Online Markov Decision Processes with Time-varying Transition Probabilities and Rewards

Yingying Li [1] Aoxiao Zhong [1] Guannan Qu [1] Na Li [1]

Abstract

We consider online Markov decision process (MDP) problems where both the transition probabilities and the rewards are time-varying or even adversarially generated. We propose an online algorithm based on an online implementation of value iterations and show that its dynamic regret, i.e. its total reward compared with that of the optimal (nonstationary) policies in hindsight, is upper bounded by the total variation of the transition probabilities and the rewards. Moreover, we show that the dynamic regret of any online algorithm is lower bounded by the total variation of the transition probabilities and the rewards, indicating that the proposed algorithm is optimal up to a constant factor. Finally, we test our algorithm in a power management problem for a data center and show that our algorithm reduces energy costs and ensures quality of services (QoS) under real electricity prices and job arrival rates.

1. Introduction

Online Markov Decision Process (online MDP) problems have found many applications in sequential decision problems (Even-Dar et al., 2009; Wei et al., 2018; Bayati, 2018; Gandhi & Harchol-Balter, 2011; Lowalekar et al., 2018; Al-Sabban et al., 2013; Goldberg & Matarić, 2003; Waharte & Trigoni, 2010). In an online MDP problem, a decision maker needs to take an action based on the current state before observing the current transition probabilities and the reward (Even-Dar et al., 2009; Abbasi et al., 2013). In most applications, the transition probabilities and rewards are *time-varying*, bringing great difficulties in making effective online decisions. For example, consider the power management problem for the data center. The goal is to reduce the energy costs while ensuring the quality of services (QoS). In reality, the workload usually changes with time, inducing time-varying transitions of job queues; and electricity prices are volatile, generating time-varying operation costs.

Despite the challenges, there has been much progress in online decision making problems. When there is no need to model states or state transitions across time, the problem is actively studied, e.g. in online convex optimization (Hazan, 2016; Shalev-Shwartz, 2012; Zinkevich, 2003; Mokhtari et al., 2016; Jadbabaie et al., 2015; Li et al., 2018; Besbes et al., 2015; Hall & Willett, 2015; 2013; Zhang et al., 2017), in prediction with expert advice (Cesa-Bianchi & Lugosi, 2006; Herbster & Warmuth, 1998; Adamskiy et al., 2012; Hazan & Seshadhri, 2007; Hutter & Poland, 2005; Daniely et al., 2015), and in multi-armed bandit (Bubeck et al., 2012; Besbes et al., 2014; Luo et al., 2017; Besbes et al., 2018). There is also a limited amount of work on online MDP trying to handle the complexity brought by the state transitions (Even-Dar et al., 2009; Neu & Gómez, 2017; Neu et al., 2010; Yu & Mannor, 2009b;a; Abbasi et al., 2013; Zimin & Neu, 2013; Yu et al., 2009). However, many issues remain under-explored in online MDP problems especially when both the transition probabilities and rewards are time-varying.

Firstly, most work in online MDP only considers time-varying rewards and proposes algorithms based on the assumption of time-invariant transition probabilities, e.g. MDP-E (Even-Dar et al., 2005; 2009), Lazy FPL (Yu et al., 2009), Q-FPL (Yu et al., 2009; Yu & Mannor, 2009a), OREPS (Zimin & Neu, 2013), bandit MDP-E (Neu et al., 2010; Dick et al., 2014), FTL-MDP (Neu & Gómez, 2017), OMDP-PI (Ma et al., 2015), etc. However, in many applications, e.g. the data center power management problem as discussed above, the transition probabilities are also changing with time (Bayati, 2018; Gandhi & Harchol-Balter, 2011; Lowalekar et al., 2018; Al-Sabban et al., 2013; Goldberg & Matarić, 2003; Waharte & Trigoni, 2010). Thus, the algorithms introduced above are not applicable.

Secondly, although there has been some work on the time-varying transition probabilities (Even-Dar et al., 2005; Yu & Mannor, 2009b;a; Abbasi et al., 2013), most analysis is centered on the *static regret*, which compares the algorithm with the optimal static policy in hindsight. The optimal static policy may perform poorly when the environment is time-varying. Hence, some other nonestatic regrets have been proposed, *dynamic regret* (Zinkevich, 2003; Mokhtari et al., 2016; Jadbabaie et al., 2015; Li et al., 2018; Besbes et al., 2014; 2015), tracking regret (Herbster & Warmuth, 1998;

[1] SEAS Harvard University, Cambridge, Massachusetts, USA. Correspondence to: Yingying Li <yingyingli@g.harvard.edu>.

Real-world Sequential Decision Making workshop at ICML 2019.

Adamskiy et al., 2012), adaptive regret (Hazan & Seshadhri, 2007; Hutter & Poland, 2005) and strongly adaptive regret (Daniely et al., 2015; Zhang et al., 2017), etc. In particular, the dynamic regret considers the optimal time-varying policies as the benchmark.

In this paper, we focus on designing online algorithms under both time-varying transition probabilities and rewards, aiming at good dynamic regrets. In details, we propose Online Value Iterations, OVIs, based on an online implementation of value iterations, and show that OVIs' dynamic regret is upper bounded by the total variation of the transition probabilities and the rewards. Moreover, we show that any online algorithm's dynamic regret is lower bounded by the total variation of the transition probabilities and rewards in the worst case scenario, indicating that our algorithm almost achieves the optimal online performance. We also test our algorithm using real data in a power management problem for the data center (Bayati, 2018; Gandhi & Harchol-Balter, 2011) and show that our algorithm ensures the quality of service at a low energy cost.

1.1. Notations

For a vector $x \in \mathbb{R}^n$, $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$; given a positive weight $v \in \mathbb{R}^n$, $\|x\|_{\infty,v} = \max_{i=1,\dots,n} |x_i/v_i|$. For a matrix $L \in \mathbb{R}^{n,m}$, $\|L\|_{\max} = \max_{i,j} |L_{ij}|$ and $\|L\|_\infty = \max_i \sum_j |L_{ij}|$. For a transition probability function $\mathcal{P}$, the infinity norm is defined as $\|\mathcal{P}\|_\infty = \max_{a\in A} \|P(a)\|_\infty$. Π_X denotes projection onto a convex set X and $|A|$ denotes the cardinality of a finite set A. $\vec{e} \in \mathbb{R}^n$ denotes the all-one vector and $M \in \mathbb{R}^{n,m}$ denote the all-one matrix. $\tilde{O}(\cdot)$ denotes the variant of $O(\cdot)$ that ignores logarithmic factors.

2. Problem Formulation

Consider finite state space $S = \{1, \dots, n\}$ and finite action space A with $|A| = m$. Our *online Markov decision process (online MDP)* problem follows the following procedure. At each time step $t = 1, 2, \dots,$

- the environment generates a reward function r_t and a transition probability function $\mathcal{P}_t = (P_t(a))_{a\in A}$,[1] where $P_t(a) = (P_{ij,t}(a))_{i,j\in S}$ is the transition probability matrix under action a,
- the decision maker observes the current state s_t and takes action a_t,
- the decision maker receives the reward function r_t and the transition probability function $\mathcal{P}_t$ and collects the reward $r_t(s_t, a_t)$, [2]
- the next state s_{t+1} is generated randomly according to the transition probabilities $\mathcal{P}_t$.

A policy is a mapping $\pi : S \to \Delta_A$, where Δ_A denotes the probability simplex on action space A. We let $\pi(a|i)$ denote the probability of taking action a given the current state i under policy π. An online MDP algorithms $\mathcal{A}$ is an algorithm that generates policy π_t only based on the *history* rewards and transitions,

$$\pi_t = \mathcal{A}(r_1, \mathcal{P}_1, \dots, r_{t-1}, \mathcal{P}_{t-1}) \tag{1}$$

and the goal of an online MDP algorithm is to maximize the total *undiscounted* reward in T time steps, where T may not be revealed to the decision maker beforehand. More formally, let $\{\pi_t\}_{t=1}^T$ denote the policies generate by $\mathcal{A}$'s, then the total reward with initial state $i \in S$ is defined as follows,

$$J_T(\mathcal{A})(i) = \mathop{\mathbb{E}}_{a_t\sim\pi_t(\cdot|s_t)}[\sum_{t=1}^T r_t(s_t, a_t) \mid s_1 = i] \tag{2}$$

In this paper, we measure the performance of an online MDP algorithm by *dynamic regret*, which compares the reward achieved by the algorithm with the optimal total reward in hindsight.

Definition 1 (Dynamic regret). *The dynamic regret of an online algorithm $\mathcal{A}$ is defined as*

$$\mathcal{D}_T(\mathcal{A}) = \|J_T(\mathcal{A}) - J_T^*\|_\infty$$

where J_T^* is the optimal total reward in T time steps after all rewards and transitions are revealed. Formally,

$$J_T^*(i) = \max_{\{\pi_t\}_{t=1}^T} \mathop{\mathbb{E}}_{a_t\sim\pi_t(\cdot|s_t)}[\sum_{t=1}^T r_t(s_t, a_t)|s_1 = i], \ \forall\, i \in S$$

The regret defined above is called *dynamic* regret because the benchmark is the optimal policies which are (possibly) time-varying. This is in contrast with the *static regret* (Even-Dar et al., 2005; Yu & Mannor, 2009b;a; Abbasi et al., 2013) which compares the algorithm performance with the optimal static policy in hindsight.

Throughout the paper, we make the following ergodicity assumption, which is a common assumption in MDP literature (Puterman, 2005; Bertsekas, 2012).

Assumption 1. *At each time step t, the transition probability function $\mathcal{P}_t$ induces an ergodic MDP, i.e., for any policy π, the induced Markov chain under π and $\mathcal{P}_t$ is ergodic.*

Lastly, we remark that we do not require bounded reward function as in other online MDP literature. This allows us to consider more general changing patterns of reward functions in time-varying environments.

[1] We make the standard assumption that the adversarial environment is oblivious, i.e. r_t and $\mathcal{P}_t$ do not depend on history states $s_1, \dots, s_{t-1}$.

[2] For simplicity, we consider the observation of reward and transition function in full and leave as future work the study of bandit feedback.

3. Algorithm Development

Our online algorithm is based on value iteration for average-reward MDP. Therefore, we first review some preliminaries for average-reward MDP and the value iteration algorithm.

3.1. Average-reward MDP and Value Iteration

In an optimal average-reward MDP problem, the transition probability function and the reward function are static, i.e. $r_t = r$ and $\mathcal{P}_t = \mathcal{P}$ for all t, and the horizon is infinite. The objective is to maximize the average of the total reward:

$$\max_{\pi_1,\pi_2,\dots} \lim_{T\to+\infty} \mathbb{E}_{a_t\sim\pi_t(\cdot|s_t)} \frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t)$$

The optimal average reward is also called as the optimal *gain* in literature, and denoted as g^*. The gain $g^* \in \mathbb{R}$ together with a vector $h^* \in \mathbb{R}^n$, called as *bias* in literature, is the unique solution to the *Bellman equation system* below.

$$h^*(i) = \max_{a\in A}(r(i,a) - g^* + \sum_{j=1}^{n} P_{ij}(a)h^*(j)),\ \forall\, i \quad (3)$$

$$h^*(\tau) = 0 \quad (4)$$

where $\tau \in S$ is an arbitrary state. The constraint (4) ensures the uniqueness of the solution to (3). The optimal policy is the corresponding solution $a^*(i)$ for the Bellman equation (3,4). To find the optimal policy, value iteration (VI) (Bertsekas, 1998) picks an arbitrary state $\tau \in S$ and updates both the bias h^k and the gain g^k at each iteration k by

$$h^{k+1}(i) = \max_{a\in A}\left(r(i,a) - g^k + \sum_{j\neq\tau} P_{ij}(a)h^k(j) \right),\ \forall\, i \quad (5)$$

$$g^{k+1} = g^k + \gamma^{k+1}h^{k+1}(\tau) \quad (6)$$

with stepsizes $\gamma^{k+1} \in (0,1)$. The τ here correponds to the τ in (4). To be more compact, we define an operator VI and write the value iteration as

$$(h^{k+1}, g^{k+1}) = \mathrm{VI}(r, \mathcal{P}, h^k, g^k, \gamma^{k+1}) \quad (7)$$

3.2. Our algorithm: Online Value Iterations

We now introduce our algorithm: Online Value Iterations (OVIs). The algorithm details are provided in 1. At each time step t, OVIs performs two steps:

Step 1. The algorithm runs K iterations of VI based on the previous reward r_{t-1} and transition probabilities $\mathcal{P}_{t-1}$, and the previous estimated gain g_{t-1} and bias h_{t-1}, and obtains a new estimation of bias h_t and gain g_t. During each VI iteration, we project g_t to a finite interval X_t to avoid the value exploding to infinity.

Step 2. The algorithm computes an estimated Q function based on current estimation of bias h_t and gain g_t and the previous reward r_{t-1} and transition probabilities $\mathcal{P}_{t-1}$. Then, we take the action that maximizes the Q function given current state s_t.

Algorithm 1 Online Value Iterations (OVIs)

```
**Inputs:** Initial environment: $(r_0, \mathcal{P}_0)$. Initial values $g_0, h_0$. Stepsize $\gamma_t^k$. Number of iterations $K$.
**for** $t = 1 : T$ **do**
    *Step 1: update bias $h_t$ and gain $g_t$ by $K$ iterations of VI*
    Set $h_t^0 = h_{t-1}$, and $g_t^0 = g_{t-1}$.
    **for** $k = 1 : K$ **do**
        Update $h_t^k$ and $g_t^k$ by VI with projection
            $(h_t^k, \bar{g}_t^k) = \mathrm{VI}(r_{t-1}, \mathcal{P}_{t-1}, h_t^{k-1}, g_t^{k-1}, \gamma_t^k)$
            $g_t^k = \Pi_{X_t}(\bar{g}_t^k)$
        where $X_t = [-\|r_{t-1}\|_{\max}, \|r_{t-1}\|_{\max}]$
    **end for**
    Set $h_t = h_t^K$, $g_t = g_t^K$.
    *Step 2: take action $a_t$ to maximize the estimated $Q$ function*
        $a_t \in \arg\max_a Q_t(s_t, a)$
    where the estimated $Q_t$ function is defined as $Q_t(s,a) = r_{t-1}(s,a) - g_t + \sum_{j=1}^n P_{ij,t-1}(a)h_t(j)$
**end for**
```

The computation complexity at each t is $O(n^2mK)$. In Section 4.2, we will show that when K is as small as $\tilde{O}(1)$, OVIs can already achieve a near-optimal dynamic regret guarantee. Therefore, OVIs has polynomial complexity $\tilde{O}(n^2m)$ which is more efficient computationally than many existing online MDP algorithms for online MDP with time-varying transition probabilities, e.g. solving an average-reward MDP problem completely as in (Yu & Mannor, 2009b) or computing the stationary distributions of all policies as in (Abbasi et al., 2013). The space complexity is $O(n)$ which is more efficient than maintaining a weight for each policy as in (Abbasi et al., 2013).

4. Regret Analysis

4.1. Preliminaries: Some Definitions

Before providing the regret bound, we introduce useful concepts in MDP and the constant factors in the regret bound and the choice of stepsize will depend on these concepts. The first is the condition number $\kappa(P)$ of Markov chains which captures the sensitivity of the stationary distribution to a change in the transition probabilities.

Definition 2 (Condition number of Markov Chain (Meyer, 1980)). *Consider a markov chain with transition probability matrix P. The* condition number *of P is defined as the maximum norm of the Drazin inverse of $I - P$, that is,*

$$\kappa(P) := \|(I-P)^{\#}\|_{\max} \tag{8}$$

where $A^{\#}$ denotes the Drazin inverse of a matrix A (Meyer, 1975).

Next, we introduce the recurrent coefficient of state τ in MDP: $\rho_\tau(\mathcal{P})$ for any $\tau \in S$. Roughly speaking, $\rho_\tau(\mathcal{P})$ represents how likely the state τ is going to be visited in n time steps with any initial state. The larger the value is, the more recurrent the state is.

Definition 3 (Recurrent coefficient of MDP (Section 7.4 in (Bertsekas, 2005)))**.** *Consider an MDP with transition probabilities $\mathcal{P} = (P(a))_{a\in A}$ and $P(a) = (P_{ij}(a))_{i,j\in S}$. Define $Y_{i\tau}(\pi)$ as the first-passage time of τ from i following stationary policy π. The* recurrent coefficient *of τ in MDP with transition function $\mathcal{P}$ is defined as the probability that the first-passage time is smaller than the number of the states n, that is,*

$$\rho_\tau(\mathcal{P}) := \min_{i,\pi} \mathbb{P}(Y_{i\tau}(\pi) \le n) \tag{9}$$

It can be shown that for any ergodic MDP and any state τ, there is a positive lower bound on the recurrent coefficients (Bertsekas, 1998; 2005).

Further, the VI update in (5) is contractive with respect to h^k ((Bertsekas, 2012) and Proposition 1 in Appendix B): if we let $F : \mathbb{R}^n \to \mathbb{R}^n$ be $F_i(h) := \max_{a\in A}\left(r(i,a) - g^k + \sum_{j\neq\tau} P_{ij}(a)h(j)\right)$ for any i, then there exists a weighted infinity norm $||\cdot||_{\infty,v}$ such that $||F(h) - F(h')||_{\infty,v} \le \alpha||h-h'||_{\infty,v}$. Moreover, the contraction factor $\alpha = \alpha(\mathcal{P})$ only depends on $\mathcal{P}$ and does not depend on reward r.

To obtain a meaningful regret bound, we consider a class of MDP $\mathfrak{M}(\kappa,\rho,\alpha)$ whose condition number, recurrent coefficient and contraction factor satisfy the following, $\forall (r,\mathcal{P}) \in \mathfrak{M}(\kappa,\rho,\alpha)$,

$$\kappa(P(\pi^*)) \le \kappa, \quad \rho_\tau(\mathcal{P}) \ge \rho, \quad \alpha(\mathcal{P}) \le \alpha$$

where π^* denotes the optimal average-reward policy of MDP $(r,\mathcal{P})$.

4.2. Regret Upper Bound of OVIs

Theorem 1 (Regret upper bound of OVIs)**.** *For any state $\tau \in S$ and any MDP sequences $\{(r_t,\mathcal{P}_t)\}_{t=1}^T \subseteq \mathfrak{M}(\kappa,\rho,\alpha)$, there exists a positive scalar $\bar{\gamma}$ and a function $\delta(\rho,\alpha)$, such that for any $0 < \gamma \le \bar{\gamma}$, and any $K \ge \delta(\rho,\alpha)\log n$, if stepsizes are $\gamma_t^1 = \gamma, \gamma_t^2 = \ldots = \gamma_t^K = 0$ for all t, and initial values g_0 h_0 satisfy the Bellman equation system of $(r_0, \mathcal{P}_0)$, then the dynamic regret of OVIs can be bounded by the total variation of reward functions and the transition probability functions, i.e.*

$$\mathcal{D}_T(OVIs) = O(\sum_{t=0}^T \|r_t - r_{t+1}\|_{\max}) + \sum_{t=0}^T R_t\|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty)$$

where $r_{T+1} = 0$, $R_t = \max(\|r_t\|_{\max}, \|r_{t-1}\|_{\max})$, $\mathcal{P}_{T+1} = \mathcal{P}_T$, and the factor hidden in the $O(\cdot)$ only depends on $\kappa,\rho,\alpha,n,\gamma$.

Next, we provide some discussion on the regret bounds.

Dependence on the variation of the environment. Theorem 1 shows that the dynamic regret of OVIs relies on the total variation of the environment which consists of two parts: i) $\sum_{t=0}^T \|r_t - r_{t+1}\|_{\max}$ and ii) $\sum_{t=0}^T R_t\|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty$. Part i) is common in literature when the problem considered does not have a Markov processe structure, e.g. (Besbes et al., 2014; 2015). Part ii) is new but intuitive: the variation of the transition probabilities should also affect the algorithm performance. Interestingly, part ii) not only contains the variation of the transition probabilities but also has a factor of the maximal reward values. This is intuitive because if the rewards of all states and actions are small, even under drastic changes in the transition probabilities, the impact on the total reward collected is still small.

Static environment. When reward functions and transition probability functions are time-invariant, OVIs reduces to VI and the regret of OVIs is bounded by a constant which does not depend on the horizon T, indicating that OVIs quickly converges to the optimal policy. In fact, in this case our regret bound holds under a relaxed assumption, cf. Appendix A.

Choice of stepsize γ_t^k. Theorem 1 uses a special stepsize rule for the K VI iterations, where $\gamma_t^1 = \gamma > 0$, and the rest $K-1$ stepsizes are set to be 0. The special stepsize rule is for the simplicity of the analysis. Our proof can be extended to more general stepsize rules such as constant stepsizes, diminishing stepsizes (Bertsekas, 1998), and diminishing stepsizes with restarts (Besbes et al., 2015), etc

Computation complexity and choice of K. In Theorem 1, $K = O(\log n)$, so the computational complexity at each time step is $O(n^2 m \log n)$. In practice, there is a tradeoff in K. When K is too small, the algorithm may not adapt to the new environment quickly enough. When K is too large, the algorithm becomes too greedy and neglects the history before time $t-1$, which might also harm the performance.

4.3. A Fundamental Regret Low Bound

Theorem 2 (Fundamental lower bound)**.** *Consider an online MDP problem with $n, m \ge 2$. For any online MDP algorithm $\mathcal{A}$ with any computation complexity, any reward variation budget $L_r > 0$ and any transition probability variation budget $L_P > 0$, there exists a sequence of rewards and* *transitions* $(r_t, \mathcal{P}_t)_{t=1}^T$ *satisfying* $\sum_{t=0}^T \|r_t - r_{t+1}\|_{\max} \leq L_r$ *and* $\sum_{t=0}^T R_t \|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty \leq L_P$*, such that the dynamic regret is lower bounded by the transition probability variation budget and the reward variation budget, that is,*

$$\mathcal{D}_T(\mathcal{A}) = \Omega(L_P + L_r).$$

Theorem 2 shows that for any online algorithm, the dynamic regret is at least $\Omega(\sum_{t=0}^T \|r_t - r_{t+1}\|_{\max} + \sum_{t=0}^T R_t \|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty)$ in the worst case scenario. This matches the upper bound of OVIs' regret up to a constant, indicating that our algorithm achieves the near-optimal dynamic regret bound. Theorem 2 also indicates that when the environment variation is $\Omega(T)$, *no* online algorithm can achieve $o(T)$ dynamic regret. Similar results have been established for other online problems without state transitions (Besbes et al., 2014; 2015; Li et al., 2018).

5. Experiments

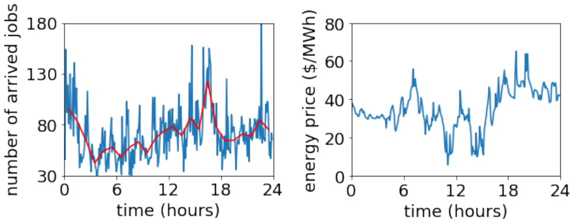

Figure 1. Left is the traffic trace: blue is the real traffic and red is the average of each hour. Right is the electricity price.

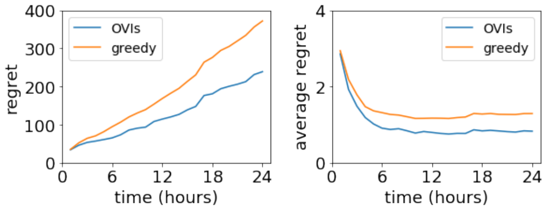

Figure 2. Left and Right respectively plot the dynamic regret and the average dynamic regret of OVIs and the greedy On/Off

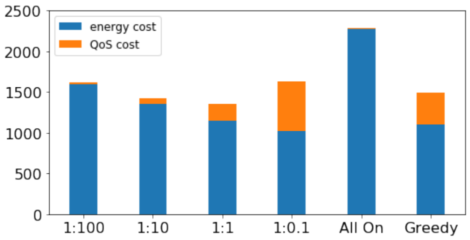

Figure 3. Cost breakdown

We test our algorithm on a data center power management problem. Overall, we follow the model in (Bayati, 2018) and (Gandhi & Harchol-Balter, 2011) where we consider a data center with heterogeneous servers: high servers with high service rate and power consumption and low servers with low service rate and power consumption. The goal is to dynamically switch on/off servers in order to reduce the energy cost while still ensuring the Quality of Service (QoS) such as a short waiting queue and few jobs lost. Switching off servers effectively reduces energy cost but comes with a risk of degrading QoS because it takes time to set up a server (Barroso & Hölzle, 2007). The electricity prices and job arrival rates are constantly changing in reality, motivating the use of our online algorithm. In our experiments, we use electricity prices from CAISO (California Independent Service Operator) and traffic trace from a Google data center (Reiss et al., 2011) for the job rates (Fig.1). For more details of the problem, we refer the reader to Appendix F. When applying our OVIs algorithm, the stepsize rule follows Theorem 1 and $\gamma = 0.2$, $K = 7$. We simulate different cases with different weights (1:100, 1:10, 1:1, 1:0.1) between the energy cost and the QoS cost in the cost/reward function to see how our OVIs algorithm balances the two costs.

Other policies for comparison: Greedy On/Off, All On. Both policies are from (Gandhi & Harchol-Balter, 2011). The Greedy On/Off switches off servers when they become idle and switches on servers when jobs arrive. When both types of servers are available, the policy switches on high servers in prior. The All On policy keeps all servers on all the time.

Observation from the plots. Figure 2 plots the dynamic regret and the average regret over time of our algorithm OVIs and the greedy policy. We don't plot the regret of the All On policy here because its regret is very large. We can see that OVIs achieves much smaller dynamic regret than Greedy On/Off. Figure 3 shows the cost breakdown of different policies. Firstly, we can see that our OVIs algorithm respond to the different weights: the larger the weight on the QoS cost, the lower QoS cost and the higher energy cost, which is consistent with our intuitions. On the other hand, the All On policy will generate a large energy cost and a negligible service cost; whereas the greedy algorithm could induce a relatively high QoS cost at a low energy cost. An interesting observation is that when we set a very high weight on the QoS cost, e.g., the case of 1:100, there is almost no QoS cost, similar to the All On policy, but the energy cost is much lower than the All On policy.

6. Conclusion

In this paper, we consider an online MDP problem with time-varying transition probabilities and rewards. We propose a computationally efficient algorithm OVIs and provide a dynamic regret bound which depends on the total variation of the transition probabilities and the rewards. Moreover, we show that the regret upper bound matches the fundamental lower bound of any online algorithm up to a constant, demonstrating that our algorithm is near optimal. Future work includes relaxing the ergodic assumption, considering the problem under noisy feedback and bandit feedback, and online MDP problems with predictions.

Acknowledgement

This work was supported by NSF Career 1553407, ARPA-E NODES, AFOSR YIP and ONR YIP programs.

References

Abbasi, Y., Bartlett, P. L., Kanade, V., Seldin, Y., and Szepesvári, C. Online learning in markov decision processes with adversarially chosen transition probability distributions. In *Advances in neural information processing systems*, pp. 2508–2516, 2013.

Adamskiy, D., Warmuth, M. K., and Koolen, W. M. Putting bayes to sleep. In *Advances in neural information processing systems*, pp. 135–143, 2012.

Al-Sabban, W. H., Gonzalez, L. F., and Smith, R. N. Wind-energy based path planning for unmanned aerial vehicles using markov decision processes. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 784–789. IEEE, 2013.

Barroso, L. A. and Hölzle, U. The case for energy-proportional computing. *Computer*, (12):33–37, 2007.

Bayati, M. Power management policy for heterogeneous data center based on histogram and discrete-time mdp. *Electronic Notes in Theoretical Computer Science*, 337: 5–22, 2018.

Bertsekas, D. P. A new value iteration method for the average cost dynamic programming problem. *SIAM journal on control and optimization*, 36(2):742–759, 1998.

Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 2005.

Bertsekas, D. P. *Dynamic programming and optimal control*, volume 2. Athena scientific Belmont, MA, 2012.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pp. 199–207, 2014.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Besbes, O., Gur, Y., and Zeevi, A. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *Available at SSRN 2436629*, 2018.

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

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 1405–1411, 2015.

Dick, T., Gyorgy, A., and Szepesvari, C. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pp. 512–520, 2014.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Experts in a markov decision process. In *Advances in neural information processing systems*, pp. 401–408, 2005.

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

Gandhi, A. and Harchol-Balter, M. How data center size impacts the effectiveness of dynamic power management. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pp. 1164–1169. IEEE, 2011.

Goldberg, D. and Matarić, M. J. Maximizing reward in a non-stationary mobile robot environment. *Autonomous Agents and Multi-Agent Systems*, 6(3):287–316, 2003.

Hall, E. C. and Willett, R. M. Dynamical models and tracking regret in online convex programming. *arXiv preprint arXiv:1301.1254*, 2013.

Hall, E. C. and Willett, R. M. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.

Hazan, E. *Introduction to Online Convex Optimization*. Foundations and Trends(r) in Optimization Series. Now Publishers, 2016. ISBN 9781680831702.

Hazan, E. and Seshadhri, C. Adaptive algorithms for online decision problems. In *Electronic colloquium on computational complexity (ECCC)*, volume 14, 2007.

Herbster, M. and Warmuth, M. K. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

Hutter, M. and Poland, J. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6(Apr):639–660, 2005.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pp. 398–406, 2015.

Li, Y., Qu, G., and Li, N. Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit. *arXiv preprint arXiv:1801.07780*, 2018.

Lowalekar, M., Varakantham, P., and Jaillet, P. Online spatio-temporal matching in stochastic and dynamic domains. *Artificial Intelligence*, 261:71–112, 2018.

Luo, H., Wei, C.-Y., Agarwal, A., and Langford, J. Efficient contextual bandits in non-stationary worlds. *arXiv preprint arXiv:1708.01799*, 2017.

Ma, Y., Zhang, H., and Sugiyama, M. Online markov decision processes with policy iteration. *arXiv preprint arXiv:1510.04454*, 2015.

Meyer, C. D. Sensitivity of the stationary distribution of a markov chain. *SIAM Journal on Matrix Analysis and Applications*, 15(3):715–728, 1994.

Meyer, Jr, C. D. The role of the group generalized inverse in the theory of finite markov chains. *Siam Review*, 17(3): 443–464, 1975.

Meyer, Jr, C. D. The condition of a finite markov chain and perturbation bounds for the limiting probabilities. *SIAM Journal on Algebraic Discrete Methods*, 1(3):273–283, 1980.

Mokhtari, A., Shahrampour, S., Jadbabaie, A., and Ribeiro, A. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. *arXiv preprint arXiv:1603.04954*, 2016.

Neu, G. and Gómez, V. Fast rates for online learning in linearly solvable markov decision processes. *arXiv preprint arXiv:1702.06341*, 2017.

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

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming (wiley series in probability and statistics). 2005.

Reiss, C., Wilkes, J., and Hellerstein, J. L. Google cluster-usage traces: format+ schema. *Google Inc., White Paper*, pp. 1–14, 2011.

Shalev-Shwartz, S. *Online Learning and Online Convex Optimization*. Foundations and Trends(r) in Machine Learning. Now Publishers, 2012. ISBN 9781601985460.

Waharte, S. and Trigoni, N. Supporting search and rescue operations with uavs. In *Emerging Security Technologies (EST), 2010 International Conference on*, pp. 142–147. IEEE, 2010.

Wei, X., Yu, H., and Neely, M. J. Online learning in weakly coupled markov decision processes: A convergence time study. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):12, 2018.

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

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

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

Zhang, L., Yang, T., Jin, R., and Zhou, Z.-H. Strongly adaptive regret implies optimally dynamic regret. *arXiv preprint arXiv:1701.07570*, 2017.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pp. 1583–1591, 2013.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 928–936, 2003.

Appendices

A. Additional results on online MDP with static transition probability function

In this section, we provide some additional results considering time-invariant transition probabilities.

When the transition probability function is not time-varying and is known to the decision maker, the dynamic regret bound in Theorem 1 holds under the following weaker assumption.

Assumption 2. *The MDP is unichain, and there exists a state that is recurrent under all policies.*

It can be verified that Assumption 1 induces Assumption 2. We provide the dynamic regret bound for the case of static transition probability function in Corollary 1.

Corollary 1. *Under Assumption 2. If the initial values h_0 and g_0 satisfy the Bellman equation system under reward function r_0. There exists a positive scalar $\bar{\gamma}$, such that for any constant stepsize $0 < \gamma \leq \bar{\gamma}$ and any $K \geq 1$, the dynamic regret of OVIs can be bounded by the total variation of reward functions, i.e.*

$$\mathcal{D}_T(OVIs) = O(\sum_{t=0}^{T} \|r_t - r_{t+1}\|_{\max})$$

where $r_{T+1} = 0$.

Corollary 1 indicates that in the simpler setting where $\mathcal{P}$ is not time-varying, OVIs achieves a similar regret bound as in Theorem 1 under a weaker assumption. Further, Corollary 1 only requires $K \geq 1$, which means the computational complexity can be $O(n^2 m)$ (when $K = 1$), which is more efficient than that of the time-varying $\mathcal{P}$ case in Theorem 1.

B. Preliminaries: some useful perturbation results for MDP and some helpful notations

In this section, we will list some useful perturbation results for the proofs and define some new notations for ease of reference.

Perturbation results for the Bellman operators. We first introduce some standard notations that succinctly represent the Bellman equation system (3). We define the Bellman operator $B^* : \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times n\times m} \times \mathbb{R}^n \to \mathbb{R}^n$ maps the reward matrix r, the transition probability $\mathcal{P}$ and the bias vector h to a vector in $\mathbb{R}^n$, given by,

$$B^*(r, \mathcal{P}, h)(i) = \max_a r(i, a) + \sum_{j=1}^{n} P_{ij}(a)h(j), \quad \forall\, i \in S. \tag{10}$$

With the notation B^*, the Bellman equation (3) can be written in a compact way,

$$h = B^*(r - gM, \mathcal{P}, h) = \max_{\pi} B(r - gM, \mathcal{P}, h, \pi) \tag{11}$$

where M is an all-one matrix in $\mathbb{R}^{n\times m}$.

Similarly, we define the Bellman operator B under a policy π as the following,

$$B(r, \mathcal{P}, h, \pi)(i) = r(i, \pi) + \sum_{j=1}^{n} P_{ij}(\pi)h(j), \quad \forall\, i \in S \tag{12}$$

where $r(i, \pi) = \mathbb{E}_{a\sim\pi(\cdot|i)}\, r(i, a)$ and $P_{ij}(\pi) = \mathbb{E}_{a\sim\pi(\cdot|i)}\, p_{ij}(a)$.

We will frequently use the following Lemma, which bounds the outputs of the operators B and B^* under perturbations to their inputs.

Lemma 1. *The following results hold.*

(a) $\|B(r, \mathcal{P}, h, \pi) - B(r', \mathcal{P}', h', \pi)\|_\infty \leq \|r - r'\|_{\max} + \|h - h'\|_\infty + \min(\|h\|_\infty, \|h'\|_\infty)\|\mathcal{P} - \mathcal{P}'\|_\infty$

(b) $\|B^*(r, \mathcal{P}, h) - B^*(r, \mathcal{P}, h')\|_\infty \leq \|h - h'\|_\infty$

Perturbation results on the solutions to the Bellman equation systems. Firstly, we define some notations to represent the solutions to the Bellman equation system.

Definition 4 (Solutions to the Bellman equations)**.** *Let* $(h^*(r, \mathcal{P}), g^*(r, \mathcal{P}))$ *be the solution to the Bellman Equation systems* (3) *and* (4) *under transition probability function* $\mathcal{P}$ *and reward function* r*.*

Next, we provide a lemma that bounds the change in the solutions in terms of change in $\mathcal{P}$ and r. The proof of Lemma 2 is postponed to Appendix E.2.

Lemma 2. *Let* $(r, \mathcal{P}), (r', \mathcal{P}') \in \mathfrak{M}(\kappa, \rho, \alpha)$*. Then,*

$$|g^*(r, \mathcal{P}) - g^*(r', \mathcal{P}')| \leq \|r - r'\|_{\max} + n\kappa \max(\|r\|_{\max}, \|r'\|_{\max})\|\mathcal{P} - \mathcal{P}'\|_\infty$$
$$\|h^*(r, \mathcal{P}) - h^*({}'r, \mathcal{P}')\|_\infty \leq 2\eta\|r - r'\|_{\max} + (n\eta\kappa + 2\eta^2) \max(\|r\|_{\max}, \|r'\|_{\max})\|\mathcal{P} - \mathcal{P}'\|_\infty$$

where $\eta = n/\rho$*.*

Recall that h_t^*, g_t^* is the solution to the Bellman equation system (3) and (4) under transition probability function $\mathcal{P}_t$ and reward function r_t. Therefore, an immediate corollary of Lemma 2 is that $\|h_t^* - h_{t-1}\|_\infty$ and $|g_t^* - g_{t-1}^*|$ can be bounded in terms of the variation in environments, $\|\mathcal{P}_t - \mathcal{P}_{t-1}\|_\infty$ and $\|r_t - r_{t-1}\|_{\max}$. This is formally stated by the following Corollary.

Corollary 2. *Under the conditions in Theorem 1, for* $t = 1, \ldots, T + 1$*,*

$$|g_t^* - g_{t-1}^*| \leq \|r_t - r_{t-1}\|_{\max} + n\kappa R_{t-1}\|\mathcal{P}_t - \mathcal{P}_{t-1}\|_\infty$$
$$\|h_t^* - h_{t-1}^*\|_\infty \leq 2\eta\|r_t - r_{t-1}\|_{\max} + (n\eta\kappa + 2\eta^2)R_{t-1}\|\mathcal{P}_t - \mathcal{P}_{t-1}\|_\infty$$

where $\mathcal{P}_{T+1} = \mathcal{P}_T$*,* $\eta = n/\rho$ *and* $R_t = \max(\|r_t\|_{\max}, \|r_{t+1}\|_{\max})$*.*

Modified Bellman equation: notations, contraction, and perturbation. VI can be viewed as the fixed point iteration of the following modified but equivalent Bellman equation system with stepsize γ^{k+1} when updating g:

$$h(i) = \max_{a \in A} \left(r(i, a) - g + \sum_{j \neq \tau} P_{ij}(a)h(j) \right), \ \forall\, i \tag{13}$$

$$g = \max_{a \in A} \left(r(\tau, a) + \sum_{j \neq \tau} P_{\tau j}(a)h(j) \right) \tag{14}$$

When stepsizes γ^k are small enough, h^k and g^k converge to the h^* and g^* geometrically (Bertsekas, 1998). The main reason for a small stepsize is shortly discussed below. Consider a fixed g^k, the update of h^k by (5) establishes a contraction property with respect to some weighted infinity norm, which is formally stated below.

Proposition 1 (Contraction of modified Bellman operator F)**.** *[(Bertsekas, 2012)] Let the operator* $F : \mathbb{R}^n \to \mathbb{R}^n$ *denote the update of* h^k *by* (5)*, that is*

$$F_i(h) = \max_{a \in A} \left(r(i, a) - g^k + \sum_{j \neq \tau} P_{ij}(a)h^k(j) \right), \quad \forall\, i$$

There exists a weighted infinity norm $\|\cdot\|_{\infty, v}$ *with a positive weight* v *and a contraction factor* α *such that for any* $h, h' \in \mathbb{R}^n$*,*

$$\|F(h) - F(h')\|_{\infty, v} \leq \alpha\|h - h'\|_{\infty, v}$$

The contraction factor $\alpha = \alpha(\mathcal{P})$ *only depends on* $\mathcal{P}$ *and does not depend on reward* r*.*

It can be shown that under a small enough stepsize γ^k, the update of (g^k, h^k) by (5) and (6) also establishes a contraction property. As a result, VI converges geometrically. For more details, we refer the reader to (Bertsekas, 2012; 1998).

Finally, we introduce a new notation for the solution to (modified) Bellman equation (13).

Definition 5 (Solution to the modified Bellman equations). *Let $\tilde{h}^*(r, g, \mathcal{P}) \in \mathbb{R}^n$ be the unique solution to the (modified) Bellman equation* (13) *with fixed gain g under reward r and probability transition function $\mathcal{P}$.*

By the equivalence of the two Bellman equation systems, we have

$$\tilde{h}^*(r, g^*(r, \mathcal{P}), \mathcal{P}) = h^*(r, \mathcal{P}) \tag{15}$$

In the following Lemma, we bound the change of $\tilde{h}^*(r, g, \mathcal{P})$ in terms of the change of $r, g, \mathcal{P}$. The proof of Lemma 3 is postponed to Appendix E.3.

Lemma 3 (A perturbation result for $\tilde{h}^*$). *For any $(r, \mathcal{P}), (r', \mathcal{P}') \in \mathfrak{M}(\kappa, \rho, \alpha)$ and any g, g',*

$$\|\tilde{h}^*(r, g, \mathcal{P}) - \tilde{h}^*(r', g', \mathcal{P}')\|_\infty \leq \eta^2\big[\max(|g|, |g'|) + \max(\|r\|_{\max}, \|r'\|_{\max})\big]\|\mathcal{P} - \mathcal{P}'\|_\infty + (\|r - r'\|_{\max} + |g - g'|)\eta$$

where $\eta = n/\rho$.

C. Proof of Theorem 1

It is usually challenging to directly compare J_T^* and $J_T(OVIs)$. In this paper, we construct an auxiliary variable W_1 to serve as the middle ground and aim to bound $\|J_T^* - W_1\|_\infty$ and $\|W_1 - J_T(OVIs)\|_\infty$ by the variation of environment. Then, the dynamic regret bound follows by using the triangle inequality:

$$\mathcal{D}_T(OVIs) = \|J_T(OVIs) - J_T^*\|_\infty \leq \|J_T(OVIs) - W_1\|_\infty + \|W_1 - J_T^*\|_\infty$$

The proof relies on dynamic programming, perturbation results and contraction properties introduced in Appendix B.

In rest of this subsection, we will first define W_1, and then bound $\|J_T^* - W_1\|_\infty$ and $\|W_1 - J_T(OVIs)\|_\infty$ by the variation of the environment.

Definition of W_1: Consider a *static* average-reward MDP problem $(r, \mathcal{P})$ with reward function $r = r_t$ and transition probability function $\mathcal{P} = \mathcal{P}_t$ for all time steps. We denote the solution to the corresponding Bellman equation system (3) and (4) as g_t^* and h_t^*, and we denote the optimal policy for this static MDP as π_t^*.[3] The Bellman equation can be written as

$$h_t^* + g_t^* \vec{e} = B^*(r_t, \mathcal{P}_t, h_t^*) \tag{16}$$

We define $W_1 \in \mathbb{R}^n$ as follows:

$$W_1(i) := h_1^*(i) + \sum_{t=1}^T g_t^*, \quad \forall\, i \in S.$$

Bound $\|J_T^* - W_1\|_\infty$. In the following, we show that $\|J_T^* - W_1\|_\infty = O(\sum_{t=1}^T \|r_t - r_{t+1}\|_{\max}) + O(\sum_{t=1}^T R_t\|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty)$. The proof relies on dynamic programming and perturbation results of B^*.

First of all, notice that the optimal total cost J_T^* can be computed by dynamic programming. Let V_k^* be the optimal reward to go from $t = k$. By dynamic programming,

$$V_k^*(i) = \max_a (r_k(i, a) + \sum_{j=1}^n P_{ij,k}(a) V_{k+1}^*(j))$$

which, using the notations in (11), can be compactly written as,

$$V_k^* = B^*(r_k, \mathcal{P}_k, V_{k+1}^*) \tag{17}$$

and $V_1^* = J_T^*,\ V_{T+1} = 0$.

In addition to W_1, we define an auxiliary variable W_k for $k \geq 1$ as

$$W_k := h_k^* + \sum_{t=k}^T g_t^* \vec{e} = B^*(r_k \mathcal{P}_k, W_{k+1} - h_{k+1}^* + h_k^*) \tag{18}$$

[3] Policy $\{\pi_t^*\}_{t=1}^T$ is *not* the optimal policies for the T-step time-varying MDP problem in hindsight.

where the second equality is by Bellman equation 16. Let $W_{T+1} = h^*_{T+1} = 0$.

In order to upper bound $\|J^*_T - W_1\|_\infty = \|V^*_1 - W_1\|_\infty$, we use backward induction to bound $\|V^*_k - W_k\|_\infty$ recursively. Specifically, we show that, for any $1 \leq k \leq T$,

$$\|V^*_k - W_k\|_\infty \leq \sum_{t=k}^{T} \|h^*_t - h^*_{t+1}\|_\infty \tag{19}$$

When $k = T$, (19) is true because by (17), (16) and Lemma 1 (b),

$$\|V^*_T - W_T\|_\infty = \|B^*(r_T, P_T, 0) - B^*(r_T, P_T, h^*_T)\|_\infty \leq \|h^*_T\|_\infty.$$

Assume (19) is true for $k+1$, Then for k, using the dynamic programming equation (17) and perturbation results of B^*, we have

$$\begin{aligned}\|V^*_k - W_k\|_\infty &= \|B^*(r_k, \mathcal{P}_k, V^*_{k+1}) - B^*(r_k, \mathcal{P}_k, W_{k+1} + h^*_k - h^*_{k+1})\|_\infty \\ &\leq \|V^*_{k+1} - W_{k+1}\|_\infty + \|h^*_k - h^*_{k+1}\|_\infty \\ &\leq \sum_{t=k}^{T} \|h^*_t - h^*_{t+1}\|_\infty\end{aligned}$$

which shows (19) is true for k, and hence the backward induction is complete.

Lastly, by the bound on $\|h^*_t - h^*_{t-1}\|_\infty$ (Corollary 2 in Appendix B), we have

$$\|J^*_T - W_1\|_\infty = \|V^*_1 - W_1\|_\infty \leq \sum_{t=1}^{T} \|h^*_t - h^*_{t+1}\|_\infty = O(\sum_{t=1}^{T}(\|r_t - r_{t+1}\|_{\max} + R_t\|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty))$$

which completes the bound of $\|J^*_T - W_1\|$.

Bound $\|J_T(OVIs) - W_1\|_\infty$**.** In the following, we will show that $\|J_T(OVIs) - W_1\|_\infty = O(\sum_{t=0}^{T} \|r_t - r_{t+1}\|_{\max}) + O(\sum_{t=0}^{T} R_t\|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty)$. The proof is by dynamic programming, perturbation results and contraction properties of the Bellman operators introduced in Appendix B.

Firstly, we define V_k as the reward-to-go of OVIs under a shifted reward function $r_t - g^*_t M$ at each time step t, where M denote the all-one matrix of the same size as r_t. By dynamic programming,

$$V_k = B(r_k - g^*_k M, \mathcal{P}_k, V_{k+1}, \pi_k)$$

As a result, $V_1 + \sum_{t=1}^{T} g^*_t \vec{e} = J_T(OVIs)$ and $\|W_1 - J_T(OVIs)\|_\infty = \|h^*_1 - V_1\|_\infty$. So it suffices to bound $\|h^*_1 - V_1\|_\infty$ in the following. The following lemma establishes a recursive relation between $\|V_k - h^*_k\|_\infty$ and $\|V_{k+1} - h^*_{k+1}\|_\infty$, which yields a bound by the telescoping summation.

Lemma 4. *For* $1 \leq k \leq T$*,*

$$\begin{aligned}\|V_k - h^*_k\|_\infty \leq \|V_{k+1} - h^*_{k+1}\|_\infty &+ (4\eta + 2)\|r_k - r_{k-1}\|_{\max} + (2\eta\kappa n + 4\eta^2 + 4\eta)R_{k-1}\|\mathcal{P}_k - \mathcal{P}_{k-1}\|_\infty \\ &+ 2\eta\|r_{k+1} - r_k\|_{\max} + (\eta\kappa n + 2\eta^2)R_k\|\mathcal{P}_{k+1} - \mathcal{P}_k\|_\infty + 2\|h_k - h^*_{k-1}\|_\infty\end{aligned}$$

where $V_{T+1} = 0$*,* $h^*_{T+1} = 0$*,* h_k *is produced by our algorithm 1 in Line 8.*

The proof is deferred to Appendix C.1.

Summing over k the both sides of the recursive relation provided in Lemma 4, we have

$$\|J_T(OVIs) - W_1\|_\infty \leq (6\eta + 2)\sum_{k=0}^{T} \|r_k - r_{k+1}\|_{\max} + (3\eta\kappa n + 6\eta^2 + 4\eta)\sum_{k=0}^{T} R_k\|\mathcal{P}_{k+1} - \mathcal{P}_k\|_\infty + 2\sum_{k=1}^{T} \|h_k - h^*_{k-1}\|_\infty$$

$$= O(\sum_{t=0}^{T} \|r_t - r_{t+1}\|_{\max}) + O(\sum_{t=0}^{T} R_t \|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty) + 2\sum_{k=1}^{T} \|h_k - h^*_{k-1}\|_\infty$$

The proof is completed if $\sum_{k=1}^T \|h_k - h^*_{k-1}\|_\infty = O(\sum_{t=0}^T \|r_t - r_{t+1}\|_{\max}) + O(\sum_{t=0}^T R_t \|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty)$, which is indeed the case and is formally stated in the following lemma and proved in the next section.

Lemma 5.

$$\sum_{k=1}^{T} \|h_k - h^*_{k-1}\|_\infty = O(\sum_{t=0}^{T} \|r_t - r_{t+1}\|_{\max}) + O(\sum_{t=0}^{T} R_t \|\mathcal{P}_t - \mathcal{P}_{t+1}\|_\infty)$$

The proof is deferred to Appendix C.2.

C.1. Proof of Lemma 4

The proof relies on the perturbation results in Appendix B.

Proof. By triangle inequality and Bellman operator perturbation results (Lemma 1 (a)), we have

$$\begin{aligned}
\|V_k - h^*_k\|_\infty &= \|B(r_k - g^*_k M, \mathcal{P}_k, V_{k+1}, \pi_k) - h^*_k\|_\infty \\
&\le \|B(r_k - g^*_k M, \mathcal{P}_k, V_{k+1}, \pi_k) - B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k)\|_\infty + \|B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k) - h^*_k\|_\infty \\
&\le \|V_{k+1} - h^*_{k+1}\|_\infty + \|h^*_k - h^*_{k+1}\|_\infty + \|B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k) - h^*_k\|_\infty \qquad (20)
\end{aligned}$$

Now, it suffices bound $\|B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k) - h^*_k\|_\infty$. Recall the definition of h^*_k and g^*_k,

$$h^*_k(i) = \max_\pi B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi)(i) = B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi^*_k)(i)$$

where π^*_k is the policy that achieves the maximization in the definition of $h^*_k(i)$. The above equation implies that

$$h^*_k(i) - B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k)(i) \ge 0$$

So to bound $\|B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k) - h^*_k\|_\infty$, we only need to provide an upper bound of $h^*_k(i) - B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k)(i)$ for all i. Using definition of $h^*_k(i)$, we have,

$$\begin{aligned}
&\|B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k) - h^*_k\|_\infty \\
&= \max_{i\in S} \left[h^*_k(i) - B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k)(i)\right] \\
&= \max_{i\in S} \Big[B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi^*_k)(i) - B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi_k)(i) \\
&\qquad + B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi_k)(i) - B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k)(i)\Big] \\
&\overset{(*)}{\le} \max_{i\in S} \Big[B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi^*_k) - B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi^*_k) \\
&\qquad + B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi_k) - B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k)\Big] \\
&\le \|B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi^*_k) - B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi^*_k)\|_\infty \qquad (21) \\
&\qquad + \|B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi_k) - B(r_k - g^*_k M, \mathcal{P}_k, h^*_k, \pi_k)\|_\infty \\
&\le 2\|r_k - r_{k-1}\|_{\max} + 2\|h_k - h^*_k\|_\infty + 2\|h^*_k\|_\infty \|\mathcal{P}_k - \mathcal{P}_{k-1}\|_\infty \qquad \text{(by Lemma 1)} \\
&\le 2\|r_k - r_{k-1}\|_{\max} + 2\|h_k - h^*_{k-1}\|_\infty + 2\|h^*_k - h^*_{k-1}\|_\infty + 2\|h^*_k\|_\infty \|\mathcal{P}_k - \mathcal{P}_{k-1}\|_\infty \qquad (22)
\end{aligned}$$

In the above derivations, step (*) is due to how π_k is selected in OVIs. Recall that, OVIs selects π_k based on maximizing the Q_k function, which is equivalent to maximizing $B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi)$ because g^*_k is a constant and does not affect the maximization. As a result,

$$B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi_k)(i) = \max_\pi B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi)(i) \ge B(r_{k-1} - g^*_k M, \mathcal{P}_{k-1}, h_k, \pi^*_k)(i)$$

which implies the (*) step. Plugging (22) into (20), we have

$$\begin{aligned}\|V_k - h_k^*\|_\infty \leq \|V_{k+1} - h_{k+1}^*\|_\infty + 2\|r_k - r_{k-1}\|_{\max} + 2\|h_k^*\|_\infty \|\mathcal{P}_k - \mathcal{P}_{k-1}\|_\infty + 2\|h_k - h_{k-1}^*\|_\infty \\ + \|h_k^* - h_{k+1}^*\|_\infty + 2\|h_k^* - h_{k-1}^*\|_\infty \qquad (23)\end{aligned}$$

To reach the desired results, we still need to control the last two terms as well as the $\|h_k^*\|_\infty$ factor in (23).

For the $\|h_k^*\|_\infty$ factor, we use the perturbation result for Bellman equation systems in Lemma 2 in Appendix B and compare $h_k^* = h^*(r_k, \mathcal{P}_k)$ with a zero reward MDP $h^*(0, \mathcal{P}_k)$, implying,

$$\|h_k^*\|_\infty \leq 2\eta \|r_k\|_{\max} \leq 2\eta R_{k-1}.$$

For the terms $\|h_k^* - h_{k+1}^*\|_\infty$ and $2\|h_k^* - h_{k-1}^*\|_\infty$, we use Corollary 2 in Appendix B,

$$\begin{aligned}2\|h_k^* - h_{k-1}^*\|_\infty &\leq 4\eta \|r_k - r_{k-1}\|_{\max} + (2n\eta\kappa + 4\eta^2) R_{k-1} \|\mathcal{P}_k - \mathcal{P}_{k-1}\|_\infty \\ \|h_k^* - h_{k+1}^*\|_\infty &\leq 2\eta \|r_{k+1} - r_k\|_{\max} + (n\eta\kappa + 2\eta^2) R_k \|\mathcal{P}_{k+1} - \mathcal{P}_k\|_\infty\end{aligned}$$

Plugging the above into (23), we get,

$$\begin{aligned}\|V_k - h_k^*\|_\infty \leq \|V_{k+1} - h_{k+1}^*\|_\infty + (4\eta + 2)\|r_k - r_{k-1}\|_{\max} + (2\eta\kappa n + 4\eta^2 + 4\eta) R_{k-1} \|\mathcal{P}_k - \mathcal{P}_{k-1}\|_\infty \\ + 2\eta \|r_{k+1} - r_k\|_{\max} + (\eta\kappa n + 2\eta^2) R_k \|\mathcal{P}_{k+1} - \mathcal{P}_k\|_\infty + 2\|h_k - h_{k-1}^*\|_\infty\end{aligned}$$

which concludes the proof of Lemma 4. □

C.2. Proof of Lemma 5

The proof will rely on the contraction property of the Bellman operator introduced in Proposition 1.

Recall that $h_{k-1}^* = h^*(r_{k-1}, \mathcal{P}_{k-1}) = \tilde{h}^*(r_{k-1}, g_{k-1}^*, \mathcal{P}_{k-1})$ by (15) and Definition 4, 5. Then, using the perturbation results on $\tilde{h}(\cdot, \cdot, \cdot)$ in Lemma 3 (Appendix B), we have

$$\begin{aligned}\|h_k - h_{k-1}^*\|_\infty &\leq \|h_k - \tilde{h}^*(r_{k-1}, g_k, P_{k-1})\|_\infty + \|\tilde{h}^*(r_{k-1}, g_k, P_{k-1}) - h_{k-1}^*\|_\infty \\ &= \|h_k - \tilde{h}^*(r_{k-1}, g_k, P_{k-1})\|_\infty + \|\tilde{h}^*(r_{k-1}, g_k, P_{k-1}) - \tilde{h}(r_{k-1}, g_{k-1}^*, \mathcal{P}_{k-1})\|_\infty \\ &\leq \|h_k - \tilde{h}^*(r_{k-1}, g_k, P_{k-1})\|_\infty + \eta |g_k - g_{k-1}^*| \qquad \text{(by Lem 3)} \\ &\leq (1 + \eta) G_k^\infty\end{aligned}$$

where

$$G_t^\infty = \max(|g_t - g_{t-1}^*|, \|h_t - \tilde{h}^*(r_{t-1}, g_t, \mathcal{P}_{t-1})\|_\infty)$$

and $\tilde{h}^*(r, g, \mathcal{P}) \in \mathbb{R}^n$ refers to the unique solution to (13) under reward r gain g and transition function $\mathcal{P}$.

The key of the proof is to show that G^∞ has some kind of contraction property with some error term:

$$G_t^\infty \leq c G_{t-1}^\infty + O(R_{t-2} \|\mathcal{P}_{t-2} - \mathcal{P}_{t-1}\|_\infty) + O(\|r_{t-2} - r_{t-1}\|_{\max}) \qquad (24)$$

where $0 < c < 1$. Then, by summing up G_t^∞ over t, we have

$$\sum_{t=1}^T G_t^\infty = G_1^\infty + \sum_{t=2}^T G_t^\infty \leq G_1^\infty + c \sum_{t=2}^T G_{t-1}^\infty + C_3 \sum_{t=2}^T R_{t-2} \|\mathcal{P}_{t-2} - \mathcal{P}_{t-1}\|_\infty n + C_4 \sum_{t=2}^T \|r_{t-2} - r_{t-1}\|_{\max}$$

Subtracting $c \sum_{t=1}^T G_t^\infty$ on both sides and dividing by $1 - c$ leads to

$$\sum_{t=1}^T G_t^\infty = O(\sum_{t=1}^T R_{t-1} \|\mathcal{P}_t - \mathcal{P}_{t-1}\|_\infty + \sum_{t=1}^T \|r_t - r_{t-1}\|_{\max})$$

The proof is concluded by noticing that $G_1^\infty = 0$ under the initialization rule provided in Theorem 1.

Now it suffices to prove (24). The proof relies on the contraction property in Proposition 1. We will use the contraction of the operator F to establish the contraction of the operator VI used in our algorithm 1. One trouble is that in the contraction of F, the norm that allows the contraction depends on the transition probabilities, which becomes time-varying in the time-varying MDP problems. Nevertheless, by iterating VI for sufficient steps as in Algorithm 1, we are able to prove the contraction with respect to a universal norm, the infinity norm. Then, the proof of (24) can be completed.

The proof is divided into three steps. In step 1, we will analyze a single VI step with projection, and establish its contraction property with respect to a weighted infinity norm. In step 2, we analyze K consecutive VI steps, and show the contraction property of K VI steps with respect to infinity norm. In step 3, we apply the contraction property established in step 2 to show the contraction property of G_t^∞.

Step 1: Contraction of a single VI step with respect to weighted infinity norm. Consider one VI step with projection for MDP $(r, \mathcal{P})$ with stepsize γ^k:

$$\begin{aligned}&(h^{k+1}, \bar{g}^{k+1}) = \mathrm{VI}(r, \mathcal{P}, h^k, g^k, \gamma^{k+1})\\ &g^{k+1} = \Pi_X(\bar{g}^{k+1})\end{aligned}$$

where $X = [-\|r\|_{\max}, \|r\|_{\max}]$. Recall that the first step in VI uses operator F to update h^k and by Proposition 1, F is constractive with respect to h^k. In the following, we will show that the above VI step with projection also has contraction property with respect to both h^k and g^k in some sense.

Proposition 2 (Contraction w.r.t. the weighted infinity norm)**.** *There exists a weighted infinity norm $\|\cdot\|_{\infty,v}$ and a positive scalar $\bar{\gamma}$, such that for any $\gamma^{k+1} \in (0, \bar{\gamma}]$, there exists a constant $c \in [0, 1)$ that depends on $\rho, n, \alpha, \gamma^{k+1}$ such that,*

$$\max(\|h^{k+1} - \tilde{h}^*(r, g^{k+1}, \mathcal{P})\|_{\infty,v}, |g^{k+1} - g^*|) \le c \max(\|h^k - \tilde{h}^*(r, g^k, \mathcal{P})\|_{\infty,v}, |g^k - g^*|)$$

where $g^* = g^*(r, \mathcal{P})$ is the solution to the Bellman Equation System (3) *and* (4) *under reward function r and probability transition function $\mathcal{P}$; $\tilde{h}^*(r, g^k, \mathcal{P})$ is the unique solution to the (modified) Bellman equation* (13) *with gain g^k under reward r and probability transition function $\mathcal{P}$. Moreover, $|g^k| \le \|r\|_{\max}$ when $k \ge 1$.*

Proof. The proof is based on the contraction property of VI without projection established in (Bertsekas, 1998), which we state below.

Proposition 3 (Proposition 1 in (Bertsekas, 1998); Proposition 3.3.1 (Bertsekas, 2012); Eq. (7.5) in (Bertsekas, 2005))**.** *There exists a positive scalar $\tilde{\gamma}$ and a weighted infinity norm $\|\cdot\|_{\infty,v}$, such that for any $0 < \gamma^{k+1} \le \tilde{\gamma}$, there exists a scalar $\tilde{c} \in (0, 1)$ that depends on $\rho, n, \alpha, \gamma^{k+1}$, such that,*

$$\max(\|h^{k+1} - \tilde{h}^*(r, \bar{g}^{k+1}, \mathcal{P})\|_{\infty,v}, |\bar{g}^{k+1} - g^*|) \le \tilde{c} \max(\|h^k - \tilde{h}^*(r, g^k, \mathcal{P})\|_{\infty,v}, |g^k - g^*|)$$

Moreover, the weight v satisfies $1 \le v_i \le n/\rho$ for all $i \in S$, and hence the weighted norm satisfies

$$\frac{1}{\eta}\|x\|_\infty \le \|x\|_{\infty,v} \le \|x\|_\infty, \forall x \in \mathbb{R}^n.$$

Due to the projection step $g^{k+1} = \Pi_X(\bar{g}^{k+1})$, we have $|g^{k+1}| \le \|r\|_{\max}$, i.e. the boundness of $|g^k|$. Let $E = \max(\|h^k - \tilde{h}^*(r, g^k, \mathcal{P})\|_{\infty,v}, |g^k - g^*|)$. To prove the proposition we need to bound $|g^{k+1} - g^*|$ and $\|h^{k+1} - \tilde{h}^*(r, g^{k+1}, \mathcal{P})\|_{\infty,v}$ by cE for some $c \in [0, 1)$ and small enough step size γ^{k+1}.

Bounding $|g^{k+1} - g^*|$. Notice that $g^* = g^*(r, \mathcal{P})$ is the optimal average reward of MDP $(r, \mathcal{P})$ and hence g lies inside interval X. By projection theorem and Proposition 3, we have

$$|g^{k+1} - g^*| \le |\bar{g}^{k+1} - g^*| \le \tilde{c}E$$

Bounding $\|h^{k+1} - \tilde{h}^*(r, g^{k+1}, \mathcal{P})\|_{\infty,v}$. Recall that F is the operator defined in Proposition 1 that updates the h vector in value iteration (5), $h^{k+1} = F(h^k)$. Here we abuse the notation and write F as $F(r, g, \mathcal{P}, h)$ to make F's dependence on $r, g, \mathcal{P}$ explicit. Then, update of h in VI can be written as,

$$h^{k+1} = F(r, g^k, \mathcal{P}, h^k)$$

and the contraction property introduced in Proposition 1 can be rewritten as,

$$\|F(r,g,\mathcal{P},h) - F(r,g,\mathcal{P},\tilde{h})\|_{\infty,v} \leq \alpha\|h-\tilde{h}\|_{\infty,v}. \tag{25}$$

The operator F can be viewed as the Bellman operator of the modified Bellman equation (13). Since $h^*(r,g^{k+1},\mathcal{P})$ is the solution to Bellman equation (13), we have

$$\tilde{h}^*(r,g^{k+1},\mathcal{P}) = F(r,g^{k+1},\mathcal{P},\tilde{h}^*(r,g^{k+1},\mathcal{P})).$$

Using the above fact, we proceed to bound $\|h^{k+1} - \tilde{h}^*(r,g^{k+1},\mathcal{P})\|_{\infty,v}$.

$$\begin{aligned}
\|h^{k+1} - \tilde{h}^*(r,g^{k+1},\mathcal{P})\|_{\infty,v} &= \|F(r,g^k,\mathcal{P},h^k) - F(r,g^{k+1},\mathcal{P},\tilde{h}^*(r,g^{k+1},\mathcal{P}))\|_{\infty,v} \\
&\leq \|F(r,g^k,\mathcal{P},h^k) - F(r,g^k,\mathcal{P},\tilde{h}^*(r,g^{k+1},\mathcal{P}))\|_{\infty,v} \\
&\quad + \|F(r,g^k,\mathcal{P},\tilde{h}^*(r,g^{k+1},\mathcal{P})) - F(r,g^{k+1},\mathcal{P},\tilde{h}^*(r,g^{k+1},\mathcal{P}))\|_{\infty,v} \\
&= \|F(r,g^k,\mathcal{P},h^k) - F(r,g^k,\mathcal{P},\tilde{h}^*(r,g^{k+1},\mathcal{P}))\|_{\infty,v} + \|(g^k - g^{k+1})\vec{e}\|_{\infty,v} \\
&\leq \alpha\|h^k - \tilde{h}^*(r,g^{k+1},\mathcal{P})\|_{\infty,v} + |g^k - g^{k+1}| \qquad \text{(by (25) and change of norm)} \\
&\leq \alpha\|h^k - \tilde{h}^*(r,g^k,\mathcal{P})\|_{\infty,v} + \alpha\|\tilde{h}^*(r,g^k,\mathcal{P}) - \tilde{h}^*(r,g^{k+1},\mathcal{P})\|_{\infty,v} + |g^k - g^{k+1}| \\
&\leq \alpha\|h^k - \tilde{h}^*(r,g^k,\mathcal{P})\|_{\infty,v} + (\alpha\eta+1)|g^k - g^{k+1}| \quad \text{(by Lemma 3 and change of norm)} \\
&\leq \alpha E + (\alpha\eta+1)|g^k - g^{k+1}|
\end{aligned} \tag{26}$$

Next, we bound $|g^k - g^{k+1}|$. By projection theorem,

$$|g^k - g^{k+1}| \leq |g^k - \bar{g}^{k+1}| = \gamma^{k+1}|h^{k+1}(\tau)|.$$

We notice,

$$h^{k+1} = F(r,g^k,\mathcal{P},h^k), \quad \tilde{h}^*(r,g^k,\mathcal{P}) = F(r,g^k,\mathcal{P},\tilde{h}^*(r,g^k,\mathcal{P}))$$

which, combined with the contraction property of F (25), shows

$$\|h^{k+1} - \tilde{h}^*(r,\mathcal{P},g^k)\|_{\infty,v} \leq \alpha\|h^k - \tilde{h}^*(r,g^k,\mathcal{P})\|_{\infty,v} \tag{27}$$

Therefore,

$$\begin{aligned}
|h^{k+1}(\tau)| &\leq |h^{k+1}(\tau) - \tilde{h}^*(r,g^k,\mathcal{P})(\tau)| + |\tilde{h}^*(r,g^k,\mathcal{P})(\tau)| \\
&= |h^{k+1}(\tau) - \tilde{h}^*(r,g^k,\mathcal{P})(\tau)| + |\tilde{h}^*(r,g^k,\mathcal{P})(\tau) - \tilde{h}^*(r,g^*,\mathcal{P})(\tau)| \quad (\text{by } \tilde{h}^*(r,g^*,\mathcal{P})(\tau)=0) \\
&\leq \|h^{k+1} - \tilde{h}^*(r,g^k,\mathcal{P})\|_\infty + \|\tilde{h}^*(r,g^k,\mathcal{P}) - \tilde{h}^*(r,g^*,\mathcal{P})\|_\infty \\
&\leq \|h^{k+1} - \tilde{h}^*(r,g^k,\mathcal{P})\|_\infty + \eta|g^k - g^*| && \text{(by Lemma 3)} \\
&\leq \eta\|h^{k+1} - \tilde{h}^*(r,g^k,\mathcal{P})\|_{\infty,v} + \eta|g^k - g^*| && \text{(change of norm)} \\
&\leq \eta\alpha\|h^k - \tilde{h}^*(r,g^k,\mathcal{P})\|_{\infty,v} + \eta|g^k - g^*|. && \text{(by (27))}
\end{aligned}$$

Therefore, we get the following bound on $|g^k - g^{k+1}|$.

$$|g^k - g^{k+1}| \leq \gamma^{k+1}\Big[\eta\alpha\|h^k - \tilde{h}^*(r,g^k,\mathcal{P})\|_{\infty,v} + \eta|g^k - g^*|\Big] \leq \gamma^{k+1}(\eta\alpha+\eta)E.$$

Plugging the above into (26), we have

$$\|h^{k+1} - \tilde{h}^*(r,g^{k+1},\mathcal{P})\|_{\infty,v} \leq [\alpha + \gamma^{k+1}(\alpha\eta+1)(\alpha+1)\eta]E$$

Therefore, if $\gamma^{k+1} \leq \tilde{\gamma}'$ and $\tilde{\gamma}'$ is small enough s.t., $\alpha + \tilde{\gamma}'(\alpha\eta+1)\,(\alpha+1)\,\eta < 1$, the desired contraction property holds.

Value of $\bar{\gamma}$ and c. Based on the above derivations, the step size upper bound is $\bar{\gamma} = \min(\tilde{\gamma}, \tilde{\gamma}')$ where $\tilde{\gamma}$ is from Proposition 3 and $\tilde{\gamma}'$ can be any positive number s.t. $\alpha + \tilde{\gamma}'(\alpha\eta+1)\,(\alpha+1)\,\eta < 1$. The contraction constant c is,

$$c = \max(\tilde{c}, \alpha + \tilde{\gamma}'(\alpha\eta+1)(\alpha+1)\eta)$$

where $\tilde{c}$ is the contraction coefficient without projection from Proposition 3, and $\tilde{c}$ depends on $\alpha, \rho, n, \gamma^{k+1}$. □

Step 2: Contraction of K VI steps with respect to infinity norm. We apply Proposition 2 to analyze K consecutive VI steps and show its contraction property in the following proposition.

Proposition 4 (Contraction w.r.t. infinity norm)**.** *For any MDP with $(r, \mathcal{P}) \in \mathfrak{M}(\kappa, \rho, \alpha)$, there exists a positive scalar $\bar{\gamma}$, a function $\delta(\rho, \alpha)$, such that for any $0 < \gamma \leq \bar{\gamma}$, and any $K \geq \delta(\rho, \alpha) \log n$, under a stepsize rule $\gamma^1 = \gamma, \gamma^2 = \ldots = \gamma^K = 0$, there exists a scalar $0 \leq c < 1$, such that the following contraction property holds*

$$\max(\|h^K - \tilde{h}^*(r, g^K, \mathcal{P})\|_\infty, |g^K - g^*|) \leq c \max(\|h^0 - \tilde{h}^*(r, g^0, \mathcal{P})\|_\infty, |g^0 - g^*|)$$

where $g^* = g^*(r, \mathcal{P})$ is the optimal average reward of MDP $(r, \mathcal{P})$.

Proof. Let $E = \max(\|h^0 - \tilde{h}^*(r, g^0, \mathcal{P})\|_\infty, |g^0 - g^*|)$. We set $\bar{\gamma}$ to be the same as in Proposition 2. Then, since $\gamma^1 = \gamma \leq \bar{\gamma}$, we apply Proposition 2 to the update from (h^0, g^0) to (h^1, g^1), and get

$$\begin{aligned} \max(\|h^1 - \tilde{h}^*(r, g^1, \mathcal{P})\|_{\infty,v}, |g^1 - g^*|) &\leq c \max(\|h^0 - \tilde{h}^*(r, g^0, \mathcal{P})\|_{\infty,v}, |g^0 - g^*|) \\ &\leq c \max(\|h^0 - \tilde{h}^*(r, g^0, \mathcal{P})\|_\infty, |g^0 - g^*|) = cE \end{aligned} \tag{28}$$

where c is from Proposition 2 and depends on α, n, ρ, γ, and in the second inequality we have used change of norm (cf. Proposition 3).

By (28), and $\gamma^2 = \cdots = \gamma^K = 0$, we have $|g^K - g^*| = |g^1 - g^*| \leq cE$. Now it suffices to show $\|h^K - \tilde{h}^*(r, g^K, \mathcal{P})\|_\infty \leq cE$.

$$\begin{aligned} \|h^K - \tilde{h}^*(r, g^K, \mathcal{P})\|_\infty &= \|h^K - \tilde{h}^*(r, g^1, \mathcal{P})\|_\infty \\ &\leq \eta \|h^K - \tilde{h}^*(r, g^1, \mathcal{P})\|_{\infty,v} && \text{(change of norm)} \\ &\leq \alpha^{K-1} \eta \|h^1 - \tilde{h}^*(r, g^1, \mathcal{P})\|_{\infty,v} && \text{(by (25))} \\ &\leq \alpha^{K-1} \eta c E && \text{(by (28))} \end{aligned}$$

Hence, as long as $\alpha^{K-1}\eta \leq 1$, we have $\|h^K - \tilde{h}^*(r, g^K, \mathcal{P})\|_\infty \leq cE$. This requires $K \geq \log(n/\rho)/\log(1/\alpha) + 1$. □

Step 3: Proof of (24)**.** Recall $G_t^\infty = \max(|g_t - g_{t-1}^*|, \|h_t - \tilde{h}^*(r_{t-1}, g_t, \mathcal{P}_{t-1})\|_\infty)$, where $g_t = g_t^K, h_t = h_t^K$, and (h_t^K, g_t^K) is obtained by applying K VI steps with initial condition $(h_t^0, g_t^0) = (h_{t-1}, g_{t-1})$, under MDP $(r_{t-1}, \mathcal{P}_{t-1})$, with step size the same as the setting in Proposition 4. Therefore, by Proposition 4,

$$\begin{aligned} G_t^\infty &= \max(\|h_t^K - \tilde{h}^*(r_{t-1}, g_t^K, \mathcal{P}_{t-1})\|_\infty, |g_t^K - g_{t-1}^*|) \\ &\leq c \max(\|h_t^0 - \tilde{h}^*(r_{t-1}, g_t^0, \mathcal{P}_{t-1})\|_\infty, |g_t^0 - g_{t-1}^*|) \\ &= c \max(\|h_{t-1} - \tilde{h}^*(r_{t-1}, g_{t-1}, \mathcal{P}_{t-1})\|_\infty, |g_{t-1} - g_{t-1}^*|) \\ &\leq c G_{t-1}^\infty + c|g_{t-1}^* - g_{t-2}^*| + c\|\tilde{h}^*(r_{t-1}, g_{t-1}, \mathcal{P}_{t-1}) - \tilde{h}^*(r_{t-2}, g_{t-1}, \mathcal{P}_{t-2})\|_\infty \\ &\leq c G_{t-1}^\infty + c|g_{t-1}^* - g_{t-2}^*| + c\eta^2(R_{t-2} + |g_{t-1}|)\|\mathcal{P}_{t-1} - \mathcal{P}_{t-2}\|_\infty + c\eta\|r_{t-1} - r_{t-2}\|_{\max} && \text{((by Lem 3))} \\ &\leq c G_{t-1}^\infty + c(n\kappa + 2\eta^2) R_{t-2} \|\mathcal{P}_{t-1} - \mathcal{P}_{t-2}\|_\infty + c(1+\eta)\|r_{t-1} - r_{t-2}\|_{\max} \\ & && \text{((by } |g_{t-1}| \in X_{t-2} \text{ and Corollary 2))} \end{aligned}$$

which concludes the proof of (24).

D. Proof of Theorem 2.

Proof sketch: Roughly speaking, we will construct four MDP problems, with very different optimal policies. We let the online MDP problem at t be randomly selected from the four MDP problems. Then we will show that no online algorithm can always perform well without knowing the current stage cost and transition probabilities, which results in a positive regret lower bound.

Proof. We fix any online algorithm $\mathcal{A}$. We set $n = m = 2$, i.e. there are only two states $\{1, 2\}$ and two actions $\{a, \tilde{a}\}$. The proof outline is as follows. In step 1, we generate an random ensemble of problem instances $(r_t, \mathcal{P}_t)_{t=1}^T$. In step 2, we calculate the expected total reward of the online algorithm $\mathcal{A}$ as well as the optimal total reward in hindsight. Here the expectation is taken with respect to both the randomness of the problem ensamble and the randomness of the Markov chain. In step 3, we get a lower bound of the expected regret of the algorithm over the problem instance ensemble, from which we can show the existence of a problem instance $(r_t, \mathcal{P}_t)_{t=1}^T$ that meets the lower bound, and we check the instance meets the variation budget.

Step 1: construct random problem ensemble. Firstly, define transition probability function $\mathcal{P}$ and $\tilde{\mathcal{P}}$ as,

$$P(a) = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}, P(\tilde{a}) = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}, \tilde{P}(a) = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}, \tilde{P}(\tilde{a}) = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

We also define two reward functions r and $\tilde{r}$.

$$r(i, a, j) = \begin{cases} R & \text{if } j = 1 \\ 0 & \text{if } j = 2 \end{cases}, \tilde{r}(i, a, j) = \begin{cases} 0 & \text{if } j = 1 \\ R & \text{if } j = 2 \end{cases}$$

where R is any positive constant satisfying $R \le \min(L_P, L_r)$.

We next randomly generate a sequence of reward functions and transition probability functions $\{(r_t, \mathcal{P}_t)\}_{t=1}^T$. To generate $\mathcal{P}_t$, We divide the time horizon $\{1, \dots, T\}$ evenly into $\ell_P = \lfloor \frac{3L_P}{2R} \rfloor \ge 1$ windows each of size $\Delta_P = \lceil \frac{T}{\ell_P} \rceil$ (the last window may be of size smaller than Δ_P). The k'th window is $\{(k-1)\Delta_P + 1, \dots, \min(k\Delta_P, T)\}$, and let $\mathcal{T}_P = \{1, \Delta_P + 1, \dots, (\ell_P - 1)\Delta_P + 1\}$ be the set of the first time index of the windows. For, each $t \in \mathcal{T}_P$, i.e. the starting index of a window, we independently generate $\mathcal{P}_t = \mathcal{P}$ w.p. $1/2$, and $\mathcal{P}_t = \tilde{\mathcal{P}}$ w.p. $1/2$. Next, for $t \notin \mathcal{T}_P$, we set $\mathcal{P}_t = \mathcal{P}_{\max\{\tau \in \mathcal{T}_P, \tau < t\}}$, i.e. $\mathcal{P}_t$ is to set to be same as the starting index of the window that t is in.

The reward functions r_t are generated in a similar way. We divide the time horizon $\{1, \dots, T\}$ evenly into $\ell_r = \lfloor \frac{L_r}{R} \rfloor \ge 1$ windows each of size $\Delta_r = \lceil \frac{T}{\ell_r} \rceil$ (the last window may be of size smaller than Δ_r). The k'th window is $\{(k-1)\Delta_r + 1, \dots, \min(k\Delta_r, T)\}$, and let $\mathcal{T}_r = \{1, \Delta_r + 1, \dots, (\ell_r - 1)\Delta_r + 1\}$ be the set of the first time index of the windows. For, each $t \in \mathcal{T}_r$, i.e. the starting index of a window, we independently generate $r_t = r$ w.p. $1/2$, and $r_t = \tilde{r}$ w.p. $1/2$. Next, for $t \notin \mathcal{T}_r$, we set $r_t = r_{\max\{\tau \in \mathcal{T}_r, \tau < t\}}$, i.e. r_t is to set to be same as the starting index of the window t is in.

Step 2.1: expected total reward of online algorithm $\mathcal{A}$. Given problem instance $\{(\mathcal{P}_t, r_t)\}_{t=1}^T$, we run the online algorithm, which generates state-policy-action sequence $\{(s_t, \pi_t, a_t)\}_{t=1}^T$. The expected reward, is

$$\mathbb{E} \sum_{t=1}^{T} r_t(s_t, a_t, s_{t+1}) = \sum_{t=1}^{T} \mathbb{E}_{s_t, \pi_t, \mathcal{P}_t, r_t} \mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t, r_t] \tag{29}$$

We have the following calculations,

$$\mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t = \mathcal{P}, r_t = r] = \frac{1}{3} R \pi_t(a|s_t) + \frac{2}{3} R \pi_t(\tilde{a}|s_t) \tag{30a}$$

$$\mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t = \mathcal{P}, r_t = \tilde{r}] = \frac{2}{3} R \pi_t(a|s_t) + \frac{1}{3} R \pi_t(\tilde{a}|s_t) \tag{30b}$$

$$\mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t = \tilde{\mathcal{P}}, r_t = r] = \frac{2}{3} R \pi_t(a|s_t) + \frac{1}{3} R \pi_t(\tilde{a}|s_t) \tag{30c}$$

$$\mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t = \tilde{\mathcal{P}}, r_t = \tilde{r}] = \frac{1}{3} R \pi_t(a|s_t) + \frac{2}{3} R \pi_t(\tilde{a}|s_t) \tag{30d}$$

Claim 1: For any t, $\mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t, r_t] \le \frac{2}{3} R$ almost surely. This is a direct concequence of (30) since in all four cases, the expectation is a convex combination of $\frac{1}{3} R$ and $\frac{2}{3} R$.

Claim 2: For $t \in \mathcal{T}_P$, we have $\mathbb{E}_{s_t, \pi_t, \mathcal{P}_t, r_t} \mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t, r_t] = \frac{R}{2}$. We note that for such t, P_t is independent from π_t, s_t, r_t because π_t, s_t is determined before the randomness of P_t is revealed. As a result, we can take expectation separately,

$$\mathbb{E}_{s_t, \pi_t, \mathcal{P}_t, r_t} \mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t, r_t] = \mathbb{E}_{s_t, \pi_t, r_t} \mathbb{E}_{\mathcal{P}_t} \mathbb{E}[r_t(s_t, a_t, s_{t+1}) | s_t, \pi_t, \mathcal{P}_t, r_t]$$

$$= \mathbb{E}_{s_t,\pi_t,r_t}\left\{\frac{1}{2}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t=\mathcal{P},r_t] + \frac{1}{2}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t=\tilde{\mathcal{P}},r_t]\right\}$$

Both the right hand side of (30a) and (30c), and the right hand side of (30b) and (30d) sums up to R, so we have

$$\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t=\mathcal{P},r_t] + \mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t=\tilde{\mathcal{P}},r_t] = R, a.s.$$

As a result, $\mathbb{E}_{s_t,\pi_t,\mathcal{P}_t,r_t}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t] = \frac{1}{2}R$ almost surely.

Claim 3: For $t \in \mathcal{T}_r$, we have $\mathbb{E}_{s_t,\pi_t,\mathcal{P}_t,r_t}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t] = \frac{R}{2}$. The reason is similar to that of Claim 2. For such t, r_t is independent from $s_t,\pi_t,\mathcal{P}_t$. Then,

$$\begin{aligned}\mathbb{E}_{s_t,\pi_t,\mathcal{P}_t,r_t}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t] &= \mathbb{E}_{s_t,\pi_t,\mathcal{P}_t}\mathbb{E}_{r_t}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t]\\ &= \mathbb{E}_{s_t,\pi_t,\mathcal{P}_t}\left\{\frac{1}{2}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t=r] + \frac{1}{2}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t=\tilde{r}]\right\}\end{aligned}$$

Both the right hand side of (30a) and (30b), and the right hand side of (30c) and (30d) sums up to R, so we have

$$\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t=r] + \mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t=\tilde{r}] = R, a.s. \tag{31}$$

As a result, $\mathbb{E}_{s_t,\pi_t,\mathcal{P}_t,r_t}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t] = \frac{1}{2}R$ almost surely.

Combining Claim 1,2,3 with (29), we have the following upper bound on the expected reward of the online algorithm,

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T} r_t(s_t,a_t,s_{t+1}) &= \sum_{t\in\mathcal{T}_P\cup\mathcal{T}_r}\mathbb{E}_{s_t,\pi_t,\mathcal{P}_t,r_t}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t] + \sum_{t\notin\mathcal{T}_P\cup\mathcal{T}_r}\mathbb{E}_{s_t,\pi_t,\mathcal{P}_t,r_t}\mathbb{E}[r_t(s_t,a_t,s_{t+1})|s_t,\pi_t,\mathcal{P}_t,r_t]\\ &\le \frac{R}{2}|\mathcal{T}_P\cup\mathcal{T}_r| + \frac{2}{3}R(T-|\mathcal{T}_P\cup\mathcal{T}_r|)\\ &\le \frac{2}{3}RT - \frac{R}{6}\max(\ell_P,\ell_r)\\ &\le \frac{2}{3}RT - \frac{R}{12}(\ell_P+\ell_r)\end{aligned} \tag{32}$$

Step 2.2: optimal total reward in hindsight. We fix a problem instance $\{(P_t,r_t)\}_{t=1}^T$. Note that, claim 1 still holds even if π_t is allowed to depend the entire problem instance $\{(P_t,r_t)\}_{t=1}^T$, including the future. As such, the optimal total reward in hindsight for problem instance $\{(P_t,r_t)\}_{t=1}^T$ is upper bounded by $\frac{2}{3}RT$. Consider the following determistic policy, in which the action at time t is set as,

$$a_t^* = \begin{cases} a & \text{if } (P_t,r_t) = (\mathcal{P},r) \text{ or } (\tilde{\mathcal{P}},\tilde{r})\\ \tilde{a} & \text{if } (P_t,r_t) = (\mathcal{P},\tilde{r}) \text{ or } (\tilde{\mathcal{P}},r)\end{cases}$$

It is easy to check that the above policy, that depends on with full knowledge of $\{(P_t,r_t)\}_{t=1}^T$, can achieve total reward $\frac{2}{3}RT$. As such, the the above policy is optimal for problem instance $\{(P_t,r_t)\}_{t=1}^T$ and achieves total reward in hindsight $\frac{2}{3}RT$.

Step 3: lower bounding regret. With the upper bound on expected total reward of algorithm $\mathcal{A}$ (32) and the optimal total reward in hindsight, we can lower bound the expected regret over the random ensemble of problem instances $\{(\mathcal{P}_t,r_t)\}_{t=1}^T$

$$\begin{aligned}\mathop{\mathbb{E}}_{\{(\mathcal{P}_t,r_t)\}_{t=1}^T}(\mathcal{D}_T(\mathcal{A})) &= \mathop{\mathbb{E}}_{\{(\mathcal{P}_t,r_t)\}_{t=1}^T}\left[\frac{2}{3}RT - \mathbb{E}\big[\sum_{t=1}^T r_t(s_t,a_t,s_{t+1})|\{(P_t,r_t)\}_{t=1}^T\big]\right]\\ &= \frac{2}{3}RT - \mathbb{E}\sum_{t=1}^T r_t(s_t,a_t,s_{t+1})\\ &\ge \frac{R}{12}(\ell_P+\ell_r)\end{aligned}$$

Therefore, there must exists a sequence of $\{(\mathcal{P}_t, r_t)\}_{t=1}^T$, s.t. the regret of the online algorithm $\mathcal{D}_T(\mathcal{A}) \geq \frac{R}{12}(\ell_P + \ell_r) = \frac{R}{12}(\lfloor \frac{3L_P}{2R} \rfloor + \lfloor \frac{L_r}{R} \rfloor) = \Omega(L_P + L_r)$. Finally, we check the instance $\{(\mathcal{P}_t, r_t)\}_{t=1}^T$ meets the variation budget. Recalling that r_t changes only every Δ_r steps, we have

$$\sum_{t=1}^T \|r_t - r_{t+1}\|_{\max} = \sum_{t \in \mathcal{T}_r} \|r_{t-1} - r_t\|_{\max} \leq \ell_r R \leq L_r$$

Similarly, noticing $R_t = R$, we have the variation of $\mathcal{P}_t$ is

$$\sum_{t=1}^T R_t \|\mathcal{P}_t - \mathcal{P}_{t+1}\|_{\max} = \sum_{t \in \mathcal{T}_P} R_{t-1} \|\mathcal{P}_{t-1} - \mathcal{P}_t\|_\infty \leq \frac{2}{3} R \ell_P \leq L_P$$

□

E. Proofs of results in Appendix B

E.1. Proof of Lemma 1

Proof. **Proof of (a).** For any i,

$$\begin{aligned}
|B(r, h, \mathcal{P}, \pi)(i) - B(r, h, \mathcal{P}', \pi)(i)| &= |r(i, \pi) + \sum_j P_{ij}(\pi) h(j) - (r(i, \pi) + \sum_j P'_{ij}(\pi) h(j))| \\
&= |\sum_j h(j)(P_{ij}(\pi) - P'_{ij}(\pi))| \\
&\leq \|P(\pi) - P'(\pi)\|_\infty \|h\|_\infty \\
&\leq \max_a \|P(a) - P'(a)\|_\infty \|h\|_\infty \\
&= \|\mathcal{P} - \mathcal{P}'\|_\infty \|h\|_\infty
\end{aligned}$$

which implies $\|B(r, h, \mathcal{P}, \pi) - B(r, h, \mathcal{P}', \pi)\|_\infty \leq \|\mathcal{P} - \mathcal{P}'\|_\infty \|h\|_\infty$. A symmetric argument shows $\|B(r, h, \mathcal{P}, \pi) - B(r, h, \mathcal{P}', \pi)\|_\infty \leq \|\mathcal{P} - \mathcal{P}'\|_\infty \|h'\|_\infty$, and hence we have

$$\|B(r, h, \mathcal{P}, \pi) - B(r, h, \mathcal{P}', \pi)\|_\infty \leq \min(\|h\|_\infty, \|h'\|_\infty) \|\mathcal{P} - \mathcal{P}'\|_\infty \tag{33}$$

Further, note

$$\begin{aligned}
\|B(r, \mathcal{P}', h, \pi) - B(r', \mathcal{P}', h', \pi)\|_\infty &= \max_i |B(r, \mathcal{P}', h, \pi)(i) - B(r', \mathcal{P}', h', \pi)(i)| \\
&= \max_i |r(i, \pi) - r'(i, \pi) + \sum_{j=1}^n P'_{ij}(\pi)(h(j) - h'(j))| \\
&\leq \|r - r'\|_{\max} + \|h - h'\|_\infty
\end{aligned}$$

Combining the above and (33), we get

$$\begin{aligned}
\|B(r, \mathcal{P}, h, \pi) - B(r', \mathcal{P}', h', \pi)\|_\infty &\leq \|B(r, \mathcal{P}, h, \pi) - B(r, \mathcal{P}', h, \pi)\|_\infty + \|B(r, \mathcal{P}', h, \pi) - B(r', \mathcal{P}', h', \pi)\|_\infty \\
&\leq \|r - r'\|_{\max} + \|h - h'\|_\infty + \min(\|h\|_\infty, \|h'\|_\infty) \|\mathcal{P} - \mathcal{P}'\|_\infty
\end{aligned}$$

which shows (a).

Proof of (b). By symmetry and the definition of the infinite norm, it suffices to show $B^*(r, \mathcal{P}, h)(i) - B^*(r, \mathcal{P}, h')(i) \leq \|h - h'\|_\infty$ holds for any $i \in S$. For any i, suppose action a attains the maximization in $B(r, \mathcal{P}, h)(i)$ and action a' attains the maximization in $B(r, \mathcal{P}, h')(i)$, then

$$B^*(r, \mathcal{P}, h)(i) - B^*(r, \mathcal{P}, h')(i) = r(i, a) + \sum_{j=1}^n P_{ij}(a) h(j) - \left(r(i, a') + \sum_{j=1}^n P_{ij}(a') h'(j) \right)$$

$$
\begin{aligned}
&\le r(i,a) + \sum_{j=1}^{n} P_{ij}(a)h(j) - \left(r(i,a) + \sum_{j=1}^{n} P_{ij}(a)h'(j) \right) \\
&= \sum_{j=1}^{n} P_{ij}(a)(h(j) - h'(j)) \\
&\le \|h - h'\|_\infty
\end{aligned}
$$

where in the first inequality, we have used that a' is the maximizer in the definition of $B(r, \mathcal{P}, h')(i)$.

□

E.2. Proof of Lemma 2

To derive the proof, we need to first introduce the perturbation result of stationary distribution of Markov chain in literature.

Proposition 5 ((1.4) in (Meyer, 1994))**.** *Consider two ergodic Markov chains. The distance between the two stationary distributions can be upper bounded by the distance between the transition probability matrices multiplied by the condition number, that is,*

$$\|d - d'\|_\infty \le \|P - P'\|_\infty \kappa(P) \tag{34}$$

where d *and* d' *denote the stationary distribution of* P *and* P' *respectively and* $\kappa(P)$ *is the condition number of* P*.*

Bounding $|g^*(r, \mathcal{P}) - g^*(r', \mathcal{P}')|$**.** We will focus on the upper bound on $g^*(r, \mathcal{P}) - g^*(r', \mathcal{P}')$; $g^*(r', \mathcal{P}') - g^*(r, \mathcal{P})$ can be bounded similarly. Let $g(r, \mathcal{P}, \pi)$ be the average reward of MDP problem with reward function r, transition probability function $\mathcal{P}$ and policy π. Let the optimal policy corresponding to MDP $(r, \mathcal{P})$ be π^*, and the optimal policy corresponding to MDP $(r', \mathcal{P}')$ be $(\pi^*)'$. Then,

$$
\begin{aligned}
g^*(r, \mathcal{P}) - g^*(r', \mathcal{P}') &= g(r, \mathcal{P}, \pi^*) - g(r', \mathcal{P}', (\pi^*)') \\
&\le g(r, \mathcal{P}, \pi^*) - g(r', \mathcal{P}', \pi^*) && \text{(((}\pi^*)' \text{ maximizes } g(r', \mathcal{P}', \cdot)\text{))} \\
&= g(r, \mathcal{P}, \pi^*) - g(r, \mathcal{P}', \pi^*) + g(r, \mathcal{P}', \pi^*) - g(r', \mathcal{P}', \pi^*) \\
&= g(r, \mathcal{P}, \pi^*) - g(r, \mathcal{P}', \pi^*) + g(r - r', \mathcal{P}', \pi^*) && \text{(by the def. of average reward)} \\
&\le g(r, \mathcal{P}, \pi^*) - g(r, \mathcal{P}', \pi^*) + \|r - r'\|_{\max} && \text{(by the def. of average reward)}
\end{aligned}
$$

Now, it suffices to bound the difference of the average reward $g(r, \mathcal{P}, \pi^*) - g(r, \mathcal{P}', \pi^*)$. Let d and d' denote the stationary distribution of $\mathcal{P}(\pi^*)$ and $\mathcal{P}'(\pi^*)$. Let $r(\pi^*) \in \mathbb{R}^n$ denote the vector of one-time-step reward by following policy π^*, that is, $r(\pi^*)(i) = \mathbb{E}_{a \sim \pi^*(\cdot|i)}\, r(i, a)$. Then,

$$
\begin{aligned}
g(r, \mathcal{P}, \pi^*) - g(r, \mathcal{P}', \pi^*) &= \langle d - d', r(\pi^*) \rangle && \text{(by def. of average reward)} \\
&\le \|d - d'\|_1 \|r(\pi^*)\|_\infty \\
&\le n\|d - d'\|_\infty \|r\|_{\max} && \text{(change norm)} \\
&\le n\kappa \max(\|r\|_{\max}, \|r'\|_{\max}) \|\mathcal{P} - \mathcal{P}'\|_\infty
\end{aligned}
$$

where the last inequality is by (34) and $\mathcal{P} \in \mathfrak{M}(\kappa, \rho, \alpha)$.

Bounding $\|h^*(r, \mathcal{P}) - h^*(r', \mathcal{P}')\|_\infty$**.** Before we proceed to upper bound $\|h^*(r, \mathcal{P}) - h^*(r', \mathcal{P}')\|_\infty$, we introduce the connection between MDP and stochastic longest path (SLP) problem, which is the same as the connection between MDP and stochastic shortest path problem introduced in (Bertsekas, 1998; 2012), except that here we consider reward instead of cost as in (Bertsekas, 1998; 2012).

Connection between MDP and SLP. For any MDP with reward r and transition probability function $\mathcal{P}$ and an arbitrarily chosen state τ, we can define g-SLP, which is a MDP problem that has a termination state q on top of the state space of the original MDP problem. The reward function of g-SLP is given by $\tilde{r}(i, a) = r(i, a) - g$ if $i \neq q$; and $\tilde{r}(q, a) = 0$. The probability transition function $\mathcal{P}^{SLP}$ of g-SLP is given by,

$$
P_{ij}^{SLP}(a) = \begin{cases} P_{ij}(a), & j \neq \tau, j \neq q \\ 0, & j = \tau \\ P_{i\tau}(a), & j = q \end{cases} \tag{35}
$$

We denote the total reward of g-SLP problem under policy π as $\tilde{h}(r, g, \mathcal{P}, \pi) \in \mathbb{R}^n$ (excluding the reward of termination state because it is trivial). It can be shown that the optimal total reward of g-SLP happens to be $\tilde{h}^*(r, g, \mathcal{P})$ defined in Appendix B because the Bellman equation for g-SLP is the same as the modified Bellman equation (13) for the original MDP problem (Bertsekas, 2012). For more details on this connection between MDP and SLP, we refer the reader to (Bertsekas, 1998; 2012). Moreover, it is proved in (Bertsekas, 2005) that the total reward $\tilde{h}(r, g, \mathcal{P}, \pi)$ is bounded by the following:

$$\|\tilde{h}(r, \mathcal{P}, g, \pi)\|_\infty \leq \frac{n}{\rho}(\|r\|_{\max} + |g|) \tag{36}$$

Lemma 6 (Perturbation results for SLP.)**.**

$$\|\tilde{h}(r, g, \mathcal{P}, \pi) - \tilde{h}(r, g, \mathcal{P}', \pi)\|_\infty \leq \left(\frac{n}{\rho}\right)^2 (|g| + \|r\|_{\max})\|\mathcal{P} - \mathcal{P}'\|_\infty$$

Proof of Lemma 6. Let $\tilde{P}(\pi)$ denote the transition probability matrix for state $\{1, \dots, n\}$ (excluding termination state). By the definition of the total reward, we have

$$\tilde{h}(r, g, \mathcal{P}, \pi) = (I + \tilde{P}(\pi) + \dots) = (I - \tilde{P}(\pi))^{-1}\tilde{r}(\pi)(1:n)$$

where $\tilde{r}(\pi)(1:n) = (r(1, \pi) - g, \dots, r(n, \pi) - g)' = \dot{r}(\pi)$. Thus, let $Y_1 = I - \tilde{P}(\pi)$, $Y_2 = I - \tilde{P}'(\pi)$.

$$\begin{aligned}\tilde{h}(r, g, \mathcal{P}, \pi) - \tilde{h}(r, g, \mathcal{P}', \pi) &= (Y_1^{-1} - Y_2^{-1})\dot{r}(\pi)\\ = Y_1^{-1}(Y_2 - Y_1)Y_2^{-1}\dot{r}(\pi) &= Y_1^{-1}(\tilde{P}(\pi) - \tilde{P}'(\pi))Y_2^{-1}\dot{r}(\pi)\end{aligned}$$

Now, consider ∞-norm, by induced-norm definition and induced-norm's submultiplicativity,

$$\begin{aligned}\|\tilde{h}(r, g, \mathcal{P}, \pi) - \tilde{h}(r, g, \mathcal{P}', \pi)\|_\infty &\leq \|Y_1^{-1}\|_\infty \|(\tilde{P}(\pi) - \tilde{P}'(\pi))\|_\infty \|Y_2^{-1}\|_\infty \|\dot{r}(\pi)\|_\infty\\ &\leq \left(\frac{n}{\rho}\right)^2 \|(\tilde{P}(\pi) - \tilde{P}'(\pi))\|_\infty \|\dot{r}(\pi)\|_\infty\end{aligned}$$

where the last inequality is by

$$\|Y_1\|_\infty = \max_i \|Y_1 e_i\|_\infty = \max_i |\tilde{h}(0, -e_i, \mathcal{P}, \pi)| \leq n/\rho(\mathcal{P})$$

by (36) where $e_i \in \mathbb{R}^n$ is the indicator vector with the ith element as one and the rest as zeros.

Finally, we will convert the bound to the bound in the lemma statement. First we bound $\|\dot{r}(\pi)\|_\infty$ by definition above.

$$\|\dot{r}(\pi)\|_\infty \leq g + \|r(\pi)\|_\infty \leq |g| + \|r\|_{\max}$$

because $|r(i, \pi)| \leq \|r\|_{\max}$.

Then, we consider P.

$$\begin{aligned}\sum_{j=1}^n |\tilde{P}_{ij}(\pi) - \tilde{P}'_{ij}(\pi))| &\leq \sum_a \pi(a|i) \sum_{j=1}^n |\tilde{P}_{ij}(a) - \tilde{P}'_{ij}(a))| && \text{((by } |\cdot| \text{ is convex))}\\ &\leq \max_a \|P(a) - P'(a)\|_\infty\end{aligned}$$

□

Bounding $\|h^*(r, \mathcal{P}) - h^*(r', \mathcal{P}')\|_\infty$. Now we are ready for the upper bound. We let $g^* = g^*(r, \mathcal{P})$ and $(g^*)' = g^*(r', \mathcal{P}')$, and let π^* and $(\pi^*)'$ denote the optimal average reward policy for MDP $(r, \mathcal{P})$ and $(r', \mathcal{P}')$ respectively. By the equivalence between the two Bellman equation systems introduced in Appendix B, we have

$$\begin{aligned}h^*(r, \mathcal{P}) &= \tilde{h}(r, g^*, \mathcal{P}, \pi^*)\\ h^*(r', \mathcal{P}') &= \tilde{h}(r', (g^*)', \mathcal{P}', (\pi^*)')\end{aligned}$$

Thus, for each $i \in S$,

$$\begin{aligned}
& h^*(r,\mathcal{P})(i) - h^*(r',\mathcal{P}')(i) \\
& \leq \tilde{h}(r,g^*,\mathcal{P},\pi^*)(i) - \tilde{h}(r',(g^*)',\mathcal{P}',\pi^*) && \text{(((}\pi^*)' \text{ maximizes } \tilde{h}(r',(g^*)',\mathcal{P}',\cdot)(i))) \\
& = \tilde{h}(r,g^*,\mathcal{P},\pi^*)(i) - \tilde{h}(r,g^*,\mathcal{P}',\pi^*)(i) + \tilde{h}(r,g^*,\mathcal{P}',\pi^*)(i) - \tilde{h}(r',(g^*)',\mathcal{P}',\pi^*)(i) \\
& = \tilde{h}(r,g^*,\mathcal{P},\pi^*)(i) - \tilde{h}(r,g^*,\mathcal{P}',\pi^*)(i) + \tilde{h}(r-r',g^*-(g^*)',\mathcal{P}',\pi^*)(i) && \text{(by def. of total reward)} \\
& \leq \tilde{h}(r,g^*,\mathcal{P},\pi^*)(i) - \tilde{h}(r,g^*,\mathcal{P}',\pi^*)(i) + \eta(\|r-r'\|_{\max} + |g^* - (g^*)'|) && \text{(by (36))} \\
& \leq \eta^2(|g^*| + \|r\|_{\max})\|\mathcal{P}-\mathcal{P}'\|_\infty + \eta(\|r-r'\|_{\max} + |g^*-(g^*)'|)
\end{aligned}$$

where the last inequality is by Lemma 6. Notice that $|g^*| \leq \|r\|_{\max}$, and also by symmetry, we have

$$\begin{aligned}
& |h^*(r,\mathcal{P})(i) - h^*(r',\mathcal{P}')(i)| \\
& \leq 2\eta^2 \max(\|r\|_{\max}, \|r'\|_{\max})\|\mathcal{P}-\mathcal{P}'\|_\infty + \eta(\|r-r'\|_{\max} + |g^*-(g^*)'|) \\
& \leq 2\eta^2 \max(\|r\|_{\max}, \|r'\|_{\max})\|\mathcal{P}-\mathcal{P}'\|_\infty + \eta\|r-r'\|_{\max} + \eta[\|r-r'\|_{\max} + n\kappa \max(\|r\|_{\max}, \|r'\|_{\max})\|\mathcal{P}-\mathcal{P}'\|_\infty] \\
& = 2\eta\|r-r'\|_{\max} + (2\eta^2 + \eta n\kappa)\max(\|r\|_{\max}, \|r'\|_{\max})\|\mathcal{P}-\mathcal{P}'\|_\infty
\end{aligned}$$

which concludes the proof.

E.3. Proof of Lemma 3

We will use the notion of $\tilde{h}(r,g,\mathcal{P},\pi)$ defined in Appendix E.2, which means the total reward for the associated g-SLP problem for MDP $(r,\mathcal{P})$ under policy π. We will use the fact that $\tilde{h}^*(r,g,\mathcal{P}) = \max_\pi \tilde{h}(r,g,\mathcal{P},\pi)$, and we let the maximizing policy be π^*; similarly, $\tilde{h}^*(r',g',\mathcal{P}') = \max_\pi \tilde{h}(r',g',\mathcal{P}',\pi)$ with the maximizing policy being $(\pi^*)'$. Then, we have for any $i \in S$,

$$\begin{aligned}
& \tilde{h}^*(r,g,\mathcal{P})(i) - \tilde{h}^*(r',g',\mathcal{P}')(i) \\
& = \tilde{h}(r,g,\mathcal{P},\pi^*)(i) - \tilde{h}(r',g',\mathcal{P}',(\pi^*)')(i) \\
& \leq \tilde{h}(r,g,\mathcal{P},\pi^*)(i) - \tilde{h}(r',g',\mathcal{P}',\pi^*)(i) && \text{(((}\pi^*)' \text{ maximizes } \tilde{h}(r',g',\mathcal{P}',\cdot)(i))) \\
& = \tilde{h}(r,g,\mathcal{P},\pi^*)(i) - \tilde{h}(r,g,\mathcal{P}',\pi^*)(i) + \tilde{h}(r,g,\mathcal{P}',\pi^*)(i) - \tilde{h}(r',g',\mathcal{P}',\pi^*)(i) \\
& = \tilde{h}(r,g,\mathcal{P},\pi^*)(i) - \tilde{h}(r,g,\mathcal{P}',\pi^*)(i) + \tilde{h}(r-r',g-g',\mathcal{P}',\pi^*) && \text{(by def. of total reward)} \\
& \leq \eta^2(|g| + \|r\|_{\max})\|\mathcal{P}-\mathcal{P}'\|_\infty + \eta(\|r-r'\|_{\max} + |g-g'|) && \text{(by Lem 6 and (36))}
\end{aligned}$$

By symmetry, we have

$$|\tilde{h}^*(r,g,\mathcal{P})(i) - \tilde{h}^*(r',g',\mathcal{P}')(i)| \leq \eta^2(\max(|g|,|g'|) + \max(\|r\|_{\max}, \|r'\|_{\max}))\|\mathcal{P}-\mathcal{P}'\|_\infty + \eta(\|r-r'\|_{\max} + |g-g'|)$$

which concludes the proof.

F. Supplemental experiment description

Problem introduction. The energy cost of data centers has increased significantly in recent years. Much efforts have been spent on the power management of data centers. One method is switching off idle servers to save energy because studies have show that an idle server consume around 70% of the server's maximum energy consumption. However, there are other factors that should be considered while switching off servers, e.g. the quality of service, the setup costs when switching servers on. Moreover, the electricity prices and job arrival rates are changing with time, so a static management policy may be far from being optimal. In contrast, our online algorithm can be naturally applied to this problem.

Experiment Setup. Consider $n = 100$ servers: 50 high servers, 50 low servers. Servers are controlled by clusters. Each cluster has 10 servers of the same type. Thus, there are $N_h = 5$ high server clusters, $N_l = 5$ low server clusters. The service rate of each server are $d_h = 3$ jobs per unit time, and $d_l = 1$ job per unit time. Consider one unit time as 5 minutes. When the server is busy, the energy compution of a high server is $e_{h,b} = 400$W, and that of a low server is $e_{l,b} = 300$W. We consider a server consumes 70% of the peak energy while being idle (i.e., the server is on but not processing any job) and consumes 80% peak energy every time when switching from off to on mode (Gandhi & Harchol-Balter, 2011). Let $e_{h,i}, e_{l,i}$ denote the energy consumption when the high and low server are idle respectively. Similarly, let $e_{h,o}$, $e_{l,o}$ denote the energy consumption when the high and low server is in the setup mode.

When jobs arrive in the data center, we consider they join a single queue and are processed by batches. Each batch consists of 10 jobs. We denote the number of batches arrived at each time as $H(t)$ and model $H(t)$ as a Poisson distribution with arrival rate $\lambda(t)$ which is estimated based on traffic trace (Reiss et al., 2011). We denote the maximum number of job batches that can be stored in the buffer as Q and let $Q = 20$, indicating that there are at most 200 jobs in the queue. When being processed, the job batch will be priorily allocated to the high-server clusters, unless all high-server clusters are busy.

The operation cost of data center consists of two parts: energy costs and QoS costs. The electricity price is based on the real data from CAISO on Jan 18 2019. The QoS costs depend on the congestion (i.e. queue length) and the job lost due to the finite buffer size. We consider the congestion cost as 0.1 cents per job waiting in the buffer/queue and job lost cost as 1 cent per job.

MDP model. Consider state as (n_h, n_l, q), where n_h is the number of high-server clusters that are on, n_l is the number of low-server clusters that are on, q is the number of job batches in the queue.

The action is (u_h, u_l), which means that the manager wants to keep exactly u_h high clusters on and u_l low clusters on.

The transition probability is

$$\begin{aligned}
&n_h(t+1) = u_h(t)\\
&n_l(t+1) = u_l(t)\\
&q(t+1) = \min(Q, \max(q(t) + H(t) - d_h n_h(t) - d_l n_l(t), 0))
\end{aligned}$$

The cost has two parts: energy cost C_e and QoS cost C_s.

For the energy cost C_e, there are three parts: i) the energy cost when the servers are busy: $p(t)(b_l(t)e_{l,b} + b_h(t)e_{h,b})$ where $b_l(t)$ and $b_h(t)$ denote the number of low and high clusters that are busy respectively; ii) the energy cost when the servers are idle: $p(t)(I_l(t)e_{l,i} + I_h(t)e_{h,i})$ where $I_l(t)$ and $I_h(t)$ denote the number of low and high clusters that are idle respectively; and setup cost: $p(t)(\max(u_l(t) - n_l(t), 0)e_{l,o} + \max(u_h(t) - n_h(t), 0)e_{h,o})$. The $b_h(t)$ and $b_l(t)$, can be computed by the following. If $q(t) + H(t) \geq d_h n_h(t) + d_l n_l(t)$, then $b_h(t) = \frac{q(t)+H(t)}{d_h}$ and $b_l(t) = 0$; otherwise, $b_h(t) = n_h(t)$, and $b_l(t) = \frac{q(t)+H(t)-d_h n_h(t)}{d_l}$. The idle cluster numbe $I_h(t), I_l(t)$ can be computed by $I_h(t) = n_h(t) - b_h(t)$ and $I_l(t) = n_l(t) - b_l(t)$.

For the QoS cost C_s, it has two parts: the cost due to queue length: $c_q q(t)$, where c_q models the cost per batch in the queue; and the cost due to lost jobs: $l(t)c_l$, where c_l denote the cost of losing one batch of jobs. For example, if losing one job is 1 cent, losing one batch is 10 cents.

Greedy On/Off. The greedy On/Off is proposed in (Gandhi & Harchol-Balter, 2011) to reduce the energy consumption of the data center by switching on and off the servers based on the current arrivals of the jobs.When a batch of jobs arrived and if there is no idle cluster, then greedy On/Off policy switches on one cluster (priorily switch on high clusters). When the cluster finishes processing the jobs and there is no job batches in the queue, the cluster of servers are switched off. For more details, we refer the reader to (Gandhi & Harchol-Balter, 2011).

Tuning the weigtht in OVIs. Under the weight $1 : w$ on energy costs and QoS costs, we run OVIs for the online MDP model introduced above with cost function $C_e + wC_s$.