We thank the reviewers for the comments and constructive feedback and we are delighted that they appreciated the clarity of the paper and the novelty of the approach. Reviewer 2 lamented the lack of experiments and Reviewer 4 the lack of new practical algorithms. While we wholeheartedly agree that these would be great additions, and we are currently working on both, we believe this current set of results to be of enough interest to the community to develop their own practical algorithms. Please find responses to specific comments below; we will update the paper to reflect this discussion and to fix all other minor points.

Reviewer 1

"In Section 2, it is assumed that the state and action spaces are finite . . ." We assume finite state-action spaces for simplicity in presentation of the proofs, but the results extend to continuous state-action spaces. The deterministic transition model is a typo, and we will update it appropriately.

"Proposition 6: I am a little confused about the notation . . . " This is a typo in Proposition 6: the remark should use the state-action improvement operator. An analogous statement does apply for the trajectory formulation (Proposition 10).

"Proposition 5: ... Are there some sufficient conditions to enforce that $Var(R) > 0$ along the whole learning process? . . . " $Var(R) > 0$ can typically be enforced by exploration strategies, which are outside the scope of this paper. However, even without these strategies, REINFORCE is guaranteed to converge to a (potentially suboptimal) stationary point using results from optimization theory.

"Proposition 2: ... There can be other fixed points?" There can indeed be other fixed points of the operator: these corresponds to suboptimal stationary points of the expected reward objective $J(\pi)$. Under certain conditions, e.g. tabular policy classes (Agarwal et al 2019), the optimal solution is known to be the only fixed point, but characterizing all fixed-points under arbitrary function approximation remains an open challenge.

Reviewer 2

We will organize the related work to be more clear, and add discussion about AlphaGo-Zero in the updated manuscript.

"It would be interesting to more empirically analyze the induced interpolated algorithm . . . on more realistic benchmarks

We agree that this would be interesting to analyze on more challenging domains, and in fact, we have several such experiments in mind. In this submission, we decided to focus on the theoretical aspects of the work and presentation of the main ideas – a more empirical work is to follow.

Reviewer 3

We are sorry to hear that you found the paper to be unclear. We appreciate your specific comments about our introduction section, and will update the paper to improve the precision of these statements and better organize the related work.

We emphasize that we are not proposing a new objective, but rather providing a new way to interpret the original policy gradient objective using ideas of policy improvement. While our work is related to the pseudo-likelihood in RL-as-inference (see discussion in Section 4.3), they differ crucially in how the objective is interpreted. Whereas RL-as-inference interprets the RL objective as inference in a graphical model, our paper re-interprets the objective as application of an improvement and a projection step. This alternative framework provides new insights about the behavior of existing policy gradient methods (e.g. Prop 5, 6) and enables the potential for new algorithms (e.g. Prop 8).

Reviewer 4

". . . the result of improvement operator cannot be implemented in a practical manner." We apologize for the misunderstanding here – a practical algorithm can in fact be created using an improvement operator that produces policies not realizable in our function class. Fundamentally, this is because we never need to explicitly represent the improved policy $\mathcal{I}(\pi)$, only the *projected* improved policy $\mathcal{P} \circ \mathcal{I}(\pi)$.

For a practical algorithm, this requirement primarily boils down to the ability to *implicitly* represent the improved policy (e.g. by weighting previously seen state-action pairs by rewards). In fact, all the existing policy gradient methods we study in this paper (REINFORCE, PPO, and MPO) use weighted implicit policies to implement improvement operators that can produce potentially unrealizable policies. We will update the manuscript to reflect this discussion.

". . . CPI also uses a surrogate function which is linear approximation term plus constant times a kind of KL divergence."

We apologize for the confusion; we will update Figure 1 with the trust-region penalty typically used with CPI. Please note though that the KL terms in the two approximations serve different purposes: in our bound, the KL term promotes closeness to the *improved* policy, whereas in CPI, it penalizes distance to the *original* policy.