We would like to thank the reviewers for their helpful feedback, and for their positive comments regarding the originality and significance of our new objective, as well as the clarity of its derivation and exposition. Please find our responses to specific reviewer questions or comments below.

Analytical derivatives of MDP value (@R1)

The probability distribution over states at time t is $p(s_t) = (P_\pi)^t P_0$, where P_π is the state transition matrix defined by the MDP's transition function and the (tabular) policy π, and P_0 is the initial state distribution. The mean reward at time t is $r_t = R^T p(s_t)$, where R is the vector of per-state rewards. Then $V^\pi = \sum_{t=0}^{\infty} \gamma^t r_t = R^T \sum_{t=0}^{\infty} (\gamma P_\pi)^t P_0 = R^T (I - \gamma P_\pi)^{-1} P_0$. This V^π is differentiable wrt π and may be easily computed with automatic differentiation packages. We will clarify this formulation in the paper.

Relationship to Conjugate MDPs (@R3)

Conjugate MDPs are an interesting framework for learning useful abstractions. In their original formulation (and in the most obvious extensions to deep actor-critic methods), only first-order derivatives would be required. However, we could imagine a similar framework in which the co-agent is optimised with some awareness of the learning process of the primary agent (e.g. is optimised such that the primary agent performs well after some additional gradient-based learning), which could make use of higher-order derivatives and therefore our new objective. We regard this as an exciting avenue for future work, and can certainly discuss further potential applications in the paper.

Figures

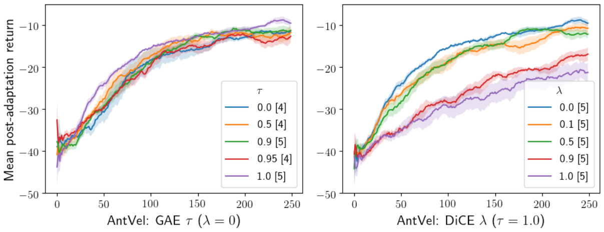
Legibility (@R3). Thank you for your feedback – we will improve the legibility of all figures with attention to B&W.
Figure 2 (@R1). "Ours" in Figure 2 is indeed using $\lambda = 1$, so as to compare against the other unbiased estimators. We will clarify this in the text.
Figure 4 (@R1). We are indeed using the best τ found for GAE when running with $\lambda = 0$. We will clarify this also.

MAML experiments

We would like to emphasise that the main contribution of our work is the derivation of an objective that produces a family of useful estimators, and that we look forward to future work to explore the full range of possible applications for higher order derivatives (which extend well beyond MAML-style meta-RL). We know of one research group already making use of Loaded DiCE for their work on multi-agent learning, and think it would be a valuable tool to share with the whole NeurIPS community. To respond directly to R3, we do not feel that wanting more (unspecified) experiments justifies a score of 4, when the review does not contain a single other substantive criticism.

That being said, we do want to evidence the practical utility of our objective, and will address some of the specific comments on the final part of our empirical study here.

Half Cheetah (@R1). Half Cheetah is indeed a somewhat limited benchmark. We chose it because it trains quickly, and the base algorithm required no additional tuning to work out of the box (unlike other environments, where we found existing implementations would perform unreliably or require very large amounts of training). We ran some additional experiments on the Ant MuJoCo domain, with results shown here. In this domain, λ is a more important factor than τ. We also note that, following existing implementations, our value function is a simple linear function of some handcrafted features. We expect that the utility of our objective will be more pronounced as researchers use stronger (and themselves meta-learned) value functions and tackle harder domains, but this will require further research in other aspects of meta-learning for RL. We also believe that such research will be facilitated by the use of our objective!



Baselines (@R1). The proposed baseline method (E-MAML) is mathematically equivalent to DiCE. DiCE is simply an objective that makes the calculation of the E-MAML estimator much simpler, and generalises to other applications of higher-order derivatives (rather than just MAML-RL). As such, our proposed method, which generalises DiCE, will have equal or better performance. Since our method also generalises LVC, we believe the range of comparisons is fairly comprehensive with respect to the choice of higher-order estimator. We will include for completeness experiments without any baseline for variance reduction (which substantially underperform in all cases – e.g. no more than -70 achieved on the AntVel task inset above).

It would certainly be interesting to explore the utility of our objective by combining it with a more elaborate setup involving more complex value functions and a more advanced base learning algorithm (like PPO). However, we wanted to keep the setup simple to focus on our key point: trading off bias and variance in estimators of higher-order derivatives is important, and our new objective gives a principled and straightforward way to do so.