Reviewer 1 We understand the reviewer concerns, but point that the *strong duality of constrained RL is in fact not a trivial result*. Indeed, though *Slater's condition* is often used in the context of convex optimization and does imply strong duality in this specific case, the constrained RL problem in (PI) is not convex, as the reviewer correctly points out. Hence, even if Slater's condition holds, it is not immediate that it has no duality gap. Moreover, what we mean by "Slater's condition" in Theorem 1 (and Proposition 1) is not that the constraints are convex, but simply that *there exists a policy* $\pi^\dagger$ *(which need not be optimal) that is strictly feasible, i.e.,* $V_i(\pi^\dagger) > s_i$ *for all* $i = 1, \dots m$. Hence, it is not an assumption on the shape of the reward but on the feasibility of the problem, which in practice is fairly mild: if the RL problem has any solution, even one that is tight, an arbitrarily small relaxation of the constraints will make this solution strictly feasible. Nevertheless, we agree with the reviewer that this and other assumptions in the paper should be better explained. For instance, *Assumption 1* simply makes sure that we are able to find good solutions to unconstrained RL problems, i.e., that the value function achieved by the learned policy is close to the optimal one. In fact, Theorem 3 implies in broad terms that *if we know how to solve unconstrained RL problems well, then we also know how to solve constrained ones*. In the case of the ϵ*-universality* assumption, note that only requires that the parametrization be able to approximate probability distributions "on average" (in TV norm), which is considerably milder than the worst case (uniform) results typically derived in the literature. Finally, the reviewer raises an excellent point that this error leads to an $\mathcal{O}(\epsilon/(1-\gamma))$ duality gap, which may be quite large for $\gamma \approx 1$. This is a fundamental result about RL that characterizes the *trade-off between the quality of the parametrization and the difficulty of the problem*. Indeed, when a lot of weight is placed on the long-term behavior of the agent, good policies must dictate actions with extreme precision as they cannot afford any mistake. Achieving good performance then requires having more precise descriptions of policies. We will clarify and make these points more explicit in the final version of the paper. As for the *constraints being defined in terms of value functions*, though other forms of probabilistic constraints can be used, the stochastic nature of MDPs obstructs the use of Theorem 1 in the presence of deterministic constraints. The reviewer is also correct in that *projections onto* $\mathbb{R}^m_+$ *involve a simple elementwise maximum.*

Reviewer 2 The reviewer is correct that the *main contribution of the work is to provide theoretical guarantees in support of existing approaches* to constrained RL, i.e., our main result is the strong duality of constrained RL for rich parametrizations. In that sense, the algorithm is indeed only described to make the results more concrete, namely, so that we can obtain Theorem 3 and show that if we know how to solve unconstrained RL problems well, we also know how to solve constrained ones. We agree that the remark in contribution 3 in the introduction does not make these points sufficiently clear and we fully intend to rewrite the abstract and introduction to make this explicit in the final version of the paper. We will also develop the discussion of existing methods more deeply in the related work section to *better give credit to existing algorithms*. If the reviewers agree that it would be appropriate to change the title, we propose to name the paper by its main contribution: "*Constrained reinforcement learning has zero duality gap*." If the reviewers have other suggestions, we would welcome their feedback.

Reviewer 3 As suggested by the reviewer, we will include a *numerical section* showcasing our results in the final version of the manuscript. Below, we display a few plots to showcase the type of numerical examples we will include. We use a simple navigation scenario for which we can obtain the optimal policy (e.g., using a Dijkstra-type algorithm) so as to illustrate the theoretical results. In this problem (Fig. 1), an agent must cross from one side of the world to another using bridges, one of which is "unsafe" (we constrain the agent to not use this bridge). In Fig. 2, we show that by solving step 4 of Algorithm 1 exactly the duality gap effectively vanishes (red curve). We also showcase a curve in which step 4 is replaced by a single policy gradient step (blue curve). Since the minimization in step 4 is done approximately, the duality gap decreases at a slower rate and will only converge to a neighborhood of zero (Theorem 3). Fig 3 displays the effect of using coarser parametrizations (i.e., larger ϵ in Definition 1) by forcing the policy to be the same over sets of adjacent states (as illustrated in Fig 1). Note that as the parametrization becomes coarser, the duality gap increases (as per Theorem 2).

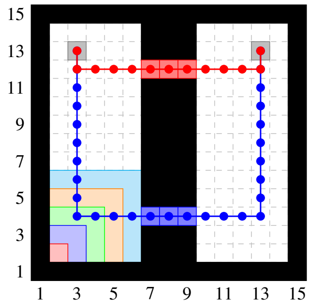

Figure 1: Safe (blue) and unsafe (red) optimal path. Coarseness of parametrization is shown on the bottom left.

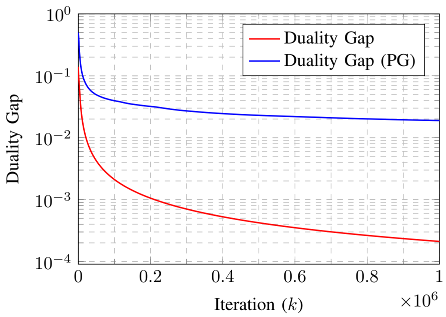

Figure 2: Gap of optimal policy and policy gradient.

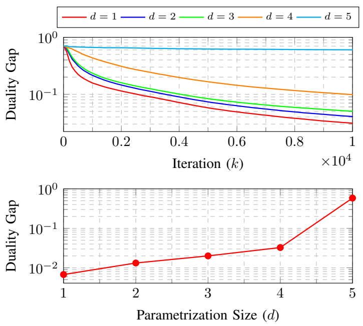

Figure 3: Effect of coarseness of the parametrization