1. Experimental details

Unless otherwise specified, in all experiments below we report the interquantile mean after 40 million environment steps, aggregated over 15 games with 5 seeds each; error bars indicate 95% stratified bootstrap confidence intervals (Agarwal et al., 2021).

2. Experiments with PPO

Based on reviewer suggestions, we have run some initial experiments with PPO and SAC on MuJoCo. We have not observed significant performance gains nor degradation with SoftMoE; with Top1-MoE we see a degradation in performance, similar to what we observed in our submission. We see a few possible reasons for the lack of improvement with SoftMoE:

1. For ALE experiments, all agents use Convolutional layers, whereas for the MuJoCo experiments (where we ran SAC and PPO) the networks only use dense layers. It is possible the induced sparsity provided by MoEs is most effective when combined with convolutional layers.
2. The suite of environments in MuJoCo are perhaps less complex than the set of experiments in the ALE, so performance with agents like SAC and PPO is somewhat saturated.

As mentioned, these experiments are rather preliminary, but we will continue exploring this, as we agree it would provide greater insights. We will add a discussion of our findings to the final version of the paper.

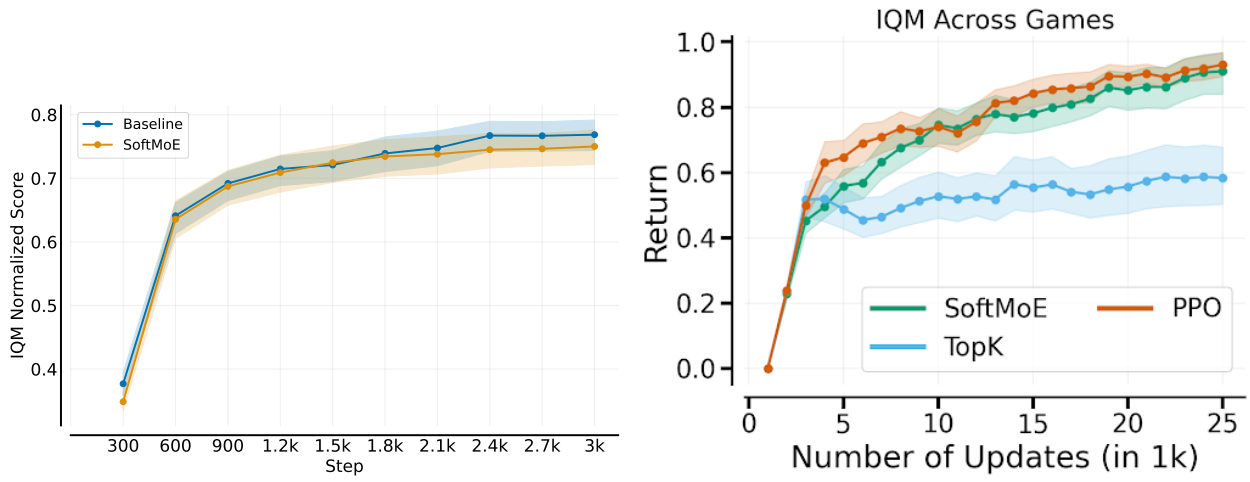

Figure 1. **Left:** Evaluating SAC with SoftMoE on 28 MuJoCo environments and **Right:** Evaluatin PPO on 9 MuJoCo-Brax environments. SoftMoEs seems to provide no gains nor degradation, whereas TopK seems to degrade performance (consistent with paper's findings). MuJoCo scores are normalized between 0 and 1000, with 5 seeds each; error bars indicate 95% stratified bootstrap confidence intervals. MuJoCo-Brax scores are normalized with respect to Jesson et al. (2023).

3. Varying Impala filter sizes

The default filter size is 3x3, and we ran experiments with and without SoftMoE using 4x4 and 6x6 filters. In both cases, SoftMoE outperforms the baseline.

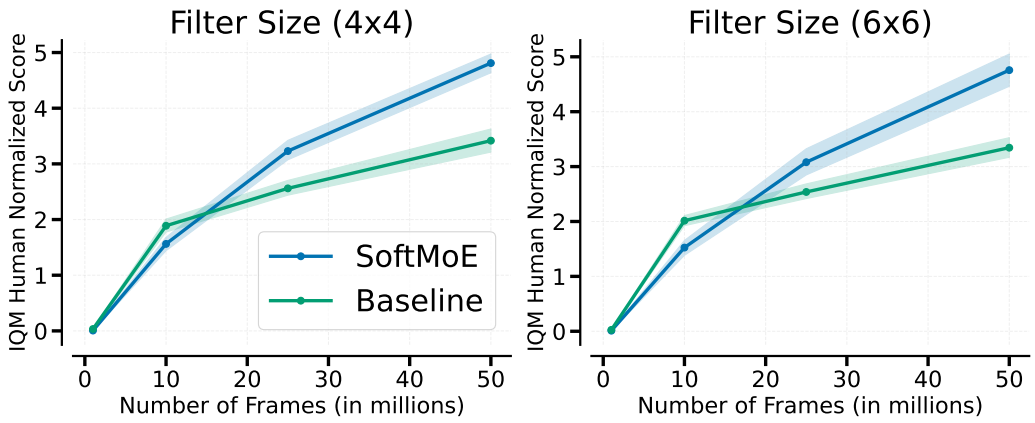

Figure 2. **Normalized performance across 20 Atari games with the ResNet architecture.** SoftMoE achieves the best results in both scenarios; default filter size (3x3) is increased to (4x4) and (6x6).

4. Measuring runtime

We plotted IQM performance against wall time, instead of the standard environment frames. SoftMoE and baseline have no noticeable difference in running time, whereas Top1-MoE is slightly faster than both.

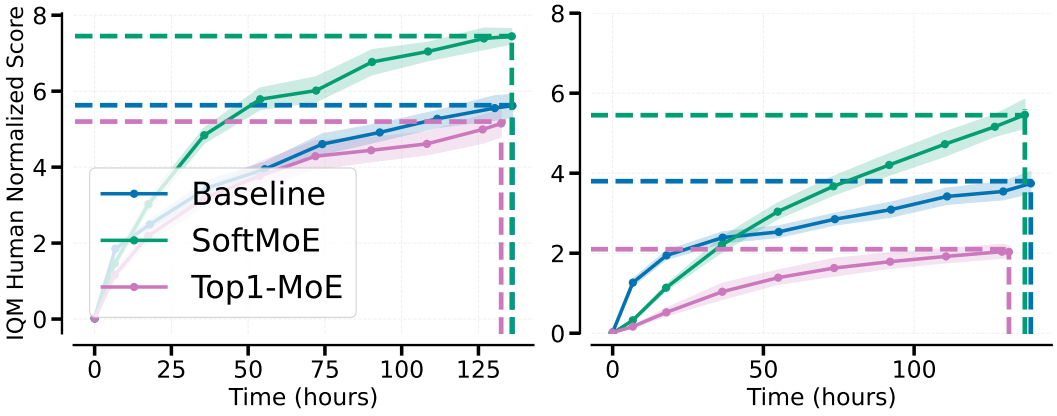

Figure 3. **Measuring wall-time versus IQM of human-normalized scores** in Rainbow over 20 games. **Left:** ImpalaCNN and **Right:** CNN network. Each experiment had 3 independent runs, and the confidence intervals show 95% confidence intervals.

References

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Jesson, A., Lu, C., Gupta, G., Filos, A., Foerster, J. N., and Gal, Y. Relu to the rescue: Improve your on-policy actor-critic with positive advantages. *arXiv preprint arXiv:2306.01460*, 2023.