Provably Sample-Efficient Model-Free Algorithm for MDPs with Peak Constraints

Qinbo Bai   BAI113@PURDUE.EDU
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907, USA

Vaneet Aggarwal   VANEET@PURDUE.EDU
School of IE and ECE
Purdue University
West Lafayette, IN 47907, USA

Ather Gattami   ATHER.GATTAMI@AI.SE
AI Sweden
Stockholm, Sweden

Editor: Laurent Orseau

Abstract

In the optimization of dynamic systems, the variables typically have constraints. Such problems can be modeled as a Constrained Markov Decision Process (CMDP). This paper considers the peak Constrained Markov Decision Process (PCMDP), where the agent chooses the policy to maximize total reward in the finite horizon as well as satisfy constraints at each epoch with probability 1. We propose a model-free algorithm that converts PCMDP problem to an unconstrained problem and a Q-learning based approach is applied. We define the concept of probably approximately correct (PAC) to the proposed PCMDP problem. The proposed algorithm is proved to achieve an (ϵ, p)-PAC policy when the episode $K \geq \Omega(\frac{I^2 H^6 S A \ell}{\epsilon^2})$, where S and A are the number of states and actions, respectively. H is the number of epochs per episode. I is the number of constraint functions, and $\ell = \log(\frac{SAT}{p})$. We note that this is the first result on PAC kind of analysis for PCMDP with peak constraints, where the transition dynamics are not known apriori. We demonstrate the proposed algorithm on an energy harvesting problem and a single machine scheduling problem, where it performs close to the theoretical upper bound of the studied optimization problem.

Keywords: Markov Decision Process, Model-free Algorithm, Peak Constraints, Reinforcement Learning

1. Introduction

Optimization of dynamic systems typically has constraints, e.g., battery capacity for robots. As an example, if a robot is powered by a battery, which is also being charged with an external power supply, the amount of energy used at each time is limited by the battery capacity. The dynamical systems are typically modeled as a Markov Decision Process (MDP), while the transition probabilities may not be known apriori (or maybe dynamic). In the absence of knowledge of transition probabilities, the MDP is modeled as a Reinforcement

Learning (RL) problem which aims to maximize the total reward in the finite horizon by making actions given the state of the process to be controlled. RL algorithms can be divided into model-based and model-free, where the model-based approaches estimate the transition probabilities, while model-free approaches do not. In this paper, we consider a model-free approach to RL in the presence of peak constraints, which is an important constraint in many dynamical systems. For instance, algorithms with peak constraints have been studied for communications Shamai and Bar-David (1995), flow-shop scheduling Fang et al. (2013), thermostatically-controlled systems Karmakar et al. (2013), economics Bailey (1972), robotics Li et al. (1997), etc. The constrained optimization problems have been considered for Markov Decision Processes Altman (1999). However, these require complete knowledge of the transition probabilities. Without such knowledge, algorithms have been proposed Geibel and Wysotzki (2005); Geibel (2006). However, to the best of our knowledge, none of the algorithms so far has considered MDPs with peak constraints and provably given PAC analysis for objective and constraint violations.

Contributions: In this work, we assume that the transition probability is unknown, the reward and constraint functions can be observed but are not known in closed form. We extend the concept of probably approximately correct (PAC) for the proposed PCMDP problem. We introduce an **Approximated PCMDP** to propose a novel model-free algorithm with stochastic policy. The proposed algorithm is shown to achieve an (ϵ, p)-PAC policy when the number of episodes are $K \geq \Omega(\frac{I^2 H^6 SA\ell}{\epsilon^2})$. Finally, the proposed algorithm is evaluated on an energy-harvesting transmitter studied in Wang et al. (2014) and a single machine scheduling problem with deadlines studied in Koulamas and Kyparisis (2001). It is found that the proposed algorithm performs close to the genie-aided upper bound for the problem.

2. Related Work

Online Convex Optimization (OCO): OCO problem is an extension of the constrained convex optimization. In this problem, we wish to optimize $\sum_{t=1}^{T} f_t(\boldsymbol{x})$ for given functions f_t, $t \in \{1, \cdots, T\}$, such that $\boldsymbol{x} \in \mathcal{K}$. In online convex optimization, we select $\boldsymbol{x}_t$ at time t, such that the regret in objective is minimized, which is defined as

$$\text{Regret}(T) = \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \min_{\boldsymbol{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\boldsymbol{x}). \tag{1}$$

Further, $\boldsymbol{x}_t$ may not satisfy constraints, and thus there will be a constraint violation. By changing the problem into an online convex-concave optimization problem, The authors of Mahdavi et al. (2011) proposed an algorithm which achieves the $O(\sqrt{T})$ bound for the regret and $O(T^{3/4})$ bound on the violation of constraints. Further, they proposed another algorithm based on the mirror-prox method (Nemirovski, 2004) that achieves $O(T^{2/3})$ bound on both regret and constraints when the domain can be described by a finite number of linear constraints. The authors of Jenatton et al. (2016) proposed an algorithm which achieves $O(T^{\max(\beta, 1-\beta)})$ objective regret and $O(T^{1-\beta/2})$ constraint violations for $\beta \in (0, 1)$. Further, the authors of (Yu and Neely, 2016) proposed an algorithm with $O(\sqrt{T})$ regret bound for objective with finite constraint violations. However, CMDP is different from OCO

because the reward function depends both on the state and action and the previous action can influence current state and thus change the reward function. Further, the functions and constraints are not known explicitly in reinforcement learning (RL). Thus, the analysis of CMDP doesn't directly follow from that of OCO.

Constrained Markov Decision Process (CMDP): When the system model (the transition probability distribution, the reward function, and constraint functions) is known, the problem is generally considered as CMDP. CMDP in the form of discounted and average reward has been studied in Altman (1999). It is well known that CMDP problem is convex and can be converted into an equivalent unconstrained MDP problem by using the method of Lagrange multipliers. Thus, when the model is known, CMDP can be solved using linear programming (LP). In addition to the LP method, Three different algorithms, WeiMDP, AugMDP, and RecMDP, are proposed to solve CMDP in different settings Geibel (2006). Constrained Upper Confidence Reinforcement Learning (C-UCRL) Zheng and Ratliff (2020) algorithm achieves $O(T^{3/4}\sqrt{\log(T/\delta)})$ regret on the reward and satisfies the constrains even during learning process with probability at least $1 - \delta$.

The key difference between these works and reinforcement learning (RL) approach is that the transition probabilities for the next state given the previous state and action are assumed to be known in CMDP approaches, while are not known apriori in RL approaches. They may be learnt in model-based RL, while not learnt at all in model-free RL approaches. Recently, the authors of Efroni et al. (2020) proposed OptCMDP, OptCMDP-bonus, OptDual-CMDP, and OptPrimalDual-CMDP algorithms to achieve both $O(\sqrt{T})$ bound on the reward and constraint violations. However, all of these four algorithms are model-based. A model-free algorithm is provided in Ding et al. (2020) and also achieve $\mathcal{O}(\sqrt{T})$ bound on both the objective and constraint violation. In this paper, we consider model-free reinforcement learning based approaches for the peak constraint problem which is different from the above paper in the average expected setting.

Regret Bounds for Reinforcement Learning: Regret Analysis for the Reinforcement Learning has been considered for both the model-based approaches Jaksch et al. (2010); Agrawal and Jia (2017); Azar et al. (2017); Kakade et al. (2018) and the model-free approaches Kearns and Singh (2002); Strehl et al. (2006); Jin et al. (2018). Our paper extends the epsiodic reinforcement learning setup with the addition of peak constraints.

Model-free Reinforcement Learning Algorithm for CMDP with Peak Constraints: Q-learning based methods with peak constraints has been studied Bouton et al. (2019); Wang et al. (2014), where the Q function in each epoch is projected to the constraint set. These algorithms involve knowledge of constraint functions explicitly (since projection to the constraint set is needed) to make decisions at each time. In contrast, we do not require knowledge of constraint function. Recently, based on the primal-dual method, The authors of Paternain et al. (2019) proposed an algorithm with policy descent and showed that the algorithm is $1 - \delta$ safe ($P(\cap_{t \geq 0}\{s_t \in \mathcal{S}_0\}|\pi_\theta) \geq 1 - \delta$, where $\mathcal{S}_0$ is the safe region). Besides, Gattami (2019) related PCMDP to unconstrained zero-sum game where the objective is the Lagrangian of the optimization problem, and applied max-min Q-learning to PCMDP to prove convergence. However, none of the works in this direction have given a PAC kind of analysis for objectives and constraints, which is the focus of our paper. To the best of our knowledge, this paper provides the first PAC analysis for model-free reinforcement learning with peak constraints.

3. Problem Formulation and Assumption

We consider an episodic setting of the PCMDP with finite state and action space, defined by PCMDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, f_i, s_1)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the set of actions with $|\mathcal{A}| = A > 1$, H is the number of epochs in each episode, and $\mathbb{P}$ is the transition matrix such that $\mathbb{P}_h(\cdot|s,a)$ gives the probability distribution over next state based on the state and action pair (s,a) at epoch h. Further, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the deterministic reward function and $f_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $i = 1, \cdots, I$, are peak constraint functions. s_1 is a fixed initial state. In the RL setting, the transition dynamics $\mathbb{P}_h$, the reward function r and constraint functions f_i are unknown to the agent but can be measured when a state action pair (s,a) is observed. If we know the model of MDP (which means that the transition dynamics, the reward and constraint functions are known), we can solve the problem by solving the optimal Bellman Equation.

$$\tilde{Q}_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h \tilde{V}_{h+1}^*(s,a)] \quad \tilde{V}_h^*(s) = \max_{a \in \mathcal{A}_h^{safe}(s)} \tilde{Q}_h^*(s,a) \tag{2}$$

where $\tilde{Q}(s,a)$ is the state-action value function for reward function $r(s,a)$ such that

$$\tilde{Q}_h^\pi(s,a) = \mathbf{E}[\sum_{h'=h+1}^H r_{h'}(s_{h'}, \pi_{h'} s_{h'})|s_h = s, a_h = a] \tag{3}$$

and $\mathcal{A}_h^{safe}(s) = \{a : f_i(s,a) \geq 0, \forall i \in [I]\}$. As a result, the problem can be considered as an unconstrained MDP problem with specified action set for each $s \in \mathcal{S}$ and $h \in [H]$. In this paper, we make the following assumptions.

Assumption 1 *The absolute values of the reward function r and constraint functions $f_i, i = 1, \cdots, I$ are strictly bounded by a constant known to the agent. Without loss of generality, we let this constant be 1.*

Assumption 2 *The values of the reward function r is non-negative, i.e., $0 \leq r(s,a) \leq 1, \forall (s,a)$.*

These assumptions on reward function are typical in reinforcement learning Jin et al. (2018); Ni et al. (2019); Azar et al. (2017), and the bound of reward function can be normalized. Further, the reward can be shifted up by adding a constant to make the reward function non-negative.

Remark 1 *(Nonidentical Initial State) Despite that the MDP model is defined with a fixed initial state for all episodes, the result in this paper still works for random initial state with a simple modification. Denote the distribution for the initial state as d such that $s_1 \sim d$. Then, we artificially add an extra state s_0 and define $r(s_0, a) = 0, \forall a \in \mathcal{A}$, $f_i(s_0, a) = 0, \forall a \in \mathcal{A}, \forall i \in [I]$ and $\mathbb{P}_0(s_1|s_0, a) = d(s_1), \forall a \in \mathcal{A}$. Also, we modify $\mathbb{P}_h(s_0|s,a) = 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.*

By this modification, s_0 is considered as a dummy state, which will not influence future epochs. Thus, the proposed algorithm can start from s_0 in this setting. Thus, a fixed initial state is without loss of generality.

We define the policy as a function that maps a state $s \in \mathcal{S}$ to a probability distribution of the actions with a probability assigned to each action $a \in \mathcal{A}$. In an episodic setting, the policy $\boldsymbol{\pi}$ is a collection of H policy functions π_h at each epoch, that is $\pi_h(s) = a$ with probability $\mathbf{Pr}(a|s, h)$. Constrained RL problem is concerned with finding the optimal policy to achieve the highest total reward subject to a set of constraints, which can be formally stated as

$$\textbf{Original PCMDP}: \quad \max_{\boldsymbol{\pi}} \quad \mathbf{E}\left[\sum_{h=1}^{H} r(s_h, \pi_h(s_h))\right] \tag{4}$$
$$\text{s.t.} \quad \mathbf{E}[f_i^-(s_h, \pi_h(s_h))] \geq 0 \quad \forall h \in [H], \forall i \in [I]$$

where the expectation is taken with respect to both the policy π and the transition probability $\mathbb{P}_h$, and $x^- \triangleq \min\{x, 0\}$. In the following parts, we use $\mathbf{E}$ instead of $\mathbf{E}_{\pi, \mathbb{P}}$ for simplicity, which is the expectation value on the randomness of policy and transition dynamics. The formulated problem in Eq. (4) is called **Original PCMDP** in this paper, which optimizes the total reward and satisfies the peak constraints simultaneously.

Remark 2 *Since this paper considers the episodic setting with a tabular MDP. For a fixed h, there exists a discrete distribution for each state action pair (s_h, a_h) where $a_h \sim \pi_h(\cdot|s_h)$ and $s_h \sim \mathbb{P}_{h-1}(\cdot|s_{h-1}, a_{h-1})$. Denote this distribution as $\lambda(s, a)$. Thus, the constraint in Eq. (4) can be expressed as*

$$\mathbf{E}[f_i^-(s_h, \pi_h(s_h))] = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \lambda(s, a) f_i^-(s, a) \tag{5}$$

*If $f_i(s, a) < 0$ with a positive probability, then $\sum_{s \in \mathcal{S}, a \in \mathcal{A}} \lambda(s, a) f_i^-(s, a) < 0$, which gives a contradiction with Eq. (4). Thus, the constraint in **Original PCMDP** can be considered as $f_i(s_h, a_h) \geq 0$ with probability 1, equivalently.*

We emphasis that the proposed PCMDP problem is a special case of Constrained MDP problem mentioned in Altman (1999). The difference is that constraint functions need to be satisfied in each epoch h in our formulation, while it is only needed to be satisfied on an average along one episode in Altman (1999). In the proof of Lemma 12, it can be seen that PCMDP can be converted to the standard CMDP with HI constraints. A standard approach for constrained MDP would use HI Q-tables, one for each constraint. However, in this work, we provide a low space-time complexity approach that uses a single Q-table.

In order to make the problem non-trivial, we assume that the problem has a feasible solution. More formally,

Assumption 3 (Feasibility) *There exists a policy π such that $\mathbf{E}[f_i^-(s_h, \pi_h(s_h))] \geq 0$ for all $h \in [H]$ and $i \in [I]$.*

We will make use of all the three assumptions in the remainder of the paper.

We note that based on the definition of $x^- = \min\{x, 0\}$, Slater Condition Lempio (1974) will not hold for Eq. (4). Thus, we introduce a new slack variable $\xi > 0$ and define $g_{i,\xi}(s, a) = f_i^-(s, a) + \xi$ to formulate the **Approximated PCMDP** as follows.

$$\textbf{Approximated PCMDP:} \quad \max_{\boldsymbol{\pi}} \quad \mathbf{E}\left[\sum_{h=1}^{H} r(s_h, \pi_h(s_h))\right] \tag{6}$$
$$\text{s.t.} \quad \mathbf{E}[g_{i,\xi}(s_h, \pi_h(s_h))] \geq 0 \;\; \forall h \in [H], \; \forall i \in [I]$$

We notice that the introduction of approximation parameter $\xi > 0$ relaxes constraints and makes the feasible region larger. In the next lemma, it is shown that the Slater Condition always holds for **Approximated PCMDP**.

Lemma 3 *There exists a set of $\gamma_{h,i}, h \in [H], i \in [I]$ satisfying $0 < \gamma_{h,i} \leq \xi, \forall h, i$ and a policy π such that*

$$\mathbf{E}[g_{i,\xi}(s_h, \pi_h(s_h))] \geq \gamma_{h,i}, \forall h, i \tag{7}$$

Proof By Assumption 3, we have

$$\mathbf{E}[g_{i,\xi}(s_h, \pi_h(s_h))] = \mathbf{E}[f_i^-(s_h, \pi_h(s_h))] + \xi \geq \xi \geq \gamma_{h,i} \tag{8}$$

$\blacksquare$

We define the state value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ at epoch h under policy π as follows

$$V_h^\pi(s) := \mathbf{E}\left[\sum_{h'=h}^{H} r(s_{h'}, \pi_{h'}(s_{h'}))|s_h = s\right] \tag{9}$$

Denote the set Π as the constraint set in which the policy satisfies the constraints in the Eq. (4). We denote an optimal policy as π^*, which gives the optimal value function for the original problem as

$$V_h^*(s) = \sup_{\pi \in \Pi} V_h^\pi(s), \tag{10}$$

for all $s \in \mathcal{S}$ and $h \in [H]$. Note that the introduction of approximation parameter ξ would change the optimal policy and thus the optimal value function for the relaxed problem is a function of ξ which is denoted by $V_{1,\xi}^*(s_1)$. In this paper, we extend the concept of PAC to define a concept of ϵ-optimal policy for both the reward function and constraint violation, which is given as follows.

Definition 4 ((ϵ, p)-PAC policy) *For any $p \in (0, 1)$, if a policy $\bar{\pi}$ satisfies the following equations with probability at least $1 - p$, then we call it an (ϵ, p)-PAC policy*

$$V_1^*(s_1) - V_1^{\bar{\pi}}(s_1) \leq \epsilon$$
$$\sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}\left|f_i^-(s_h, \bar{\pi}_h(s_h))\right| \leq \epsilon \tag{11}$$

By this definition, an (ϵ, p)-PAC policy is a policy for which the value function is ϵ close to the optimal and the total constraint violation is less than ϵ. Moreover, due to the definition of x^- and the absolute notation in Eq. (11), we know that the peak constraint violation is less than ϵ for each epoch h. Notice that the (ϵ, p)-PAC policy is defined with respect to **Original PCMDP**.

4. Proposed Algorithm

For any state action pair (s, a), we define a modified reward function as

$$R_\xi(s, a) = r(s, a) + \frac{\eta}{I} \sum_{i=1}^{I} g_{i,\xi}^-(s, a) \tag{12}$$

where $g_{i,\xi}^- := \min\{g_{i,\xi}, 0\}$, $\eta = \frac{2HI}{\gamma}$ and $\gamma = \min_{h,i} \gamma_{h,i}$. This modified reward function gives nearly the original reward function $r(s, a)$ when all constraints function are satisfied because $f_i^-(s, a) = 0$ in this case. Further, this provides a penalty function when any of the constraints is larger than ξ. Based on the modified reward function, we define a counterpart of the value function $W_{h,\xi}^\pi(s)$ as

$$W_{h,\xi}^\pi(s) := \mathbf{E}\left[\sum_{h'=h}^{H} R_\xi(s_{h'}, \pi_{h'}(s_{h'}))|s_h = s\right] \tag{13}$$

Further, let $W_{1,\xi}^*(s_1) = \max_\pi W_{1,\xi}^\pi(s_1)$. W with the notation $[\mathbb{P}_h V_{h+1}](s, a) := \mathbf{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} V_{h+1}(s')$, we define a counterpart of the state-action function $Q_h^\pi(s, a)$ as

$$Q_{h,\xi}^\pi(s, a) := R_\xi(s, a) + \mathbf{E}\left[\sum_{h'=h+1}^{H} R_\xi(s_{h'}, \pi_{h'}(s_{h'}))\bigg|s_h = s, a_h = a\right] = (R_\xi + \mathbb{P}_h W_{h+1,\xi}^\pi)(s, a) \tag{14}$$

With these notations, we are able to define a modified unconstrained MDP problem, **Modified MDP**, as

$$\textbf{Modified MDP}: \quad \max_\pi \quad \mathbf{E}\left[\sum_{h=1}^{H} R_\xi(s_h, \pi_h(s_h))\right] \tag{15}$$

Recalling the assumption that the original reward function $r(\cdot)$ is bounded, we show the absolute value of the modified reward function R_ξ is also bounded.

Lemma 5 *If $\gamma < \min\{\xi, 2HI(1 - \xi)\}$. the absolute value of the modified reward function $R(s, a)$ is bounded. Formally,*

$$|R(s, a)| \leq \eta \quad \forall(s, a) \in \mathcal{S} \times \mathcal{A} \tag{16}$$

Proof If $f_i(s, a) \geq -\xi$, then $R_\xi(s, a) = r(s, a)$ and thus $0 \leq R_\xi(s, a) \leq 1$. Otherwise, we have from (12) that

$$R_\xi(s, a) = r(s, a) + \frac{\eta}{I} \sum_{i=1}^{I} \left[f_i^-(s, a) + \xi\right] \tag{17}$$

Algorithm 1 Constrained Q-Learning Algorithm

1: Initialize $Q_h(s,a) \leftarrow \eta H$, $W_h(s,a) \leftarrow \eta H$, $N_h(s,a) \leftarrow 0$, $\mu_h(s,a) \leftarrow 0$, $\sigma_h(s,a) \leftarrow 0$ and $\beta_0(s,a,h) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Initial parameter ξ and η

2: **for** episode $k = 1,...K$ **do**

3:     Observe s_1

4:     **for** step $h = 1,...H$ **do**

5:         Take action $a_h \leftarrow arg \max_{a'} Q_h(s_h, a')$ and observe s_{h+1}

6:         $t = N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$

7:         $\mu_h(s_h, a_h) \leftarrow \mu_h(s_h, a_h) + W_{h+1}(s_{h+1})$

8:         $\sigma_h(s_h, a_h) \leftarrow \sigma_h(s_h, a_h) + (W_{h+1}(s_{h+1}))^2$

9:         $\beta_t(s_h, a_h, h) \leftarrow \min\{c_1(\sqrt{\frac{H}{t}(\frac{\sigma_h(s_h,a_h)-(\mu_h(s_h,a_h))^2}{t} + \eta H)\ell} + \eta\frac{\sqrt{H^7 SA\ell}}{t}), c_2\eta\sqrt{\frac{H^3\ell}{t}}\}$

10:         $b_t \leftarrow \frac{\beta_t(s_h,a_h,h)-(1-\alpha_t)\beta_{t-1}(s_h,a_h,h)}{2\alpha_t}$

11:         $Q_h(s_h, a_h) \leftarrow (1-\alpha_t)Q_h(s_h, a_h) + \alpha_t[R_h(s_h, a_h) + W_{h+1}(s_{h+1}) + b_t]$ (R_h as defined in Eq. (12))

12:         $W_h(s_h) \leftarrow \min\{\eta H, \max_{a' \in \mathcal{A}} Q_h(s_h, a')\}$

13:     **end for**

14: **end for**

Notice that $-\eta \leq \frac{\eta}{I}\sum_{i=1}^{I} f_i^-(s,a) \leq 0$. Thus, we have $-\eta + \eta\xi \leq R_\xi(s,a) \leq 1 + \eta\xi$. Since $\gamma < \min\{\xi, 2HI(1-\xi)\}$, we have $-\eta \leq R_\xi(s,a) < \eta$. Then, the two cases together provide the result in the statement of the Lemma. ∎

We use the modified reward function to provide a Q-learning based algorithm as described in Algorithm 1. The basic steps of Q-learning follow from that in Jin et al. (2018), while are adapted to incorporate constraints. In line 1, the agent initializes the Q-table and $N_h(s,a)$, which is the notation for the number of times that the state-action pair is taken at epoch h. In line 3, the agent is given an initial state at the beginning of each episode. Then, in line 5, the agent takes an action to maximize the current state-value function $Q_h(s_h, a_h)$ and observes the next state. $N_h(s,a)$ is updated in line 6. Line 7 to line 10 gives an efficient way to compute a Bernstein type UCB. Q-table and the W-table are then updated according to the line 11 and line 12, where b_t is the upper confidence bound and α_t is the learning rate defined as $\alpha_t := \frac{H+1}{H+t}$.

Using Algorithm 1, we find the policy π_h^k at the step h in episode k is deterministic, and is defined as

$$\pi_h^k(a|s_h^k) = \begin{cases} 1 & a = \arg\min_{a'} Q_h^k(s_h^k, a') \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

Given a Markov Decision Problem with peak constraints, this paper shows that an ϵ-optimal policy can be extracted from Algorithm 1. The guarantees of the proposed algorithm will be analyzed in the next section.

5. PAC Analysis

In this section, we will show that the proposed algorithm gives an ϵ-optimal policy for K large enough. In order to prove the result, we first provide connections among the original problem **Original PCMDP**, the approximated problem **Approximated PCMDP** and modified unconstrained problem **Modified MDP**. Then, we derive the sub-linear result for **Modified MDP**. The main result can be derived by the guarantees for **Modified MDP** and its relation with **Original PCMDP** and **Approximated PCMDP**

Two following results, Lemma 6 and 7, describe the relationship of optimal value function between **Modified MDP** and **Original PCMDP/Approximated PCMDP**, respectively.

Lemma 6 *The optimal value function V_1^* for **Original PCMDP** is equal to the optimal value function $W_{1,\xi}^*$ for **Modified MDP**. More formally,*

$$V_1^*(s_1) = W_{1,\xi}^*(s_1) \tag{19}$$

Proof Considering the optimal policy π^* in **Original PCMDP** in Eq. (4), we have

$$V_1^*(s_1) = \mathbf{E}\left[\sum_{h=1}^{H} r(s_h, \pi_h^*(s_h))\right] \overset{(a)}{=} \mathbf{E}\left[\sum_{h=1}^{H} R_\xi(s_h, \pi_h^*(s_h))\right] \leq W_{1,\xi}^*(s_1) \tag{20}$$

Step (a) holds because with feasible optimal policy in **Original PCMDP**, $f_i(s_h, a_h) \geq 0$ for any possible trajectories and thus $g_{i,\xi}(s_h, a_h) = f_i^-(s_h, a_h) + \xi = \xi$, which means

$$R_\xi(s_h, a_h) = r(s_h, a_h) + \frac{\eta}{I}\sum_{i=1}^{I}\left[g_{i,\xi}^-(s_h, a_h)\right] = r(s_h, a_h) \tag{21}$$

Moreover, the final inequality holds because the optimal policy for **Modified MDP** may be different from the **Original PCMDP** and any other policy will achieve less reward. For the other direction, consider the optimal policy π^{W*} in the **Modified MDP** and follow the same step,

$$W_{1,\xi}^*(s_1) = \mathbf{E}\left[\sum_{h=1}^{H} R_\xi(s_h, \pi_h^{W*}(s_h))\right] \leq \mathbf{E}\left[\sum_{h=1}^{H} r(s_h, \pi_h^{W*}(s_h))\right] \overset{(a)}{\leq} V_1^*(s_1) \tag{22}$$

The first inequality holds by $g_{i,\xi}^- \leq 0$ in Eq. (21). Thus, this gives the result as in the statement of the Lemma. ∎

Lemma 7 *The optimal value function $V_{1,\xi}^*$ for **Approximated PCMDP** and the optimal value function $W_{1,\xi}^*$ for **Modified MDP** have the following relation:*

$$W_{1,\xi}^*(s_1) \leq V_{1,\xi}^*(s_1) \leq W_{1,\xi}^*(s_1) + \eta H\xi \tag{23}$$

9

Proof Considering the optimal policy π^* in **Approximated PCMDP** in Eq. (6), define a distribution $d_h^{\pi^*,\mathbb{P}}$ such that $(s_h, a_h) \sim d_h^{\pi,\mathbb{P}}$ with policy π^* and transition dynamics $\mathbb{P}$, then

$$
\begin{aligned}
\mathbf{E}[g_{i,\xi}^-(s_h, \pi_h^*(s_h))] &= \sum_{s,a} g_{i,\xi}^-(s,a) d_h^{\pi^*,\mathbb{P}}(s,a) \\
&= \sum_{g_{i,\xi}(s,a)>0} g_{i,\xi}^-(s,a) d_h^{\pi^*,\mathbb{P}}(s,a) + \sum_{g_{i,\xi}(s,a)<0} g_{i,\xi}^-(s,a) d_h^{\pi^*,\mathbb{P}}(s,a) \\
&= \sum_{g_{i,\xi}(s,a)<0} g_{i,\xi}(s,a) d_h^{\pi^*,\mathbb{P}}(s,a) \\
&= \mathbf{E}[g_{i,\xi}(s_h, \pi_h^*(s_h))] - \sum_{g_{i,\xi}(s,a)>0} g_{i,\xi}(s,a) d_h^{\pi^*,\mathbb{P}}(s,a) \geq -\xi
\end{aligned}
\tag{24}
$$

where the last step holds because $\mathbf{E}[g_{i,\xi}(s_h, \pi_h(s_h))] \geq 0$ by the formulation (6) and $g_{i,\xi} \leq \xi$ by the definition. Thus,

$$
\mathbf{E}[R_\xi(s_h, \pi_h^*(s_h))] = \mathbf{E}[r(s_h, \pi_h^*(s_h))] + \frac{\eta}{I} \sum_{i=1}^{I} \mathbf{E}[g_{i,\xi}^-(s_h, \pi_h^*(s_h))] \geq \mathbf{E}[r(s_h, \pi_h^*(s_h))] - \eta\xi
\tag{25}
$$

Finally, by the definition of value function in (9), we have

$$
V_{1,\xi}^*(s_1) = \mathbf{E}\left[\sum_{h=1}^{H} r(s_h, \pi_h^*(s_h))\right] \leq \mathbf{E}\left[\sum_{h=1}^{H} R_\xi(s_h, \pi_h^*(s_h))\right] + \eta H\xi \leq W_{1,\xi}^*(s_1) + \eta H\xi
\tag{26}
$$

where the final inequality holds because the optimal policy for **Modified MDP** may be different from the **Approximated PCMDP** and any other policy will achieve less reward. For the other direction, recall that **Approximated PCMDP** is a relaxed version of **Original PCMDP** and thus

$$
V_{1,\xi}^*(s_1) \geq V_1^*(s_1) = W_{1,\xi}^*(s_1)
\tag{27}
$$

∎

The following lemma describes the relationship of value function with a certain policy between **Modified MDP** and **Original PCMDP**.

Lemma 8 *The value function with policy π in **Modified MDP**, $W_{1,\xi}^\pi$, can be expressed by the value function with policy π in **Original PCMDP**, V_1^π, with a gap term describing the violation of constraints. More formally,*

$$
W_{1,\xi}^\pi(s_1) = V_1^\pi(s_1) + \frac{\eta}{I} \sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}[g_{i,\xi}^-(s_h, \pi_h(s_h))]
\tag{28}
$$

Proof According to the definition of function W, we expand it as follows.

$$
\begin{aligned}
W_{1,\xi}^{\pi_k}(s_1^k) &= \mathbf{E}\bigg[\sum_{h=1}^{H} R_\xi(s_h^k, \pi_h^k(s_h^k))\bigg] \\
&= \mathbf{E}\bigg[\sum_{h=1}^{H} r(s_h^k, \pi_h^k(s_h^k))\bigg] + \frac{\eta}{I}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\big[g_{i,\xi}^{-}(s_h^k, \pi_h^k(s_h^k))\big] \qquad (29) \\
&= V_{1,\xi}^{\pi_k}(s_1^k) + \frac{\eta}{I}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\big[g_{i,\xi}^{-}(s_h^k, \pi_h^k(s_h^k))\big]
\end{aligned}
$$

which is the result as in the statement of the Lemma. ∎

The following lemma gives the sub-linear regret for **Modified MDP**.

Lemma 9 *For any $p \in (0,1)$, let $\ell = \log(SAT/p)$, where $T = KH$. Then, for $K \geq H^5 S^2 A^2 \ell^3$, the bound on the regret for **Modified MDP** with Algorithm 1 is given as*

$$
\sum_{k=1}^{K}[W_{1,\xi}^*(s_1) - W_{1,\xi}^{\pi^k}(s_1)] \leq O(\eta\sqrt{H^3 SAT\ell}) \qquad (30)
$$

with probability at least $1 - p$.

Proof The proof follows from Theorem 2 in (Jin et al., 2018), noting that the modified reward is bounded by η and not 1 as in (Jin et al., 2018). ∎

Combining results of the above lemmas, the next theorem shows that the proposed algorithm is ϵ-optimal.

Theorem 10 *For any $p \in (0,1)$, $H \geq 2$, and any $\epsilon > 0$, take $K = \Omega(\frac{I^2 H^6 SA\ell}{\epsilon^2})$, $\xi = \frac{\epsilon}{6HI}$, and $\ell = \log(SAT/p)$, where $T = KH$. Further, define $\bar{\pi}$ as a policy which is given as*

$$
\bar{\pi}(s) = \begin{cases} \pi^1(s) & \text{with probability } 1/K \\ \dots & \dots \\ \pi^k(s) & \text{with probability } 1/K \\ \dots & \dots \\ \pi^K(s) & \text{with probability } 1/K \end{cases} \qquad (31)
$$

Note that $\bar{\pi}$ chooses the different policies π^k for $k \in [K]$ uniformly at random. If $H \geq 2$, then, the policy $\bar{\pi}$ is an (ϵ, p)-PAC policy.

Proof The detailed proof is provided in the Appendix A. ∎

Remark 11 *The above result shows that we obtain an (ϵ, p)-PAC policy if the number of episodes satisfies $K \geq \Omega(\frac{I^2 H^6 SA\ell}{\epsilon^2})$. Thus, $K \geq \Omega(1/\epsilon^2)$, which is the best result for the problem even without constraints Jin et al. (2018).*

We also note that even though the Theorem statement mentions $H \geq 2$, the result also holds when $H = 1$, where we obtain 2ϵ-optimal strategy. Thus, the (ϵ, p)-PAC policy will hold by replacing ϵ with $\epsilon/2$.

6. Simulations
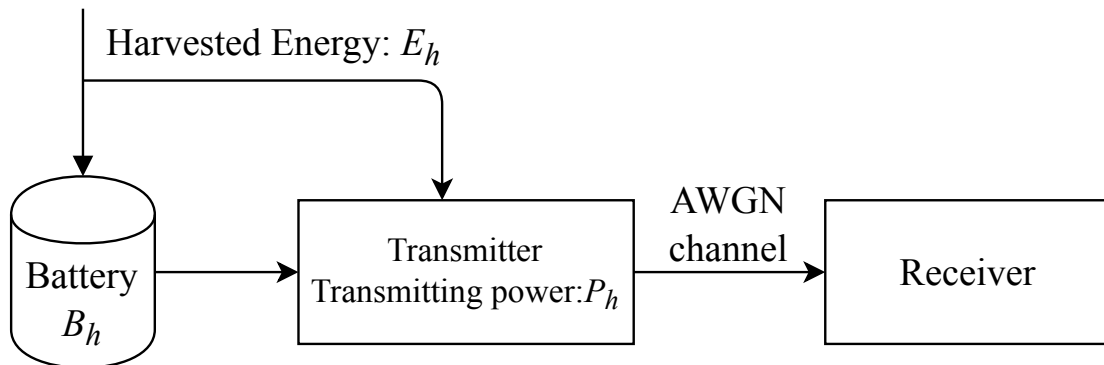
6.1 Energy Harvesting Communication System

Figure 1: Energy Harvesting Communication System

In this section, we evaluate the proposed algorithm on a communication channel, where the transmitter is powered by renewable energy. Such a model has been studied widely in communication systems Tutuncuoglu et al. (2015); Blasco et al. (2013); Yang and Ulukus (2012); Wang et al. (2015); Wang et al. (2014). In this model, we assume that the time is divided into time-slots. As shown in Fig. 1, in each time-slot, the transmitter can send data over an Additive Gaussian White Noise (AWGN) channel, where the signal transmitted by the transmitter gets added by a noise given by complex normal with zero mean and unit variance $\mathbb{CN}(0, 1)$ at each time instance within the time-slot. We assume that the transmitter can use a power of P_h in time-slot h, where the transmission is limited by a maximum power of $\bar{P}$.

We assume that the transmitter is powered by a renewable energy source, where energy E_h arrives during time-slot $h-1$ and can be used for time-slot h. Further, the transmitter is attached to a battery, which has a capacity of $\bar{B}$. The transmitter can use the energy from the existing battery capacity at the start of time-slot h, B_h, or the new energy arrival E_h. The energy from E_h that is not utilized is stored in the battery. Thus, the battery state evolves as

$$B_{h+1} = \min\{\bar{B}, B_h + E_h - P_h\}. \tag{32}$$

We wish to optimize an upper bound on the reliable transmission rate Wang et al. (2015), given as $\mathcal{C} = \sum_h \log(1 + P_h)$. We note that the transmission constraints can be modeled as

peak constraints. Thus, the overall optimization problem is given as

$$\max_{P_h, h=1,\cdots,H} \quad \mathbb{E}\left[\sum_{h=1}^{H} \log(1 + P_h)\right] \tag{33}$$

$$s.t. \quad B_{h+1} = \min\{\bar{B}, B_h + E_h - P_h\} \tag{34}$$

$$B_h \geq 0 \tag{35}$$

$$P_h \leq \bar{P} \quad w.p.1 \tag{36}$$

We note that the expectation in the above is over the energy arrivals E_h, which makes the choice of P_h stochastic. If the energy arrivals E_h are known non-causally (known at $h = 1$ for the entire future), the problem is convex and can be solved efficiently using the dynamic water-filling algorithm proposed in Wang et al. (2015) or dynamic programming based solutions Blasco et al. (2013); Wang et al. (2014). However, in realistic systems, E_h is only known at time-slot h.

   We will now model the problem as an MDP. The state at time-slot h is given as $S_h = (B_h, E_h)$, which are the current battery level and the energy arrival. The energy E_h is known causally, and the distribution is unknown. The action is the transmission power P_h. Eq. (34) gives the state evolution and the E_h may evolve based on some Markov process in general. Eq. (35) restricts battery level must be positive and indirectly gives the action space for P_h such that $\mathcal{A} = \{0, 1, ..., B_h + E_h\}$. Finally, the objective is given in (33) and the peak power constraint is given in Eq. (36). Notice that the the constraint function is not known in advance.

   We set the distribution of E_h as truncated Gaussian with mean μ and standard deviation σ, where the truncation levels are 0 and $\bar{E}$, and we let it be independent across episodes. The problem is discretized to integers in order to apply the proposed algorithm. According to the selection of the parameters in Wang et al. (2014), we set the horizon $H = 20$ time-slots, battery capacity $\bar{B} = 20$, power constraint $\bar{P} = 8$, maximal harvest energy $\bar{E} = 20$, mean and standard deviation $\mu = 10$, $\sigma = 5$, respectively.

   In the simulation, 1000 trajectories are generated by the above MDP. In Fig. 2, we plot the mean and variance for the sum of the transmission rate $V_1^{\bar{\pi}^k}$ and constraint violations defined in Eq. (11) and compare the learning speed for different choice of $\xi = 0.1, 0.01, 0.001$. The Slater parameter for γ is chosen to be $\frac{\xi}{2}$. Note that in each episode k, we evaluate the policy $\bar{\pi}^k$, which is the 'averaged' policy by $\pi^1, \pi^2, \cdots, \pi^k$ as defined in Theorem 10. It can be seen that the sum of transmission rate converges to the optimal and the constraint violation converge to 0 as the number of episodes increases. Moreover, we can find that larger γ gives a higher convergence speed while the difference between convergence speed for different γ is quite small.

   In order to compare the proposed algorithm, we consider three other baseline algorithms: the greedy policy, the balanced policy, and the optimal non-causal algorithm. The greedy policy tries to consume the harvested energy as much as possible in each slot, as calculated by $P_h = \min(\bar{P}, B_h + E_h)$. We also consider a balanced policy that consumes the fixed amount of energy in each slot if available, where the fixed value is calculated by $\sum_{h=1}^{H} E_h / H$, while that is limited by the available energy at each time. However, the balanced algorithm uses the future energy arrivals and is not a causal strategy. Further, the optimal strategy when all future energy arrivals are non-causally known is also used to show the performance
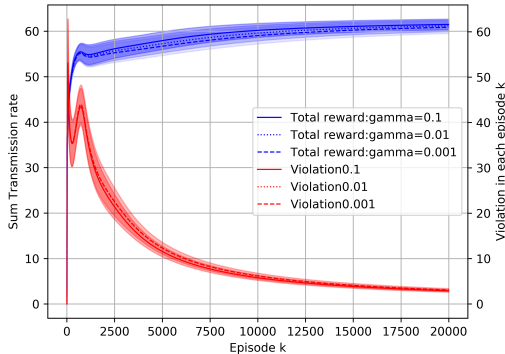
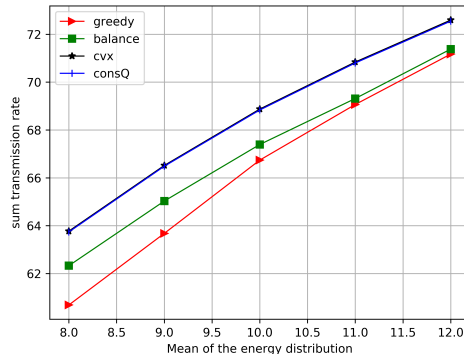Figure 2: Convergence of Reward Function and Constraint Violations for the Proposed Algorithm

Figure 3: Performance Comparison Between Different Algorithms

of the proposed algorithm. We note that the proposed algorithm only assumes that the constraint function in state s and action a can be queried, but the function is not explicitly known, thus, the algorithms that project to the constraint function are not considered as they require complete knowledge of the function.

In Fig. 3, we set the mean of the harvested energy as $8, 9, 10, 11, 12$ and compare the performance of different algorithms. In order to illustrate the policy convergence to the optimal, we choose $K = 50000$ for the proposed algorithm. $\bar{P}$ is set to 15 in this comparison. We see the balanced policy achieves higher performance than greedy policy because the energy can be allocated more reasonably while requiring non-causal information of energy arrival. The performance of the non-causal convex solver achieves the highest reward since it is an upper bound on the performance. However, the proposed algorithm achieves nearly the same performance as the upper bound, which shows that our algorithm is able to achieve the optimal solution. Furthermore, the proposed algorithm doesn't need any prior knowledge of the harvested energy and the constraint functions, which is a great advantage over the convex solver.

6.2 Single machine scheduling problem with deadlines

In this subsection, we evaluate our proposed algorithm on another problem, single machine scheduling with deadline. We adopt the problem formulated in Koulamas and Kyparisis (2001). In this scheduling problem, we assume there are a total of n jobs and each job is released at the beginning. Each job i has a process time p_i, due time d_i, and deadline $\bar{d}_i$. Notice that the deadline is different from due time in the sense that the deadline cannot be violated. The tardiness of each job is computed by $T_i = \max\{0, C_i - d_i\}$, where C_i is the completion time of job i. The preemption is not allowed in this problem. We want to minimize the maximal tardiness $T_{max} = \max_{i \in N} T_i$ and satisfy each deadline at the same time. This problem is written as $1|\bar{d}_i|T_{max}$ in the standard scheduling notation. If the

Table 1: Information of jobs in the 1st example of Single Machine Scheduling

Processing time (p_i)	3	5	7	9	10
Due Time (d_i)	22	30	33	15	18
Deadline ($\bar{d}_i$)	30	28	35	18	21

Table 2: Information of jobs in 2nd example of Single Machine Scheduling

Processing time (p_i)	2	3	5	8	13	21	34	17	19
Due Time (d_i)	75	70	65	60	88	35	59	100	100
Deadline ($\bar{d}_i$)	70	70	70	100	90	40	60	130	110

information of this problem, the process time p_i, the due time d_i, and deadline $\bar{d}_i$ are given in advance, there exists an optimal algorithm DT proposed in Koulamas and Kyparisis (2001) and thus the DT algorithm can be considered as an offline baseline.

The scheduling problem with the deadline which cannot be violated is an important setting in the field of scheduling. This problem has several applications in practice such as the chemical industry Koulamas and Kyparisis (2001), energy efficient packet transmission Shan et al. (2014), workflow scheduling Abrishami and Naghibzadeh (2012), etc. In order to solve this problem with the proposed algorithm, it needs to be first formulated as a PCMDP.

The state of CMDP s in this problem consists of three parts, time t, job states js, and maximal tardiness $maxT$ such that $s = [t, js, maxT]$. The job states js are made of state of each job such that $js = [js_1, js_2, \cdots, js_N]$ where $js_i = 0$ means the job i has not been started and $js_i = 1$ means the job i has been completed. $maxT$ here stands for the current maximal tardiness which can be defined as $maxT = \max_{i \in A_t} T_i$ and A_t is the set including all completed jobs until time t. The action space of CMDP is the list of uncompleted jobs. Denote a as the action, t', js', $maxT'$ as the state of CMDP in next time. The dynamics of this PCMDP is written as follow

$$
\begin{aligned}
t' &= t + p_a \\
js' &= [js_1, \cdots, 1, \cdots, js_N] (\text{where 1 is at position a}) \\
maxT' &= \max\{maxT, \max\{0, t + p_a - d_a\}\}
\end{aligned}
\tag{37}
$$

Besides, according to the objective of problem setting which is to minimize the maximum of tardiness, the reward function and constraint function are designed as

$$
\begin{aligned}
r(s, a) &= -\max\{0, t + p_a - d_a - maxT\} \\
f(s, a) &= \max\{0, t + p_a - \bar{d}_a\}
\end{aligned}
\tag{38}
$$

It can be seen that the dynamics, reward, and constraint functions are totally decided by current state and action. Thus, the above formulation can be considered as a CMDP.

To test our algorithm and compare with the offline DT algorithm, we manually design two setups with 5 jobs and 9 jobs, respectively. With 5 jobs, we consider two scenarios, where the processing time of the jobs are deterministic or stochastic. For 9 jobs, we consider deterministic processing times. We label the three examples here as Example 1: 5

Table 3: Information of jobs in the 3rd example of Single Machine Scheduling

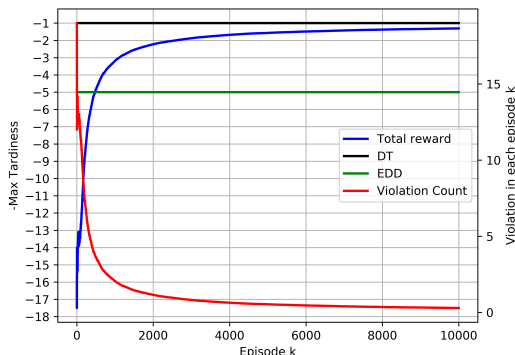| Processing time (p_i)	$\sim U(2,4)$	$\sim U(4,6)$	$\sim U(3,8)$	$\sim U(8,11)$	$\sim U(8,11)$
Due Time (d_i)	22	30	33	15	12
Deadline $(\bar{d}_i)$	30	28	35	18	23

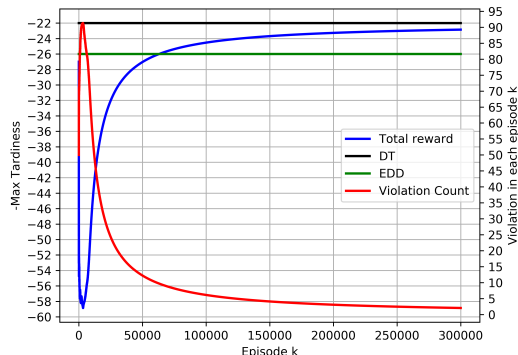Figure 4: Comparison of proposed algorithm with DT and EDD in 1st example

Figure 5: Comparison of proposed algorithm with DT and EDD in 2nd example

jobs, deterministic processing time, Example 2: 9 jobs, deterministic processing time, and Example 3: 5 jobs, stochastic processing time. The details for these examples are listed in Tables 1, 2 and 3, respectively.

The comparison among the proposed algorithm, offline DT algorithm and Earliest Deadline Date (EDD) algorithm for three examples are shown in Fig. 4, 5, and 6, respectively. In Fig. 6, we run 100 independent tests and plot the average maximal tardiness and the standard variance. It can be seen that EDD algorithm always selects the job which has the earliest deadline at each time.Further, EDD algorithm in this setting is sub-optimal. Furthermore, the total reward of the proposed algorithm converges to optimal DT value and the constraint violation converges to 0, which matches the theoretical result. Notice that the proposed algorithm is an online algorithm and thus doesn't need any information of job in advance, which is a great advantage over the offline DT algorithm.

Finally, it is noteworthy to see that the proposed algorithm can also be applied in the other settings with deadline such as the total tardiness Koulamas and Kyparisis (2001), total weight late work Chen et al. (2019) which has been shown as NP-hard and lack an optimal algorithm. However, the proposed algorithm can achieve ϵ-optimal in finite time.

7. Conclusion

In this paper, we formulate a constrained MDP problem with a set of peak constraints. By using a modified reward function, we convert the problem into an unconstrained RL problem
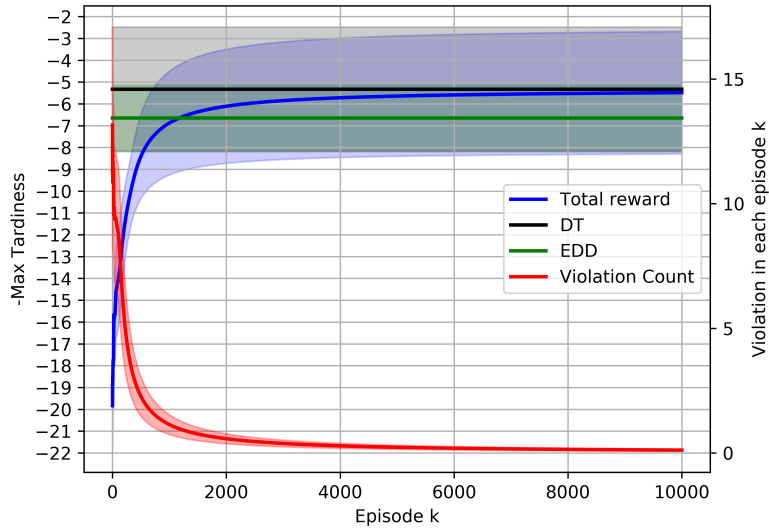
Figure 6: Performance comparison between different algorithms in 3rd example

and propose a novel model-free algorithm. This paper proves our algorithm can achieve ϵ-optimal objective and constraint violations. We note this is the first result of PAC analysis for PCMDP when the state evolution and the constraint functions are unknown. The result is applied to an energy harvesting communication link and a single machine scheduling with deadline, and the proposed algorithm is shown to be close to the non-causal optimal solution.

Appendix A. Proof of Theorem 10

A.1 Bounding the Regret

Combining the results in Lemmas 6, 8, and 9, for large enough T, we have

$$\sum_{k=1}^{K}[V_1^*(s_1) - V_1^{\pi^k}(s_1)] - \frac{\eta}{I}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\big[g_{i,\xi}^-(s_h^k, \pi_h^k(s_h^k))\big] \le \sum_{k=1}^{K}W_{1,\xi}^*(s_1) - W_{1,\xi}^{\pi^k}(s_1) \le O(\eta\sqrt{H^3SATl})$$

(39)

We note that the sum of terms $g_{i,\xi}^-$ (Accumulation of constraint violations) on the left hand side of (39) is non-positive, which gives

$$\frac{1}{K}\sum_{k=1}^{K}[V_{1,\xi}^*(s_1) - V_1^{\pi^k}(s_1)] \le O(\frac{\eta}{K}\sqrt{H^3SAT\ell})$$

(40)

We define a new policy $\bar{\pi}$ which uniformly chooses the policy π^k for $k \in [K]$. By the occupancy measure method, $V_1^{\pi^k}(s_1)$ is linear in terms of an occupancy measure induced by policy π^k and initial state s_1, thus:

$$\frac{1}{K}\sum_{k=1}^{K}V_1^{\pi^k}(s_1) = V_1^{\bar{\pi}}(s_1)$$

(41)

Combining Eq. (40) with (41) and recalling our choice for $\eta = \frac{2HI}{\gamma}$, we have

$$V_1^*(s_1) - V_1^{\bar{\pi}}(s_1) \le O(I\sqrt{\frac{H^6SA\ell}{K}})$$

(42)

A.2 Bounding the Constraint Violations

Similar to Eq. (39), however, instead of Lemma 6, 8 and 9, using Lemma 7, 8 and 9, we have

$$\frac{1}{K}\sum_{k=1}^{K}[V_{1,\xi}^*(s_1) - V_1^{\pi^k}(s_1)] - \frac{1}{K}\frac{\eta}{I}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\big[g_{i,\xi}^-(s_h^k, \pi_h^k(s_h^k))\big] \le O(\frac{\eta}{K}\sqrt{H^3SATl}) + \eta H\xi$$

(43)

Due to the concavity of the function $x^- = \min\{x, 0\}$ and by Jensen Inequality, we have

$$\frac{\eta}{I}\frac{1}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\big[g_{i,\xi}^-(s_h^k, \pi_h^k(s_h^k))\big] \le \frac{\eta}{I}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\big[g_{i,\xi}^-(s_h, \bar{\pi}_h(s_h))\big]$$

(44)

Combining (43) with Eq. (41) and (44), we have

$$V_{1,\xi}^*(s_1) - V_1^{\bar{\pi}}(s_1) - \frac{\eta}{I}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\big[g_{i,\xi}^-(s_h, \bar{\pi}_h(s_h))\big] \le O(\frac{\eta}{K}\sqrt{H^3SATl}) + \eta H\xi$$

(45)

Notice that $\eta = \frac{2HI}{\gamma}$ satisfies the condition that $C^* \ge 2Y^*$ in Lemma 15, thus,

$$\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{E}\Big|g_{i,\xi}^-(s_h, \bar{\pi}_h(s_h))\Big| \le O(\frac{I}{K}\sqrt{H^3SATl}) + 2HI\xi = O(I\sqrt{\frac{H^4SA\ell}{K}}) + 2HI\xi$$

(46)

18

A.3 ϵ-Optimal Policy

Choosing $K = \Omega(\frac{I^2 H^6 SA\ell}{\epsilon^2})$, by Eq. (42) and (46), we have

$$V_1^*(s_1) - V_1^{\bar{\pi}}(s_1) \leq \epsilon$$

$$\sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E} \left| g_i^-(s_h, \bar{\pi}_h(s_h)) \right| \leq \frac{\epsilon}{H} + 2HI\xi \tag{47}$$

Recalling the definition of $g_{i,\xi} = f_i^-(s,a) + \xi$, we have

$$\sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E} \left| f_i^-(s_h, \bar{\pi}_h(s_h)) \right| \leq \frac{\epsilon}{H} + 3HI\xi \tag{48}$$

Since $\xi = \frac{\epsilon}{6HI}$, the above shows that we obtain an ϵ-optimal strategy.

Appendix B. Supporting Lemmas from Optimization

We collect some standard results from the literature for readers' convenience. The following lemmas are mainly from (Ding et al., 2020), while are extended to the peak constraint setting rather than average constraint in the prior work. First, we rewrite our constrained problem (6) as,

$$\max_{\pi} \ \mathbf{E} \left[\sum_{h=1}^{H} r_h(s_h, \pi_h(s_h)) \right] \ s.t \ \mathbf{E} \big[g_{i,\xi}(s_h, \pi_h(s_h)) \big] \geq 0 \quad \forall h, i \tag{49}$$

Let the optimal solution be π^* such that

$$V_{1,\xi}^{\pi^*}(s_1) \ = \ \max_{\pi} \ \{ V_1^{\pi}(s_1) \, | \, \mathbf{E} \big[g_{i,\xi}(s_h, \pi_h(s_h)) \big] \geq 0 \quad \forall h, i \}. \tag{50}$$

Let the Lagrangian be $\mathcal{L}_{\xi}(\pi, \boldsymbol{\lambda}) := V_1^{\pi}(s_1) + \sum_{h=1}^{H} \sum_{i=1}^{I} \lambda_{h,i} \mathbf{E} \big[g_{i,\xi}(s_h, \pi_h(s_h)) \big]$, where $\lambda_{h,i} \geq 0$ is the Lagrange multipliers or dual variables. The associated dual function is defined as

$$\mathcal{D}_{\xi}(\boldsymbol{\lambda}) := \max_{\pi} \mathcal{L}_{\xi}(\pi, \boldsymbol{\lambda}) := V_1^{\pi}(s_1) + \sum_{h=1}^{H} \sum_{i=1}^{I} \lambda_{h,i} \mathbf{E} \big[g_{i,\xi}(s_h, \pi_h(s_h)) \big] \tag{51}$$

and the optimal dual is $\boldsymbol{\lambda}^* := \arg\min_{\boldsymbol{\lambda} \geq 0} \mathcal{D}_{\xi}(\boldsymbol{\lambda})$,

$$\mathcal{D}_{\xi}(\boldsymbol{\lambda}^*) := \min_{\boldsymbol{\lambda} \geq 0} \mathcal{D}_{\xi}(\boldsymbol{\lambda}) \tag{52}$$

We recall that the problem (6) enjoys strong duality under the standard Slater condition. The proof is a special case of Ding et al. (2020) in finite-horizon for peak constraint.

Assumption 4 (Slater Condition) *There exists $\gamma_{h,i} > 0, \forall h, i$ and $\hat{\pi}$ such that $\mathbf{E} \big[g_{i,\xi}(s_h, \hat{\pi}_h(s_h)) \big] \geq \gamma_{h,i}$.*

Lemma 12 *[Strong Duality]Ding et al. (2020) If the Slater condition holds, then the strong duality holds for Eq. (6),*

$$V_{1,\xi}^{\pi^*}(s_1) \ = \ \mathcal{D}_{\xi}(\boldsymbol{\lambda}^*).$$

Proof This is an extension of Lemma 7 in Ding et al. (2020) to the peak constraint case. To prove the strong duality still holds for peak constraint, we construct an equivalent problem to Eq. (6) in the average constraint setting. Consider, for each fixed $h \in [H]$ and $i \in [I]$, we define a new function $\tilde{g}_{h',i,\xi}$ in the following way.

$$\tilde{g}_{h',i,\xi}(s,a) = \begin{cases} g_{i,\xi}(s,a) & \text{if } h' = h \\ 0 & \text{otherwise} \end{cases} \tag{53}$$

By this way, the original constraint function can be written as

$$\mathbf{E}[g_{i,\xi}(s_h, \pi_h(s_h))] = \mathbf{E}\left[\sum_{h'=1}^{H} \tilde{g}_{h',i,\xi}(s_{h'}, \pi_{h'}(a_{h'})) \right] =: V^{\pi}_{1,\tilde{g}_{h',i,\xi}}(s_1) \tag{54}$$

Thus, an equivalent problem to Eq. (6) is

$$\max_{\pi} \; \mathbf{E}\left[\sum_{h=1}^{H} r_h(s_h, \pi_h(s_h)) \right] \;\; s.t \;\; V^{\pi}_{1,\tilde{g}_{h',i,\xi}}(s_1) \geq 0 \quad \forall h', i \tag{55}$$

It can be seen that Eq. (55) is the MDP problem with averaged constraints which is the same setting as Ding et al. (2020). The strong duality holds for Eq. (55) and thus holds for the Eq. (6) ∎

It is implied by the strong duality that the optimal solution to the dual problem: $\min_{\boldsymbol{\lambda} \geq 0} \mathcal{D}(\boldsymbol{\lambda}, \xi)$ is obtained at λ^*. Denote the set of all optimal dual variables as Λ^*.

Under the Slater condition, an useful property of the dual variable is that the sub-level sets are bounded Ding et al. (2020).

Lemma 13 (Boundedness of Sublevel Sets of the Dual Function) *Let the Slater condition hold. Fix $C \in \mathbb{R}$. For any $\boldsymbol{\lambda} \in \{\boldsymbol{\lambda} \geq 0 \,|\, \mathcal{D}_\xi(\boldsymbol{\lambda}) \leq C\}$, define $Y := \max_{h,i}\{\lambda_{h,i}\}$ and $\gamma = \min_{h,i} \gamma_{h,i}$, it holds that*

$$Y \leq \frac{1}{\gamma}\left[C - V_1^{\hat{\pi}}(s_1) \right].$$

Proof

$$C \geq \mathcal{D}_\xi(\boldsymbol{\lambda}) \geq V_1^{\hat{\pi}}(s_1) + \sum_{h=1}^{H}\sum_{i=1}^{I} \lambda_{h,i} \mathbf{E}\left[g_{i,\xi}(s_h, \hat{\pi}_h(s_h)) \right] \geq V_r^{\hat{\pi}}(s_1) + Y\gamma \tag{56}$$

where we utilize the Slater point $\hat{\pi}$ and $Y \leq \sum_{h,i} \lambda_{h,i}$ in the last inequality. We complete the proof by noting $\gamma > 0$. ∎

Corollary 14 (Boundedness of $Y^\star$) *Let the Slater condition hold, if we take $C = V_{1,\xi}^{\pi^*}(x_1)$ and by Strong Duality $C = \mathcal{D}_\xi(\boldsymbol{\lambda}^*)$, then $\Lambda^\star = \{\boldsymbol{\lambda} \geq 0 \,|\, \mathcal{D}_\xi(\boldsymbol{\lambda}) \leq C\}$. Thus, for any $\boldsymbol{\lambda} \in \Lambda^\star$, define $Y^* := \max_{h,i}\{\lambda_{h,i}\}$*

$$Y^* \leq \frac{1}{\gamma}\left[V_{1,\xi}^{\pi^*}(s_1) - V_1^{\hat{\pi}}(s_1) \right] \leq \frac{H}{\gamma}.$$

Another useful theorem from the optimization is given as follows. It describes that the constraint violation $\sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}|g_{i,\xi}^-(s_h, \pi_h(s_h))|$ can be bounded if we have some weak bound. We next state and prove it for our problem, which is used in our constraint violation analysis in Appendix A.

Lemma 15 (Constraint Violation) *Let the Slater condition hold and $\boldsymbol{\lambda}^\star \in \Lambda^\star$. If $C^\star \geq 2Y^\star$ and for any $\tilde{\pi}$ we have*

$$V_{1,\xi}^{\pi^*}(s_1) - V_1^{\tilde{\pi}}(s_1) - C^\star \sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}\big[g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\big] \leq \delta \tag{57}$$

Then,

$$\sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}\left|g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\right| \leq \frac{2\delta}{C^*} \tag{58}$$

where $x^- = \min(x, 0)$.

Proof Define $\boldsymbol{\tau} := \{\tau_{h,i}\}_{h \in [H], i \in [I]}$ and let

$$v_\xi(\boldsymbol{\tau}) \triangleq \max_\pi \{V_1^\pi(s_1) | \mathbf{E}\big[g_{i,\xi}(s_h, \pi_h(s_h))\big] \geq \tau_{h,i} \quad \forall h, i\}. \tag{59}$$

By definition, $v_\xi(\mathbf{0}) = V_{1,\xi}^{\pi^*}(s_1)$. By the Lagrangian and the strong duality,

$$\mathcal{L}_\xi(\pi, \boldsymbol{\lambda}^*) \leq \max_\pi \mathcal{L}_\xi(\pi, \boldsymbol{\lambda}^*) = \mathcal{D}_\xi(\boldsymbol{\lambda}^*) = V_{1,\xi}^{\pi^*}(s_1) = v_\xi(\mathbf{0}) \tag{60}$$

For any $\tau_{h,i} \in \mathbb{R}$ and $\pi \in \{\pi | \mathbf{E}\big[g_{i,\xi}(s_h, \pi_h(s_h))\big] \geq \tau_{h,i} \quad \forall h, i\}$, we have

$$
\begin{aligned}
v_\xi(\mathbf{0}) - \sum_{h=1}^{H} \sum_{i=1}^{I} \tau_{h,i} \lambda_{h,i}^* &\geq \mathcal{L}_\xi(\pi, \boldsymbol{\lambda}^\star) - \sum_{h=1}^{H} \sum_{i=1}^{I} \tau_{h,i} \lambda_{h,i}^* \\
&= V_1^\pi(s_1) + \sum_{h=1}^{H} \sum_{i=1}^{I} \lambda_{h,i}^* \mathbf{E}\big[g_{i,\xi}(s_h, \pi_h(s_h))\big] - \sum_{h=1}^{H} \sum_{i=1}^{I} \tau_{h,i} \lambda_{h,i}^* \\
&= V_1^\pi(s_1) + \sum_{h=1}^{H} \sum_{i=1}^{I} \lambda_{h,i}^* [\mathbf{E}\big[g_{i,\xi}(s_h, \pi_h(s_h))\big] - \tau_{h,i}] \\
&\geq V_1^\pi(s_1).
\end{aligned}
$$

If we maximize the right-hand side of above inequality over $\{\pi | \mathbf{E}\big[g_{i,\xi}(s_h, \pi_h(s_h))\big] \geq \tau_{h,i} \quad \forall h, i\}$, we have

$$v_\xi(\mathbf{0}) - \sum_{h=1}^{H} \sum_{i=1}^{I} \tau_{h,i} \lambda_{h,i}^* \geq v_\xi(\boldsymbol{\tau}) \tag{61}$$

On the other hand, if we define $\tilde{\boldsymbol{\tau}} = \{\tilde{\tau}_{h,i}\}$ and $\tilde{\tau}_{h,i} := \mathbf{E}\big[g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\big]$, notice that the policy $\tilde{\pi} \in \{\pi | \mathbf{E}\big[g_{i,\xi}(s_h, \pi_h(s_h))\big] \geq \tilde{\tau}_{h,i}, \forall h, i\}$. Thus,

$$V_1^{\tilde{\pi}}(s_1) \leq v_\xi(\tilde{\boldsymbol{\tau}}). \tag{62}$$

Combining (61) and (62), we have

$$V_1^{\tilde{\pi}}(s_1) - V_{1,\xi}^{\pi^*}(s_1) \leq v_\xi(\tilde{\boldsymbol{\tau}}) - v_\xi(\mathbf{0}) \leq -\sum_{h=1}^{H} \sum_{i=1}^{I} \tilde{\tau}_{h,i} \lambda_{h,i}^* \tag{63}$$

By the assumption in the statement of the lemma, we have

$$V_{1,\xi}^{\pi^*}(s_1) - V_1^{\tilde{\pi}}(s_1) - C^\star \sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}\big[g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\big] \leq \delta \tag{64}$$

Combined with Eq. (63), we have

$$-C^\star \sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}\big[g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\big] \leq \delta - \sum_{h=1}^{H} \sum_{i=1}^{I} \tilde{\tau}_{h,i}\lambda_{h,i}^* \tag{65}$$

Recalling the definition of $\tilde{\tau}_{h,i} = \mathbf{E}\big[g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\big]$ above Eq. (62), we have

$$\sum_{h=1}^{H} \sum_{i=1}^{I} (\lambda_{h,i}^* - C^*)\mathbf{E}\big[g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\big] \leq \delta \tag{66}$$

Due to $g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h)) \leq 0$, we have

$$\sum_{h=1}^{H} \sum_{i=1}^{I} (C^* - \lambda_{h,i}^*)\mathbf{E}\big|g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\big| \leq \delta \tag{67}$$

Finally, due to $C \geq 2Y^* = 2\max_{h,i} \lambda_{h,i}^*$, we have

$$\sum_{h=1}^{H} \sum_{i=1}^{I} \mathbf{E}\left|g_{i,\xi}^-(s_h, \tilde{\pi}_h(s_h))\right| \leq \frac{2\delta}{C^*} \tag{68}$$

$\blacksquare$

References

S. Abrishami and M. Naghibzadeh. Deadline-constrained workflow scheduling in software as a service cloud. *Scientia Iranica*, 19(3):680–689, 2012. ISSN 1026-3098. doi: https://doi.org/10.1016/j.scient.2011.11.047. URL https://www.sciencedirect.com/science/article/pii/S1026309812000351.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NeurIPS*, pages 1184–1194, 2017.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, pages 263–272, 2017.

Elizabeth E Bailey. Peak-load pricing under regulatory constraint. *Journal of Political Economy*, 80(4):662–679, 1972.

P. Blasco, D. Gunduz, and M. Dohler. A learning theoretic approach to energy harvesting communication system optimization. *IEEE Transactions on Wireless Communications*, 12(4):1872–1882, April 2013. ISSN 1558-2248. doi: 10.1109/TWC.2013.030413.121120.

M. Bouton, J. Karlsson, A. Nakhaei, K. Fujimura, M. Kochenderfer, and J. Tumova. Reinforcement learning with probabilistic guarantees for autonomous driving. *CoRR*, abs/1904.07189, 2019. URL http://arxiv.org/abs/1904.07189.

Rubing Chen, Jinjiang Yuan, C.T. Ng, and T.C.E. Cheng. Single-machine scheduling with deadlines to minimize the total weighted late work. *Naval Research Logistics (NRL)*, 66(7):582–595, 2019. doi: https://doi.org/10.1002/nav.21869. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.21869.

D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanović. Provably efficient safe exploration via primal-dual policy optimization. *arXiv preprint arXiv:2003.00534*, 2020.

Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained mdps. *arXiv:2003.02189*, 2020.

Kan Fang, Nelson A Uhan, Fu Zhao, and John W Sutherland. Flow shop scheduling with peak power consumption constraints. *Annals of Operations Research*, 206(1):115–145, 2013.

Ather Gattami. Reinforcement learning of markov decision processes with peak constraints. *arXiv:1901.07839*, 2019.

Peter Geibel. Reinforcement learning for MDPs with constraints. In *European Conference on Machine Learning*, pages 646–653. Springer, 2006.

Peter Geibel and Fritz Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *JMLR*, 11(Apr):1563–1600, 2010.

Rodolphe Jenatton, Jim C Huang, and Cedric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *ICML*, pages 402–411, 2016.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *NeurIPS*, pages 4863–4873. Curran Associates, Inc., 2018.

Sham Kakade, Mengdi Wang, and Lin F Yang. Variance reduction methods for sublinear reinforcement learning. *arXiv:1802.09184*, 2018.

Gopinath Karmakar, Ashutosh Kabra, and Krithi Ramamritham. Coordinated scheduling of thermostatically controlled real-time systems under peak power constraint. In *RTAS*, pages 33–42. IEEE, 2013.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

Christos Koulamas and George J Kyparisis. Single machine scheduling with release times, deadlines and tardiness objectives. *European Journal of Operational Research*, 133(2):447–453, 2001. ISSN 0377-2217. doi: https://doi.org/10.1016/S0377-2217(00)00207-1. URL `https://www.sciencedirect.com/science/article/pii/S0377221700002071`. Multiobjective Programming and Goal Programming.

Frank Lempio. A note on the slater-condition. In *Numerische Methoden bei Optimierungsaufgaben*, pages 165–165. Springer, 1974.

Degao Li, Andrew A Goldenberg, and Jean W Zu. A new method of peak torque reduction with redundant manipulators. *IEEE Transactions on Robotics and Automation*, 13(6):845–853, 1997.

Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: Online convex optimization with long term constraints. *arXiv:1111.6082*, 2011.

Arkadi Nemirovski. Prox-method with rate of convergence o $(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

C. Ni, L. Yang, and M. Wang. Learning to control in metric space with optimal regret. In *Allerton*, page 726–733. IEEE Press, 2019. doi: 10.1109/ALLERTON.2019.8919864. URL `https://doi.org/10.1109/ALLERTON.2019.8919864`.

Santiago Paternain, Miguel Calvo-Fullana, Luiz F. O. Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *arXiv:1911.09101*, 2019.

Shlomo Shamai and Israel Bar-David. The capacity of average and peak-power-limited quadrature gaussian channels. *IEEE Transactions on Information Theory*, 41(4):1060–1071, 1995.

Feng Shan, Junzhou Luo, and Xiaojun Shen. Optimal energy efficient packet scheduling with arbitrary individual deadline guarantee. *Computer Networks*, 75:351–366, 2014. ISSN 1389-1286. doi: https://doi.org/10.1016/j.comnet.2014.10.022. URL `https://www.sciencedirect.com/science/article/pii/S1389128614003843`.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *ICML*, pages 881–888, 2006.

K. Tutuncuoglu, A. Yener, and S. Ulukus. Optimum policies for an energy harvesting transmitter under energy storage losses. *IEEE JSAC*, 33(3):467–481, 2015.

Z. Wang, V Aggarwal, and X. Wang. Power allocation for energy harvesting transmitter with causal information. *IEEE Transactions on Communications*, 62(11):4080–4093, 2014.

Z. Wang, V. Aggarwal, and X. Wang. Iterative dynamic water-filling for fading multiple-access channels with energy harvesting. *IEEE JSAC*, 33(3):382–395, March 2015. ISSN 1558-0008. doi: 10.1109/JSAC.2015.2391571.

J. Yang and S. Ulukus. Optimal packet scheduling in an energy harvesting communication system. *IEEE Transactions on Communications*, 60(1):220–230, January 2012. ISSN 1558-0857. doi: 10.1109/TCOMM.2011.112811.100349.

Hao Yu and Michael J. Neely. A low complexity algorithm with $o(\sqrt{T})$ regret and finite constraint violations for online convex optimization with long term constraints. *arXiv:1604.02218*, 2016.

L. Zheng and L. Ratliff. Constrained upper confidence reinforcement learning. *arXiv:2001.09377*, 2020.