We thank all reviewers for the valuable advice and questions. Our responses are provided below.

Reviewer #1: Thank you for your valuable suggestions. We are sorry for the typo causing confusions in Assumption 2.2. We will add the missing factor $1/T$ and change the summation to $\sum_{t=0}^{T-1}$. Furthermore, we will follow the suggestions to polish our paper and add the empirical comparisons.

Reviewer #2: **(Technical intuition behind the dual multiplier Q.)** At round t, we have obtained the constraints $\langle g_i^{t-1}, \theta \rangle - c_i \leq 0$ for all $i \in [I]$. By Lagrangian duality theory, with these constraints and the loss function $\langle f^{t-1}, \theta \rangle$, we can have a Lagrange function whose dual variables are Q_i for the i-th constraint for all $i \in [I]$ and should be non-negative. Then, the dual multiplier updating step in Eq. (5) can be viewed as a one-step dual ascent in an online setting. The operation $\max\{\cdot, 0\}$ is to guarantee the dual multiplier always non-negative. This is the main intuition behind the dual multiplier updating step. We will add this discussion in our paper.

(Importance of the loop-free assumption.) In our paper, this loop-free assumption is a standard assumption for episodic MDP. Technically, this assumption is a important for Lemma 5.1 to hold. Lemma 5.1 essentially gives the upper bound of the distance between the chosen occupancy measure and the true occupancy measure. Without the loop-free assumption, we need to develop new techniques to bound such distance. The loop-free assumption can potentially be weakened to the assumption that the underlying MDP has a fixed renewal state under any policy. We leave this as our future work.

(Typo.) In line 203, $f^{(t,\tau)}(\theta)$ and $g_i(\theta)$ should be corrected as $\langle f^{(t,\tau)}, \theta \rangle$ and $\langle g_i, \theta \rangle$. Thanks for pointing out the typo.

Reviewer #3: **(Techniques from existing works.)** We remark that the goal of this paper is to provide theoretical analysis of the constrained MDP scenario which are not fully studied before and present new bounds for this problem. In general, our paper provides a novel high probability bound for the mirror descent algorithms which is not a trivial extension of the paper Yu et al. 2017, as their paper only studied the online gradient algorithm in Euclidean space. On the other hand, our paper studies the problem without knowing the transition model, and involves the exploration step to deal with this challenge. Thus, we cannot directly apply Yu et al. 2017. Moreover, our work is beyond the simple constrained online learning setting and focuses on a more challenging constrained MDP problem. On the other hand, comparing to the previous work Rosenberg et al. 2019, we fully exploit the doubling of epoch length to obtain a sublinear constraint violation bound. This also provides a new insight on the application of epoch length doubling.

(The hyper-parameters of the algorithm require the knowledge of $\overline{\vartheta}$, B, and $\overline{\sigma}$.) As shown in Theorem 4.3, the settings of hyper-parameters α, V, λ in our algorithm do not need $\overline{\vartheta}$, B, and $\overline{\sigma}$. The constants $\overline{\vartheta}$, B, and $\overline{\sigma}$ are mainly for the purpose of theoretical analysis. Here $\overline{\sigma}$ is the only constant that are associated with a hyper-parameter ζ, namely, $\zeta \in (0, 1/(4 + 8L/\overline{\sigma})]$. And ζ corresponds to the confidence interval ε_ℓ^ζ defined as Eq. (3). In practice, we can set ζ sufficiently small, which will further guarantee that the probability $1 - 4\zeta$ in Theorem 4.3 is large. This will not affect the value of ε_ℓ^ζ too much as it only depends on a factor of $\log^{1/2}(1/\zeta)$.

(Is the dependence on parameters L, S, A tight?) As discussed in Jaksch et al., 2010, the lower bound of the regret for learning the *unconstrained* episodic MDP is $\Omega(\sqrt{L|\mathcal{S}||\mathcal{A}|T})$. The best known upper bound of the regret for the *unconstrained* episodic MDP is $\widetilde{\mathcal{O}}(L|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ (Rosenberg and Mansour, 2019a) with a gap $\widetilde{\mathcal{O}}(\sqrt{L|\mathcal{S}|})$ to the lower bound. Different from the aforementioned works, in this paper, we study a constrained MDP. Intuitively, solving the constrained problem is more challenging than solving the unconstrained problem, as the class of the unconstrained MDPs is a subset of the constrained MDPs (since the constrained problem can be reduced to the unconstrained problem when the feasible set is the whole space.). For the constrained MDP, our paper can still obtain an $\widetilde{\mathcal{O}}(L|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ regret, which matches the best known upper bound for the unconstrained MDP. But whether this is an optimal result remains to be explored. We leave the rigorous proof of the lower bound for the constrained MDP as our future work.

Reviewer #4: **(Comparisons to the result in Ding et al.,2020.)** In order to compare with the result in Ding et al.,2020, we introduce a new notation $|X|$ to denote the upper bound of the number of states at each layer. Thus, $|\mathcal{S}|$ in our paper is upper bounded by $L|X|$. Then, our regret bound and constraint violation are equivalently $\widetilde{\mathcal{O}}(\sqrt{L^4|X|^2|\mathcal{A}|T})$. In Ding et al.,2020, for the tabular case, the dimension d is $|X||\mathcal{A}|$, H is equivalently L, and K is equivalent to T in our paper. Thus, their results on regret bound and constraint violation can be rewritten as $\widetilde{\mathcal{O}}(\sqrt{L^8|X|^3|\mathcal{A}|^3T})$, which has worse dependencies on the factors $L, |X|, |\mathcal{A}|$.

(More details of solving the constrained sub-problem.) Solving the constrained sub-problem is basically in two steps: (1) perform an *unconstrained* mirror descent step, which admits a closed-form solution (Rosenberg and Mansour, 2019a); (2) project the iterate in the last step to the feasible set formed by the constraints. Note that the projection step 2 is another constrained minimization problem, whose objective is a KL divergence. Furthermore, we can reformulate this latter constrained minimization into its dual form, which is a convex optimization with *only non-negativity constraints*. Now this new problem can be efficiently solved as the constraints are much simpler than before. This algorithm is described in detail in Section 4.2 of Rosenberg and Mansour, 2019a. We will add this algorithm to our paper.