, at each time step, to choose a policy which maximizes the previously observed reward (on average) and, at the same time, is not too far from the previous choice. We present the complete theory underlying the algorithm and, at first, show how to apply it to the both versions of the online stochastic shortest path problem. We show that it achieves the regret of order $\mathcal{O}(L\sqrt{T})$ in both cases, which is a vast improvement over previously known result in terms of the dependence on L and α. Next, we present the O-REPS algorithm, the application of the Proximal Point Algorithm to the online learning problem in more general MDPs. We prove the optimal $\mathcal{O}(\sqrt{T})$ order of the regret, but with smaller additional terms than in the previously known results.

The rest of the thesis is structured as follows. In Chapter 1 we present the theory underlying the Proximal Point Algorithm and show how it is applied to the online linear optimization and multi-armed bandit problems. In Chapter 2 we present the problems described above formally. Then we show how the algorithm works in episodic problems and prove the corresponding regret bounds. Finally, we derive the closed form of O-REPS for the online learning in MDPs and show its performance guarantees.

Chapter 1

Online linear optimization

The problem we consider in this chapter comes as a natural generalization of a multi-armed bandit problem. Bandit online linear optimization is a very general framework that is used not only in applications, but also as a building block in algorithms for more complex problems. We are following the exposition presented in Szepesvári et al. (2011) and Cesa-Bianchi and Lugosi (2006).

1.1 The problem description

The problem is described as follows: there is a learner (algorithm) that interacts with some environment. We do not make any assumptions on the nature of the environment, except that it is non-adaptive, but it can be adversarial in the game theoretic sense. Interaction goes in rounds up to some prescribed finite horizon T. At each round, the environment chooses a loss function $\ell_t(\cdot) : \mathbb{R}^d \to \mathbb{R}$, but does not reveal it to the learner. The algorithm then chooses a vector $d_t \in D$, where $D \subseteq \mathbb{R}^d$ is called a decision set. Then the learner suffers a loss $\ell_t(d_t)$. The goal of the learner is to achieve a small cumulative loss

$$\hat{L}_T = \sum_{t=1}^{T} \ell_t(d_t).$$

We measure our performance by the so-called *regret*. First, let $L_T(p)$ be a cumulative loss of a fixed point $p \in D$

$$L_T(p) = \sum_{t=1}^{T} \ell_t(p).$$

Then we define the regret $\mathcal{R}_T(p)$ with respect to some fixed point p

$$\mathcal{R}_T(p) = \hat{L}_T - L_T(p).$$

Therefore, the ultimate goal is to minimize the following notion of the regret, which is equivalent to minimizing the learner's cumulative loss

$$\mathcal{R}_T = \sup_{p \in D} \mathcal{R}_T(p) = \hat{L}_T - \inf_{p \in D} L_T(p).$$

The important issue is the information available to the learner at the end of each round. We distinguish between two important cases of the problem. The first one is the full-information case, when the learner receives the loss function $\ell_t(\cdot)$ itself. The second one

Parameters: decision set D, finite horizon T.
Environment chooses points $f_1, ..., f_T \in \mathbb{R}^d$.
for $t = 1$ **to** T:

1. Learner chooses point $d_t \in D$.
2. f_t (full information) or $\langle d_t, f_t \rangle$ (bandit information) is revealed.
3. Learner suffers loss $\langle d_t, f_t \rangle$.

Figure 1.1: (Bandit) linear online optimization problem

is the bandit case, when the algorithm learns only the loss in the point chosen, that is, $\ell_t(d_t)$.

The problem described above is a (bandit) online optimization problem. We will be interested in the particular case of it when $\ell_t(d) = \langle d, f_t \rangle$, where $f_t \in \mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ denotes a dot product ($\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i$, where x_i and y_i are the components of the corresponding vectors).

We already mentioned that we assume that the environment is non-adaptive. In other words, we disallow the environment to choose loss functions at round t based on the choices of the algorithm in the previous rounds, i.e. on $d_1, .., d_{t-1}$. In this case, we can assume that all the loss functions are chosen in advance, before the actual interaction starts. The summary of the resulting problem is given in Figure 1.1.

In the next sections we will describe a Proximal Point Algorithm that solves this problem and which will be a core for our approaches for solving MDP problems. This method was originally discovered in the context of convex optimization by Martinet (1970). To define it and to derive the regret bounds, we will need some definitions and results from convex analysis.

1.2 Bregman divergences

The Proximal Point Algorithm (PPA) is based on the notion of *Bregman divergences*. They were introduced in Bregman (1967) as a base for the method of finding common points of convex sets. First, we need a definition of a *Legendre function*.

Definition 1.1 (Legendre function)**.** A function $R : A \to \mathbb{R}$ is called a Legendre function if it satisfies the following conditions.

1. $A \subseteq \mathbb{R}^d$, $A \neq \emptyset$ and A° is convex (A° denotes the interior of A)
2. R is strictly convex
3. partial derivatives $\frac{\partial R}{\partial x_i}$ exist and are continuous for all $i = 1, .., d$
4. any sequence $(y_n) \in A$ converging to a boundary point of A satisfies

$$\lim_{n \to \infty} ||\nabla R(y_n)|| = \infty.$$

In the classical text of Rockafellar (1970) such functions are called essentially smooth. The strict convexity and "blowing up" of the gradient at the border ensures the existence

and uniqueness of the minimum of R inside A, a property that will be of big importance for PPA. In the definition we do not specify the norm used, since the convergence in one norm implies the convergence in any other norm due to the norm equivalence in $\mathbb{R}^d$.

Definition 1.2 (Bregman divergence)**.** Let $R : A \to \mathbb{R}$ be a Legendre function, then the Bregman divergence corresponding to R is a function $D_R : A \times A^\circ \to \mathbb{R}$ defined by

$$D_R(u, v) = R(u) - R(v) - \langle \nabla R(v), u - v \rangle.$$

In other words, the Bregman divergence is just the difference between function $R(u)$ and its first-order Taylor expansion around point v. It can also be seen as a generalization of a notion of a "distance" to arbitrary convex functions. In fact, the squared euclidean distance is a particular case of a Bregman divergence.

Example 1.1. Let us consider the simplest case when $d = 1$. If we take $R(u) = u^2$ and $A = \mathbb{R}$, then the corresponding divergence is $D_R(u, v) = u^2 - v^2 - 2v(u - v) = (u - v)^2$.

Example 1.2. Now turn to d-dimensional space. Take $R(u) = ||u||_2^2$ and $A = \mathbb{R}^d$, then $D_R(u, v) = ||u||_2^2 - ||v||_2^2 - \langle 2v, u - v \rangle = ||u - v||_2^2$.

Example 1.3. The next divergence will be of great importance for us. We start with $R(u) = \sum_{i=1}^d u_i \ln u_i - \sum_{i=1}^d u_i$. This R is called *un-normalized negative entropy* and it is a Legendre function on $A = \mathbb{R}^d_+ = \left\{x \in \mathbb{R}^d : x_i > 0, i = 1, .., d\right\}$. The corresponding Bregman divergence is the *un-normalized Kullback-Leibler divergence*.

$$\begin{aligned} D_R(u, v) &= \sum_{i=1}^d u_i \ln u_i - \sum_{i=1}^d u_i - \sum_{i=1}^d v_i \ln v_i + \sum_{i=1}^d v_i - \sum_{i=1}^d \ln v_i (u_i - v_i) \\ &= \sum_{i=1}^d u_i \ln \frac{u_i}{v_i} + \sum_{i=1}^d (v_i - u_i). \end{aligned}$$

As we already mentioned, Bregman divergences enjoy some similar properties as metrics. These properties are summarized in the following proposition.

Proposition 1.1. *Let $R : A \to \mathbb{R}$ be a Legendre function and $D_R(u, v)$ is the corresponding Bregman divergence. Then the following holds*

1. *$D_R(u, v) \geq 0$ for all $u \in A, v \in A^\circ$*
2. *$D_R(u, v) = 0 \iff u = v$*

Proof. From its definition, the Bregman divergence is the difference between $R(u)$ and its first-order Taylor expansion of R around v. Then the first point follows from the convexity of R.

Since R is also strictly convex, it is equal to its linear approximation around v only in that very point, which is the content of the second claim. □

However, the Bregman divergence is not a metric, since it is not symmetric, and the triangle inequality does not hold. The next proposition establishes a connection between the divergences of three arbitray points. This can be thought of as a generalization of "law of cosines".

Proposition 1.2 (Generalized law of cosines)**.** *Let $R : A \to \mathbb{R}$ be a Legendre function. Then for all $u \in A$, and $v, w \in A^\circ$*

$$D_R(u, v) + D_R(v, w) = D_R(u, w) + \langle \nabla R(w) - \nabla R(v), u - v \rangle.$$

Proof. The statement can be proved by the following direct computation:

$$\begin{aligned} D_R(u, v) + D_R(v, w) &= R(u) - R(v) - \langle \nabla R(v), u - v \rangle \\ &\quad + R(v) - R(w) - \langle \nabla R(w), v - w \rangle \\ &= R(u) - R(w) - \langle \nabla R(w), u - w \rangle + \langle \nabla R(w), u - w \rangle \\ &\quad - \langle \nabla R(v), u - v \rangle - \langle \nabla R(w), v - w \rangle \\ &= D_R(u, w) + \langle \nabla R(w), u - v \rangle - \langle \nabla R(v), u - v \rangle \\ &= D_R(u, w) + \langle \nabla R(w) - \nabla R(v), u - v \rangle. \end{aligned}$$

□

An interesting property of a Legendre function is that if we add a linear function to it, the resulting function will also be Legendre with the same divergence. In particular, this implies that Bregman divergences of Legendre functions are Legendre functions themselves.

Proposition 1.3. *Let $R : A \to \mathbb{R}$ be a Legendre function. For any $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$ define $\hat{R}(u) = R(u) + \langle \alpha, u \rangle + \beta$. Then the following holds*

1. *$\hat{R}$ is a Legendre function*
2. *$D_R(u, v) = D_{\hat{R}}(u, v)$*

Proof. The domain of $\hat{R}$ is the same as the domain of R, hence the first condition in the definition of Legendre function is satisfied. $\hat{R}$ is a combination of strictly convex and linear functions, hence it is strictly convex. From the definition of $\hat{R}$ we can compute

$$\nabla \hat{R}(u) = \nabla R(u) + \alpha.$$

We conclude that the continuity of partial derivatives is preserved.

For the proof of the fourth condition take a sequence $(y_n) \in A$ that approaches the boundary of A. Then we have

$$\begin{aligned} ||\nabla \hat{R}(y_n)|| &= ||\nabla R(y_n) + \alpha|| \\ &\geq ||\nabla R(y_n)|| - ||\alpha||, \end{aligned}$$

and because $||\alpha||$ is a constant, the $\lim_{n\to\infty} ||\nabla R(y_n)|| = \infty$ implies $\lim_{n\to\infty} ||\nabla \hat{R}(y_n)|| = \infty$. This concludes the proof that $\hat{R}$ is Legendre.

The second claim follows from direct computation:

$$\begin{aligned} D_{\hat{R}}(u, v) &= \hat{R}(u) - \hat{R}(v) - \langle \nabla \hat{R}(v), u - v \rangle \\ &= R(u) + \langle \alpha, u \rangle + \beta - R(v) - \langle \alpha, v \rangle - \beta - \langle \nabla R(u) + \alpha, u - v \rangle \\ &= R(u) - R(v) - \langle \nabla R(v), u - v \rangle \\ &= D_R(u, v). \end{aligned}$$

□

Corollary 1.1. *Let $R : A \to \mathbb{R}$ be a Legendre function. Then $D_R(\cdot, v)$ is a Legendre function for any fixed $v \in A^\circ$.*

The next important notion for PPA is a *Bregman projection*. We start with a definition.

Definition 1.3 (Bregman projection)**.** Let $R : A \to \mathbb{R}$ be a Legendre function and $K \subseteq \mathbb{R}^d$ be a closed convex set, such that $K \cap A \neq \emptyset$. Then a Bregman projection corresponding to R and K is a function $\Pi_{R,K} : A^\circ \to K \cap A$ defined by

$$\Pi_{R,K}(w) = \operatorname*{argmin}_{u \in K \cap A} D_R(u, w).$$

It is not clear from the definition if is it well-defined or not. This is the content of the following lemma.

Lemma 1.1. *For all Legendre functions $R : A \to \mathbb{R}$ with bounded partial level sets, i.e. with $\{u \in A : D_R(u, v) \leq \alpha\}$ bounded for all $v \in A^\circ$ and all $\alpha \in \mathbb{R}$, for all closed convex sets $K \subseteq \mathbb{R}^d$ such that $A \cap K \neq \emptyset$, and for all $w \in A^\circ$, the Bregman projection of w onto K exists and unique.*

As noted in Bauschke and Borwein (1997), the boundedness of partial level sets is not required for this lemma. However, in this case the proof is complex and based on some advanced tools from convex analysis. We present the simpler version from Censor and Zenios (1998), which is sufficient for our needs, since our functions of interest satisfy the stated requirement.

Proof of Lemma 1.1. We start with the proof of existence. First, we observe that $D_R(\cdot, w)$ is a continuous function, since it is a combination of continuous and linear functions. Then fix any $v \in A \cap K$. The set

$$L = \{u \in A : D_R(u, w) \leq D_R(v, w)\}$$

is bounded by the assumption. It is also closed because of the continuity of $D_R(\cdot, w)$.

Now we define another set $B = A \cap K \cap L$. From the facts that $v \in A \cap K$ and $v \in L$ it follows that $v \in B$. So B is non-empty. B is also bounded, because it is a subset of bounded set L and B is closed as an intersection of closed sets. We conclude that B is compact, therefore, by the extreme value theorem, the continuous function $D_R(\cdot, w)$ attains its infimum over B in $w' = \operatorname{argmin}_{u \in B} D_R(u, w)$.

For any $u \in A \cap K$ outside B, i.e. such that $u \notin L$, $D_R(w', w) < D_R(u, w)$ by the definition of L. Hence, $w' = \operatorname{argmin}_{u \in A \cap K} D_R(u, w)$.

The next thing we need to show is uniqueness. Assume there are $u, v \in A \cap K$, such that

$$D_R(u, w) = D_R(v, w) = \min_{x \in A \cap K} D_R(x, w)$$

and $u \neq v$. Then the point $\frac{u+v}{2} \in A \cap K$ by the convexity of $A \cap K$. From strict convexity of R we get

$$\begin{aligned} D_R(\frac{u+v}{2}, w) &= R(\frac{u+v}{2}) - R(w) - \langle \nabla R(w), \frac{u+v}{2} - w \rangle \\ &< \frac{1}{2}R(u) + \frac{1}{2}R(v) - R(w) - \frac{1}{2}\langle \nabla R(w), u - w \rangle - \frac{1}{2}\langle \nabla R(w), v - w \rangle \\ &= \frac{1}{2}D_R(u, w) + \frac{1}{2}D_R(v, w) \\ &= \min_{x \in A \cap K} D_R(x, w). \end{aligned}$$

Thus, we arrived at contradiction, since the value of $D_R(\cdot, w)$ in $\frac{u+v}{2}$ can not be smaller than the minimum over the whole set. $\square$

Example 1.4. Of course, $R(u) = ||u||_2^2$ has bounded partial level sets. In this case these sets take the next form for some $v \in \mathbb{R}^d$

$$\left\{u \in A : ||u - v||_2^2 \leq \alpha\right\},$$

and these sets are just balls in $\mathbb{R}^d$.

Example 1.5. In this example we will show that un-normalized negative entropy also satisfies conditions of Lemma 1.1. For this we need log sum inequality, which can be found in Cover and Thomas (1991), for example. It states that for any $u, v \in \mathbb{R}^d_+$

$$\sum_{i=1}^{d} u_i \ln \frac{u_i}{v_i} \geq \left(\sum_{i=1}^{d} u_i\right) \ln \frac{\sum_{i=1}^{d} u_i}{\sum_{i=1}^{d} v_i}$$

Just for reminder, $R(u) = \sum_{i=1}^{d} u_i \ln u_i - \sum_{i=1}^{d} u_i$ and $A = \mathbb{R}^d_+$. Since we need to show boundedness for some norm (which would imply boundedness in any norm), we will use ℓ_1-norm and note that on $\mathbb{R}^d_+$: $||u||_1 = \sum_{i=1}^{d} |u_i| = \sum_{i=1}^{d} u_i$. Again we are proving the boundedness of sets in the following form (for some $v \in A^\circ$)

$$\left\{u \in A : \sum_{i=1}^{d} u_i \ln \frac{u_i}{v_i} + \sum_{i=1}^{d} (v_i - u_i) \leq \alpha\right\}.$$

To prove boundedness, we can prove that if $||u||_1 \to \infty$, then $D_R(u, v) \to \infty$. Using the log sum inequality

$$\begin{aligned} D_R(u, v) &= \sum_{i=1}^{d} u_i \ln \frac{u_i}{v_i} + \sum_{i=1}^{d} (v_i - u_i) \\ &\geq ||u||_1 \ln \frac{||u||_1}{||v||_1} + ||v||_1 - ||u||_1 \\ &= ||u||_1 \left(\ln \frac{||u||_1}{||v||_1} - 1\right) + ||v||_1. \end{aligned} \tag{1.1}$$

And when $||u||_1 \to \infty$, (1.1) goes to infinity.

Since Lemma 1.1 holds in general case, further we will omit the requirement for bounded partial level sets. In the next few lemmas we present the important properties of Bregman projections.

Lemma 1.2 (Generalized pythagorean inequality)**.** *Let $R : A \to \mathbb{R}$ be a Legendre function. For all closed and convex sets $K \subseteq \mathbb{R}^d$, such that $K \cap A \neq \emptyset$*

$$D_R(u, w) \geq D_R(u, \Pi_{R,K}(w)) + D_R(\Pi_{R,K}(w), w)$$

for all $u \in K$ and $w \in A^\circ$.

Proof. Denote $w' = \Pi_{R,K}(w)$. Let $F(v) = D_R(v, w) - D_R(v, w')$. Then we have

$$\begin{aligned} F(v) &= D_R(v, w) - D_R(v, w') \\ &= R(v) - R(w) - \langle \nabla R(w), v - w \rangle - R(v) + R(w') + \langle \nabla R(w'), v - w' \rangle \\ &= R(w') - R(w) + \langle \nabla R(w'), v - w' \rangle - \langle \nabla R(w), v - w \rangle. \end{aligned}$$

And we observe that $F(v)$ is a linear function. Hence, if we fix $u \in K$ and take $h(\alpha) = \alpha u + (1 - \alpha)w'$ for $\alpha \in [0, 1]$, then $F(h(\alpha)) = \alpha F(u) + (1 - \alpha)F(w')$. In other words,

$$D_R(h(\alpha), w) - D_R(h(\alpha), w') = \alpha(D_R(u, w) - D_R(u, w')) + (1 - \alpha)D_R(w', w).$$

For $\alpha \neq 0$, this is equivalent to

$$D_R(u, w) - D_R(u, w') - D_R(w', w) = \frac{D_R(h(\alpha), w) - D_R(h(\alpha), w') - D_R(w', w)}{\alpha}. \quad (1.2)$$

By the definition of w', $D_R(w', w) \leq D_R(x, w)$ for all $x \in K$. From the facts that $u \in K$, $w' \in K$ and convexity of K it follows that $h(\alpha) \in K$ for $\alpha \in [0, 1]$. Thus, $D_R(w', w) \leq D_R(h(\alpha), w)$ for $\alpha \in [0, 1]$. Combining with (1.2)

$$D_R(u, w) - D_R(u, w') - D_R(w', w) \geq \frac{-D_R(h(\alpha), w')}{\alpha}. \quad (1.3)$$

Let us define $G(x) = D_R(x, w')$. Then the right-hand side of the last inequality is $\frac{-G(h(\alpha))}{\alpha}$. Now we take a limit and rewrite (without minus sign, for the moment)

$$\begin{aligned} \lim_{\alpha \to 0^+} \frac{G(h(\alpha))}{\alpha} &= \lim_{\alpha \to 0^+} \frac{G(h(\alpha)) - G(w')}{\alpha} \\ &= \lim_{\alpha \to 0^+} \frac{G(w' + \alpha(u - w')) - G(w')}{\alpha}. \end{aligned} \quad (1.4)$$

Observe that (1.4) is just a definition of the directional derivative of G in the direction $u - w'$ in the point w'. Denote it as $\nabla_{u-w'} G(w')$. We need to ensure that it exists. First, we note that $G(x)$ is continuously differentiable in A°, because R is Legendre. Therefore, $\nabla_{u-w'} G(x)$ exists for every point $x \in A^\circ$. Second, $w' \in A^\circ$, again, because R is Legendre and the condition 4 of its definition ensures that w' does not belong to the boundary. Now we use the relation between the directional derivative and the gradient of a function:

$$\nabla_{u-w'} G(w') = (u - w')\nabla G(w').$$

Therefore, we need to compute $\nabla G(w')$. $G(x)$ is non-negative function and $G(w') = 0$, in other words, it attains its minimum in the point w'. Since it is differentiable, all this implies that $\nabla G(w') = 0$. Hence, (1.4) is zero. Substituting this into (1.3) we obtain the desired inequality. □

We also would like to note that if K is a hyperplane in $\mathbb{R}^d$, then inequality in Lemma 1.2 holds with equality. However, we do not need this result further, so we will not prove it.

The next lemma is useful for the implementation and analysis of the Proximal Point Algorithm. It says that if we want to compute the minimum of the Legendre function over some set, we can first compute the unconstrained minimizer and then project it to the desired set.

Lemma 1.3 (Projection lemma)**.** *Let $R : A \to \mathbb{R}$ be a Legendre function. Let $K \subseteq \mathbb{R}^d$ be convex, closed and $A \cap K \neq \emptyset$. Then*

$$\Pi_{R,K}(\operatorname*{argmin}_{u \in A} R(u)) = \operatorname*{argmin}_{u \in A \cap K} R(u).$$

Proof. Denote $w' = \operatorname{argmin}_{u \in A} R(u)$. Since it is an unconstrained minimizer of R, $\nabla R(w') = 0$. Combining this with the definition of the Bregman divergence gives us

$$\begin{aligned}
\Pi_{R,K}(w') &= \operatorname*{argmin}_{u \in A \cap K} D_R(u, w') \\
&= \operatorname*{argmin}_{u \in A \cap K} (R(u) - R(w') - \langle \nabla R(w'), u - w' \rangle) \\
&= \operatorname*{argmin}_{u \in A \cap K} (R(u) - R(w')) \\
&= \operatorname*{argmin}_{u \in A \cap K} R(u).
\end{aligned}$$

□

1.3 Strong convexity and dual norms

To state the main result for the Proximal Point Algorithm we need a few more definitions and results from functional analysis. The first is more strong notion of convexity. While for convex functions we require a linear lower bound, the strong convexity strengthen this to a quadratic lower bound. More precisely, we have the following definition.

Definition 1.4 (α-strongly convex function)**.** Let $R : A \to \mathbb{R}$ be a Legendre function. Then R is called strongly convex with respect some fixed norm $||\cdot||$, if for any $u, v \in A$

$$R(u) \geq R(v) + \langle \nabla R(v), u - v \rangle + \alpha ||u - v||^2.$$

Note that this holds for a more general class of functions, but we only need this property for Legendre functions now. In addition, the definition depends on the norm used and it implies strict convexity (independently of the norm). We will need the notion of the dual norm and the Hölder's inequality.

Definition 1.5 (Dual norm)**.** Let $||\cdot||$ be a norm on $\mathbb{R}^d$. The dual norm $||\cdot||_*$ for it is defined by

$$||u||_* = \sup_{v \in \mathbb{R}^d : ||v|| = 1} \langle u, v \rangle.$$

The basic results from functional analysis tell us that $||\cdot||_*$ is indeed a norm and that the dual of $||\cdot||_*$ is again $||\cdot||$.

Lemma 1.4 (Hölder's inequality)**.** *Let $||\cdot||$ be a norm on $\mathbb{R}^d$. Then*

$$\langle u, v \rangle \leq ||u|| \cdot ||v||_*$$

for any $u, v \in \mathbb{R}^d$.

Proof. The proof follows trivially from the definition:

$$\begin{aligned}\langle u, v\rangle &\leq ||u||\langle \frac{u}{||u||}, v\rangle \\ &\leq ||u|| \sup_{u\in\mathbb{R}^d} \langle \frac{u}{||u||}, v\rangle \\ &= ||u|| \sup_{u\in\mathbb{R}^d:||u||=1} \langle u, v\rangle \\ &= ||u|| \cdot ||v||_*.\end{aligned}$$

□

1.4 Proximal Point Algorithm

Now we are ready to present the Proximal Point Algorithm for the linear online optimization problem. Let $R : A \to \mathbb{R}$ be a Legendre function. Then the algorithm computes the point d_{t+1} using the following rule

$$d_{t+1} = \operatorname*{argmin}_{d\in D\cap A} \left(\eta\langle d, f_t\rangle + D_R(d, d_t)\right), \tag{1.5}$$

where η is a tuning parameter, and it starts with $d_1 = \operatorname{argmin}_{d\in D\cap A} R(d)$. Intuitively, the algorithm tries to minimize the previous loss and also not to deviate too much from the previously chosen point, since it minimized the losses on the rounds before the last one. One can formulate the same algorithm in a more convenient way. First, we compute unconstrained minimizer $\tilde{d}_{t+1}$ and then project it to the decision space. This formulation simplifies the analysis and is also more suitable for the implementation. Formally,

$$\tilde{d}_{t+1} = \operatorname*{argmin}_{d\in A} \left(\eta\langle d, f_t\rangle + D_R(d, d_t)\right) \tag{1.6}$$

$$d_{t+1} = \Pi_{R,D}(\tilde{d}_{t+1}) \tag{1.7}$$

and the algorithm starts with $\tilde{d}_1 = \operatorname{argmin}_{d\in A} R(d)$ and $d_1 = \Pi_{R,D}(\tilde{d}_1)$.

Proposition 1.4. *The two formulations of PPA are equivalent in the sense that they produce the same sequence of points.*

Proof. Introduce function $F_t(u) = \eta\langle u, f_t\rangle + D_R(u, d_t)$. Expanding the D_R by its definition, we get

$$\begin{aligned}F_t(u) &= \eta\langle u, f_t\rangle + D_R(u, d_t) \\ &= \eta\langle u, f_t\rangle + R(u) - R(d_t) - \langle \nabla R(d_t), u - d_t\rangle.\end{aligned}$$

Hence, by Proposition 1.3, F_t is Legendre with the same divergence as R has, i.e. $D_R(u, v) = D_{F_t}(u, v)$. Now, using the projection lemma, we can rewrite

$$\begin{aligned}\tilde{d}_{t+1} &= \operatorname*{argmin}_{d\in A} \left(F_t(d)\right) \\ d_{t+1} &= \Pi_{F_t,D}(\tilde{d}_{t+1}).\end{aligned}$$

The only difference from (1.7) is that we make a projection with respect to F_t, but these projections are equivalent because of the equivalence of corresponding divergences. □

Algorithm 1: Proximal Point Algorithm for linear online optimization

Parameters: *D - decision space, finite horizon T, η, R - Legendre function*
Compute $\tilde{d}_1 = \operatorname{argmin}_{d \in A} R(d)$;
Compute $d_1 = \Pi_{R,D}(\tilde{d}_1)$;
Output d_1 as a decision;
Receive f_1;
for $t = 2, \ldots, T$ **do**
  Compute $\tilde{d}_t = \operatorname{argmin}_{d \in A} (\eta \langle d, f_{t-1} \rangle + D_R(d, d_{t-1}))$;
  Compute $d_t = \Pi_{R,D}(\tilde{d}_t)$;
  Output d_t as a decision;
  Receive f_t;
end

1.4.1 Analysis of the Proximal Point Algorithm

To prove the regret bound for PPA, we start with a small proposition.

Proposition 1.5. *The sequence of points generated by PPA satisfies*

$$\nabla R(\tilde{d}_{t+1}) - \nabla R(d_t) = -\eta f_t.$$

Proof. Since in (1.6) we compute an unconstrained minimizer, we have

$$\nabla \left(\eta \langle d, f_t \rangle + D_R(d, d_t) \right) |_{d = \tilde{d}_{t+1}} = 0. \tag{1.8}$$

Now we just compute the gradient

$$\begin{aligned} \nabla \left(\eta \langle d, f_t \rangle + D_R(d, d_t) \right) &= \eta f_t + \nabla D_R(d, d_t) \\ &= \eta f_t + \nabla (R(d) - R(d_t) - \langle \nabla R(d_t), d - d_t \rangle) \\ &= \eta f_t + \nabla R(d) - \nabla R(d_t). \end{aligned}$$

Then evaluate it at $\tilde{d}_{t+1}$ and the claim follows

$$(\eta f_t + \nabla R(d) - \nabla R(d_t))|_{d = \tilde{d}_{t+1}} = \eta f_t + \nabla R(\tilde{d}_{t+1}) - \nabla R(d_t) = 0.$$

□

Corollary 1.2. *For any $t = 1, \ldots, T$*

$$D_R(d_t, \tilde{d}_{t+1}) + D_R(\tilde{d}_{t+1}, d_t) = \eta \langle f_t, d_t - \tilde{d}_{t+1} \rangle.$$

Proof.

$$\begin{aligned} D_R(d_t, \tilde{d}_{t+1}) + D_R(\tilde{d}_{t+1}, d_t) &= R(d_t) - R(\tilde{d}_{t+1}) - \langle \nabla R(\tilde{d}_{t+1}), d_t - \tilde{d}_{t+1} \rangle \\ &\quad + R(\tilde{d}_{t+1}) - R(d_{t+1}) - \langle \nabla R(d_t), \tilde{d}_{t+1} - d_t \rangle \\ &= \langle \nabla R(d_t) - \nabla R(\tilde{d}_{t+1}), d_t - \tilde{d}_{t+1} \rangle \\ &= \eta \langle f_t, d_t - \tilde{d}_{t+1} \rangle \text{ (Proposition 1.5).} \end{aligned}$$

□

The next lemma will be useful not only as an intermediate step towards the main result of this section, but also as a tool to prove the bound for the bandit version of the algorithm. Basically, this lemma tells us that to prove a good bound it is enough to bound $\langle d_t, f_t\rangle - \langle \tilde{d}_{t+1}, f_t\rangle$ in each time step.

Lemma 1.5. *For any point $p \in D \cap A$ and any $\eta > 0$*

$$\hat{L}_T - L_T(p) \leq \sum_{t=1}^{T}(\langle d_t, f_t\rangle - \langle \tilde{d}_{t+1}, f_t\rangle) + \frac{D_R(p, d_1)}{\eta}.$$

Proof. We start with bounding the differences in losses in each time step:

$$\begin{aligned}
\langle d_t, f_t\rangle - \langle p, f_t\rangle &= \langle d_t - p, f_t\rangle \\
&= \frac{1}{\eta}\langle p - d_t, \nabla R(\tilde{d}_{t+1}) - \nabla R(d_t)\rangle \text{ (Proposition 1.5)} \\
&= \frac{1}{\eta}\left(D_R(p, d_t) + D_R(d_t, \tilde{d}_{t+1}) - D_R(p, \tilde{d}_{t+1})\right) \text{ (Proposition 1.2)} \\
&\leq \frac{1}{\eta}\left(D_R(p, d_t) + D_R(d_t, \tilde{d}_{t+1}) - D_R(p, d_{t+1}) - D_R(d_{t+1}, \tilde{d}_{t+1})\right) \text{ (Lemma 1.2)} \\
&\leq \frac{1}{\eta}\left(D_R(p, d_t) + D_R(d_t, \tilde{d}_{t+1}) - D_R(p, d_{t+1})\right).
\end{aligned}$$

The last line follows from the non-negativity of the divergence. Now we sum up the inequalities obtained from 1 to T. The terms $D_R(p, d_t)$ form a telescoping sequence, hence we have the following

$$\begin{aligned}
\hat{L}_T - L_T(p) &\leq \frac{1}{\eta}\left(D_R(p, d_1) - D_R(p, d_{T+1}) + \sum_{t=1}^{T} D_R(d_t, \tilde{d}_{t+1})\right) \\
&\leq \frac{1}{\eta}\left(D_R(p, d_1) + \sum_{t=1}^{T} D_R(d_t, \tilde{d}_{t+1})\right). \qquad (1.9)
\end{aligned}$$

Where we again used the non-negativity of the divergence. It remains to bound the second term in (1.9). Again we do it for every t separately

$$\begin{aligned}
D_R(d_t, \tilde{d}_{t+1}) &\leq D_R(d_t, \tilde{d}_{t+1}) + D_R(\tilde{d}_{t+1}, d_t) \\
&= \eta\langle f_t, d_t - \tilde{d}_{t+1}\rangle \text{ (Corollary 1.2).}
\end{aligned}$$

We finish the proof by substituting the obtained inequality in (1.9). □

Now we are ready to state and prove the main bound for the Proximal Point Algorithm in the full-information case.

Theorem 1.1. *Let $R : A \to \mathbb{R}$ be a Legendre and α-strongly convex function with respect to some norm $||\cdot||$. If the proximal point algorithm is run using R and $\eta > 0$, then for any $p \in D \cap A$*

$$\hat{L}_T - L_T(p) \leq \frac{\eta}{2\alpha}\sum_{t=1}^{T}||f_t||_*^2 + \frac{R(p) - R(d_1)}{\eta}.$$

Proof. Our starting point is Lemma 1.5:

$$\hat{L}_T - L_T(p) \leq \sum_{t=1}^{T}(\langle d_t, f_t\rangle - \langle \tilde{d}_{t+1}, f_t\rangle) + \frac{D_R(p, d_1)}{\eta}. \tag{1.10}$$

First, we deal with the second term. By the projection lemma, d_1 is a minimum of R in $D \cap A$. This implies that $\langle \nabla R(d_1), p - d_1\rangle \geq 0$ for all $p \in D \cap A$, because otherwise we could decrease R by taking small step in the direction $p - d_1$. Using this fact

$$\begin{aligned} D_R(p, d_1) &= R(p) - R(d_1) - \langle \nabla R(d_1), p - d_1\rangle \\ &\leq R(p) - R(d_1). \end{aligned}$$

Now we turn our attention to the first term in (1.10). For each t the Hölder inequality gives us

$$\langle d_t, f_t\rangle - \langle \tilde{d}_{t+1}, f_t\rangle \leq ||d_t - \tilde{d}_{t+1}|| \cdot ||f_t||_*.$$

It remains to prove that $||d_t - \tilde{d}_{t+1}|| \leq \frac{\eta}{2\alpha}||f_t||_*$. If $d_t = \tilde{d}_{t+1}$, then it is trivial. Otherwise, the strong convexity of R implies

$$\begin{aligned} D_R(d_t, \tilde{d}_{t+1}) &\geq \alpha||d_t - \tilde{d}_{t+1}||^2 \\ D_R(\tilde{d}_{t+1}, d_t) &\geq \alpha||\tilde{d}_{t+1} - d_t||^2. \end{aligned}$$

Summing up these inequalities and using Corollary 1.2 and Hölder inequality

$$\begin{aligned} 2\alpha||d_t - \tilde{d}_{t+1}||^2 &\leq D_R(d_t, \tilde{d}_{t+1}) + D_R(\tilde{d}_{t+1}, d_t) \\ &= \eta\langle f_t, d_t - \tilde{d}_{t+1}\rangle \\ &\leq \eta||f_t||_* \cdot ||d_t - \tilde{d}_{t+1}||. \end{aligned}$$

We conclude the proof dividing both sides by the $2\alpha||d_t - \tilde{d}_{t+1}|| > 0$. □

Corollary 1.3. *If $||f_t||_* \leq 1$ for any $t = 1, \ldots, T$ and $\eta = \sqrt{2\alpha\frac{(R(p)-R(d_1))}{T}}$ we obtain the following regret bound for any $p \in D \cap A$*

$$\hat{L}_T - L_T(p) \leq 2\sqrt{2\alpha T(R(p) - R(d_1))}$$

Rephrasing the result we just proved, PPA is able to achieve the regret of order $\mathcal{O}(\sqrt{T})$. Actually, this is a tight result, since there is a matching lower bound of order $\Omega(\sqrt{T})$ for discrete prediction problems, which are special cases of online linear optimization (Cesa-Bianchi and Lugosi (2006)).

1.4.2 Exponentially Weighted Average algorithm

In this subsection we are presenting the Exponentially Weighted Average algorithm (EWA). It is just an instance of the Proximal Point Algorithm that uses un-normalized negative entropy. In addition, we restrict ourselves to the problems where the decision space D is a probability simplex in $\mathbb{R}^d$. The purpose of this section is to derive the closed form formulas for the updates and also to prove a bound on the regret of EWA. Just to recall, throughout this section we will work with the following functions

$$R(u) = \sum_{i=1}^{d} u_i \ln u_i - \sum_{i=1}^{d} u_i$$

$$D_R(u,v) = \sum_{i=1}^{d} u_i \ln \frac{u_i}{v_i} - \sum_{i=1}^{d} (u_i - v_i).$$

We will also need the gradients of the functions $R(\cdot)$ and $D_R(\cdot, v)$

$$\nabla_i R(u) = \ln u_i$$

$$\nabla_i D_R(u,v) = \ln u_i - \ln v_i.$$

First, we derive the formula for $\tilde{d}_t$. Recall that it is computed using the following rule

$$\tilde{d}_t = \operatorname*{argmin}_{d \in A} \left(\eta \langle d, f_{t-1} \rangle + D_R(d, d_{t-1}) \right).$$

Since this is unconstrained minimization, we can just compute the gradient and set it to zero (we denote by $u_{t,i}$ the i-th component of the vector u_t)

$$\eta f_{t-1,i} + \ln d_i - \ln d_{t-1,i} = 0.$$

Therefore, $\tilde{d}_{t,i} = d_{t-1,i} e^{-\eta f_{t-1,i}}$. The projection step is similar, with the difference that we have a constrained minimization

$$d_t = \operatorname*{argmin}_{d \in D \cap A} D_R(d, \tilde{d}_t).$$

First we write the Lagrangian (recall that D is a probability simplex)

$$\mathcal{L} = D_R(d, \tilde{d}_t) + \lambda \sum_{i=1}^{d} d_i,$$
$$\nabla_i \mathcal{L} = \ln d_i - \ln \tilde{d}_{t,i} + \lambda = 0.$$

We conclude that $d_i = \tilde{d}_{t,i} e^{\lambda}$ and the λ should be chosen such that $\sum_{i=1}^{d} d_i = 1$. Thus, $e^{\lambda} = \frac{1}{\sum_{i=1}^{d} \tilde{d}_{t,i}}$ and, finally, the projection corresponds to the following simple normalization

$$\begin{aligned} d_{t,i} &= \frac{\tilde{d}_{t,i}}{\sum_{i=1}^{d} \tilde{d}_{t,i}} \\ &= \frac{d_{t-1,i} e^{-\eta f_{t-1,i}}}{\sum_{j=1}^{d} d_{t-1,j} e^{-\eta f_{t-1,j}}}. \end{aligned}$$

When we try to prove the bound on the regret for EWA we can not use Theorem 1.1 directly. The reason is that un-normalized negative entropy is not strongly convex with respect to any norm on A, but it is possible to exploit the closed form of the updates to prove the same bound.

Theorem 1.2. *If EWA algorithm is run using $\eta > 0$, then for any $p \in D \cap A$*

$$\hat{L}_T - L_T(p) \leq \eta \sum_{t=1}^{T} ||f_t||_{\infty}^2 + \frac{R(p) - R(d_1)}{\eta}$$

Algorithm 2: Exponentially Weighted Average algorithm for linear online optimization

Parameters: *finite horizon* T, η
Set $d_1 = \left(\frac{1}{d}, .., \frac{1}{d}\right)^T$;
Output d_1 as a decision;
Receive f_1;
for $t = 2, \ldots, T$ **do**
  Compute d_t component-wise as $d_{t,i} = \frac{d_{t-1,i}e^{-\eta f_{t-1,i}}}{\sum_{j=1}^{d} d_{t-1,j}e^{-\eta f_{t-1,j}}}$;
  Output d_t as a decision;
  Receive f_t;
end

Proof. We start with Lemma 1.5:

$$\hat{L}_T - L_T(p) \leq \sum_{t=1}^{T}(\langle d_t, f_t\rangle - \langle \tilde{d}_{t+1}, f_t\rangle) + \frac{D_R(p, d_1)}{\eta}. \tag{1.11}$$

The second term in (1.11) is bounded in the same way as in the Theorem 1.1. Thus, we focus our attention on the first one. Let us introduce notation: for $u, v \in \mathbb{R}^d$ $u \circ v = (u_1 v_1, \ldots, u_d v_d)^T$. Then for any time step t

$$\begin{aligned}\tilde{d}_{t+1,i} &= d_{t,i}e^{-\eta f_{t,i}} \\ &\geq d_{t,i} - \eta d_{t,i} f_{t,i},\end{aligned}$$

where we used the fact that $e^x \geq 1 + x$. Hence, we can rewrite

$$\begin{aligned}\langle d_t, f_t\rangle - \langle \tilde{d}_{t+1}, f_t\rangle &\leq \langle d_t, f_t\rangle - \langle d_t, f_t\rangle + \eta\langle d_t \circ f_t, f_t\rangle \\ &= \eta\langle d_t \circ f_t, f_t\rangle \\ &\leq \eta||d_t \circ f_t||_1||f_t||_\infty \text{ (Hölder's inequality).}\end{aligned}$$

The only thing left to deal with is $||d_t \circ f_t||_1$. First we note that $||d_t \circ f_t||_1 = \sum_{i=1}^{d} |d_{t,i} f_{t,i}| = \langle d_t, |f_t|\rangle$, where $|f_t|$ denotes vector $(|f_{t,1}|, .., |f_{t,d}|)^T$. Hence, we can again use Hölder's inequality. Note that the max-norm of f_t equals to the max-norm of $|f_t|$

$$\begin{aligned}||d_t \circ f_t||_1 &= \langle d_t, |f_t|\rangle \\ &\leq ||d_t||_1||f_t||_\infty \\ &= ||f_t||_\infty.\end{aligned}$$

Where the last step follows since $d_t \in D$. □

Corollary 1.4. *If* $||f_t||_\infty \leq 1$ *for any* $t = 1, \ldots, T$ *and* $\eta = \sqrt{\frac{\ln d}{T}}$ *we obtain the following regret bound for any* $p \in D \cap A$

$$\hat{L}_T - L_T(p) \leq 2\sqrt{T \ln d}.$$

Proof. The only non-trivial step is to bound $R(p) - R(d_1)$:

$$\begin{aligned} R(p) - R(d_1) &= \sum_{i=1}^{d} p_i \ln p_i - \sum_{i=1}^{d} p_i - \sum_{i=1}^{d} d_{1,i} \ln d_{1,i} + \sum_{i=1}^{d} d_{1,i} \\ &= \sum_{i=1}^{d} p_i \ln p_i - 1 - \sum_{i=1}^{d} d_{1,i} \ln d_{1,i} + 1 \\ &\leq -\sum_{i=1}^{d} d_{1,i} \ln d_{1,i} \\ &\leq \ln d. \end{aligned}$$

Where the last two lines follow from the properties of the usual Shannon entropy $H(u) = -\sum_{i=1}^{d} u_i \ln u_i$. Specifically, that $0 \leq H(u) \leq \ln d$ (see Cover and Thomas (1991)).

Using the proven inequality and the assumption that $||f_t||_\infty \leq 1$ we can rewrite the bound of Theorem 1.2 as

$$\hat{L}_T - L_T(p) \leq \eta T + \frac{\ln d}{\eta}.$$

Finally, we prove the corollary optimizing over η. □

The bound on the regret for EWA is tight not only in T, what is the case for general PPA, but also in N, since there is a matching lower bound of order $\Omega(\sqrt{T \ln N})$ (see Cesa-Bianchi and Lugosi (2006), Theorem 3.7).

1.4.3 Exponentially Weighted Average algorithm for the multi-armed bandit

In this subsection we consider the simplified version of bandit online linear optimization. It is called a *multi-armed bandit* problem, which dates back to the work of Hannan (1957). Fitting it into the framework of online linear optimization, we can formulate the problem as follows. At each time step we should choose a vector $e_{a_t} \in \Delta = \{e_1, .., e_d\}$, where $\{e_i\}_{i=1}^{d}$ is a basis in $\mathbb{R}^d$ and $a_t \in \{1, \ldots, d\}$. Then the loss incurred is the same as before: $\ell_t(e_{a_t}) = \langle e_{a_t}, f_t \rangle = f_{t,a_t}$. This is exactly what we observe at the end of each round. It is easy to see that for every deterministic strategy there exists a sequence of points f_t such that the strategy's regret is linear in T. Hence, we need to consider algorithms that make random choices. Formally, at each time step an algorithm should choose a distribution $\mathbf{d}_t \in D$, where $D \subset \mathbb{R}^d$ is a probability simplex and then it samples $\mathbf{a}_t \in \{1, \ldots, d\}$ according to this distribution. If we would consider the following notion of pseudo-regret as a measure of performance

$$\hat{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} f_{t,\mathbf{a}_t}\right] - \inf_{d \in D} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{a} \sim d}\left[f_{t,\mathbf{a}}\right]$$

then this problem is equivalent to the online linear optimization problem with decision space D with the only difference that after each time step we observe a component $f_{t,\mathbf{I}_t}$ of vector f_t, where $\mathbf{I}_t$ is drawn from $\mathbf{d}_t$. Let us denote by $\mathbf{u}_t$ the information, observed by the learner up to time t, i.e.

$$\mathbf{u}_t = \{f_{1,\mathbf{I}_1}, f_{2,\mathbf{I}_2}, \ldots, f_{t,\mathbf{I}_t}\}$$

Parameters: finite horizon T.
Environment chooses points $f_1, ..., f_T \in \mathbb{R}^d$.
for $t = 1$ **to** T:

1. Learner chooses a distribution $\mathbf{d}_t$ over $\{e_1, .., e_d\}$.
2. Learner observes a component $f_{t,\mathbf{I}_t}$, where $\mathbf{I}_t$ is drawn from $\mathbf{d}_t$.
3. Learner suffers loss $\langle \mathbf{d}_t, f_t \rangle$.

Figure 1.2: Multi-armed bandit as a linear online optimization problem

Then the equivalence of the problems follows from the fact that

$$\begin{aligned}
\mathbb{E}\left[f_{t,\mathbf{a}_t}\right] &= \mathbb{E}\left[\mathbb{E}\left[f_{t,\mathbf{a}_t} \middle| \mathbf{u}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} f_{t,i} \mathbb{P}\left[\mathbf{a}_t = i \middle| \mathbf{u}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} f_{t,i} \mathbf{d}_{t,i}\right] \\
&= \mathbb{E}\left[\langle \mathbf{d}_t, f_t \rangle\right].
\end{aligned}$$

Similarly, $\mathbb{E}_{\mathbf{a}\sim d}\left[f_{t,\mathbf{a}}\right] = \langle d, f_t \rangle$. And the pseudo-regret transforms to

$$\hat{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbf{d}_t, f_t \rangle\right] - \inf_{d \in D} \sum_{t=1}^{T} \langle d, f_t \rangle$$

Note that $\mathbf{d}_t$ are now random due to the dependence on the observed information. This is the final version of a problem that we consider in this and the next chapters. The summary is given in Figure 1.2.

The main idea for solving this problem is to construct an estimate $\tilde{f}_t$ for the vector f_t knowing only one component of it. Then we feed the estimate to the EWA as if it was the actual loss vector. The estimate that we are going to use is of the following form:

$$\tilde{f}_{t,i} = \begin{cases} \frac{f_{t,i}}{\mathbf{d}_{t,i}} & \text{if } \mathbf{I}_t = i \\ 0 & \text{otherwise} \end{cases}.$$

This can be equivalently written as $\tilde{f}_{t,i} = \frac{f_{t,i}}{\mathbf{d}_{t,i}} \mathbb{I}\left[\mathbf{I}_t = i\right]$. The reason why we chose this estimator is that it is unbiased in the following sense

$$\begin{aligned}
\mathbb{E}\left[\tilde{f}_{t,i} \middle| \mathbf{u}_{t-1}\right] &= \frac{f_{t,i}}{\mathbf{d}_{t,i}} \mathbb{E}\left[\mathbb{I}\left[\mathbf{I}_t = i\right] \middle| \mathbf{u}_{t-1}\right] \\
&= \frac{f_{t,i}}{\mathbf{d}_{t,i}} \mathbb{P}\left[\mathbf{I}_t = i \middle| \mathbf{u}_{t-1}\right] \\
&= \frac{f_{t,i}}{\mathbf{d}_{t,i}} \mathbf{d}_{t,i} \\
&= f_{t,i}.
\end{aligned}$$

The resulting algorithm that we call Exp3 algorithm (following Auer et al. (2002)) is constructed by feeding the estimate to EWA using it as a black-box.

The following theorem provides the bound on the regret of Exp3.

Algorithm 3: Exp3 algorithm for the multi-armed bandit

Parameters: *finite horizon* T, η
Set $d_1 = \left(\frac{1}{d}, .., \frac{1}{d}\right)^T$;
Output $\mathbf{d}_1$ as a decision;
Receive $f_{\mathbf{I}_t}$;
Compute $\tilde{f}_1$ component-wise as $\tilde{f}_{1,i} = \frac{f_{1,i}}{\mathbf{d}_{1,i}}\mathbb{I}\left[\mathbf{I}_1 = i\right]$;
for $t = 2, \ldots, T$ **do**
  Compute $\mathbf{d}_t$ component-wise as $\mathbf{d}_{t,i} = \frac{\mathbf{d}_{t-1,i}e^{-\eta\tilde{f}_{t-1,i}}}{\sum_{j=1}^{d}\mathbf{d}_{t-1,j}e^{-\eta\tilde{f}_{t-1,j}}}$;
  Output $\mathbf{d}_t$ as a decision;
  Receive $f_{t,\mathbf{I}_t}$;
  Compute $\tilde{f}_t$ component-wise as $\tilde{f}_{t,i} = \frac{f_{t,i}}{\mathbf{d}_{t,i}}\mathbb{I}\left[\mathbf{I}_t = i\right]$;
end

Theorem 1.3. *If* $||f_t||_\infty \leq 1$ *for any* $t = 1, \ldots, T$ *and Exp3 algorithm is run using* $\eta = \sqrt{\frac{\ln d}{dT}}$*, we obtain the following regret bound for any* $p \in D \cap A$

$$\hat{\mathcal{R}}_T \leq 2\sqrt{Td\ln d}.$$

Proof. We use the unbiased property of the estimate $\tilde{f}_t$ to show that the loss in the game that outputs $\tilde{f}_t$ is equal to the expected loss in the game with the actual vectors. More precisely, for every $p \in D \cap A$

$$\begin{aligned}
\langle p, f_t\rangle &= \sum_{i=1}^{d} p_i f_{t,i} \\
&= \sum_{i=1}^{d} p_i \mathbb{E}\left[\tilde{f}_{t,i}\middle|\, \mathbf{u}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} p_i \tilde{f}_{t,i}\middle|\, \mathbf{u}_{t-1}\right] \\
&= \mathbb{E}\left[\langle p, \tilde{f}_t\rangle\middle|\, \mathbf{u}_{t-1}\right].
\end{aligned}$$

This equality also holds for $p = \mathbf{d}_t$, hence, we can rewrite

$$\langle \mathbf{d}_t, f_t\rangle - \langle p, f_t\rangle = \mathbb{E}\left[\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle p, \tilde{f}_t\rangle\middle|\, \mathbf{u}_{t-1}\right]. \tag{1.12}$$

Since the expression under the expectation is exactly what black-box EWA is minimizing, we can use the Lemma 1.5 to upper bound this term. Recall that Lemma 1.5 states

$$\sum_{t=1}^{T}\left(\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle p, \tilde{f}_t\rangle\right) \leq \sum_{t=1}^{T}\left(\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle \tilde{\mathbf{d}}_{t+1}, \tilde{f}_t\rangle\right) + \frac{D_R(p, \mathbf{d}_1)}{\eta}.$$

Now we take an expectation on both sides. The left-hand side becomes

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left(\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle p, \tilde{f}_t\rangle\right)\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle p, \tilde{f}_t\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle p, \tilde{f}_t\rangle \middle| \mathbf{u}_{t-1}\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\langle \mathbf{d}_t, f_t\rangle - \langle p, f_t\rangle\right] \\
&= \hat{\mathcal{R}}_T.
\end{aligned}$$

Now we deal with the expectation of the right-hand side of (1.4.3):

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left(\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle \tilde{\mathbf{d}}_{t+1}, \tilde{f}_t\rangle\right)\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle \tilde{\mathbf{d}}_{t+1}, \tilde{f}_t\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle \tilde{\mathbf{d}}_{t+1}, \tilde{f}_t\rangle \middle| \mathbf{u}_{t-1}\right]\right]. \qquad (1.13)
\end{aligned}$$

Using the argument similar to Theorem 1.2

$$\mathbb{E}\left[\langle \mathbf{d}_t, \tilde{f}_t\rangle - \langle \tilde{\mathbf{d}}_{t+1}, \tilde{f}_t\rangle \middle| \mathbf{u}_{t-1}\right] \leq \mathbb{E}\left[\eta\langle \mathbf{d}_t \circ \tilde{f}_t, \tilde{f}_t\rangle \middle| \mathbf{u}_{t-1}\right]. \qquad (1.14)$$

Now we take a closer look at $\mathbf{d}_t \circ \tilde{f}_t$. For each component

$$\begin{aligned}
\mathbf{d}_{t,i}\tilde{f}_{t,i} &= \mathbf{d}_{t,i}\frac{f_{t,i}}{\mathbf{d}_{t,i}}\mathbb{I}\left[\mathbf{I}_t = i\right] \\
&= f_{t,i}\mathbb{I}\left[\mathbf{I}_t = i\right] \\
&\leq ||f_t||_\infty \\
&\leq 1 \text{ (By assumption).}
\end{aligned}$$

Substitute this into (1.14):

$$\begin{aligned}
\mathbb{E}\left[\eta\langle \mathbf{d}_t \circ \tilde{f}_t, \tilde{f}_t\rangle \middle| \mathbf{u}_{t-1}\right] &= \mathbb{E}\left[\eta\sum_{i=1}^{d}\mathbf{d}_{t,i}\tilde{f}_{t,i}\tilde{f}_{t,i} \middle| \mathbf{u}_{t-1}\right] \\
&\leq \eta\sum_{i=1}^{d}\mathbb{E}\left[\tilde{f}_{t,i} \middle| \mathbf{u}_{t-1}\right] \\
&= \eta\sum_{i=1}^{d} f_{t,i} \\
&\leq \eta d||f_t||_\infty \\
&\leq \eta d.
\end{aligned}$$

If we combine this with (1.14) and plug into (1.13), upper bound $D_R(p, \mathbf{d}_1)$ by $\ln d$ as in the Corollary 1.4, we obtain the following bound

$$\hat{\mathcal{R}}_T \leq \eta dT + \frac{\ln d}{\eta}.$$

Optimizing over η yields the statement of the theorem. □

We can see that even in the case of bandit information we can obtain optimal $\sqrt{T}$ dependence of the regret. However, we have an additional multiple of $\sqrt{d}$ which can be thought of as a price for not observing f_t in each round. Actually, the result is not tight in d, since the existing lower bound is $\Omega(\sqrt{nd})$ (Cesa-Bianchi and Lugosi (2006), Theorem 6.11).

Chapter 2

Markovian Decision Processes

In this chapter we present different learning situations that can be formulated with Markovian Decision Processes. This framework is used to model a lot of different problems and has applications in various areas. We focus on the problem of online learning in MDPs. First, we consider the simpler one: learning in episodic loop-free MDPs. Similar to online linear optimization, we investigate two important cases: full-information and bandit feedback. Next, we turn our attention to the unichain MDP. In both cases we show how the application of the Proximal Point Algorithm improves the previously known results.

2.1 The problem description

We start with a definition of Markovian Decision Process.

Definition 2.1 (Markovian Decision Process)**.** MDP is a tuple $\langle \mathcal{X}, \mathcal{A}, P, P_1 \rangle$. Where $\mathcal{X}$ denotes a state space, $\mathcal{A}$ is an action space, $P : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$ is a transition function and P_1 is an initial state distribution.

In its usual formulation MDP is supplied with a fixed loss function $\ell : \mathcal{X} \times \mathcal{A} \to [0, 1]$, however, in the problems that we consider we assume that there is a sequence of loss functions $\{\ell_t\}_{t=1}^T$ for some fixed horizon T where each $\ell_t : \mathcal{X} \times \mathcal{A} \to [0, 1]$.

MDPs model interactions between an agent and a stochastic environment. At each time step $t = 1, \ldots, T$ the agent observes the current state $\mathbf{x}_t$, chooses an action $\mathbf{a}_t \in \mathcal{A}$ and suffers a loss $\ell_t(\mathbf{x}_t, \mathbf{a}_t)$. Then the next state $\mathbf{x}_{t+1} \in \mathcal{X}$ is drawn from the distribution $P(\cdot|\mathbf{x}_t, \mathbf{a}_t)$. The process starts in state $\mathbf{x}_1 \in \mathcal{X}$ drawn from P_1. The goal of an agent is to minimize its expected cumulative loss

$$\hat{L}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\mathbf{x}_t, \mathbf{a}_t)\right].$$

We denote by $\mathbf{u}_t$ the history of the interaction up to step t

$$\mathbf{u}_t = \{\mathbf{x}_1, \mathbf{a}_1, \ell_1(\mathbf{x}_1, \mathbf{a}_1),\ ..,\ \mathbf{x}_t, \mathbf{a}_t, \ell_t(\mathbf{x}_t, \mathbf{a}_t)\}$$

for $t = 1, \ldots, T$, where $\{(\mathbf{x}_t, \mathbf{a}_t)\}_{t=1}^T$ is a random trajectory generated by the agent. And we set $\mathbf{u}_0 = \emptyset$. Using this notation, an agent is defined by a sequence of *policies* $\{\boldsymbol{\pi}_t\}_{t=1}^T$, where each policy $\boldsymbol{\pi}_t : \mathcal{A} \times \mathcal{X} \to [0, 1]$ such that $\boldsymbol{\pi}_t(\cdot|x)$ defines a distribution over $\mathcal{A}$ for every $x \in \mathcal{X}$. Note that $\boldsymbol{\pi}_t$ can be random and depend on the past. Formally,

$$\boldsymbol{\pi}_t(a|x) = \mathbb{P}\left[\mathbf{a}_t = a \middle| \mathbf{x}_t = x, \mathbf{u}_{t-1}\right].$$

Parameters: MDP $\langle \mathcal{X}, \mathcal{A}, P, P_1 \rangle$, finite horizon T.
Environment chooses loss functions $\ell_1, ..., \ell_T : \mathcal{X} \times \mathcal{A} \to [0, 1]$.
Initial state $\mathbf{x}_1$ is drawn from P_1.
for $t = 1$ **to** T:

1. Learner chooses a policy $\boldsymbol{\pi}_t$.
2. Action $\mathbf{a}_t$ is drawn from $\boldsymbol{\pi}_t(\cdot|\mathbf{x}_t)$.
3. ℓ_t (full information) or $\ell_t(\mathbf{x}_t, \mathbf{a}_t)$ (bandit information) is revealed.
4. Learner suffers loss $\ell_t(\mathbf{x}_t, \mathbf{a}_t)$.
5. The next state $\mathbf{x}_{t+1}$ is drawn from $P(\cdot|\mathbf{x}_t, \mathbf{a}_t)$.

Figure 2.1: Online Learning in Markovian Decision Processes

Hence, our goal is to design an algorithm that chooses this sequence of policies and aims to minimize its cumulative loss for every possible sequence of loss functions. Recall that in the framework of linear online optimization this corresponded to the model of oblivious adversary. Since, in general, this aim is not achievable, we measure the performance of the algorithm by the regret with respect to some reference class of policies. To move further we need the following definition.

Definition 2.2 (Stationary stochastic policy)**.** An agent is said to follow a stationary stochastic policy $\pi : \mathcal{A} \times \mathcal{X} \to [0, 1]$ if at each time step t it chooses an action $\mathbf{a}_t$ drawn from $\pi(\cdot|\mathbf{x}_t)$.

We identify such agents with the policies they follow and let Γ denote the class of all such policies (agents). If $\{(\mathbf{x}'_t, \mathbf{a}'_t)\}_{t=1}^T$ is a random trajectory generated by policy π then we define the expected cumulative loss of this policy as

$$L_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\mathbf{x}'_t, \mathbf{a}'_t)\right].$$

The classical result of Puterman (1994), Theorem 4.4.2, tells us that the class of stationary stochastic policies is a reasonable class to compete with, since there exist a stationary and deterministic policy π^* such that

$$L_T(\pi^*) = \inf_{\pi \in \Gamma} L_T(\pi).$$

Therefore, in fact, the inf can be replaced by min. Finally, we measure the performance of an algorithm by means of the regret

$$\mathcal{R}_T = \hat{L}_T - \min_{\pi \in \Gamma} L_T(\pi).$$

As in the case of online linear optimization we consider two types of problems. The first one is a full-information case, where the full loss function ℓ_t becomes available to the algorithm at the end of each round. The second type is a bandit case, where the learner observes only $\ell_t(\mathbf{x}_t, \mathbf{a}_t)$. The resulting problem is summarized in Figure 2.1.

2.2 Stationary distributions

In this section we introduce the concept of stationary distributions which will be used in the formulation of the algorithm for solving MDP problem. Furthermore, we introduce the first assumption on the MDP that we require for our results to hold.

Definition 2.3 (Stationary state-action distribution)**.** A distribution q over $\mathcal{X} \times \mathcal{A}$ is called a stationary state-action distribution if it satisfies

$$\sum_a q(x,a) = \sum_{x',a'} q(x',a')P(x|x',a') \tag{2.1}$$

for all $x \in \mathcal{X}$

Note that throughout the chapter we will omit the action and state spaces while writing sums in all places where no confusion can occur.

The set of all stationary state-action distribution is denoted by Δ.

Definition 2.4. We say that a policy $\pi \in \Gamma$ generates a stationary distribution $q \in \Delta$ if

$$\pi(a|x) = \frac{q(x,a)}{\sum_b q(x,b)} \tag{2.2}$$

for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.

We need a strong correspondence between stationary distributions and stationary stochastic policies. The one direction always holds, i.e. for every $q \in \Delta$ there is a policy $\pi \in \Gamma$ that generates it. But this is not true for the other direction, so we need to make our first assumption.

Assumption 1. *Every policy $\pi \in \Gamma$ generates a unique stationary state distribution $q_\pi \in \Delta$ such that (2.1) holds.*

This assumption motivates the following definition.

Definition 2.5 ((Global) stationary state distribution)**.** Given a policy $\pi \in \Gamma$, the distribution μ_π over $\mathcal{X}$ such that

$$\mu_\pi(x) = \sum_a q_\pi(x,a) \tag{2.3}$$

for all $x \in \mathcal{X}$ is called the (global) stationary state distribution generated by π.

Note that the definition implies that for all $\pi \in \Gamma$

$$q_\pi(x,a) = \mu_\pi(x)\pi(a|x).$$

In what follows it will be useful to rewrite (2.3) as

$$\begin{aligned} \mu_\pi(x) &= \sum_{x',a'} q_\pi(x',a')P(x|x',a') \\ &= \sum_{x',a'} \mu_\pi(x')\pi(a'|x')P(x|x',a'), \end{aligned} \tag{2.4}$$

which holds for all $x \in \mathcal{X}$.

2.3 Idealized setting

In this section we consider the simplified version of the problem introduced and show how to solve it efficiently. The Assumption 1 allows us to define the *average loss* ρ_t^π of a policy π at time step t as

$$\begin{aligned}\rho_t^\pi &= \mathbb{E}_{x\sim\mu_\pi, a\sim\pi(\cdot|x)}\left[\ell_t(x,a)\right] \\ &= \sum_{x,a}\mu_\pi(x)\pi(a|x)\ell_t(x,a) \\ &= \sum_{x,a} q_\pi(x,a)\ell_t(x,a) \\ &= \langle q_\pi, \ell_t\rangle.\end{aligned}$$

Where we adopted the view of function ℓ_t and distribution q_π as a vectors over product space $\mathcal{X}\times\mathcal{A}$ equipped with an inner product $\langle u, v\rangle = \sum_{x,a} u(x,a)v(x,a)$.

We would call a learning problem an *idealized setting* of MDP if at each time step the performance of the algorithm is measured by $\rho_t^{\pi_t}$ instead of $\ell_t(\mathbf{x}_t, \mathbf{a}_t)$. This setting provides us a big simplification because, actually, it rules out the notion of state from the problem, since the average loss does not depend on the actual state of MDP and the action chosen in this state. We are still aiming at minimizing the corresponding regret

$$\begin{aligned}\bar{\mathcal{R}}_T &= \sum_{t=1}^T \rho_t^{\pi_t} - \min_{\pi\in\Gamma}\sum_{t=1}^T \rho_t^\pi \\ &= \sum_{t=1}^T \langle q_{\pi_t}, \ell_t\rangle - \min_{q\in\Delta}\sum_{t=1}^T \langle q, \ell_t\rangle.\end{aligned}$$

In fact, the problem becomes an online linear optimization problem with the decision set Δ, which is a subset of a probability simplex over $\mathcal{X}\times\mathcal{A}$. So we can apply the machinery developed in the previous chapter. In particular, we are going to apply the Exp3 algorithm, which in this context we would call Online Relative Entropy Policy Search (O-REPS) following the paper of Peters et al. (2010) that introduced it. The algorithm chooses the sequence of vectors $\{q_t\}_{t=1}^T$ according to the following procedure:

$$\tilde{q}_{t+1} = \operatorname*{argmin}_{q\in A}\left(\eta\langle q, \ell_t\rangle + D_R(q, q_t)\right)$$

$$q_{t+1} = \Pi_{R,\Delta}(\tilde{q}_{t+1}).$$

Where, as before, R is un-normalized negative entropy and D_R is un-normalized Kullback-Leibler divergence. The starting vector is the same as in EWA, $\tilde{q}_1(x,a) = \frac{1}{|\mathcal{X}||\mathcal{A}|}$ for all $x\in\mathcal{X}, a\in\mathcal{A}$ and $q_1 = \Pi_{R,\Delta}(\tilde{q}_1)$. After choosing q_t we can extract π_t using (2.2).

The only difference from the EWA we considered in the subsection 1.4.2 of Chapter 1 is that it used the probability simplex as a decision set and O-REPS uses Δ. From this we can conclude that $\tilde{q}_t$ will have the same form as in EWA, but the projected vector q_t will differ. If we take a closer look at the proof of Theorem 1.2, we see that it uses only the exact form of $\tilde{q}_t$. This argument shows that O-REPS in the idealized setting will have the same regret bound as EWA has. We summarize this in the next theorem and corollary which are just repetitions of the Theorem 1.2 the Corollary 1.4 in the new setting.

Theorem 2.1. *If we run O-REPS with $\eta > 0$, then for any policy $\pi \in \Gamma$*

$$\sum_{t=1}^{T} \rho_t^{\pi_t} - \sum_{t=1}^{T} \rho_t^{\pi} \leq \eta \sum_{t=1}^{T} ||\ell_t||_\infty^2 + \frac{R(p) - R(d_1)}{\eta}.$$

Corollary 2.1. *If we run O-REPS with $\eta = \sqrt{\frac{\ln |\mathcal{X}||\mathcal{A}|}{T}}$ then we have for any policy $\pi \in \Gamma$*

$$\sum_{t=1}^{T} \rho_t^{\pi_t} - \sum_{t=1}^{T} \rho_t^{\pi} \leq 2\sqrt{T \ln |\mathcal{X}||\mathcal{A}|}$$

2.4 Online loop-free stochastic shortest path problems

In this section we introduce a problem of online learning in an *episodic* MDP, which sometimes is called a online stochastic shortest path problem. We present a loop-free assumption which allows us to solve the problem efficiently, using the algorithm presented in the previous section.

2.4.1 Episodic Markovian Decision Processes

Definition 2.6 (Episodic loop-free Markovian Decision Process, E-MDP)**.** MDP $\langle \mathcal{X}, \mathcal{A}, P, P_1 \rangle$ is called an episodic loop-free MDP if it satisfies the following assumptions:

- The state space $\mathcal{X}$ can be decomposed into non-intersecting layers, i.e. $\mathcal{X} = \bigcup_{i=0}^{L} \mathcal{X}_i$ where $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for $i \neq j$.
- $\mathcal{X}_0$ and $\mathcal{X}_L$ are singletons, i.e. $\mathcal{X}_0 = \{x_0\}$ and $\mathcal{X}_L = \{x_L\}$.
- The transitions are possible only between the layers. Formally, if $P(x'|x, a) > 0$, then $x' \in \mathcal{X}_{k+1}$ and $x \in \mathcal{X}_k$ for some $0 \leq k \leq L - 1$.
- The agent always starts in the state x_0, i.e. $P_1(x_0) = 1$.

The interaction of an agent with E-MDP goes in episodes, where each episode ends when the agent reaches the state x_L. Similarly to the ordinary Markovian Decision Process, E-MDP is supplied with a sequence of loss functions $\{\ell_t\}_{t=1}^{T}$ and the functions change only between episodes (Actually, there is no sense for them to change within episodes since no state can be visited twice).

Similarly to the usual MDP, we introduce the history of interaction in the episode t

$$\mathbf{h}_t = \left\{\mathbf{x}_0^t, a_0^t, \ell_t(\mathbf{x}_0^t, \mathbf{a}_0^t),\ ..,\ \mathbf{x}_{L-1}^t, \mathbf{a}_{L-1}^t, \ell_t(\mathbf{x}_{L-1}^t, \mathbf{a}_{L-1}^t), \mathbf{x}_L^t\right\}$$

for $t = 1, \ldots, T$, where $\{(\mathbf{x}_l^t, \mathbf{a}_l^t)\}_{l=0}^{L-1}$ is a random trajectory generated by the agent in the episode t. The full history up to the episode t is

$$\mathbf{u}_t = \{\mathbf{h}_1,\ ..,\ \mathbf{h}_t\}.$$

Since each state is visited only once during an episode and a loss function changes only after the episode, there is no sense for the agent to switch policies within an episode.

Parameters: E-MDP $\langle \mathcal{X}, \mathcal{A}, P \rangle$, finite horizon T.
Environment chooses loss functions $\ell_1, ..., \ell_T : \mathcal{X} \times \mathcal{A} \to [0, 1]$.
for $t = 1$ **to** T:

1. Learner chooses a policy $\boldsymbol{\pi}_t$.
2. The current state is set to $\mathbf{x}_0$.
3. **for** $k = 0$ **to** L:
   (a) Action $\mathbf{a}_k$ is drawn from $\boldsymbol{\pi}_t(\cdot|\mathbf{x}_k)$.
   (b) Learner suffers loss $\ell_t(\mathbf{x}_k, \mathbf{a}_k)$.
   (c) The next state $\mathbf{x}_{k+1}$ is drawn from $P(\cdot|\mathbf{x}_k, \mathbf{a}_k)$.
4. ℓ_t (full information) or $\{\ell_t(\mathbf{x}_k, \mathbf{a}_k)\}_{k=0}^{L-1}$ (bandit information) is revealed.

Figure 2.2: Online Loop-free Stochastic Shortest Path Problem

Hence, we specify an agent by a sequence of policies $\{\boldsymbol{\pi}_t\}_{t=1}^T$, where $\boldsymbol{\pi}_t$ is a policy to follow in the episode t. Using the notation introduced

$$\boldsymbol{\pi}_t(a|x) = \mathbb{P}\left[\mathbf{a} = a \middle| \mathbf{x} = x, \mathbf{u}_{t-1}\right].$$

The expected cumulative loss of an agent in the episode t is

$$c_t(\boldsymbol{\pi}_t) = \mathbb{E}\left[\sum_{k=0}^{L-1} \ell_t(\mathbf{x}_k^t, \mathbf{a}_k^t) \middle| \mathbf{u}_{t-1}\right].$$

The total expected cumulative loss of an agent is

$$\hat{L}_T = \mathbb{E}\left[\sum_{t=1}^{T} c_t(\boldsymbol{\pi}_t)\right].$$

If $\{(\mathbf{x}'_k, \mathbf{a}'_k)\}_{t=1}^T$ is a random trajectory generated by a stationary stochastic policy π, then we define the expected cumulative loss of a policy π in the episode t as

$$c_t(\pi) = \mathbb{E}\left[\sum_{k=0}^{L-1} \ell_t(\mathbf{x}'_k, \mathbf{a}'_k)\right].$$

The total expected cumulative loss of a policy π is

$$L_T(\pi) = \sum_{t=1}^{T} c_t(\pi).$$

Finally, we measure the performance of an algorithm by means of its regret with respect to the class of stationary stochastic policies.

$$\mathcal{R}_T = \hat{L}_T - \min_{\pi \in \Gamma} L_T(\pi)$$

The resulting problem is described in Figure 2.2.

2.4.2 Stationary distributions in episodic Markovian Decision Processes

In this subsection we introduce the concept of *local stationary state distribution* and show how it is connected to the notions introduced in the previous chapters. This connection will allow us to use O-REPS to solve the problem.

In section 2.2 we showed that if we fix a policy π then it induces a stationary state distribution μ_π over the whole state space $\mathcal{X}$ (under Assumption 1). This is not true in E-MDP, instead every policy induces a probability distribution over each layer $\mathcal{X}_k$. Actually, we can compute these distributions explicitly. Denote by ν_k^π a distribution over k-th layer induced by an agent that follows a policy π, i.e. $\nu_k^\pi(x) = \mathbb{P}[\mathbf{x}_k = x]$. Since the 0-th layer is a singleton, $\nu_0^\pi(x_0) = 1$. We are going to compute these distributions recursively, for the layer $k \geq 1$ and for $x \in \mathcal{X}_k$

$$\begin{aligned}
\nu_k^\pi(x) &= \mathbb{P}[\mathbf{x}_k = x] \\
&= \sum_{x' \in \mathcal{X}_{k-1}} \mathbb{P}[\mathbf{x}_k = x | \mathbf{x}_{k-1} = x'] \mathbb{P}[\mathbf{x}_{k-1} = x'] \\
&= \sum_{x' \in \mathcal{X}_{k-1}} \mathbb{P}[\mathbf{x}_k = x | \mathbf{x}_{k-1} = x'] \nu_{k-1}^\pi(x') \\
&= \sum_{x' \in \mathcal{X}_{k-1}} \sum_{a \in \mathcal{A}} \mathbb{P}[\mathbf{x}_k = x | \mathbf{x}_{k-1} = x', \mathbf{a}_{k-1} = a] \mathbb{P}[\mathbf{a}_{k-1} = a] \nu_{k-1}^\pi(x') \\
&= \sum_{x' \in \mathcal{X}_{k-1}, a \in \mathcal{A}} P(x|x', a)\pi(a|x')\nu_{k-1}^\pi(x').
\end{aligned}$$

And we will make this a definition.

Definition 2.7 ((local) stationary state distributions)**.** A family of distributions $\{\nu_k^\pi\}_{k=0}^L$ is called a family of local stationary state distributions generated by π if $\nu_0^\pi(x_0) = 1$ and

$$\nu_k^\pi(x) = \sum_{x' \in \mathcal{X}_{k-1}, a \in \mathcal{A}} P(x|x', a)\pi(a|x')\nu_{k-1}^\pi(x') \tag{2.5}$$

for all $x \in \mathcal{X}_k$.

Let us introduce a notation l_x, which denotes the number of the layer that x belongs to, i.e. $l_x = k$ iff $x \in \mathcal{X}_k$. In the next two lemmas we are establishing the connection between two notions of stationary state distribution. The intuition behind this connection is following: imagine we fix a policy and we let the agent run through the MDP forever. At some moment you stop it and you want to compute the probability that the agent is in some particular state x. If you knew the layer, you could just say that this is $\nu_{l_x}^\pi(x)$. Therefore, we need some prior over the layers to use this way of reasoning. Since we have no notion of a time spent in a particular layer, the most suitable prior is the uniform one. As we will see, this is indeed the case.

Lemma 2.1 (From local to global distributions)**.** *For any policy $\pi \in \Gamma$, if we define $\mu_\pi(x) = \nu_{l_x}^\pi(x)\frac{1}{L}$ for all $x \in \mathcal{X}$, then μ_π is a global stationary state distribution generated by π.*

Proof. Using the definition of a global stationary state distribution in the form of (2.4), we need to prove that for all $x \in \mathcal{X}$

$$\mu_\pi(x) = \sum_{x',a'} \mu_\pi(x')\pi(a'|x')P(x|x',a')$$

Fix some $x \in \mathcal{X}$. Let us start with the right-hand side

$$\begin{aligned}
\sum_{x',a'} \mu_\pi(x')\pi(a'|x')P(x|x',a') &= \sum_{k=0}^{L} \sum_{x'\in\mathcal{X}_k, a'\in\mathcal{A}} \mu_\pi(x')\pi(a'|x')P(x|x',a') \\
&= \sum_{x'\in\mathcal{X}_{l_x-1}, a'\in\mathcal{A}} \mu_\pi(x')\pi(a'|x')P(x|x',a') \\
&= \frac{1}{L} \sum_{x'\in\mathcal{X}_{l_x-1}, a'\in\mathcal{A}} \nu^\pi_{l_x-1}(x')\pi(a'|x')P(x|x',a') \\
&= \frac{1}{L}\nu_{l_x}(x) \\
&= \mu_\pi(x),
\end{aligned}$$

where we used the fact that $P(x|x',a') = 0$ for $x' \notin \mathcal{X}_{l_x-1}$ in the second line, the definition of $\mu_\pi(x')$ in the third and the equation 2.5 in the fourth. □

Definition 2.8 (Localizable distribution)**.** The distribution μ over $\mathcal{X}$ is called *localizable* if

$$\sum_{x\in\mathcal{X}_k} \mu(x) = \frac{1}{L} \tag{2.6}$$

for $k = 1, \ldots, L$.

Lemma 2.2 (From global to local distributions)**.** *Under Assumption 1, for any policy $\pi \in \Gamma$, let μ_π be a global stationary state distribution generated by it. Define a family of distributions $\{\nu^\pi_k\}_{k=0}^L$ by $\nu^\pi_{l_x}(x) = L\mu_\pi(x)$ for all $x \in \mathcal{X}$. If μ_π is localizable, then $\{\nu^\pi_k\}_{k=0}^L$ are local stationary state distributions.*

Proof. First, note that each of ν^π_k is indeed a distribution over $\mathcal{X}_k$, by the localizable property of μ_π. Let us check the 0-th layer. Since x_0 is the only state in $\mathcal{X}_0$, $\mu_\pi(x_0) = \frac{1}{L}$ and $\nu^\pi_0(x_0) = L\mu_\pi(x_0) = 1$. What is left is to check (2.5). Again, we fix a state x. Since $P(x|x',a') = 0$ if $x' \notin \mathcal{X}_{l_x-1}$, we can add zeros to the right-hand side of (2.5) and write it as

$$\begin{aligned}
\sum_{x'\in\mathcal{X}_{l_x-1}, a\in\mathcal{A}} P(x|x',a)\pi(a|x')\nu^\pi_{l_x-1}(x') &= \sum_{x'\in\mathcal{X}, a\in\mathcal{A}} P(x|x',a)\pi(a|x')\nu^\pi_{l_{x'}}(x') \\
&= L\sum_{x',a} P(x|x',a)\pi(a|x')\mu_\pi(x') \\
&= L\mu_\pi(x) \ \ (2.4) \\
&= \nu^\pi_{l_x}(x).
\end{aligned}$$

□

2.4.3 Episodic O-REPS for learning with full information

Lemmas 2.1 and 2.2 provide us the machinery to use O-REPS in episodic MDP. The idea is that since O-REPS is aiming to minimize the average loss, we would like to find a way to connect the average loss and the cumulative loss in the episode. Now we are ready to do it.

Lemma 2.3.

$$c_t(\pi) = L\rho_t^\pi.$$

Proof.

$$\begin{aligned}
c_t(\pi) &= \mathbb{E}\left[\sum_{k=0}^{L-1} \ell_t(\mathbf{x}'_k, \mathbf{a}'_k)\right] \\
&= \sum_{k=0}^{L-1} \mathbb{E}\left[\ell_t(\mathbf{x}'_k, \mathbf{a}'_k)\right] \\
&= \sum_{k=0}^{L-1} \sum_{x\in\mathcal{X}_k, a\in\mathcal{A}} \ell_t(x,a)\mathbb{P}\left[\mathbf{x}'_k = x, \mathbf{a}'_k = a\right] \\
&= \sum_{x\in\mathcal{X}, a\in\mathcal{A}} \ell_t(x,a)\mathbb{P}\left[\mathbf{x}'_{l_x} = x, \mathbf{a}'_{l_x} = a\right] \\
&= \sum_{x,a} \ell_t(x,a)\mathbb{P}\left[\mathbf{a}'_{l_x} = a \middle| \mathbf{x}'_{l_x} = x\right]\mathbb{P}\left[\mathbf{x}'_{l_x} = x\right] \\
&= \sum_{x,a} \ell_t(x,a)\pi(a|x)\nu_{l_x}^\pi(x) \\
&= L\sum_{x,a} \ell_t(x,a)\pi(a|x)\mu_\pi(x) \text{ (Lemma 2.1)} \\
&= L\rho_t^\pi.
\end{aligned}$$

□

Recall that O-REPS chooses a vector $q \in \Delta$ and then extracts a policy to follow using (2.2). In the episodic case we can not do this for every $q \in \Delta$ (we can extract a policy, but we will lose the connection between losses), but if the corresponding stationary state distribution μ_π is localizable, this is possible. Denote the corresponding space by Θ. Combining (2.3) and definition 2.8, we can write the space Θ as

$$\Theta = \left\{ q \in \Delta : \sum_{x\in\mathcal{X}_k, a\in\mathcal{A}} q(x,a) = \frac{1}{L} \text{ for } k = 1..L \right\}.$$

Note that for every policy $\pi \in \Gamma$, vector $q(x,a) = \frac{1}{L}\pi(a|x)\nu_{l_x}^\pi(x) \in \Theta$. Finally, episodic O-REPS chooses a sequence of points $\{q_t\}_{t=1}^T$ using the following steps

$$\tilde{q}_{t+1} = \operatorname*{argmin}_{q\in A} \left(\eta\langle q, \ell_t\rangle + D_R(q, q_t)\right)$$

$$q_{t+1} = \Pi_{R,\Theta}(\tilde{q}_{t+1}).$$

Where D_R is, as before, an un-normalized Kullback-Leibler divergence and R is un-normalized negative entropy. Now we are going to express these two steps explicitly. Actually, the formula for $\tilde{q}_{t+1}$ remains unchanged:

$$\tilde{q}_{t+1}(x,a) = q_t(x,a)e^{-\eta\ell_t(x,a)}.$$

The projection step requires more care. Beforehand, let us introduce some more notations. For any function $v(x) : \mathcal{X} \to R$ (which are called value functions in the reinforcement learning literature) and loss function $\ell : \mathcal{X} \times \mathcal{A} \to [0,1]$ we define a corresponding Bellmann-error function

$$\delta(x,a|v,\ell) = -\eta\ell(x,a) - \sum_{x'\in\mathcal{X}} v(x')P(x'|x,a) + v(x). \tag{2.7}$$

Now let us write the projection step as an optimization problem with all constraints:

$$\min_q D_R(q,\tilde{q}_{t+1})$$

$$\sum_a q(x,a) = \sum_{x',a'} q(x',a')P(x|x',a')$$

$$\sum_{x\in\mathcal{X}_k,a\in\mathcal{A}} q(x,a) = \frac{1}{L} \text{ for } k = 1,\dots,L.$$

Actually, there is one more requirement that $\sum_{x,a} q(x,a) = 1$, but it is included in the second constraint. To solve the problem we write the Lagrangian:

$$\begin{aligned}\mathcal{L} &= D_R(q,\tilde{q}_{t+1}) + \sum_{k=0}^{L}\lambda_k\left(\sum_{x\in\mathcal{X}_k,a\in\mathcal{A}} q(x,a) - \frac{1}{L}\right)\\ &+ \sum_x v(x)\left(\sum_{x',a'} q(x',a')P(x|x',a') - \sum_a q(x,a)\right)\\ &= D_R(q,\tilde{q}_{t+1}) + \sum_{x,a} q(x,a)\left(\lambda_{l_x} + \sum_{x'} v(x')P(x'|x,a) - v(x)\right) - \frac{1}{L}\sum_{k=0}^{L}\lambda_k.\end{aligned}$$

Where $\{\lambda_k\}_{k=0}^L$ and $\{v(x)\}_{x\in\mathcal{X}}$ are Lagrange multipliers. We differentiate and set the derivatives to zero:

$$\nabla_{q(x,a)}\mathcal{L} = \ln q(x,a) - \ln\tilde{q}_{t+1}(x,a) + \lambda_{l_x} + \sum_{x'} v(x')P(x'|x,a) - v(x) = 0.$$

Hence, we obtain the formula for $q(x,a)$

$$q(x,a) = \tilde{q}_{t+1}(x,a)e^{-\lambda_{l_x} - \sum_{x'} v(x')P(x'|x,a) + v(x)}.$$

Substituting the formula for $\tilde{q}_{t+1}(x,a)$

$$q(x,a) = q_t(x,a)e^{-\lambda_{l_x} + \delta(x,a|v,\ell_t)}.$$

Using the second constraint, we have for every $k = 1,\dots,L$

$$\sum_{x\in\mathcal{X}_k,a\in\mathcal{A}} q_t(x,a)e^{-\lambda_k + \delta(x,a|v,\ell_t)} = \frac{1}{L}.$$

Let us introduce notation for every $k = 1, \dots, L$

$$N(v, k, \ell) = \sum_{x \in \mathcal{X}_k, a \in \mathcal{A}} q_t(x, a) e^{\delta(x,a|v,\ell)}.$$

Then

$$e^{-\lambda_k} = \frac{1}{LN(v, k, \ell_t)}.$$

So the final expression for the $q_{t+1}(x, a)$ is

$$q_{t+1}(x, a) = \frac{q_t(x, a) e^{\delta(x,a|v_{t+1},\ell_t)}}{LN(v_t, l_x, \ell_t)}.$$

where v_{t+1} is determined solving the dual problem. If we substitute back the equation for q in the Lagrangian, then the dual function is

$$\sum_{x,a} \tilde{q}_{t+1}(x, a) - 1 - \frac{1}{L} \sum_{k=0}^{L} \lambda_k.$$

And we need to maximize it. We can drop the constants and substitute the formula for λ_k

$$-\sum_{k=0}^{L} \ln N(v, k, \ell_t).$$

So the final equation for v_{t+1} is

$$v_{t+1} = \operatorname*{argmin}_v \sum_{k=0}^{L} \ln N(v, k, \ell_t).$$

This last minimization is a convex optimization problem (see, for example, Boyd and Vandenberghe (2004)) and can be performed numerically.

Algorithm 4: O-REPS for episodic loop-free MDP

Parameters: *finite horizon* T, η
Compute $v_1 = \operatorname{argmin}_v \sum_{k=0}^{L} \ln N(v, k, 0)$;
Compute q_1 component-wise as $q_1(x, a) = \frac{e^{\delta(x,a|v_{t+1},0)}}{L \sum_{x' \in \mathcal{X}_{l_x}, a'} e^{\delta(x',a'|v,0)}}$;
Output π_1 as a policy to follow in the episode;
Receive ℓ_1;
for $t = 2, \dots, T$ **do**
  Compute $v_t = \operatorname{argmin}_v \sum_{k=0}^{L} \ln \sum_{x \in \mathcal{X}_k, a} q_{t-1}(x, a) e^{\delta(x,a|v,\ell_t)}$;
  Compute q_t component-wise as $q_t(x, a) = \frac{q_{t-1}(x,a) e^{\delta(x,a|v_t,\ell_t)}}{L \sum_{x' \in \mathcal{X}_{l_x}, a'} q_{t-1}(x', a') e^{\delta(x',a'|v_t,\ell_t)}}$;
  Compute π_t using $\pi_t(a|x) = \frac{q_t(x,a)}{\sum_b q_t(x,b)}$;
  Output π_t as a policy to follow in the episode;
  Receive ℓ_t;
end

The following theorem states the regret bound for this version of O-REPS.

Theorem 2.2. *If we run episodic O-REPS with parameter* $\eta = \sqrt{\frac{\ln|\mathcal{X}|\mathcal{A}|}{T}}$*, then for any* $\pi \in \Gamma$

$$\hat{L}_T - L_T(\pi) \leq 2L\sqrt{T \ln |\mathcal{X}|\mathcal{A}|}$$

Proof. The proof is just a combination of two previously proven results: Corollary 2.1 and Lemma 2.3:

$$\begin{aligned}\hat{L}_T - L_T(\pi) &= \sum_{t=1}^{T} c_t(\boldsymbol{\pi}_t) - \sum_{t=1}^{T} c_t(\pi) \\ &= L(\sum_{t=1}^{T} \rho_t^{\boldsymbol{\pi}_t} - \sum_{t=1}^{T} \rho_t^{\pi}) \\ &\leq 2L\sqrt{T \ln |\mathcal{X}|\mathcal{A}|}.\end{aligned}$$

□

2.4.4 Episodic O-REPS for learning with bandit information

In this subsection we turn our attention to the bandit version of the learning problem presented in the previous sections. Similarly to the subsection 1.4.3, we introduce the estimator of the loss and use the full-information algorithm as a black-box.

The estimates $\hat{\ell}_t$ that we are going to use are built on the same idea as the estimators for f_t in multi-armed bandit problem:

$$\hat{\ell}_t(x,a) = \begin{cases} \frac{\ell_t(x,a)}{\mathbf{q}_t(x,a)} & \text{if } (x,a) \text{ was visited in the episode } t \\ 0 & \text{otherwise} \end{cases}. \tag{2.8}$$

We denote the event $\{(x,a)$ was visited in the episode $t\}$ by $\{(x,a) \in \mathbf{h}_t\}$. Using this, we write (2.8) as

$$\hat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{\mathbf{q}_t(x,a)} \mathbb{I}\left[(x,a) \in \mathbf{h}_t\right]. \tag{2.9}$$

The interesting issue with this estimator is that the sample state (x,a) comes not from the distribution $\mathbf{q}_t(x,a)$ as it was in the multi-armed bandit case. Logically, we should have used $\pi_t(a|x)\nu_t^{\pi_t}(x)$ in the denominator, but this would lead to the worse dependence on L in the final bound. Generally, the problem of samples that come not from the chosen distribution is a big issue in online learning problems and, for example, this is what will prevent the direct extension of the results for online learning in MDP to the bandit case (we will present these results in the next sections).

The next theorem states the regret bound for the bandit episodic O-REPS. The idea of the proof is exactly the same as in the multi-armed bandit case.

Theorem 2.3. *If we run bandit episodic O-REPS in the E-MDP with parameter* $\eta = \frac{1}{L}\sqrt{\frac{\ln|\mathcal{X}||\mathcal{A}|}{T|\mathcal{X}||\mathcal{A}|}}$*, then for any* $\pi \in \Gamma$

$$\hat{L}_T - L_T(\pi) \leq 2L\sqrt{T|\mathcal{X}||\mathcal{A}| \ln |\mathcal{X}||\mathcal{A}|}.$$

Algorithm 5: O-REPS for bandit episodic loop-free MDP

Parameters: *finite horizon* T, η

Set $\boldsymbol{\pi}_1(a|x) = \frac{1}{|\mathcal{A}|}$ for all $x \in \mathcal{X}, a \in \mathcal{A}$;

Compute $\{\nu_k^{\pi_1}(x)\}_{k=0}^{L}$ recursively and set $\mathbf{q}_1(x,a) = \frac{\nu_{l_x}^{\pi_1}(x)}{L|\mathcal{A}|}$;

Output $\boldsymbol{\pi}_1$ as a policy to follow in the episode;

Receive $\{\ell_1(\mathbf{x}_k^1, \mathbf{a}_k^1)\}_{k=0}^{L-1}$;

Compute $\hat{\ell}_1$ as $\hat{\ell}_1(x,a) = \frac{\ell_1(x,a)}{\mathbf{q}_1(x,a)}\mathbb{I}\left[(x,a) \in \mathbf{h}_1\right]$;

for $t = 2, \ldots, T$ **do**

Compute $v_t = \operatorname{argmin}_v \sum_{k=0}^{L} \ln \sum_{x \in \mathcal{X}_k, a} \mathbf{q}_{t-1}(x,a) e^{\delta(x,a|v,\hat{\ell}_t)}$;

Compute $\mathbf{q}_t$ component-wise as $\mathbf{q}_t(x,a) = \frac{\mathbf{q}_{t-1}(x,a)e^{\delta(x,a|v_t,\hat{\ell}_t)}}{L\sum_{x' \in \mathcal{X}_k, a'} \mathbf{q}_{t-1}(x',a')e^{\delta(x',a'|v_t,\hat{\ell}_t)}}$;

Compute $\boldsymbol{\pi}_t$ using $\boldsymbol{\pi}_t(a|x) = \frac{\mathbf{q}_t(x,a)}{\sum_b \mathbf{q}_t(x,b)}$;

Output $\boldsymbol{\pi}_t$ as a policy to follow in the episode;

Receive $\{\ell_t(\mathbf{x}_k^t, \mathbf{a}_k^t)\}_{k=0}^{L-1}$;

Compute $\hat{\ell}_t$ as $\hat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{\mathbf{q}_t(x,a)}\mathbb{I}\left[(x,a) \in \mathbf{h}_t\right]$;

end

Proof. As usual, our starting point is Lemma 1.5. Denote by $\hat{\rho}_t^{\pi}$ the average estimated loss of a policy π. Recall that O-REPS is an instance of PPA algorithm and it is run on the sequence of losses $\left\{\hat{\ell}_t\right\}_{t=1}^{T}$, hence the lemma takes the following form

$$\begin{aligned}\sum_{t=1}^{T} \hat{\rho}_t^{\boldsymbol{\pi}_t} - \sum_{t=1}^{T} \hat{\rho}_t^{\pi} &= \sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{q}_t \rangle - \sum_{t=1}^{T} \langle \hat{\ell}_t, q \rangle \\ &\leq \sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{q}_t \rangle - \sum_{t=1}^{T} \langle \hat{\ell}_t, \tilde{\mathbf{q}}_{t+1} \rangle + \frac{D_R(q, \mathbf{q}_1)}{\eta}. \end{aligned} \tag{2.10}$$

Using the same argument that was used in Theorem 1.2 and Theorem 1.3

$$\sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{q}_t \rangle - \sum_{t=1}^{T} \langle \hat{\ell}_t, \tilde{\mathbf{q}}_{t+1} \rangle \leq \eta \sum_{t=1}^{T} \langle \mathbf{q}_t \circ \hat{\ell}_t, \hat{\ell}_t \rangle.$$

We use the exact form of the estimator to upper bound $\mathbf{q}_t \circ \hat{\ell}_t$

$$\begin{aligned}\mathbf{q}_t \circ \hat{\ell}_t &= \sum_{x,a} \mathbf{q}_t(x,a) \frac{\ell_t(x,a)}{\mathbf{q}_t(x,a)} \mathbb{I}\left[(x,a) \in \mathbf{h}_t\right] \\ &= \sum_{x,a} \ell_t(x,a) \mathbb{I}\left[(x,a) \in \mathbf{h}_t\right] \\ &\leq \sum_{x,a} \mathbb{I}\left[(x,a) \in \mathbf{h}_t\right] \\ &\leq L.\end{aligned}$$

Hence,

$$\langle \mathbf{q}_t \circ \hat{\ell}_t, \hat{\ell}_t \rangle \leq L \sum_{x,a} \hat{\ell}_t(x,a).$$

Combining this with (2.10), we get

$$\sum_{t=1}^{T} \hat{\rho}_t^{\boldsymbol{\pi}_t} - \sum_{t=1}^{T} \hat{\rho}_t^{\pi} \leq \eta L \sum_{t=1}^{T} \sum_{x,a} \hat{\ell}_t(x,a) + \frac{D_R(q, \mathbf{q}_1)}{\eta}. \tag{2.11}$$

Next, we are going to take an expectation on the both sides. Let us compute the terms separately. First, the right-hand side:

$$\begin{aligned}
\mathbb{E}\left[\hat{\ell}_t(x,a) \middle| \mathbf{u}_{t-1}\right] &= \frac{\ell_t(x,a)}{\mathbf{q}_t(x,a)} \mathbb{P}\left[(x,a) \in \mathbf{h}_t \middle| \mathbf{u}_{t-1}\right] \\
&= \frac{\ell_t(x,a)}{\boldsymbol{\pi}_t(a|x)\mu_{\boldsymbol{\pi}_t}(x)} \mathbb{P}\left[\mathbf{x}_{l_x} = x, a_{l_x} = a \middle| \mathbf{u}_{t-1}\right] \\
&= \frac{\ell_t(x,a)}{\frac{1}{L}\boldsymbol{\pi}_t(a|x)\nu_{l_x}^{\boldsymbol{\pi}_t}(x)} \mathbb{P}\left[a_{l_x} = a \middle| \mathbf{u}_{t-1}, \mathbf{x}_{l_x} = x\right] \mathbb{P}\left[\mathbf{x}_{l_x} = x \middle| \mathbf{u}_{t-1}\right] \\
&= \frac{L\ell_t(x,a)}{\boldsymbol{\pi}_t(a|x)\nu_{l_x}^{\boldsymbol{\pi}_t}(x)} \boldsymbol{\pi}_t(a|x)\nu_{l_x}^{\boldsymbol{\pi}_t}(x) \\
&= L\ell_t(x,a).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \sum_{x,a} \hat{\ell}_t(x,a)\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{x,a} \hat{\ell}_t(x,a)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\sum_{x,a} \hat{\ell}_t(x,a) \middle| \mathbf{u}_{t-1}\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{x,a} \mathbb{E}\left[\hat{\ell}_t(x,a) \middle| \mathbf{u}_{t-1}\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{x,a} L\ell_t(x,a)\right] \\
&\leq TL|\mathcal{X}||\mathcal{A}|.
\end{aligned}$$

Now we deal with the expectations on the left-hand side

$$\begin{aligned}
\mathbb{E}\left[\hat{\rho}_t^{\pi}\right] &= \sum_{x,a} \mathbb{E}\left[\hat{\ell}_t(x,a) q_\pi(x,a)\right] \\
&= \sum_{x,a} \mathbb{E}\left[\mathbb{E}\left[\hat{\ell}_t(x,a) \middle| \mathbf{u}_{t-1}\right] q_\pi(x,a)\right] \\
&= \sum_{x,a} \mathbb{E}\left[L\ell_t(x,a) q_\pi(x,a)\right] \\
&= L \sum_{x,a} \ell_t(x,a) \frac{1}{L} \pi(a|x) \nu^{\pi}(x) \\
&= c_t(\pi).
\end{aligned}$$

After taking expectation on the both sides and bounding $D_R(q, \mathbf{q}_1)$ by $\ln|\mathcal{X}||\mathcal{A}|$, the (2.11) becomes

$$\hat{L}_T - L_T(\pi) \leq \eta L^2 T |\mathcal{X}||\mathcal{A}| + \frac{\ln|\mathcal{X}||\mathcal{A}|}{\eta}.$$

Optimizing over the η yields the result. $\square$

2.5 Unichain Markovian Decision Processes

In this section we present an approach to solve the main problem of this chapter, an online learning problem in MDPs as described in the Section 2.1.

2.5.1 Mixing times

Actually, the solution was implicitly described in the Section 2.3. We build an algorithm to minimize the average loss instead of the actual one. The logical question: is there any connection between two? The answer turns out to be yes, under some assumptions. All our arguments are based on the existence of so-called *uniform mixing time*. Before giving a meaning to this notion, we need a few definitions.

Definition 2.9 (Transition matrix)**.** The matrix P^π is called a transition matrix induced by π, if each component $(P^\pi)_{x,x'}$ is a probability of getting into the state x' from x under policy π.

Note that the definition implies

$$\begin{aligned}(P^\pi)_{x,x'} &= \mathbb{P}\left[\mathbf{x}_{t+1} = x' \middle| \mathbf{x}_t = x\right] \\ &= \sum_a \mathbb{P}\left[\mathbf{x}_{t+1} = x' \middle| \mathbf{x}_t = x, \mathbf{a}_t = a\right] \mathbb{P}\left[\mathbf{a}_t = a \middle| \mathbf{x}_t = x\right] \\ &= \sum_a P(x'|x,a)\pi(a|x).\end{aligned}$$

Furthermore, if we treat the distribution μ over the states as a row vector, then μP^π is a one step distribution from μ under π. Note that the usual definition of the stationary state distribution μ_π is the following: it is the distribution that satisfies

$$\mu P^\pi = \mu. \tag{2.12}$$

In fact, the two definitions are equivalent, which can be seen by writing the (2.12) component-wise and observing that in this way it becomes (2.4). Now we are ready to define the mixing time.

Definition 2.10 (Mixing time)**.** τ_π is called a mixing time for a policy $\pi \in \Gamma$ if for any two distributions over the state space μ and μ'

$$||\mu P^\pi - \mu' P^\pi||_1 \leq e^{-1/\tau_\pi} ||\mu - \mu'||_1 \tag{2.13}$$

As we already mentioned, our second assumption is the existence of the uniform mixing time.

Assumption 2. *There exists a fixed positive uniform mixing time τ, such that $\tau_\pi \leq \tau$ for all $\pi \in \Gamma$.*

Note that this assumption implies

$$\sup_{\pi\in\Gamma} ||\mu P^{\pi} - \mu' P^{\pi}||_1 \leq e^{-1/\tau} ||\mu - \mu'||_1 \tag{2.14}$$

for any two distribution μ and μ' over the states.

The Markovian Decision Process that satisfies the Assumption 2 is called *unichain*.

Under the Assumption 2 we can find a connection between the average and the expected losses. Beforehand, we define by $\mu_{\pi,t}$ the distribution over states at time step t resulting in following policy π from the very beginning, i.e. from distribution P_1:

$$\mu_{\pi,t} = P_1 (P^{\pi})^{t-1}.$$

The following proposition establishes the rate of convergence of $\mu_{\pi,t}$ to μ_{π}.

Proposition 2.1. *For any policy* $\pi \in \Gamma$

$$||\mu_{\pi,t} - \mu_{\pi}||_1 \leq 2e^{-(t-1)/\tau_{\pi}}.$$

Proof. The proof is based on the definition of stationary state distribution in the form (2.12) and the definition of mixing time (2.13).

$$\begin{aligned} ||\mu_{\pi,t} - \mu_{\pi}||_1 &= ||\mu_{\pi,t-1} P^{\pi} - \mu_{\pi} P^{\pi}||_1 \\ &\leq e^{-1/\tau_{\pi}} ||\mu_{\pi,t-1} - \mu_{\pi}||_1 \\ &\leq e^{-(t-1)/\tau_{\pi}} ||P_1 - \mu_{\pi}||_1 \\ &\leq 2e^{-(t-1)/\tau_{\pi}} \end{aligned}$$

where in the last step we used the fact that for any two distributions μ and μ', we have $||\mu - \mu'||_1 \leq 2$. □

Corollary 2.2. *Under Assumption 2*

$$||\mu_{\pi,t} - \mu_{\pi}||_1 \leq 2e^{-(t-1)/\tau}.$$

for any $\pi \in \Gamma$.

Note that this corollary shows us that the Assumption 2 implies the Assumption 1. This is because we can define the μ_{π} as the limiting distribution that results if we let the agent that follows policy π run in the MDP forever. From this, the corresponding state-action stationary distribution q_{π} is just $\pi(a|x)\mu_{\pi}(x)$ for all x, a.

Hereafter, we state all the result under the Assumption 2. Now, using the Corollary 2.2, we can connect the average loss of a fixed policy π to the expected loss of it.

Lemma 2.4. *Assume that we follow a policy* π *from step 1. Then for any* t

$$\rho_t^{\pi} - \mathbb{E}\left[\ell_t(\mathbf{x}_t, \mathbf{a}_t)\right] \leq 2e^{-(t-1)/\tau}.$$

Proof.

$$\begin{aligned} \rho_t^{\pi} - \mathbb{E}\left[\ell_t(\mathbf{x}_t, \mathbf{a}_t)\right] &= \sum_{x,a} \mu_{\pi}(x)\pi(a|x)\ell_t(x,a) - \sum_{x,a} \mu_{\pi,t}(x)\pi(a|x)\ell_t(x,a) \\ &= \sum_{x} (\mu_{\pi}(x) - \mu_{\pi,t}(x)) \sum_{a} \pi(a|x)\ell_t(x,a) \\ &\leq \sum_{x} (\mu_{\pi}(x) - \mu_{\pi,t}(x)) \\ &\leq ||\mu_{\pi} - \mu_{\pi,t}||_1 \\ &\leq 2e^{-(t-1)/\tau}. \end{aligned}$$

□

Corollary 2.3.

$$\sum_{t=1}^{T} \rho_t^\pi - L_T(\pi) \leq 2(1+\tau).$$

Proof.

$$\begin{aligned}
\sum_{t=1}^{T} \rho_t^\pi - L_T(\pi) &\leq \sum_{t=1}^{T} (\rho_t^\pi - \mathbb{E}\left[\ell_t(\mathbf{x}_t, \mathbf{a}_t)\right]) \\
&\leq \sum_{t=1}^{T} 2e^{-(t-1)/\tau} \\
&\leq 2(1 + \int_1^\infty e^{-(t-1)/\tau} dt) \\
&\leq 2(1+\tau).
\end{aligned}$$

□

We see from Corollary 2.3 that if we do not change the policy, then our average loss is not far from the expected one. Hence, the desired property of the algorithm is not changing policies too often or change them in such a way that they are close to each other in some sense. As we will see further, this is indeed the case for the O-REPS.

2.5.2 O-REPS for learning with full information

As we mentioned, we use the O-REPS defined as in section 2.3. Recall that at every time step it chooses a state-action distribution q_t according to the following procedure

$$\tilde{q}_{t+1} = \operatorname*{argmin}_{q \in A} \left(\eta \langle q, \ell_t \rangle + D_R(q, q_t)\right)$$

$$q_{t+1} = \Pi_{R,\Delta}(\tilde{q}_{t+1}).$$

We have already obtained the closed form for the $\tilde{q}_{t+1}$, that is,

$$\tilde{q}_{t+1}(x,a) = q_t(x,a) e^{-\eta \ell_t(x,a)}.$$

To obtain the closed form for the q_{t+1}, we need to solve the constrained optimization problem:

$$\begin{aligned}
&\min_q D_R(q, \tilde{q}_{t+1}) \\
&\sum_a q(x,a) = \sum_{x',a'} q(x',a') P(x|x',a') \\
&\sum_{x,a} q(x,a) = 1.
\end{aligned}$$

We proceed as in the subsection 2.4.3. First, we write the Lagrangian

$$\begin{aligned}
\mathcal{L} &= D_R(q, \tilde{q}_{t+1}) + \lambda \left(\sum_{x,a} q(x,a) - 1 \right) + \sum_x v(x) \left(\sum_{x',a'} q(x',a') P(x|x',a') - \sum_a q(x,a) \right) \\
&= D_R(q, \tilde{q}_{t+1}) + \sum_{x,a} q(x,a) \left(\lambda + \sum_{x'} v(x') P(x'|x,a) - v(x) \right) - \lambda.
\end{aligned}$$

Where λ and $\{v(x)\}_{x\in\mathcal{X}}$ are Lagrange multipliers. We differentiate and set the derivatives to zero:

$$\nabla_{q(x,a)}\mathcal{L} = \ln q(x,a) - \ln \tilde{q}_{t+1}(x,a) + \lambda + \sum_{x'} v(x')P(x'|x,a) - v(x) = 0.$$

Hence, we obtain the formula for $q(x,a)$

$$q(x,a) = \tilde{q}_{t+1}(x,a)e^{-\lambda-\sum_{x'} v(x')P(x'|x,a)+v(x)}.$$

Substituting the formula for $\tilde{q}_{t+1}(x,a)$

$$q(x,a) = q_t(x,a)e^{-\lambda+\delta(x,a|v,\ell_t)}.$$

Where we used the Bellmann-error function introduced in the subsection 2.4.3, see (2.7) . Using the second constraint, we have

$$\sum_{x,a} q_t(x,a)e^{-\lambda+\delta(x,a|v,\ell_t)} = 1.$$

Hence,

$$e^{-\lambda} = \frac{1}{\sum_{x,a} q_t(x,a)e^{\delta(x,a|v,\ell_t)}},$$

and the final formula for $q_{t+1}(x,a)$ is

$$q_{t+1}(x,a) = \frac{q_t(x,a)e^{\delta(x,a|v_{t+1},\ell_t)}}{\sum_{x,a} q_t(x,a)e^{\delta(x,a|v_{t+1},\ell_t)}},$$

where v_{t+1} is determined solving the dual problem. If we substitute back the equation for q in the Lagrangian, then the dual function is

$$\sum_{x,a} \tilde{q}_{t+1}(x,a) - 1 - \lambda.$$

And we need to maximize it. We can drop the constants and substitute the formula for λ

$$-\ln \sum_{x,a} q_t(x,a)e^{\delta(x,a|v,\ell_t)}.$$

The final equation for v_{t+1} is

$$v_{t+1} = \operatorname*{argmin}_{v} \left(\ln \sum_{x,a} q_t(x,a)e^{\delta(x,a|v,\ell_t)} \right).$$

Again, as in the previous derivations, the last minimization is a convex optimization problem and can be performed numerically.

Now, when we have the full description of the algorithm, we can prove that, if we follow the policies generated by O-REPS, the expected losses suffered by the algorithm can be bounded by the average losses. First, we prove that the algorithm changes the policies slowly.

Algorithm 6: Online Relative Entropy Policy Search

Parameters: *finite horizon* T, η

Compute $v_1 = \operatorname{argmin}_v \left(\ln \sum_{x,a} q_t(x,a) e^{\delta(x,a|v,0)} \right)$;

Compute $q_1(x,a) = \frac{e^{\delta(x,a|v_1,0)}}{\sum_{x,a} e^{\delta(x,a|v_1,0)}}$;

Compute π_1 as $\pi_1(a|x) = \frac{q_1(x,a)}{\sum_b q_1(x,b)}$;

Output π_1 as a policy to follow in the episode;

Receive ℓ_1;

for $t = 2, \ldots, T$ **do**

  Compute $v_t = \operatorname{argmin}_v \left(\ln \sum_{x,a} q_{t-1}(x,a) e^{\delta(x,a|v,\ell_t)} \right)$;

  Compute q_t component-wise as $q_t(x,a) = \frac{q_{t-1}(x,a) e^{\delta(x,a|v_t,\ell_t)}}{\sum_{x,a} q_{t-1}(x,a) e^{\delta(x,a|v_t,\ell_t)}}$;

  Compute π_t using $\pi_t(a|x) = \frac{q_t(x,a)}{\sum_b q_t(x,b)}$;

  Output π_t as a policy to follow in the episode;

  Receive ℓ_t;

end

Proposition 2.2. *Let* $\{q_t\}_{t=1}^T$ *be a sequence of state-action distributions generated by O-REPS, then*

$$||q_{t+1} - q_t||_1 \leq \eta.$$

Proof. Our starting point is the Pinsker's inequality (Csiszár (1967); Kemperman (1969); Kullback (1967)):

$$||q_{t+1} - q_t||_1 \leq \sqrt{2 D_R(q_t, q_{t+1})}. \tag{2.15}$$

Where $D_R(\cdot,\cdot)$ is a usual Kullback-Leibler divergence. We denoted it by the same symbol as an un-normalized version since they coincide on the probability simplex. Now we turn our attention to this divergence:

$$\begin{aligned} D_R(q_t, q_{t+1}) &= \sum_{x,a} q_t(x,a) \ln \frac{q_t(x,a)}{q_{t+1}(x,a)} \\ &= \sum_{x,a} q_t(x,a) \ln \frac{\sum_{x',a'} q_t(x',a') e^{\delta(x',a'|v_{t+1},\ell_t)}}{e^{\delta(x,a|v_{t+1},\ell_t)}} \\ &= \ln \sum_{x',a'} q_t(x',a') e^{\delta(x',a'|v_{t+1},\ell_t)} - \sum_{x,a} q_t(x,a) \delta(x,a|v_{t+1},\ell_t). \end{aligned} \tag{2.16}$$

To upper bound the first term in (2.16) we use the fact that when we compute q_{t+1} we minimize the dual function (which is exactly this term). Since it is minimized by v_{t+1}, we can substitute any other function to make an upper bound. Let us substitute $v'(x) = 0$ for all x. Then $\delta(x,a|v',\ell_t) = -\eta \ell_t(x,a)$. And we have

$$\begin{aligned} \ln \sum_{x',a'} q_t(x',a') e^{\delta(x',a'|v_{t+1},\ell_t)} &\leq \ln \sum_{x',a'} q_t(x',a') e^{\delta(x',a'|v',\ell_t)} \\ &= \ln \sum_{x',a'} q_t(x',a') e^{-\eta \ell_t(x',a')} \\ &\leq \ln \left(1 - \eta \sum_{x',a'} q_t(x',a') \ell_t(x',a') + \sum_{x',a'} q_t(x',a') \frac{(-\eta \ell_t(x',a'))^2}{2} \right) \end{aligned}$$

Where in the last step we used the fact that $e^s \leq 1+s+\frac{s^2}{2}$ for $s \leq 0$ and that $\ell_t(x,a) \geq 0$. Now, since $\log(1+s) \leq s$

$$\ln \sum_{x',a'} q_t(x',a') e^{\delta(x',a'|v_{t+1},\ell_t)} \leq -\eta \sum_{x',a'} q_t(x',a')\ell_t(x',a') + \frac{\eta^2}{2} \sum_{x',a'} q_t(x',a')\ell_t^2(x',a').$$

The second term in (2.16) is rewritten as follows

$$\begin{aligned}\sum_{x,a} q_t(x,a)\delta(x,a|v_{t+1},\ell_t) = -\eta \sum_{x,a} q_t(x,a)\ell(x,a) + \sum_{x,a} q_t(x,a)v(x) \\ - \sum_{x,a} q_t(x,a) \sum_{x'} v(x')P(x'|x,a).\end{aligned}$$

The last term can be transformed using (2.1) as

$$\begin{aligned}\sum_{x,a} q_t(x,a) \sum_{x'} v(x')P(x'|x,a) &= \sum_{x'} v(x') \sum_{x,a} q_t(x,a)P(x'|x,a) \\ &= \sum_{x'} v(x') \sum_{a} q_t(x,a) \\ &= \sum_{x,a} v(x) q_t(x,a).\end{aligned}$$

Hence,

$$\sum_{x,a} q_t(x,a)\delta(x,a|v_{t+1},\ell_t) = -\eta \sum_{x,a} q_t(x,a)\ell(x,a).$$

Combining together the obtained expressions for the terms in (2.16), we have

$$D_R(q_t, q_{t+1}) \leq \frac{\eta^2}{2} \sum_{x,a} q_t(x,a)\ell_t^2(x,a).$$

By the fact that $\ell_t(x,a) \leq 1$, the square root is also bounded by 1

$$D_R(q_t, q_{t+1}) \leq \frac{\eta^2}{2}.$$

Therefore, inserting this into Pinsker's inquality, we obtain

$$||q_{t+1} - q_t||_1 \leq \eta.$$

□

The next step is to see how far is the actual distributions over the states from the stationary ones for the policies chosen. Let us first denote the actual distribution over the states at time step t by ϑ_t, i.e.

$$\vartheta_t(x) = \mathbb{P}\left[\mathbf{x}_t = x\right].$$

Note that it can be computed recursively as

$$\vartheta_{t+1} = \vartheta_t P^{\pi_t}.$$

starting from $\vartheta_1 = P_1$. The reason for introducing this distribution is that the expected loss can be written using it:

$$\mathbb{E}\left[\ell_t(\mathbf{x}_t, \mathbf{a}_t)\right] = \sum_{x,a} \ell_t(x,a)\vartheta_t(x)\pi_t(a|x).$$

The following proposition contains the advertised bound.

Proposition 2.3.

$$||\vartheta_t - \mu_{\pi_t}||_1 \leq \eta(1+\tau) + 2e^{-(t-1)/\tau}.$$

Proof. First, we use the triangle inequality, then the definition of stationary state distribution and the inequality (2.14)

$$\begin{aligned}
||\vartheta_t - \mu_{\pi_t}||_1 &\leq ||\vartheta_t - \mu_{\pi_{t-1}}||_1 + ||\mu_{\pi_{t-1}} - \mu_{\pi_t}||_1 \\
&= ||\vartheta_{t-1}P^{\pi_{t-1}} - \mu_{\pi_{t-1}}P^{\pi_{t-1}}||_1 + ||\mu_{\pi_{t-1}} - \mu_{\pi_t}||_1 \\
&\leq e^{-1/\tau}||\vartheta_{t-1} - \mu_{\pi_{t-1}}||_1 + ||\mu_{\pi_{t-1}} - \mu_{\pi_t}||_1.
\end{aligned}$$

The second term can be bounded using Proposition 2.2

$$\begin{aligned}
||\mu_{\pi_{t-1}} - \mu_{\pi_t}||_1 &= \sum_x |\mu_{\pi_{t-1}}(x) - \mu_{\pi_t}(x)| \\
&= \sum_x |\mu_{\pi_{t-1}}(x)\sum_a \pi_{t-1}(a|x) - \mu_{\pi_t}(x)\sum_a \pi_t(a|x)| \\
&\leq \sum_{x,a} |\mu_{\pi_{t-1}}(x)\pi_{t-1}(a|x) - \mu_{\pi_t}(x)\pi_t(a|x)| \\
&= \sum_{x,a} |q_{t-1}(x,a) - q_t(x,a)| \\
&= ||q_{t-1} - q_t||_1 \\
&\leq \eta.
\end{aligned}$$

Hence,

$$||\vartheta_t - \mu_{\pi_t}||_1 \leq e^{-1/\tau}||\vartheta_{t-1} - \mu_{\pi_{t-1}}||_1 + \eta.$$

Iterating the above step, we get

$$\begin{aligned}
||\vartheta_t - \mu_{\pi_t}||_1 &\leq e^{-(t-1)/\tau}||P_1 - \mu_{\pi_1}||_1 + \eta\sum_{t=0}^{t-2} e^{-t/\tau} \\
&\leq 2e^{-(t-1)/\tau} + \eta(1 + \int_0^\infty e^{-t/\tau}dt) \\
&\leq 2e^{-(t-1)/\tau} + \eta(1+\tau).
\end{aligned}$$

□

Finally, we are ready to bound the cumulative expected loss of O-REPS by its cumulative average loss. This is a direct consequence of the previous proposition.

Lemma 2.5. *If $\{\pi_t\}_{t=1}^T$ is a sequence of policies generated by O-REPS, then*

$$\hat{L}_T - \sum_{t=1}^T \rho_t^{\pi_t} \leq \eta T(1+\tau) + 2(1+\tau).$$

Proof.

$$\begin{aligned}
\hat{L}_T - \sum_{t=1}^{T} \rho_t^{\pi_t} &= \sum_{t=1}^{T} (\mathbb{E}\left[\ell_t(\mathbf{x}_t, \mathbf{a}_t)\right] - \rho_t^{\pi_t}) \\
&= \sum_{t=1}^{T} \left(\sum_{x,a} \ell_t(x,a)\vartheta_t(x)\pi_t(a|x) - \sum_{x,a} \ell_t(x,a)\mu_{\pi_t}(x)\pi_t(a|x) \right) \\
&= \sum_{t=1}^{T} \left(\sum_{x,a} \ell_t(x,a)\pi_t(a|x)(\vartheta_t(x) - \mu_{\pi_t}(x)) \right) \\
&\leq \sum_{t=1}^{T} \left(\sum_{x} (\vartheta_t(x) - \mu_{\pi_t}(x)) \sum_{a} \pi_t(a|x) \right) \\
&= \sum_{t=1}^{T} ||\vartheta_t - \mu_{\pi_t}||_1 \\
&\leq \eta T(1+\tau) + 2\sum_{t=1}^{T} e^{-(t-1)/\tau} \\
&\leq \eta T(1+\tau) + 2(1+\tau).
\end{aligned}$$

□

To this point, we can bound the average regret of the algorithm, the difference between the average loss of a fixed policy and its expected loss, the difference between the average loss of the algorithm and its expected loss. All this easily transforms into bound on the expected regret for O-REPS.

Theorem 2.4. *If we run O-REPS with* $\eta = \sqrt{\frac{\ln|X||A|}{T(2+\tau)}}$*, then for any policy* $\pi \in \Gamma$

$$\hat{L}_T - L_T(\pi) \leq \sqrt{T(2+\tau)\ln|X||A|} + 4(1+\tau).$$

Proof. We just decompose the regret into three terms, for which we already know the bounds from Lemma 2.5, Theorem 2.1 and Corollary 2.3.

$$\begin{aligned}
\hat{L}_T - L_T(\pi) &= \left(\hat{L}_T - \sum_{t=1}^{T} \rho_t^{\pi_t} \right) + \left(\sum_{t=1}^{T} \rho_t^{\pi_t} - \sum_{t=1}^{T} \rho_t^{\pi} \right) + \left(\sum_{t=1}^{T} \rho_t^{\pi} - L_T(\pi) \right) \\
&\leq \eta T(1+\tau) + 2(1+\tau) + \eta \sum_{t=1}^{T} ||\ell_t||_\infty^2 + \frac{R(p) - R(d_1)}{\eta} + 2(1+\tau) \\
&\leq \eta T(1+\tau) + 2(1+\tau) + \eta T + \frac{\ln|X||A|}{\eta} + 2(1+\tau).
\end{aligned}$$

Optimizing over η we obtain the claim. □

Conclusion

We studied the theoretical properties of the Relative Entropy Policy Search algorithm. We explored that it is an instance of the Proximal Point Algorithm and, using this fact, developed the applications to different learning problems that can be formulated using Markovian Decision Processes.

First, we surveyed the theory underlying the Proximal Point Algorithm and showed how it is used in the context of online linear optimization.

Second, we applied the algorithm to the full-information and the bandit cases of the online stochastic shortest path problem. We showed that this approach vastly improves the previously known results.

Finally, we introduced O-REPS, a version of REPS applied to the online learning in unichain MDPs in the full-information case. We proved that it enjoys an optimal bound on the regret with smaller additional terms than previously known bounds.

Bibliography

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

H.H. Bauschke and J.M. Borwein. Legendre functions and the method of random bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

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

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967. ISSN 0041-5553.

S. Bubeck. Introduction to online optimization. *Lecture Notes*, 2011.

Y. Censor and S. Zenios. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press, USA, 1998.

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

T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley, 1991.

I. Csiszár. Information-type measures of divergence of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

E. Even-Dar, S.M. Kakade, and Y. Mansour. Experts in a markov decision process. *Advances in Neural Information Processing Systems*, 17:401–408, 2004.

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

A. György, T. Linder, and G. Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:704, 2007.

J. Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

J.H.B. Kemperman. An optimum rate of transmitting information. *The Annals of Mathematical Statistics*, 40:2156–2177, 1969.

S. Kullback. A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127, 1967.

B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Franaise Informat. Recherche Opérationnelle*, (4):154–158, 1970.

G. Neu, A. György, and C. Szepesvári. The online loop-free stochastic shortest-path problem. *COLT-10*, pages 231–243, 2010a.

G. Neu, A. György, C. Szepesvári, and A. Antos. Online markov decision processes under bandit feedback. *Advances in Neural Information Processing Systems*, 23:1804–1812, 2010b.

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

J. Peters, K. Mülling, and Y. Altun. Relative entropy policy search. In *National Conference on Artificial Intelligence (AAAI)*, 2010.

M.L. Puterman. *Markov decision processes: Discrete dynamic stochastic programming*. John Wiley, 1994.

R.T. Rockafellar. *Convex analysis*, volume 28 of *Princeton Mathematics Series*. Princeton university press, 1970.

R.S. Sutton and A.G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.

C. Szepesvári. Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010.

C. Szepesvári, G. Bartók, D. Pál, and I. Szita. Online learning. *Lecture Notes*, 2011.

J.Y. Yu and S. Mannor. Online learning in markov decision processes with arbitrarily changing rewards and transitions. In *Game Theory for Networks, 2009. GameNets' 09. International Conference on*, pages 314–322. IEEE, 2009a.

J.Y. Yu and S. Mannor. Arbitrarily modulated markov decision processes. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 2946–2953. IEEE, 2009b.

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