Online Learning in Markov Decision Processes with Changing Cost Sequences

Travis Dick TDICK@UALBERTA.CA
András György GYORGY@UALBERTA.CA
Csaba Szepesvári SZEPESVA@UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8

Abstract

In this paper we consider online learning in finite Markov decision processes (MDPs) with changing cost sequences under full and bandit-information. We propose to view this problem as an instance of online linear optimization. We propose two methods for this problem: MD^2 (mirror descent with approximate projections) and the continuous exponential weights algorithm with Dikin walks. We provide a rigorous complexity analysis of these techniques, while providing near-optimal regret-bounds (in particular, we take into account the computational costs of performing approximate projections in MD^2). In the case of full-information feedback, our results complement existing ones. In the case of bandit-information feedback we consider the online stochastic shortest path problem, a special case of the above MDP problems, and manage to improve the existing results by removing the previous restrictive assumption that the state-visitation probabilities are uniformly bounded away from zero under all policies.

1. Introduction

We consider the problem of online learning in discrete-time finite Markov decision processes (MDPs) with arbitrarily changing cost functions. It is assumed that a learner moves in a finite state space $\mathcal{X}$. Occupying a state x_t at time instant t, the learner takes an action $a_t \in \mathcal{A}(x_t)$, where $\mathcal{A}(x_t)$ denotes the finite set of actions available at state x_t. Then the agent moves to some new random state x_{t+1}, where the distribution of x_{t+1}, given x_t and a_t is determined by a Markov transition kernel $P(\cdot|x_t, a_t)$. Simultaneously, the agent receives some immediate cost $\ell_t(x_t, a_t)$, where the cost function $\ell_t : \mathcal{U} \to \mathbb{R}$ is assumed to be bounded (in fact, we will assume that the costs lie in $[0, 1]$) and $\mathcal{U} = \{(x, a) : x \in \mathcal{X}, a \in \mathcal{A}(x)\}$. The goal of the learner, who is assumed to know P before the interaction starts but not $\{\ell_t\}_t$, is to minimize its total cost. We assume here that the cost function ℓ_t can change in an arbitrary manner between time instants. The performance of the learner is measured against the best stationary policy in hindsight, giving rise to the expected regret:

$$\mathbf{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(x_t, a_t)\right] - \min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(x_t^\pi, a_t^\pi)\right].$$

Here, for a given stationary policy π (*i.e.*, π is such that $\pi(x, \cdot)$ is a probability distribution over $\mathcal{A}(x)$ for any $x \in X$), (x_t^π, a_t^π) denotes the state-action pair that policy π would visit in time step t if this policy was used from $t = 1$ (we may assume that $x_1^\pi = x_1$). Note that a sublinear regret-growth, $\mathbf{R}_T = o(T)$ $(T \to \infty)$ means that the *average* cost collected by the learning agent approaches that of the best (stationary) policy in hindsight. Naturally, a smaller growth-rate is more desirable.

Motivated by the desire to design robust learning algorithms, this problem has been studied under various conditions by numerous authors (see, e.g., Even-Dar et al., 2009; Yu et al., 2009; Neu et al., 2010; 2011; 2013) and the reader can consult these papers for examples and extra motivation.

We consider two variants of the above model with respect to what observations are available to the learner. In both models the learner can observe its actual state, x_t. In the *full information* feedback model, the learner can observe the full cost function ℓ_t at the end of time instant t, while in the *bandit* feedback model the learner only observes the cost $\ell_t(x_t, a_t)$ it receives.

Treating the online MDP problem as a huge but standard online learning problem, it is not hard to obtain algorithms that enjoy good regret bounds but whose computational complexity is huge. Therefore, earlier work in the literature concentrated on obtaining *computationally efficient* algorithms that also achieve near-optimal regret rates. These results either concern the (stochastic) shortest path problem (SSP, an episodic MDP), or uniformly ergodic MDPs. Several methods achieve near-optimal regret rates by running an independent expert algorithm for each $x \in \mathcal{X}$, see Even-Dar et al. (2005; 2009); Neu et al. (2013) for the full

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

information case and Neu et al. (2010; 2011; 2013) for the bandit case. Yu et al. (2009) gave other low-complexity methods with inferior performance guarantees.

The disadvantage of these methods is that, although they achieve optimal $O(\sqrt{T})$ regret rate in terms of the time horizon T, they often scale suboptimally in other problem parameters, such as the mixing time in the uniformly ergodic case or the length of the paths in the SSP case. Furthermore, the optimal-order bounds in the literature for the bandit setting require all states in $\mathcal{X}$ to be visited with positive probability under *any* deterministic policy, and the inverse of this, potentially very small probability appears in the regret bounds. In this paper we alleviate this problem and obtain optimal-order bounds that do not deteriorate with the minimum visitation probability.

To achieve this, we treat the MDP problem as an online linear optimization problem and show that the resulting methods can be implemented efficiently (we note that the same idea was applied successfully to the deterministic shortest path problem (György et al., 2007), where the minimum visitation probability can also be zero). We rigorously analyze the regret and the computational complexity of our online linear optimization methods, which are approximate versions of the mirror descent and the continuous exponential weights algorithms; we believe that these results are also of independent interest. The mirror descent algorithm (see, e.g., Beck & Teboulle 2003) has a usually computationally expensive projection step which we perform approximately using another mirror descent algorithm, while our results for the continuous exponential weights algorithm are based on the Dikin-walk approximation of Narayanan & Rakhlin (2011).

Recently, Zimin & Neu (2013) have independently obtained similar reductions for the SSP case, using the mirror descent algorithm, achieving essentially the same bounds. However, they do not consider the implementation issues of the projection step of the mirror descent algorithm (i.e., the computational complexity of obtaining sufficiently good approximate projections and the effect of this approximation in their final bound).

The rest of the paper is organized as follows. Section 2 introduces the classes of OMDPs we study and reduces them to online linear optimization problems. Section 3 provides the analysis of the impact of approximation errors on the mirror descent and the continuous exponential weights algorithms. In Section 4 we obtain our algorithms for the OMDP problems by applying the methods of Section 3 to the online linear optimization problems from Section 2.

2. Online Markov Decision Processes as Online Linear Optimization Problems

In this section we give a formal description of online Markov decision processes (OMDPs) and show that two classes of OMDPs can be reduced to online linear optimization. The idea behind the reduction, which goes back to Manne (1960) (for a modern account, see Borkar (2002)), is to represent policies by their stationary (occupation) measures over the set of state-action pairs. Under this representation, the map from policies to their expected cost turns out to be (approximately) linear. Further, when the transition model of the environment is known, it is easy to convert between policies and their stationary occupation measures.

First, let us introduce some notation. Let Δ_S denote the set of probability measures over S.[1] Note that for S finite, we can also view Δ_S as the unit simplex in $\mathbb{R}^{|S|}$: $\Delta_S = \{v \in [0,1]^{|S|} : \sum_{i=1}^{|S|} v_i = 1\}$. The standard inner product of Euclidean spaces will be denoted by $\langle \cdot, \cdot \rangle$. For $p \geq 1$, the p-norm of vector v is denoted by $\|v\|_p$.

The structure of an online MDP is given by a finite state space $\mathcal{X}$, finite action spaces $\mathcal{A}(x), x \in \mathcal{X}$, with $\mathcal{U} = \{(x,a) : x \in \mathcal{X}, a \in \mathcal{A}(x)\}$, and probability transition kernel $P : \mathcal{U} \times \mathcal{X} \to [0,1]$ satisfying $\sum_{x \in \mathcal{X}} P(x|u) = 1$ for all $u \in \mathcal{U}$ where $P(x|u) \overset{\text{def}}{=} P(u,x)$. The learner's starting state, x_1, is distributed according to some distribution μ_0 over $\mathcal{X}$. At each time instant $t = 1, 2, \ldots$, based on its previous observation, state, and action sequences, the learner chooses an action $a_t \in \mathcal{A}(x_t)$, possibly in a random manner. Extending this notion, we can think of a learning agent as if it chose a (randomized) *Markov policy* $\pi_t : \mathcal{U} \to [0,1]$, $\sum_{a \in \mathcal{A}(x)} \pi_t(x,a) = 1$ so that a_t is chosen according to the distribution $\pi_t(x_t, \cdot)$. If $\pi_t = \pi$ independently of t, we say that the agent's strategy is stationary and we identify such a control strategy with π. The set of such stationary Markov policies will be denoted by Π.

2.1. Online Linear Optimization

In the following subsections we reduce two special classes of OMDPs to online linear optimization, which we briefly review now. Let K be a convex and compact subset of a Hilbert space V. In most cases, we take $V = \mathbb{R}^d$ equipped with the standard inner product. In online linear optimization, an adversary selects a sequence of loss vectors $\ell_1, \ldots, \ell_T \in \mathcal{F} \subset V$ and the learner's goal is to choose a sequence of vectors $w_t \in K$ so as to keep the regret

$$\mathbf{R}_T = \sum_{t=1}^{T} \langle \ell_t, w_t \rangle - \min_{w \in K} \sum_{t=1}^{T} \langle \ell_t, w \rangle \qquad (1)$$

small. Naturally, the choice of w_t should only depend on the history of earlier choices and losses. As with OMDPs, we say that there is full-information feedback if the learner observes the entire vector ℓ_t and bandit-information feedback if only the actual loss $\langle \ell_t, w_t \rangle$ is observed. In the

[1] We assume that S is equipped with the necessary σ-algebra, and will not discuss similar trivial measurability issues in the paper.

semi-bandit case, the situation most related to our OMDP problems, $V = \mathbb{R}^d$ and only those components of ℓ_t are observed for which the corresponding components of w_t are non-zero.

2.2. Loop-Free Stochastic Shortest Path (LF-SSP) Problems

Here we assume that $\mathcal{X}$ has a layered structure, that is, $\mathcal{X}$ can be partitioned into disjunct sets $\mathcal{X}_1, \dots, \mathcal{X}_L$ ($L \geq 1$) such that if $P(x'|x,a) > 0$ then $x \in \mathcal{X}_l$ and $x' \in \mathcal{X}_{l+1}$ for some $l = 1, \dots, L-1$, or $x \in \mathcal{X}_L$, $x' \in \mathcal{X}_1$, and $P(x'|x,a) = \mu_0(x')$ for any $a \in \mathcal{A}(x)$. This assumption means that starting in $\mathcal{X}_1$ (μ_0 is concentrated on $\mathcal{X}_1$), the learner moves through $\mathcal{X}_2, \mathcal{X}_3, \dots$ to reach $\mathcal{X}_L$, after which the whole process returns to $\mathcal{X}_1$ and is restarted (we assume without loss of generality that each $x \in \mathcal{X}$ is achievable by following a suitable policy). The sequence of transitions from a state in $\mathcal{X}_1$ back to some other state in $\mathcal{X}_1$ is an episode of the MDP, and in this case t will index the episodes in the process. Since each episode starts from the same distribution, the episodes are memoryless, and any policy π introduces an "occupation measure" μ^π over $\mathcal{U}$, such that for any stage index l, $\sum_{u \in \mathcal{U}_l} \mu^\pi(u) = 1$, where $\mathcal{U}_l = \{(x,a) : x \in \mathcal{X}_l, a \in \mathcal{A}(x)\}$. Furthermore, for any $x \in \mathcal{X}_1$, $\sum_{a \in \mathcal{A}(x)} \mu^\pi(x,a) = \mu_0(x)$. With this we can view $K = \{\mu^\pi : \pi \in \Pi\}$ as a subset of $\times_{l=0}^{L} \Delta_{\mathcal{U}_l} \subset \times_{l=0}^{L} \mathbb{R}^{|\mathcal{U}_l|} = \mathbb{R}^{|\mathcal{U}|}$. Let $d = |\mathcal{U}|$. Note that K is a convex polytope in $\mathbb{R}^d$, since it can be described by a set of linear constraints:

$$K = \Big\{ \mu \in [0,1]^{\mathcal{U}} : \sum_{a' \in \mathcal{A}(x')} \mu(x',a') = \sum_{u \in \mathcal{U}} \mu(u) P(x'|u), x' \in \mathcal{X} \Big\}.$$

These constraints guarantee that the probability "flowing into" each state is equal to the probability "flowing out" of it. It is unnecessary to explicitly require the probability assigned to each layer to sum to one, thanks to the assumption that the transition probability kernel P agrees with μ_0 on the first layer.

Furthermore, with an immediate cost function $\ell : \mathcal{U} \to [0,1]$, the expected total cost of policy π in an episode can be written as $\langle \ell, \mu^\pi \rangle$. Note that with this the problem of finding the stationary policy with the smallest per episode expected cost can be written as the linear optimization problem of $\arg\min_{\mu \in K} \langle \ell, \mu \rangle$: Once the solution of this problem is found, a Markov policy π_μ is extracted from the optimizing measure μ by $\pi_\mu(x,a) = \mu(x,a) / \sum_{a \in \mathcal{A}(x)} \mu(x,a)$. Then, by construction, $\mu^{\pi_\mu} = \mu$.

The above description implies that all paths from the starting layer $\mathcal{X}_1$ back to itself are of the same length. This assumption is not restrictive, though, as any layered MDP can be modified without loss of generality to satisfy this assumption at the price of moderately increasing the state space (see György et al., 2007). For convenience, for online learning with changing costs in LF-SSPs we redefine the regret to be the regret of the first T episodes and use ℓ_t to be the cost function effective in episode t. With this,

$$\mathbf{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell_t, \mu^{\pi_t} \rangle \right] - \min_{\mu \in K} \sum_{t=1}^{T} \langle \ell_t, \mu \rangle, \qquad (2)$$

where $\pi_t \in \Pi$ is the Markov policy used *in the tth episode*. The problem of keeping the regret low is thus viewed as an instance of *online linear optimization* over the convex set K. Note that when π_t is a deterministic function of the past then the expectation can be removed.

2.3. Uniformly Ergodic MDPs

Without loss of generality, we assume that $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}$. Here, following previous works, we assume the so-called uniform mixing condition: There exists a number $\tau \geq 0$ such that under any policy π and any pair of distributions μ and μ' over $\mathcal{X}$, $\|(\mu - \mu')P^\pi\|_1 \leq e^{-1/\tau} \|\mu - \mu'\|_1$, where we use the convention of viewing distributions over $\mathcal{X}$ as *row* vectors of $\mathbb{R}^{|\mathcal{X}|}$ and $P^\pi \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is the transition probability matrix underlying π: $(P^\pi)_{x,x'} = \sum_{a \in \mathcal{A}(x)} \pi(x,a) P(x'|x,a)$. As Even-Dar et al. (2009), we call the smallest τ satisfying this assumption the *mixing time* of the transition probability kernel P, and call a resulting MDP problem uniformly ergodic. This assumption is not unrestrictive, but relaxing it would further complicate the paper and hence we leave this for future work. As for LF-SSPs, for a Markov policy π, let μ^π be its stationary distribution over $\mathcal{U}$. Under the assumption of $\tau < \infty$, μ^π is uniquely determined. Introduce $K = \{\mu^\pi : \pi \in \Pi\} \subset \Delta_{\mathcal{U}}$. Again, K is a convex polytope in $\mathbb{R}^{|\mathcal{U}|}$, since it can be described by a set of linear constraints:

$$K = \Big\{ \mu \in [0,1]^{\mathcal{U}} : \sum_{u \in \mathcal{U}} \mu(u) = 1, \sum_{a' \in \mathcal{A}(x')} \mu(x',a') = \sum_{u \in \mathcal{U}} \mu(u) P(x'|u),\ x' \in \mathcal{X} \Big\}.$$

Again, we will take $d = |\mathcal{U}|$ as the "dimension" of the problem. In this case, we are concerned with finding a sequence of policies whose expected total cost up to time T is not much larger than that of the best policy in hindsight. Similarly to Neu et al. (2011; 2013), we can bound this expected additional cost as shown in the following result:[2]

Lemma 1. *Consider a uniformly ergodic OMDP with mixing time $\tau < \infty$ and losses $\ell_t \in [0,1]^d$. Then the regret of an agent following policies $\pi_1, \dots, \pi_T$ through the trajectory $(x_t, a_t)_t$ relative to a fixed policy π can be bounded*

[2] The proof of this result can be found in the appendix along with several other proofs omitted from the main text.

as

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(x_t,a_t)\right]-\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(x_t^\pi,a_t^\pi)\right]$$
$$\leq\sum_{t=1}^{T}\mathbb{E}\left[\langle\ell_t,\mu^{\pi_t}-\mu^\pi\rangle\right]+T(\tau+1)k+4\tau+4,$$

for any $k\geq\mathbb{E}\left[\|\mu^{\pi_t}-\mu^{\pi_{t-1}}\|\right]$, $t=2,\ldots,T$.

Since we can recover a policy from a stationary distribution (as in the LF-SSP case), it is enough to find a slowly changing sequence $\mu_1,\ldots,\mu_T\in K$ such that the first term of the bound is small. This is again an online linear optimization problem.

We have now mapped online learning in MDPs, under both sets of assumptions, to online linear optimization, which is a well-studied problem in online learning (Cesa-Bianchi & Lugosi, 2006; Shalev-Shwartz, 2012). In the next section, we discuss two general algorithms designed to attack this problem and how they apply to our case.

3. Online Linear Optimization

In this section we consider the challenges of implementing two standard algorithms for online linear optimization: mirror descent (MD) and the continuous exponential weights algorithm (CEWA). When the feasible set K is complicated, some operations from both algorithms have no closed form and need to be approximated iteratively. We analyze the impact of the approximation errors on the regret analysis for both methods. With an understanding of how the regret scales with the approximation errors, we are able to determine the necessary precision, which will lead to bounds on the computational complexity of the approximate versions of these methods.

Recall that the goal of online linear optimization is to choose a sequence of vectors $w_t\in K\subset V$ in order to keep the regret $\mathbf{R}_T=\sum_{t=1}^T\langle\ell_t,w_t\rangle-\min_{w\in K}\sum_{t=1}^T\langle\ell_t,w\rangle$ small, no matter how the sequence of loss vectors $\ell_t\in\mathcal{F}$ is chosen.

3.1. Mirror Descent with Approximate Projections

Mirror descent is a well-known strategy for achieving low regret in online linear optimization problems. It has two parameters: a step size $\eta>0$ and a Legendre function[3] $R:A\to\mathbb{R}$, called a regularizer. We assume that A is a superset of K. Starting from $w_1\in K$, MD makes a sequence of predictions w_t defined by

$$w_{t+1}=\operatorname*{argmin}_{w\in K}\eta\langle\ell_t,w\rangle+D_R(w,w_t),$$

where $D_R(u,v)=R(u)-R(v)-\langle\nabla R(v),u-v\rangle$ denotes the Bregman divergence induced by R. As is well known, w_{t+1} can be obtained in the following two-step process:

$$\tilde{w}_{t+1}=\operatorname*{argmin}_{w\in A}\eta\langle\ell_t,w\rangle+D_R(w,w_t)$$
$$w_{t+1}=\Pi_R(\tilde{w}_{t+1}),$$

where $\Pi_R(\tilde{w})=\operatorname{argmin}_{w\in K}D_R(w,\tilde{w}):A\to K$ denotes the Bregman projection associated with R. Often $\tilde{w}_{t+1}$ can be expressed in closed form and computed in constant time; then the main challenge of applying MD is in computing the Bregman projection.

Unless the set K is very simple, there is no closed form for the Bregman projection and we must use inexact iterative techniques. Hence, the next iterate w_{t+1} will be different from $\Pi_R(\tilde{w}_{t+1})$. The following theorem analyzes the regret of MD with c-approximate projections when $\|w_{t+1}-\Pi_R(\tilde{w}_{t+1})\|\leq c$.[4]

Theorem 2. *Let R be convex and K be a convex set such that ∇R is λ-Lipschitz on K with respect to (wrt) $\|\cdot\|$. Let $D=\sup_{u,v\in K}\|u-v\|_*$ be the diameter of K wrt the dual norm of $\|\cdot\|$. Then the regret of MD, with c-approximate projections, step size η, and regularizer R satisfies*

$$\sum_{t=1}^{T}\langle\ell_t,w_t-w\rangle\leq\sum_{t=1}^{T}\langle\ell_t,w_t-\tilde{w}_{t+1}\rangle+\frac{D_R(w,w_1)}{\eta}+\frac{c\lambda DT}{\eta},$$

for any $w\in K$ and losses $\{\ell_t\}_{t=1}^T$, where ℓ_t can depend on $\{(w_s,\ell_s)\}_{s=1}^{t-1}$. When $c=0$, the result remains true even when λ and/or D are unbounded, in which case we interpret $c\lambda D=0$.

When the regularizer R is σ-strongly convex wrt the norm $\|\cdot\|$, i.e., if $R(w)\geq R(w')+\langle\nabla R(w'),w-w'\rangle+\frac{\sigma}{2}\|w-w'\|^2$ for any $w,w'\in A$, we can use the following lemma to bound the sum $\sum_t\langle\ell_t,w_t-\tilde{w}_{t+1}\rangle$ from Theorem 2.

Lemma 3. *Let $R:A\to\mathbb{R}$ be a σ-strongly convex Legendre function wrt the norm $\|\cdot\|$, $\eta>0$, $w\in\mathcal{A}$, $\ell\in\mathbb{R}^d$, and define $\tilde{w}\in A$ to be the unconstrained MD update: $\tilde{w}=\operatorname{argmin}_{u\in A}\eta\langle\ell,u\rangle+D_R(u,w)$. Then $\langle\ell,w-\tilde{w}\rangle\leq\frac{\eta}{\sigma}\|\ell\|_*^2$, where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.*

3.1.1. MD^2: Efficient Online Linear Optimization for Subsets of $[\beta,1]^d$.

In this section we present a particular implementation of MD with approximate projections that is of interest for the optimization problems presented in Section 2. When the

[3] $R:A\to\mathbb{R}$ is Legendre if $A\neq\emptyset$ is convex, R is strictly convex on its interior A° where ∇R exists, and $\|\nabla R(w)\|\to\infty$ as w approaches the boundary of A from inside A.

[4] The results and their proofs in this and the next section are folklore in the online learning literature. We have learned the proof techniques mostly from Cesa-Bianchi & Lugosi (2006); Rakhlin (2009); Shalev-Shwartz (2012), though some of our proof steps might be different. The reader can consult György et al. (2013) for the proofs of "general" online learning results whose proofs are omitted.

constraint set K is a subset of the unit cube $[0,1]^d \subset \mathbb{R}^d$, to obtain a regret bound that scales with the logarithm of the dimension d, we use MD with the unnormalized negentropy regularizer $R(w) = \sum_u w_u \ln(w_u) - w_u$. Unfortunately, $\nabla R(w) = (\ln(w_1), \dots, \ln(w_d))^\top$ is unbounded when any component of w approaches zero. Thus R violates the condition of Theorem 2 that requires ∇R to be Lipschitz continuous and makes it challenging to design efficient methods for computing c-approximate projections. Therefore, for the rest of this section we assume that the elements of K have components that are uniformly bounded away from zero by $\beta > 0$. In other words, we assume that $K \subset [\beta, 1]^d \subset \mathbb{R}^d$.

In order to apply MD, we need to provide a method for computing c-approximate projections onto the set K. We propose to use a second instance of MD with the squared 2-norm regularizer $R'(w) = \frac{1}{2}\|w\|_2^2$. The motivation for this choice of regularizer is that the induced Bregman divergence equal to the Euclidean distance and the associated Bregman projection is the Euclidean projection. Then the inner instance of MD will also use approximate projections, but they can be calculated as the solutions to quadratic programming problems, which can be solved efficiently by interior point methods. We call this algorithm MD^2, since it has two instances of MD running. The regret of MD^2 can be bounded as follows:

Corollary 4. *Let $\beta > 0$ and $K \subset [\beta, 1]^d$. Let $w_1, \dots, w_T \in K$ be the sequence of predictions made by MD^2 on K with losses $\ell_1, \dots, \ell_T \in [0, \infty)^d$, c-approximate projections where accuracy is measured by $\|\cdot\|_1$, step size $\eta > 0$, and the unnormalized negentropy regularizer R. Then, for any $w \in K$, we have*

$$\sum_{t=1}^T \langle \ell_t, w_t - w\rangle \le \sum_{t=1}^T \langle \ell_t, w_t - \tilde{w}_{t+1}\rangle + \frac{D_R(w, w_1)}{\eta} + \frac{cT}{\beta\eta}$$

where $\tilde{w}_{t+1}$ is defined component-wise by $\tilde{w}_{t+1,u} = w_{t,u} e^{-\eta \ell_{t,u}}$. Let $\epsilon = c/\sqrt{d}$. The per-step complexity is

$$O\left(\frac{H}{\sqrt{\beta}}\left(2\ln\left(\frac{2}{\epsilon}\right) + \ln(W_2 + 2\epsilon)\right) + d\right),$$

where $W_2 = \sup_{w,w' \in K} \|w - w'\|_2$, H is the cost of the projection step used in the inner MD instance when computed with an accuracy of $c' = \frac{1}{4}\sqrt{\beta}\epsilon$ wrt $\|\cdot\|_2$.

When $K \subset [\beta, 1]^d$ is a polytope in the positive quadrant described by $m \le d$ linear equality constraints, interior point methods can be used to achieve $H = O(d^{3.5}\ln(d/c')) = O(d^{3.5}\ln(d/(c\beta)))$ (e.g., Section 2.4.2 of den Hertog 1994, or Section 4.3.2 of Nesterov 2004). Finally, we note in passing that instead of using mirror-descent to implement the approximate projection, one could also use an interior point algorithm for this purpose. Building on the results of Potra & Ye (1993), it appears possible to compute an ϵ-approximate projection to K as above in time $O(d^{3.5}(1 + \ln(1/\beta) + \ln\ln(d/\epsilon)))$, resulting in a modest improvement of the total complexity. Fang et al. (1997) discuss more methods, but with no complexity analysis.

3.2. Continuous Exponential Weights Algorithm with Dikin Walks

Consider an online linear optimization problem over a convex closed set $K \subset \mathbb{R}^d$. Let the sequence of loss vectors be $\ell_1, \dots, \ell_T$ and let p_1 be some positive density over K (*i.e.*, $p_1 \ge 0$ and $\int_K p_1(x)\,dx = 1$). Then, the continuous exponential weights algorithm (CEWA) (see, *e.g.*, Narayanan & Rakhlin, 2011) at time $t+1$ predicts X_{t+1}, where $X_{t+1} \sim p_{t+1}(\cdot)$ and p_{t+1} is a density over K proportional to $p_1(x)\exp(-\eta \sum_{s=1}^t \langle \ell_s, x\rangle)$, where $\langle f, g\rangle = \int_x f(x)g(x)\,dx$. Here $\eta > 0$ is the learning rate of the algorithm.

When $p_1 \in L^2(K)$ (*i.e.*, $\int p_1^2(x)dx < \infty$), the continuous exponential weights algorithm can be interpreted as an instance of mirror descent with the unnormalized negentropy regularizer $R(p) = \int p(x)\ln(p(x)) - p(x)\,dx$. Indeed, it is easy to see that in this case $p_{t+1} = \arg\min_{p \in \mathcal{D}(K)} \{\eta \langle \ell'_t, p\rangle + D_{\mathrm{KL}}(p, p_t)\}$ for $t \ge 1$, where $\mathcal{D}(K)$ is the set of densities over K ($\mathcal{D}(K) = \{p : K \to [0,\infty) | \int_K p(x)dx = 1\}$), $\ell'_t(u) = \langle \ell_t, u\rangle$ for any $u \in K$, the inner product over $\mathcal{D}(K)$ is defined as where $\langle \ell', p\rangle = \int_{u \in K} \ell'(u)p(u)\,du$, and $D_{\mathrm{KL}}(p, p') = \int_K p(x)\ln(p(x)/p'(x))dx$ is the Kullback-Leibler divergence between $p, p' \in \mathcal{D}(K)$, which is also the Bregman divergence induced by R. As such, with a straightforward generalization of Theorem 2 with $c = 0$, we get that the expected regret $\mathbf{R}_T = \mathbb{E}\left[\sum_{t=1}^T \langle \ell'_t, X_t - U\rangle\right]$ against any random variable U with density p_U supported on K is bounded by

$$\mathbf{R}_T \le \sum_{t=1}^T \langle \ell'_t, p_t - q_{t+1}\rangle + \frac{D_{\mathrm{KL}}(p_U, p_1)}{\eta} \qquad (3)$$

where $q_{t+1}(x) = p_t(x)e^{-\eta \ell_t(x)}$.

The advantage of CEWA is that it avoids the usual projection step associated with MD (or similar algorithms, like “Follow the Regularized Leader”). The complexity is pushed to sampling from p_{t+1}, which, as we will see, leads to a different tradeoff. The question of how to efficiently sample from p_{t+1}, or a distribution sufficiently close to p_{t+1}, was addressed by Narayanan & Rakhlin (2011). They proposed a Markov Chain Monte-Carlo (MCMC) method to implement sampling. The stationary distribution of the Markov chain they design at time step $t+1$ is p_{t+1}. However, since the chain is only run for finitely many steps, the distribution that X_{t+1} follows may differ from p_{t+1}. In fact, in their paper they proposed to make only one step with the Markov chain, that is, to use $X_{t+1} = P_{t+1}(\cdot|X_t)$, where $P_{t+1}(\cdot|x)$ is the Markov kernel underlying the chain they designed. They argue that this is sufficient, since p_t is close to p_{t+1}. Indeed, they prove a regret bound that shows the usual $\sqrt{T}$ behavior, but the price of making only one step with the Markov chain is that the regret blows up by a factor that is proportional to d^5. Since we wish to avoid this increase of the regret, we propose to run the chain for more time steps. By following the analysis of (Narayanan & Rakhlin, 2011), we get the following result:

Proposition 5. *Assume that for any $t \geq 1$, $x \in K$, the losses satisfy $\langle \ell_t, x \rangle \in [0, B]$. Let $P_{t+1}(\cdot|x)$ be the Markov kernel underlying the "Dikin walk" of Narayanan & Rakhlin (2011) at time step $t+1$. Assume that $\eta B \leq 1$ and fix an arbitrary parameter $0 < r \leq 1$. If $X_1 \sim p_1$[5] and for $t = 1, 2, \ldots$ in time step $t+1$, $X_{t+1} = Z^{(k)}_{t+1}$, where $Z^{(i+1)}_{t+1} \sim P_{t+1}(\cdot|Z^{(i)}_{t+1})$, $i = 1, 2, \ldots$ with $Z^{(1)}_{t+1} = X_t$ then if $k \geq C\nu^2 d^3 \ln((1 + \eta(e-1)B)^2 + 2\eta(e-1)B/r)$ then $\left\|p'_{t+1} - p_{t+1}\right\|_1 = \int |p'_{t+1}(x) - p_{t+1}(x)|dx \leq r$, where p'_{t+1} is the distribution of X_{t+1}, $C > 0$ is a universal constant and $\nu = O(d)$ is a parameter that depends on the shape of K.*

The main work in making a move with the Markov chain is to sample from a Gaussian random variable. Note that the covariance of this distribution depends on the current state. Thus, the cost of one step is dominated by the $O(d^3)$ cost of computing the Cholesky factorization of this covariance matrix once the matrix is computed. Hence, the total cost of sampling at one time step is $O(d^8 \ln(1 + \eta B/r)$, where we assumed that computing the covariance matrix can also be done in $O(d^3)$ step.

Finally note that if $\left\|p'_t - p_t\right\|_1 \leq r$ in each time step t then the additional regret due to the "imprecise" implementation up to time T is bounded by rTB. Consider now the case when B is constant (*i.e.*, independent of T). Then, setting $r = 1/\sqrt{T}$ we see that the increase of the regret is bounded by $B\sqrt{T}$. Now, remember that to get a $\sqrt{T}$ regret one should use $\eta = O(1/\sqrt{T})$ (e.g., see the bound of Theorem 2). Hence, in this case the cost of sampling per time step can be kept constant independently of the time horizon with essentially no increase of the regret.

3.3. Bandit Information

The purpose of this section is to briefly consider bandit online linear optimization. The difference between bandit online linear optimization and the setting considered above is that at the end of time step t the only information received is the scalar loss of the vector chosen in that time step, that is $\langle \ell_t, \hat{w}_t \rangle$, while ℓ_t is not revealed to the learner. For a recent survey of the literature see the review paper by Bubeck & Cesa-Bianchi (2012). The approach followed by existing algorithms is to construct an estimate $\widetilde{\ell}_t$ (usually unbiased) of ℓ_t and use this in place of ℓ_t in a "full-information algorithm", like MD of the previous section. The question then is how to construct $\widetilde{\ell}_t$ and how to control the regret. Our main tool in this latter respect is going to be Theorem 2. Indeed, if MD is run with $\widetilde{\ell}_{t-1}$ in place of ℓ_{t-1}, then from Theorem 2 we see that, as far as the expected regret is concerned, it suffices to control $\mathbb{E}\left[\left\langle \widetilde{\ell}_{t-1}, \hat{w}_t - \tilde{w}_{t-1} \right\rangle\right]$. For this, we have the following result extracted from Abernethy et al. (2008):[6]

Lemma 6. *Let $w, \ell \in V = \mathbb{R}^d$, and define $\tilde{w}_u = w_u e^{-\eta \ell_u}$ for all $u = 1, \ldots, d$. Then $\langle \ell, w - \tilde{w} \rangle \leq \eta \sum_u w_u \ell_u^2$.*

Note that the lemma continues to be true even if $V = L^2(K)$, is the space of square-integrable functions over K, in which case the sum should be replaced by an integral over K wrt the Lebesgue measure.

4. Learning in Online Markov Decision Processes

In this section, we consider online learning in MDPs in the so-called full-information setting. In the case of LF-SSPs this means that ℓ_t is observed at the end of episode t, while in the case of ergodic MDPs ℓ_t is observed at the end of each time step. We only consider full-information algorithms based on MD^2; solutions using CEWA are only provided for the bandit case (to save space).

Consider first LF-SSP problems. In order to apply MD^2, we need the components of the (occupation) measures to be bounded away from zero. This will not be the case generally, since policies may choose actions with arbitrarily low probabilities. Without loss of generality we can assume that there exists a $\beta > 0$ and a policy π_{exp} such that the corresponding (occupation) measure $\mu_{\text{exp}} = \mu^{\pi_{exp}}$ satisfies $\mu_{\text{exp}}(x, a) \geq \beta$ for all $(x, a) \in \mathcal{U}$. By the convexity of K, $\mu_\delta \doteq (1-\delta)\mu + \delta\mu_{\text{exp}} \in K$ for any $0 < \delta < 1$ and $\mu \in K$ (*i.e.*, there exists a policy inducing μ_δ), and for any loss function ℓ we have

$$|\langle \ell, \mu_\delta \rangle - \langle \ell, \mu \rangle| = \delta |\langle \ell, \mu_{exp} - \mu \rangle|. \tag{4}$$

Therefore, we do not loose much if we use MD^2 with

$$K_{\delta\beta} = \{\mu \in K : \mu(x, a) \geq \delta\beta \text{ for all } x, a\}$$

instead of K, since $\mu_\delta \in K_{\delta\beta}$.

First we consider the simple case of the LF-SSP problem. By (4) and since $\ell_t(u) \in [0, 1]$ for all $u \in \mathcal{U}$,

$$\left| \sum_{t=1}^{T} \langle \ell_t, \mu \rangle - \sum_{t=1}^{T} \langle \ell_t, \mu_\delta \rangle \right| \leq \delta L T. \tag{5}$$

(Recall that L is the number of layers in the state space.) Now let us run MD^2 on $K_{\delta\beta}$ with the unnormalized negative entropy regularizer $R(\mu) = \sum_{l=0}^{L-1} R_l(\mu_i)$, where

[5] This assumption is not necessary, just simplifies the analysis.

[6] The proof can be found in (György et al., 2013).

$\mu = (\mu_0, \dots, \mu_{L-1})$, $\mu_l \in [0,\infty)^{\mathcal{U}_l}$ and R_l is the unnormalized negative entropy regularizer over $[0,\infty)^{\mathcal{U}_l}$. Since it follows from Pinsker's inequality that R is $1/L$ strongly convex wrt the $\|\cdot\|_1$-norm (see also Example 2.5 of Shalev-Shwartz 2012), combining Corollary 4, with Lemma 3 and (5), we obtain the following result:

Theorem 7 (MD^2 on LF-SSP, full information)**.** *Let π be any policy, $\mu_1 \in K$, $\delta \in (0,1]$ and $D_{\max} \geq \sup_{\mu\in K_{\delta\beta}} D_R(\mu,\mu_1)$. Run MD^2 with parameters $c = \frac{\beta\delta\eta}{\sqrt{T}}$ and $\eta = \sqrt{\frac{D_{\max}}{LT}}$ on $K_{\delta\beta}$ with the sequence of loss functions $\ell_1,\dots,\ell_T$. Let μ_t be the output of MD^2 on round t and define $\pi_t = \pi_{\mu_t}$ (i.e., the state-conditional action probabilities). Then the regret of the agent that follows policy π_t at time t relative to policy π can be bounded as*

$$\mathbf{R}_T \leq 2\sqrt{LTD_{\max}} + \sqrt{T} + L\delta T,$$

and the per-time-step computational cost is bounded by

$$O\left(\frac{d^{3.5}\mathcal{L}}{\sqrt{\beta\delta}}\left(\mathcal{L} + \ln(L + D_{\max})\right)\right), \text{ where } \mathcal{L} = \ln(\tfrac{dTL}{\beta\delta}).$$

The proof follows from the arguments preceding the theorem combined with Corollary 4 and the remark after it. Also, the next theorem has an almost identical proof, hence we decided to omit this proof.

Note that $D_{\max} = \Theta\left(L \ln \frac{1}{\pi_0}\right)$, where $\pi_0 = \min_{(x,a)\in\mathcal{U}} \pi_{exp}(x,a)$ (notice that $\pi_{exp}(x,a) \geq \beta$ since $\mu_{\text{exp}}(x,a) \geq \beta$). If, for example, $\pi_{exp}(x,\cdot)$ is selected to be the uniform distribution over $\mathcal{A}(x)$, then $\beta > 0$ and $\pi_0 = 1/\max_x |\mathcal{A}(x)|$, making the regret scale with $O(L\sqrt{T\ln(\max_x|\mathcal{A}(x)|)})$ when $\delta = 1/\sqrt{T}$. Also, this makes the computational cost $\tilde{O}(d^{3.5}T^{1/4}/\sqrt{\beta})$, where $\tilde{O}(\cdot)$ hides log-factors. Neu et al. (2010) gave an algorithm that achieves $O(L^2\sqrt{T\ln(\max_x|\mathcal{A}(x)|)})$ regret with $O(d)$ computational complexity per time-step. Thus, our regret bound scales better in the problem parameters than that of Neu et al. (2010), at the price of increasing the computational complexity. It is an interesting (and probably challenging) problem to achieve the best of the two results.

Consider now the case of uniformly ergodic MDPs. In order to apply MD^2, we need to obtain a regret bound for online linear optimization on the corresponding set K and show that the sequence of policies does not change too quickly. By (4) and because $\ell_t \in [0,1]^d$, we have

$$\left|\sum_{t=1}^{T}\langle \ell_t,\mu\rangle - \sum_{t=1}^{T}\langle \ell_t,\mu_\delta\rangle\right| \leq \delta\sum_{t=1}^{T}|\langle \ell_t,\mu_{exp}-\mu\rangle| \leq \delta T. \tag{6}$$

Therefore, running MD^2 on $K_{\delta\beta}$ with the negentropy regularizer $R(\mu) = R_d(\mu)$ gives the following result:

Theorem 8 (MD^2 on Ergodic MDPs, full information)**.** *Let π be any policy, $\mu_1 \in K$, $\delta \in (0,1]$ and $D_{\max} \geq \sup_{\mu\in K_{\delta\beta}} D_R(\mu,\mu_1)$. Run MD^2 with parameters $c = \frac{\beta\delta\eta}{\sqrt{T}}$ and $\eta = \sqrt{\frac{D_{\max}}{T(2\tau+3)}}$ on $K_{\delta\beta}$ with the sequence of loss functions $\ell_1,\dots,\ell_T$. Let μ_t be the output of MD^2 on round t, and define $\pi_t = \pi_{\mu_t}$. Then the regret of the agent that follows policy π_t at time t relative to policy π can be bounded as*

$$\mathbf{R}_T \leq 2\sqrt{(2\tau+3)TD_{\max}} + \sqrt{T} + \delta T + 4\tau + 4,$$

and the per-time-step computational cost is bounded by

$$O\left(\frac{d^{3.5}\mathcal{L}}{\sqrt{\beta\delta}}\left(\mathcal{L} + \ln(D_{\max})\right)\right), \ \mathcal{L} = \max(1, \ln(\tfrac{dT\tau}{D_{\max}\beta\delta})).$$

As far as the dependence on τ is concerned, by choosing $\delta = 1/\sqrt{T}$, we can thus improve the previous state-of-the-art bound (Neu et al., 2013) that scales as $O(\tau^{3/2}\sqrt{T\ln|\mathcal{A}|})$ to $O(\sqrt{\tau T\ln|\mathcal{A}|})$. The update cost of the algorithm of Neu et al. (2013) is $O(|\mathcal{X}|^3 + |\mathcal{X}|^2|\mathcal{A}|)$, while here the cost of the MD^2 is $\tilde{O}(T^{1/4}d^{3.5}/\sqrt{\beta})$.

5. Learning under Bandit Information in LF-SSP

The purpose of this section is to consider online learning in the LF-SSP problem under bandit feedback, that is, when at time t, the only information received is $\ell_t(x_t,a_t)$, the cost of the current transition. Based on the previous sections, we see that to control the regret, an MDP learning algorithm has to control the regret in an online linear bandit problem with decision set K.

According to Bubeck et al. (2012), for online bandit linear optimization over a compact action set $K \subset \mathbb{R}^d$, it is possible to obtain a regret of order $O(d\sqrt{T\log T})$ regardless the shape of the decision set K, which, in our case would translate into a regret bound of order $O(|\mathcal{U}|\sqrt{T\log T})$. Whether the algorithm proposed in this paper can be implemented efficiently depends, however, on the particular properties of K: Designing the exploration distribution needed by this algorithm requires the computation of the minimum volume ellipsoid containing K and this problem is in general NP-hard even when considering a constant factor approximation (Nemirovski, 2007).

In this section, focussing on LF-SSPs, we design computationally efficient bandit algorithms based on MD and the continuous exponential weights algorithm. In both cases, the immediate costs will be estimated in the same manner:

$$\widetilde{\ell}_t(x,a) = \frac{\mathbb{I}\{x_t^{(l)} = x, a_t^{(l)} = a\}}{\mu^{\pi_t}(x,a)}\ell_t(x,a). \tag{7}$$

Note that in each stage l, $\widetilde{\ell}_t(x,a)$ is nonzero only for the state-action pair visited in $\mathcal{U}_l$; hence, $\widetilde{\ell}_t$ is available to the learner. It is easy to see that as long as *(B) $\mu^{\pi_t}(x,a)$ is bounded away from zero for each state-action pair (x,a),* the above estimate is unbiased. In particular, denoting by $\mathcal{F}_t$ the σ-algebra generated by the history up to the beginning of episode t, $\mathbb{E}\left[\widetilde{\ell}_t(x,a)|\mathcal{F}_t\right] = \ell_t(x,a)$ holds for all $(x,a) \in \mathcal{U}$.

First, let us consider the application of MD^2 with the unnormalized negentropy regularizer on $K_{\delta\beta}$ to this problem. Note that the restriction to $K_{\delta\beta}$ is now used to ensure both that the projection step can be implemented efficiently and that estimates in (7) are well-defined. In particular, this implies that (B) will be satisfied. Using Lemma 6 then gives the following result:

Theorem 9 (MD^2 on Bandit LF-SSP)**.** *Let π be any policy, $\mu \in K$, $\delta \in (0,1]$ and $D_{\max} \geq \sum_{\mu \in K_{\delta\beta}} D_R(\mu,\mu_1)$. Run MD^2 with parameters $c = \frac{\beta\delta\eta}{\sqrt{T}}$ and $\eta = \sqrt{\frac{D_{\max}}{dT}}$ on $K_{\delta\beta}$ with the sequence of estimated loss functions $\widetilde{\ell}_1,\ldots,\widetilde{\ell}_T$, defined in (7). Let μ_t be the output of MD^2 on round t, and define $\pi_t = \pi_{\mu_t}$. Then the regret of the agent that follows policy π_t at time t relative to policy π can be bounded as*

$$\mathbf{R}_T \leq 2\sqrt{dTD_{\max}} + \sqrt{T} + L\delta T,$$

and the computational cost is bounded as in Theorem 7.

Selecting $\pi_{exp}(x,\cdot)$ to be the uniform distribution, $\beta > 0$ and $D_{\max} \leq L\ln(\max_x|\mathcal{A}(x)|)$, results in a $O(\sqrt{dLT\ln(\max_x|\mathcal{A}(x)|)})$ bound on the regret for $\delta = 1/\sqrt{T}$, while the time-complexity of the algorithm is still $\tilde{O}(d^{3.5}T^{1/4}/\sqrt{\beta})$ as in the full-information case. Neu et al. (2010) considered the same problem under the assumption that any policy π visits any state with probability at least α for some $\alpha > 0$, that is, $\inf_\pi \sum_{a\in\mathcal{A}(x)} \mu^\pi(x,a) \geq \alpha > 0$. They provided an algorithm with $O(d)$ per round complexity whose regret is $O(L^2\sqrt{T\max_x(\mathcal{A}(x)\ln(\mathcal{A}(x)))/\alpha})$. Compared to their result, we managed to lift the assumption $\alpha > 0$, and also improved the dependence on the size of the MDP, while paying a price in terms of increased computational complexity.

Let us now consider applying CEWA with the Dikin-walk to the same problem. As in the MD^2 case, we run the algorithm on $K_{\delta\beta}$ with $\delta > 0$. As in the full information case, we let $\ell'_t(\mu) = \langle \ell_t, \mu \rangle$, $\mu \in \mathbb{R}^{\mathcal{U}}$, $\widetilde{\ell}'_t(\mu) = \left\langle \widetilde{\ell}'_t, \mu \right\rangle$, where $\widetilde{\ell}_t$ is obtained using (7). Let $\hat{p}_t(\mu) = p_1(\mu)\exp(-\eta\sum_{s=1}^{t-1}\widetilde{\ell}'_s(\mu))$, while $p_t = \hat{p}_t/Z_t$, $Z_t = \int_{K_{\delta\beta}}\hat{p}_t(\mu)d\mu$. Let $\mu_t(= X_t) \in K_{\delta\beta}$ be the output of the Dikin-walk at time step t when the walk is run for k steps and let $\pi_t = \pi_{\mu_t}$ be the corresponding policy. Since each coordinate of $\mu_t \in K_{\delta\beta}$ is bounded away from zero, $\widetilde{\ell}_t$ is well-defined and $\mathbb{E}\left[\widetilde{\ell}_t|\mathcal{F}_t\right] = \ell_t$ and $\mathbb{E}\left[\widetilde{\ell}'_t|\mathcal{F}_t\right] = \ell'_t$. Combining (3) with Proposition 5 and Lemma 6, we get the following result:

Theorem 10 (CEWA on Bandit LF-SSP)**.** *Let $\delta = \sqrt{\frac{dL\ln(\beta T/d)}{2\beta(2L+1)T}}$, and assume that π_t is obtained by running the algorithm of Section 3.2 on $K_{\delta\beta}$ with the estimated losses $\{\widetilde{\ell}_t\}$, started from the uniform distribution, and with parameters $r = 2\delta$, $\eta = \delta\beta/L$ and $k \geq Cd^5\ln(\beta T/d)$ for some universal constant $C > 0$. Then, for any $T > 2d/\beta$, the regret against any $\mu \in K$ is bounded by*

$$\mathbf{R}_T \leq 3\sqrt{\frac{dL(L+1/2)T\ln(\beta T/d)}{\beta}}$$

while the per-step computational complexity is bounded by $O(d^3k)$.

Notice that the regret bound is $O(L\sqrt{dT\ln(T)/\beta})$, while the per-step computational complexity (choosing the smallest k) is $O(d^8\ln T)$. Thus, this algorithm does not achieve the performance of MD^2; the scaling of the regret bound with $1/\sqrt{\beta}$ is especially not nice. On the other hand, the computational cost of the algorithm is better in T than that of MD^2. The regret bound of the algorithm could be improved if in (7) we divided by $\int \mu(x,a)p_t(\mu)d\mu$ (instead of μ_t), the probability of visiting the state-action pair (x,a). However, we cannot compute this probability in a closed form, and its estimation would require additional sampling, further increasing the computational cost (see Neu & Bartók 2013 for a similar approach). We finally note that based on the proof it is obvious that for the full-information setting, CEWA would achieve a regret competitive with MD^2, and again, its cost would be lower in T, but higher in d.

6. Conclusions

In this paper, viewing online learning in MDPs as online linear optimization, we have proposed novel algorithms based on variants of mirror-descent. We proposed efficient solutions, based on approximate projections and MCMC sampling, to overcome the computational difficulty of the projection step in MD arising from the OMDP structure. We rigorously analyzed the complexity of these algorithms. Our results improve upon the state-of-the-art by improving the regret bounds and lifting some restrictive assumptions that were used earlier in the literature. The price we pay is a somewhat increased computational complexity, though our algorithms still enjoy polynomial computational complexity in the problem parameters. It is an interesting (and probably challenging) problem to find out whether the tradeoff exposed is "real". Extending our results to the bandit-information feedback case for uniformly ergodic OMDPs is also an important problem. One promising approach in this direction is to combine our methods with the reward-estimation technique of Neu et al. (2013).

Acknowledgements

This work was supported by the Alberta Innovates Technology Futures and NSERC.

References

Abernethy, J., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. In Servedio, Rocco A. and Zhang, Tong (eds.), *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pp. 263–274. Omnipress, 2008.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Borkar, V. S. Convex analytic methods in Markov Decision Processes. In Feinberg, E.A. and Shwartz, A. (eds.), *Handbook of Markov Decision Processes*, chapter 11. Kluwer Academic Publishers, 2002.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. *Journal of Machine Learning Research - Proceedings Track*, 23:41.1–41.14, 2012.

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

den Hertog, D. *Interior Point Approach to Linear, Quadratic and Convex Programming: Algorithms and Complexity*. Springer, 1994.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Experts in a Markov decision process. In Saul, Lawrence K., Weiss, Yair, and Bottou, Léon (eds.), *Advances in Neural Information Processing Systems 17*, pp. 401–408, Cambridge, MA, USA, 2005. MIT Press.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online Markov decision processes. *Mathematics of Operations Research*, 34 (3):726–736, 2009. ISSN 0364-765X. doi: http://dx.doi.org/10.1287/moor.1090.0396.

Fang, S.-C., Rajasekera, J. R., and Tsao, H. *Entropy Optimization and Mathematical Programming*. Springer, 1997.

György, A., Linder, T., Lugosi, G., and Ottucsák, Gy. The online shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:2369–2403, 2007. ISSN 1532-4435.

György, A., Pál, D., and Szepesvári, Cs. Online learning: Algorithms for big data. Lecture Notes, 2013. URL https://www.dropbox.com/s/bd38n4cuyxslh1e/online-learning-book.pdf.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning Journal*, 69(2-3):169–192, 2007.

Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., and Culotta, A. (eds.). *Advances in Neural Information Processing Systems 23*, 2011.

Manne, A.S. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

Narayanan, H. and Rakhlin, A. Random walk approach to regret minimization. In Lafferty et al. (2011), pp. 1777–1785.

Nemirovski, A. Advances in convex optimization: Conic programming. In *Proceedings of International Congress of Mathematicians*, volume 1, pp. 413–444, 2007.

Nesterov, Y. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

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

Neu, G., György, A., Szepesvári, Cs., and Antos, A. Online Markov decision processes under bandit feedback. In Lafferty et al. (2011), pp. 1804–1812.

Neu, G., György, A., Szepesvári, Cs., and Antos, A. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 2013. URL http://www.szit.bme.hu/~gya/publications/NeGySzAn13.pdf. (accepted for publication).

Neu, Gergely and Bartók, Gábor. An efficient algorithm for learning with semi-bandit feedback. In Jain, Sanjay, Munos, Rémi, Stephan, Frank, and Zeugmann, Thomas (eds.), *Algorithmic Learning Theory - 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings*, volume 8139 of *Lecture Notes in Computer Science*, pp. 234–248, 2013.

Potra, F. and Ye, Y. A quadratically convergent polynomial algorithm for solving entropy optimization problems. *SIAM J. Control and Optimization*, 3(4):843–860, 1993.

Rakhlin, A. Lecture notes on online learning. Lecture Notes, 2009.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. ISSN 1935-8237. doi: 10.1561/2200000018. URL http://dx.doi.org/10.1561/2200000018.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009. ISSN 0364-765X. doi: http://dx.doi.org/10.1287/moor.1090.0397.

Zimin, A. and Neu, G. Online learning in episodic Markovian decision processes by relative entropy policy search. In Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 1583–1591. 2013.