We thank the reviewers for careful reading and valuable comments.

Additional experiments: We emphasize that our main contributions are theoretical rather than empirical, and the contributions are novel and substantive. As the reviewers acknowledged, there are practical implications and we are undertaking systematic experiments as follow-up. For example, Figure 1 gives an additional experiments on an OpenAI Gym Algorithmic task, showing notable improvements of escort over softmax in a more complicated sampled-action setting. Here, REINFORCE PG is used and policies are parameterized by recurrent neural networks with 256 hidden LSTM units.

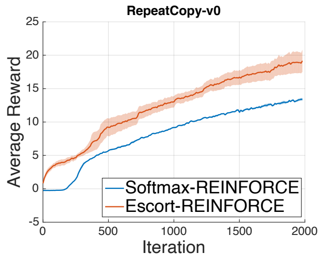

Figure 1: *Gym RepeatCopy-v0*

R1: (**i**) Note that the escort is differentiable for $p \geq 2$. (**ii**) The y-axis is $\log T$ such that $\pi_{\theta_T}(a^*) \geq 0.99$. (**iii**) In Thms 2&3, the constant $K^{1/p}$ becomes worse when $p \to 1$.

R2: (**i**) To our knowledge, our results are novel in both RL and SL, although RL is our main motivation and focus. (**ii**) For $p \geq 2$ we do not observe jitter with the escort, due to smoothness. (**iii**) It is not our intent to claim escort is uniformly better than softmax (or the best possible), but by showing its provable benefits we reveal an under-explored opportunity that can hopefully inspire future work. That said, we will add a discussion to potential benefits of softmax in the final version. (**iv**) Uniform initialization to $1/K$ is common both in the RL literature and beyond. We appreciate it if the reviewer has a suggestion for reference. While in Fig. 6(a)-(b), SPG plateaus even if the initialization is nearly uniform.

R3: *NRD paper:* Thanks for pointing it out, which we will cite and discuss. That paper makes a similar observation that PG has an action-dependent scaling factor dependent on the current policy, which can slow down update dynamics. They focus on the interesting but different multi-agent setting, and no convergence rate analysis for PG is provided. Furthermore, (**i**) the theoretical overlap between that paper and ours is minimal; (**ii**) MD is discussed in Remark 2, and Fig. 3(b,c) show MD is similar to $p = 2$ in that case; (**iii**) the NRD paper did not contemplate the escort transform, which requires non-trivial technical novelty to analyze. We will give a detailed comparison and discussion in the final version.

ICML2020 paper of Mei et al.: That paper only makes an observation, without a deeper look at the causes, analysis or solutions to fix it. These are the main contributions of the our submission.

Experiments: We tried EPG in sampled-action versions of the experiments, with or without Adam, and it outperformed SPG. These results and also details for hyper-parameter search (mainly for learning rate and batch-size) will be added.

Why the escort transform: This transform is a natural choice due to its simplicity, and has a history in the physics literature [2]. Empirical evidence in SL shows that escort with $p = 2$ performs better than softmax in MNIST and CIFAR-10 [5]. We will add more intuition and motivation to Sec. 3 in the final version.

Łojasiewicz coefficient: Given current analysis, we believe that for any function that satisfies properties like Lemmas 2–3, a conclusion like Thm. 2 follows by a similar derivation, so a simple answer to the question is yes. We can discuss a more general characterization, but this is of independent interest and is outside the scope of of the paper.

Learning rate intuitions: We appreciate the attempts to understand the effect of learning rate on convergence of SPG and EPG and we went through a similar process. We will be happy to discuss these in the paper. Unfortunately, none of the alternatives suggested appear to be viable in their current forms:

Learning rate analysis for SPG: (**i**) A large/unbounded η is not guaranteed to produce monotonic improvement, which is a basic convergence requirement; e.g., [1, Thm. 5.1] requires $\eta < 1$. (**ii**) Thm. 1 applies to a provable update, since SPG with $\eta = 0.4$ has $O(1/t)$ rate [11, Thm. 2]. (**iii**) For MDPs, the coefficient is $\min_{s \in \mathcal{S}} \pi_{\theta_t}(a^*(s)|s)$ (Line 498 in the appendix, also [11, Thm. 4, Eq. (317)]), which cannot be calculated even from the true Q^π-values since a^* is determined by Q^* not Q^π. (**iv**) For the alternative learning rate approaches suggested ($1/\pi_{\theta_t}(a^*)$ and ℓ_2-norm normalized SPG), we conducted experiments similar to Fig. 5(a). Both suggestions fail for $K = 50$ or 100 (plateaus after 10^5 iterations).

Learning rate analysis for EPG: It is easy to establish that $\|\theta_t\|_p$ is finitely bounded from above and below. First, $\|\theta_t\|_p \geq |\theta_t(a^*)|$, and Eq. (5) implies $|\theta_t(a^*)| \geq |\theta_{t-1}(a^*)| \geq |\theta_1(a^*)|$. Second, $\|\theta_t\|_p^p$ keeps decreasing after $\pi_{\theta_t}^\top r > c$, where $c < r(a^*)$ depends on reward and initialization. Therefore the EPG learning rate cannot be "arbitrarily high". The 4-room and MNIST experiments work reasonably well using constant learning rates like 0.01.

Clarity & correctness: (**i**) We will fix the typos and clarify the descriptions. (**ii**) At a (sub-)goal state, the agent can step away then step back to receive rewards. "Sub-goal" just means goals with lower rewards. (**iii**) For escort initialization, we use $\theta_1(a) = \pi_{\theta_1}(a)^{1/p}$ for all a. (**iv**) ξ is defined in [11, Def. 1], which impacts the rates [11, Lemma 8, 16].

R4: (**i**) In 4-room and MNIST, we use Eq. (3), where θ is the output of the last hidden layer. (**ii**) Escort becomes closer to softmax when p increases (Remark 1). In Fig. 3(b), $p = 2$ is far from the sub-optimal corner than $p = 4$, and in Fig. 3(c), $p = 4$ has a short "plateau" due to getting close to the sub-optimal corner. We will discuss further in the final version.