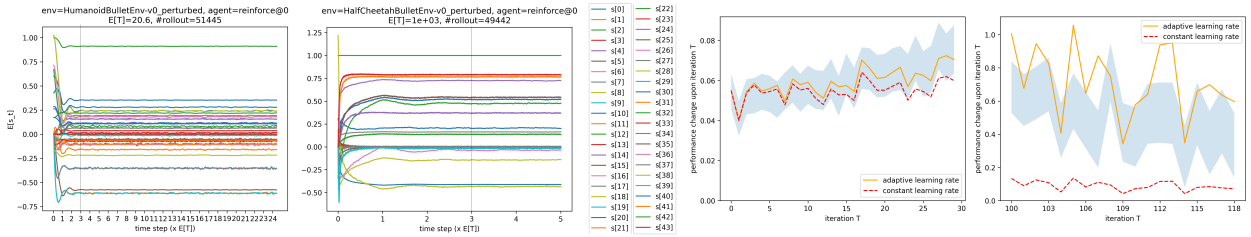
We thank all reviewers for the many helpful comments to our paper. A common suggestion from most reviewers is to further elucidate how our theory helps develop better RL algorithms. We highlight two such algorithmic ideas below.

Recursive perturbation. Figure 1(e) in the appendix showed that in the Humanoid environment, the marginal distributions of s_t may take as many as $20 \cdot \mathbb{E}[T]$ steps to converge to the steady state (where $\mathbb{E}[T]$ is the average episode length), while Figure 1(f) showed that the marginal distributions of HalfCheetah do not converge at all (because $\mathbb{E}[T] \equiv 1000$ in HalfCheetah, a situation where the marginal distributions will be strongly phased, as discussed in Section 4). These observations illustrate the *necessity* to apply the ϵ-perturbation trick in real-world RL practice.

In Section F.2 of the appendix we introduced an augmented variant of ϵ-perturbation. The idea is to recursively apply the one-shot perturbation to already perturbed models, which in the limit is equivalent to a perturbation with self-looping probability ϵ. We empirically found that recursive perturbation with $\epsilon = 1 - 1/\mathbb{E}[T]$ appears to force convergence of the marginal distribution at $t = 3\,\mathbb{E}[T]$ in *various* environments. Figure 2(b) in the appendix already showed such a result on a synthetic but challenging environment. Figure (A) and (B) below further demonstrate the same observation in Humanoid and HalfCheetah. Comparing Figure (A) with Figure 1(e), and Figure (B) with Figure 1(f), we can see clearly the effectiveness of the recursive perturbation on these two popular RL environments.



(A) (B) (C) (D)

Episode-length-adaptive policy gradient. Our steady-state policy gradient theorem (i.e. Theorem 5) showed a proportionality factor of $\mathbb{E}_\pi[T] - 1$ between the true policy gradient ∇J_{epi} and the classic policy gradient estimator $F(\theta) = \mathbb{E}_{s,a\sim\pi}[Q_\pi(s,a)\nabla \log \pi(s,a;\theta) | s \notin \mathcal{S}_\perp]$, and we have argued in Section 5 (line 320-326) that this proportionality factor can change dramatically in practical RL training, and that while it will not change the gradient direction, using $F(\theta)$ to estimate the policy gradient is equivalent to applying a learning rate $\beta = \frac{\alpha}{\mathbb{E}_\pi[T]-1}$ to the truly unbiased estimator, where β is dynamically drifting along with the changes of episode length during the training.

Figure (C) and (D) above revealed how quickly the drifting episode length can hurt the quality of gradient estimations in the Hopper Environment. As with Figure 3 in appendix, the shaded area gives 90% confidence intervals of the true changes of policy performance in iterations with $\alpha = 5 \times 10^{-4}$. The red dotted curve is the predicted performance change by assuming the proportionality factor as a constant absorbed in the learning rate, which quickly leads to bias after only tens of iterations, as Figure C shows. The bias becomes quite significant after 100 iterations, as Figure (D) shows. On the other hand, the orange curves are the estimated performance change by following the unbiased estimator given by Theorem 5, which leads to much less bias (see Section F.3 for more details about this experiment).

Meanwhile, we try to give in this response a better **summary of the theoretical implications** of our paper:

1. **(a)** We identified two formal properties (Definition 1) in the learning environments of finite-horizon decision tasks. **(b)** We proved that these properties imply existence of nondegenerate stationary distribution *in finite-horizon tasks*.
2. **(a)** We proposed a perturbation trick (one-shot version in Definition 2, recursive version in Section F.2) and proved that the perturbed model (of learning environment of finite-horizon task) is ergodic. **(b)** The *guaranteed* (instead of assumed) ergodicity implies that few-step sampling is sound in *the perturbed model of* all finite-horizon tasks.
3. **(a)** We analyzed the Bellman equation under the steady state, which connects, for *any* function, its mean values over the state space and over the episode space. **(b)** As two special cases of 3(a), we showed $J_{epi}(\pi) = J_{avg}(\pi) \cdot \mathbb{E}_\pi[T]$ and $\rho_\pi = \mu^1_{epi}$, which in turn help unify RL formulations between continual and episodic tasks (Table 1).
4. **(a)** We analyzed the differentiated Bellman equation under the steady state, which leads to $\nabla J_{epi} = (\mathbb{E}_\theta[T] - 1) \cdot \mathbb{E}_{s,a\sim\rho_\theta}[\dots]$. **(b)** 4(a) implies two sources of bias in existing policy gradient algorithms as discussed above, i.e. the fluctuating marginal distribution in tasks like HalfCheetah and the drifting proportionality factor.

To Reviewer 3: We fully agree that continual and episodic RL are equally important, yet please note that the eight results of this particular paper (see the list above) apply only to episodic RL with finite horizon. Although we did use infinite-horizon MDP as an *approach*, the *problems* we solved seem to have limited overlap with infinite-horizon RL.

To Reviewer 4: We hope the summary above better explained what the paper accomplished. The positioning of the paper is actually general, with 6 out of the 8 result items applicable to both policy-based and value-based methods (e.g. parallel few-step sampling was used in both methods [14]). Please see Section B for more discussion on this.