Response to Reviewer #2:

(C1) *Theorem 3 holds in tabular cases, and requires the policy class be convex. It's not clear if there are any other meaningful examples.* Theorem 3 doesn't require the policy class π_θ be convex, but it requires Θ and $\lambda(\Theta)$ being convex and a bijection between Θ and $\lambda(\Theta)$. With proper regularity condition on the loss function, one could still manage to prove the result for soft-max policy, which would require a case-specific proof and limiting argument (since it does not satisfy AS1). However, in this paper, we do not wish to complicate the simple and clear form of Theorem 3 as well as its proof.

(C2) *Assumption 1 requires the Jacobian has bounded eigenvalues, thus it fails for softmax policies (the derivatives can vanish). The authors should emphasize this.* Yes, we agree with reviewer's comment. More precisely, we will add the following remark in the revised paper:"It is worth noting that the AS1 implicity requires the minimum singular value of the Jacobian matrix $\nabla\lambda(\cdot)$ to be bounded away from 0 and the convex parameter set Θ to be compact. The result does hold for tabular soft-max policy if Θ is restricted to a compact subset of the orthogonal complement of the all-one vectors, but it doesn't hold for general soft-max parameterization unless there is additional regularization. It remains future work to understand the behavior of PG method under a broader family of policy parameterizations."

(C3) *How Theorem 1 would work with an extremely large state space.* Regardless of how large the state space is, the convergence rate of gradient estimates is only determined by properties of F (Theorem 2). To make solving (13) more computationally efficient, one could handle the high-dimensional z using additional/compatible function approximation, though this will induce approximation error (depending on specific choices of F) and will require further analysis.

(C4,5) *Cite "Reinforcement Learning via FR Duality." Describe why the paper's approach offers advantages over [18].* Thanks, we will cite the duality paper with discussions. The max-entropy method [18] alternates between density estimation and a planning oracle, and it seems limited to tabular problems and hard to directly work with large state space. In contrast, we focus on understanding the impact of policy parameterization which offers the potential to handle a larger state space, and our work provide a complementary alternative. See also response to Reviewer #3 (C2).

Response to Reviewer #3:

(C1) *How to leverage this work to develop more efficient RL algorithms? Or the intended outcome is a deeper understanding of the setting without particular practical upsides?* Both. While we focus on the fundamental optimization theory for RL with general utility, our approach can also yield simpler algorithms for a broad range of RL tasks such as efficient exploration or risk-sensitive policy search. Developing more practically efficient algorithms will require a case-by-case investigation for specific utilities in future work (for example entropy and barrier risk have different properties and probably will need to be handled slightly differently).

(C2) *Compare with CVaR policy optimization [e.g, C&G 2014] or MaxEnt policy optimization [Hazan et al., 2019]?*

C&G 2014 considers cost minimization subject to a CVaR constraint and follows a primal-dual gradient method that uses three timescales. In comparison, our approach exploits the hidden convexity of the CVaR constraint in λ and offers an alternative approach. Compared to [Hazan et al., 2019] which focused on tabular MDP and requires a planner oracle, we propose a method of direct policy search that allows parametrization for handling large-scale problems. This makes the setting we consider, as well as our algorithms, more suitable for practical use. See also response to Reviewer #2 (C5).

(C3) *Interpretation of the reward term z.* The entity z, instead of being observed directly from the environment, may be interpreted as the "shadow reward" derived via the Fenchel conjugate in Theorem 1. We use the term shadow reward because it plays the algorithmic role of a reward function although it is not. This is similar in spirit to shadow prices in constrained optimization/resource allocation. In a way, our PG estimation algorithm is learning the shadow reward simultaneously while it estimates the gradient.

(C4) *The reported results are restricted to stationary Markovian policies.* This is a common choice for cumulative rewards objective, since it is well-known that this policy space suffices. Is it also the case for general utilities?

Excellent question! Stationary policies are indeed sufficient, because the set of occupancy measures generated by any policy is the same as that of generated by stationary policies [Put14, Hazan etal 19]. Another way to show this is to note that $\max_\pi R(\pi_\theta) = \max_\pi \min_z V(\pi; z) - F^*(z)$. By leveraging the hidden convexity in the λ space, strong duality holds between λ, z, thus one can swap "min" and "max". Then for any fixed z the best-response policy always solves a standard MDP, therefore it suffices to focus on stationary policies and there's zero gap. We will be happy to add this argument to the final paper.

(C5) *Figure 1: the curves are function of the number of samples, episodes, or iterations? Also estimating the gradient for the entropy objective seems quite inefficient in practice.* Thanks for catching. The x-axis is the number of episodes. As for the entropy objective, entropy estimation by drawing samples from a distribution is known to be hard, and even the best estimate converges slowly, therefore it is expected that gradient estimation for this objective also converges slowly.

(C6,7) *Relations to [1,2,3].* Thank you for suggesting these papers. We will add discussions about them. Compared to O-REPS, our algorithm enjoys the fact that they are policy gradient methods and can be implemented flexibly with parametrization, having the potential of being applicable to a larger state space.

(C8) *How gamma and delta terms can be avoided when computing the gradients of z and x in equation (18,19)?*

In (18,19), γ is subsumed into terms involving $F^*(z)$ and Q^{π_θ}, and δ is let to go to zero.

Response to Reviewer #4: We thank the reviewer for the positive feedback and recommending the paper for acceptance.