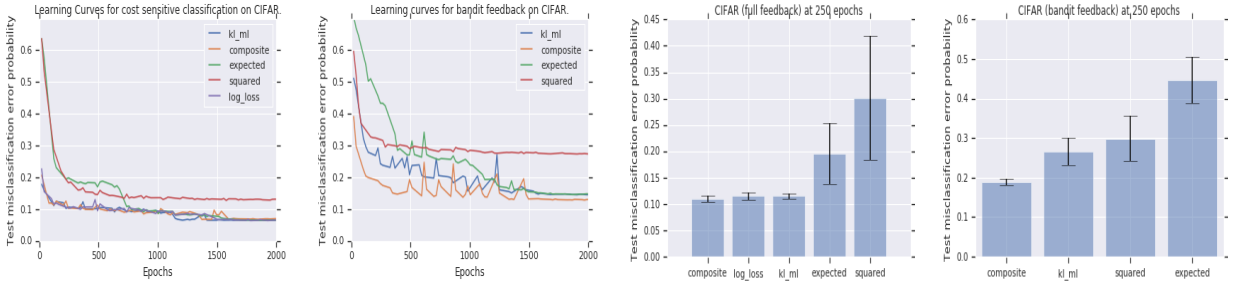
Reviewer #1 Thanks for the comments! $\hat{\mathcal{R}}(\boldsymbol{\pi})$ was defined in Equation (1), but we will make the reference clearer.

We should clarify that the theoretical results already consider out-of-sample generalization. Theorem 8 gives a proper form of generalization bound: with high probability, achieving a small value of the empirical surrogate $\hat{L}(\boldsymbol{q}, \mathcal{D})$ guarantees a bound on the **true** suboptimality gap $\mathcal{G}_\tau(\boldsymbol{f} \circ \boldsymbol{q})$. The suboptimality gap $\mathcal{G}_\tau$ in the smoothed risk $\mathcal{S}_\tau$, defined in (14), is defined in terms of full expectations over the underlying distributions $(x, \boldsymbol{r}) \sim p(x, \boldsymbol{r})$ and $a \sim \boldsymbol{\pi}(x)$. In particular, the proof of Theorem 6 explicitly accounts for how the estimator $\hat{\boldsymbol{r}}$ behaves under such a full expectation.

There are connections between this work and entropy regularized RL, but there are also distinctions. To be brief: In our scenario, entropy regularization leads to the smoothed risk (5). The connection between smoothed risk and the KL divergence (8) is known. Reversing the KL, as in (9) (without the quadratic), yields a version of maximum entropy inverse RL. A key novelty in this paper is to augment the maxent inverse RL objective with the quadratic, to achieve an *upper bound* on the original KL (8) that remains *calibrated*. In fact, (9) is the key objective we analyze, which is different from (12). This allows us to prove new generalization bounds in the form of Theorem 8. We are also able to achieve successful empirical results for **batch** policy optimization, which remains a challenge in the sequential case (Fujimoto et al. ICML-19). Finally, most work on entropy regularized RL uses split actor-critic models (or considers policy improvement in full MDP planning), whereas we achieve success with a single model that serves as both.

Reviewer #3 Thanks for the comments! We appreciate the criticism of the experimental exposition, and will improve it as suggested, including adding learning curves and standard deviations. Below are the more detailed experimental results (in terms of test misclassification) for CIFAR10, in both the fully observed (§2.2) and partially observed (§3.6) cases. We also have the same suite of results prepared for MNIST, and standard deviations for Table 1.



(a) CIFAR10 fully observed (b) CIFAR10 partially observed (c) CIFAR10 fully observed (d) CIFAR10 partially observed

"It is unclear to me if the reward estimation algorithm is actually evaluated in the experiments." Yes, Section 3.6 used reward estimation (10), with λ explained in Line 257 and other hyperparameter choices explained in Appendix 4.

"Can you comment on the increased variance demonstrated by Composite on Table 2?" To produce Table 2, hyperparameters were only chosen to maximize estimated validation reward. We do not yet know whether explicitly considering variance on validation will lead to greater statistical separation between methods.

"I find curious that [...] all the experiments consists of classification tasks "reworked" [...]." This is inaccurate: the Criteo dataset is a benchmark in this area, which has been extracted from a real online advertising challenge.

"It might also be valuable [...] to run experiments with different data size and distributions other than uniform." Yes we agree, but to clarify: Table 1 already gives results on CIFAR10 with different training set sizes and non-uniform data collection [12]. The Criteo results in Table 2 are also based on data gathered by a non-uniform logging policy [16].

"Code" Yes, we plan to release the remaining code. *"Compare to off-policy RL & BayesOpt"* The paper is already performing off-policy RL, albeit for single-step tasks given batch data. The batch scenario implies no exploration, only exploitation. A connection to general RL is given above. We will try to squeeze something in about BayesOpt.

Reviewer #4 Thanks for the comments! The baseline v is chosen as a hyperparameter on validation. It has no effect on the global optima under full expectation, but affects the variance of the empirical estimates.

F was actually defined in Line 61, and its use in Proposition 2 was consistent with this definition and subsequent occurrences. We can see how this definition was easily missed however, and we will seek to better highlight it.

Thanks for catching the typo in (8). *Line 127:* You are right, the statement should be "$\hat{\mathcal{R}}(\boldsymbol{\pi})$ yields the highest training error on both MNIST and CIFAR10, the highest test error on CIFAR10, and the second highest test error on MNIST."

As far as we are aware, *ImageNet* has never been used as a testbed in this area. It is quite expensive and would only add another "classification" data set to the evaluation. A non-classification based domain might be more interesting.