# Response to the paper "Large Scale Markov Decision Processes with Changing Rewards"

We thank the reviewers for their careful review and constructive comments. Due to a lack of space we will discuss the major issues below. We will fix all minor issues and typos in the paper as suggested.

**Reviewer #1**: (*Reward Bound*) Yes, an upper bound on the rewards is necessary, which is standard in online convex optimization problems. If the rewards were unbounded, the adversary could select an arbitrarily large reward for an action that is not taken by the decision maker, leading to arbitrarily bad regret bound. (*Discount Factor*) A discount factor can easily be incorporated. Note that we consider a finite horizon problem of $T$ rounds. If the decision maker needed to incorporate a discount factor $\gamma$, it could just scale the rewards for round $t$ by a factor $\gamma^t$. (*Description of Applications and Motivation*) Due to page limit, we had to shorten out descriptions of important applications that fit our framework (line 41-43 in the paper). We note that all of the applications mentioned in our paper are proposed and discussed in the previous literature. Detailed description or references to these applications will be given in the revised version. (*Numerical Evaluation*) We can add numerical evaluation if space permits. (*Reference*) We appreciate the references and will make sure to add them into the paper.

**Reviewer #2**: (*Comparison with [13] (Dick et al. 2014)*): Our approach and that of [13] are similar on a high level as we both cast Online MDPs as Online Linear Optimization problems. However, there are main differences in how the two papers tackle large state-action spaces. The paper [13] aims to efficiently solve subproblems in each time period by using approximate projections onto subsets of the simplex. This approach requires solving quadratic problems with $O(|S||A|)$ variables and constraints. In [13], the authors suggested (paragraph before corollary 4) solving them using interior point methods, but it is well known that interior point methods not scale well in practice for extremely large scale problems. In contrast, our algorithm tries to leverage stochastic subgradients, which are much cheaper to compute. We will include a more detailed comparison with [13] in the revised paper to clarify our contribution. We also thank the review for clarifying the proof issue in [13] with the author. To be specific, we could not understand the proof of Lemma 1 in [13]: in equation (8), since $l_t$ is of dimension $|S||A|$, vectors $\mu^\pi$ and $\nu_t$ should also be of the same dimension. However, later in the same proof, they write $\|\nu_t - \mu^{\pi_t}\|_1 \leq \|\nu_t - \mu^{\pi_{t-1}}\|_1 + \|\mu^{\pi_t} - \mu^{\pi_t}\|_1 = \|(\nu_{t-1} - \mu^{\pi_{t-1}})P^{\pi_{t-1}}\|_1 + \|\mu^{\pi_t} - \mu^{\pi_t}\|_1$. In section 2.3, they specify $P^\pi \in \mathbb{R}^{|S|\times|S|}$, so $\nu_t$ must also have dimension $|S|$, which contradicts the earlier definition that $\nu_t$ was of dimension $|S||A|$. (*Motivation of problem setting.*) We believe a more general problem involves large state-action spaces, time-varying rewards, bandit feedback, and time-varying and unknown transition dynamics. This general problem is obviously very challenging. The focus on this paper was to address the first two points rigorously. In addition, in the introduction (line 41-43) we provide several applications where the rewards are unknown but the system dynamics are known, which fits to the framework of our model. Due to the space limit, we did not expand on these applications, but we will add details and clarifications in the revised paper. (*Motivation for working with occupancy measures.*) We will include a more detailed discussion to motivate our methodology. In particular, we work with occupancy measures as opposed to the $Q$ function or value functions due to the nice properties of the linear program (Eq. 2) in section 3.1. First, the feasible set is a polytope which is a subset of the probability simplex. The diameter of the $n$-dimensional simplex (measured with the negative entropy) is $\ln(n)$, which allows us to get regret bounds that depend on $\ln(|S||A|)$. Second, when the objective in that linear program is time-varying, it is intuitive to resort to existing machinery from Online Convex Optimization, in the expense that some extra work has to be done to bound the MDP-Regret.

**Reviewer #3**: We appreciate your thorough review. (*Definition of $K(t)$*) You are correct that $K(t)$, the number of PSGA iterations, is not negligible. Fortunately, each iteration is relatively cheap as they do not depend on $|S|$, $|A|$. We would like to point out that, the specified $K(t)$ bound as a polynomial function of $t$ and $T$ is an upper bound on the number of PSGA iterations. By noticing that from round $t$ to round $t+1$ the solution to (7) may not change much (only one term is added to the summation), one may be able to derive a tighter bound. We did not pursue this as we thought it would obscure the main ideas of the proof of Theorem 2. (*Definition of constant $W$*) Thank you for pointing out the inconsistency. We can fix the issue as follows. The constraint set $\Theta$ will be defined as $\{\theta \in \mathbb{R}^d_+ : \|\theta\|_\infty \leq W\}$ (notice the + in $\mathbb{R}^d_+$). Since $\|\theta\|_2 \leq \sqrt{d}\|\theta\|_\infty$, it holds that $\|\theta\|_2 \leq \sqrt{d}W$ on line 555, and we can still apply Theorem 3, adding a factor of $\sqrt{d}$ in the bound of Lemma 15. Notice that in the proof of Lemma 11, the recasting $\theta \in \Theta$ as $-\theta(i) \geq -W \; \forall i = 1, ..., d$, $\theta(i) \geq 0 \; \forall i = 1, ..., d$ is now consistent. We are also being consistent with lines 571 and 578. (*Proof of Lemma 16*) Thank you for pointing out the error in Lemma 16. It can indeed be fixed as you suggested: since $\theta \geq 0$, $\Phi$ has probability distributions as it columns, and $1^\top \Phi\theta = 1$ defines the set $\Theta^\Phi$, we indeed have $R(\Phi\theta) \leq 0$. By construction $R^\delta(\Phi\theta) \leq R(\Phi\theta) \forall \theta \in \Theta^\Phi$, it holds that $R^\delta(\Phi\theta) \leq 0$. Therefore, the term $\frac{1}{\eta}(1 + \ln(|S||A|))$ in Lemma 16 can be replaced with the tighter bound $\frac{1}{\eta}(\ln(|S||A|))$. (*Proof of Lemma 13*) Thank you, we agree with you on the proof of Lemma 13. We can stop the bounding terms earlier so that we use $\|\nabla_\mu F_t(\mu)\|_\infty \leq t + \frac{t}{\eta}\max\{|1 + \ln(\delta)|, |1 + \ln(Wd)|\}$. This will, as you point out, improve the dependence with respect to $d, W$ in Theorem 2. (*Typos and minor issues*) Thank you very much for the comments and we will include all of them in the revised paper.