R#1: Thank you for your comments. We believe that we have addressed all of your concerns and hope that you would be open to re-evaluating quality of our contribution.
(1). **somewhat trivial*, *remains unclear**: See the review from R#3 regarding our contributions/strengths. In particular, (i) Compared to minimax optimization, e.g., [24], we exploit the problem structure (e.g., Lemma 4) to prove global convergence and incorporate function approximation error in Section 5. This approach has not been previously utilized in RL; (ii) We are the first to establish global convergence rate. Our guarantees are much stronger than the folklore stationary-point results [34, 25] and the rate for the softmax case matches the best-known one [27, 45]; (iii) Addressing the following challenges required the development of new methods that are far from being trivial: lack of monotonic improvement [3], constraint coupling, losing strong duality for the general case, and estimating gradients via samples.
(2). **natural gradient for the primal**: Our natural primal-dual method employs a *first principle thinking* to solve saddle-point problems; see, e.g., [R1]. The primal update has nice features: (i) It is well-known that NPG is closely related to popular RL algorithms, e.g., TRPO [36] and PPO [37]. Our method enables extending such RL algorithms to constrained problems; (ii) In the softmax case, the update has a concise form with multiplicative weights, which is free of any state distributions because of Fisher preconditioning; (iii) In the general case, the update naturally takes the compatible function approximation [3, 19], thereby enabling approximation of the primal update via regression with samples.
(3). **projection for the dual variable**: Since we ensure the optimal dual variable $\lambda^\star$ to be in the projection interval Λ (see Lemma 2), the projection does not impact the constraint violation guarantees. Such a projection is commonly used and it can be found in [R1] for constrained convex optimization and in [15] for CMDPs in standard LP forms. In (9b), $[\cdot]_+$ reflects the left bound of Λ and the right one is captured via ξ. Since the right bound of Λ is a constant and $\lambda^\star \in \Lambda$, the violation in Theorem 1 still vanishes sublinearly. In Theorem 2, the projection is only weakened to the left side.
(4). **Slater condition –> strong duality**: Proof of strong duality for (1) using the Slater condition can be found in [34]. In Lemma 2, we prove that the optimal dual variable is bounded under the Slater condition. These properties for CMDPs are similar to familiar properties in constrained convex optimization [R1, 7].
(5). **η's*, *dimension-free**: They are dependent. Given $T > 0$, we use η's in the theorems as multiplying positive constants. This choice does not affect the dimension-free rate and a wide range of stepsizes appears to work in experiments.
(6). **ergodicity**: We use it implicitly as in the PG literature [3, 9] which is standard. We will add clarifications.

R#2: Thank you for your positive comments and constructive suggestions.
(1). **softmax parameterized polices**: Reasons using the softmax class: (i) it served as a warm-up for studying RL algorithms; e.g., see [3, 28, R2] and it provided a lens to interpret the compatible function approximation error [3]; (ii) it has nice analytical properties: completeness and differentiability; (iii) it induces a natural update; see (2)(ii, iii) to R#1.
(2). **Dual-LP**: Reasons why PG method avails Dual-LP: (i) it's simple to apply with theoretical guarantees, e.g., [3, 9]; (ii) it's easy to deal with large state-action spaces via policy parametrization, e.g., neural nets [36, 25]; (iii) it directly optimizes/targets the value functions of interest; (iv) it's handy for estimating gradients via simulations of the policy.
(3). **Theorem 6 –> the main text*, *discounted sum**: We will add Theorem 6 before Theorem 1. Randomly terminating with probability $1-\gamma$ gives an unbiased estimate of the infinite discounted reward; see Appendix F or [3, 48] for proof.

R#3: Thank you for recognizing the contributions/strengths of our paper and for providing valuable comments.
(1). **distribution mismatch**: It is expected: Theorem 1 is free of distribution mismatch ratios because any state distribution shifts are canceled by Fisher preconditioning; see Lemma 4 or NPG [3]. In contrast, such a cancellation fails for PG [28] or regularized NPG [R2]. Theorem 2 generalizes NPG via compatible function approximation (FA) that yields a state-action distribution shift. Non-zero FA errors lead to a natural dependence on distribution ratios.
(2). **policy smoothness**: Assumption 2 also holds for the linear softmax policy with bounded feature mappings.
(3). **non-uniform PL**: Our NPG results are independent of the non-uniform PL in PG analysis [28]. We agree with your assessment that it is useful and promising. Our ongoing work proves the convergence of PG for CMDPs in LP forms.

R#4: Thank you for your positive comments and insightful questions.
(1). **global convergence**: When the policy class has limited expressiveness, Theorem 2 establishes convergence up to a function approximation (FA) error. Such an error reflects the expressiveness of the parametric policy class. When the parametrization is rich enough, e.g., tabular softmax policy, the associated FA error can be zero, and Theorem 2 establishes the usual 'global convergence.' Even though FA errors enter into the bounds, our theory is still stronger than the folklore stationary-point results in the FA setting; see, e.g., [25].
(2). **empirical and theoretical result comparison**: We provide plots for comparison in Appendix J. Sample-based algorithms converge slower, i.e., more gradient evaluations are needed to reach the same accuracy or stationarity. This verifies the effect of parameters K and ϵ_{approx} in Theorems 3 and 4 and yields loose bounds on optimality/violation gaps.
(3). **different dependency**: In contrast to optimality gaps in Theorems 2 and 3, the effect of ϵ_{approx} vanishes in constraint violations for a large value of T. In Appendix J, we also observe that constraint violations usually converge quickly. We will add this observation in the revised version and provide explanations on other dependencies.

[R1] Nedić, A. and Ozdaglar, A. "Subgradient methods for saddle-point problems." *JOTA*. (2009).
[R2] Cen, S., Cheng, C., Chen, Y., Wei, Y. and Chi, Y. "Fast global convergence of natural policy gradient methods with entropy regularization." *arXiv preprint arXiv:2007.06558*. (2020).