We thank the reviewers for their helpful comments and address the individual comments below.

Reviewer #1. Sufficiency in contribution. 1) BestDICE *is* novel and outperforms all existing DICE estimators, as shown in Fig. 1 below (variants of this figure appeared in the original submission where legends are regularization configurations rather than estimator names). 2) We also derived a comprehensive bias analysis for an expanded family of DICE estimators (Theorem 2 in the main text and Table 1 in the appendix), whereas previous DICE papers only show a particular algorithm being (almost accidentally) unbiased. Theorem 2 and Table 1 present a foundation for future distribution-based OPE research by providing theoretical guarantees for the choices of estimators and regularizers. **Objectives in Theorem 1.** The objectives connect the LP solution (Q^π or d^π) to the policy value $\rho(\pi)$, which is what OPE ultimately cares about.

Reviewer #2. Specialization. While this work focuses entirely on OPE, we believe it is also a strength, given the widely recognized importance of the OPE problem and the current proliferation of proposed algorithms. Indeed, our regularized Lagrangian formulation provides a novel unification, which shows that many of these algorithms are actually obtained simply by choosing alternative regularizations. **Direct and recovered implementation.** The current ecosystem of open-sourced DICE implementations is unfortunately fragmented and incomplete. A key empirical contribution of this work is indeed to provide a unified implementation of all DICE algorithms, where we have also verified that our implementation reproduces the results reported in previous DICE papers. (Our open-sourced code has already been released, but we need to suppress any links to preserve review anonymity.)

Reviewer #3. There are several misunderstandings and inaccuracies in this review. **1)** *"This paper proposes an off-policy evaluation method based on offline historical trajectory data."* — The paper's goal is to provide a unified view of DICE estimators, covering both existing and new methods, and understanding the impact of various algorithmic choices. **2)** *"The experimental part verifies the effectiveness of the method."* — The experiments are not to verify any method, but to analyze the impact of regularization on solution biases and optimization stability. **3)** *"The biggest problem with this article is that innovation and contribution are not enough. For example, most of the content and formulas of the Section 2, off-policy evaluation are basically the same as those in the DualDICE paper."* — Kindly observe that Section 2 is the background section intended to set up the problem formulation and notation, and has nothing to do with the work's novelty. **4)** *" the objective function ... is identical to ... DualDICE"* — This assertion is false, since objective we consider contains $R(s,a)$ and $f(Q)$, which never appeared in the DualDICE objective. DualDICE and other recent algorithms (Sec. 3.3) can be seen as particular ablations (see, e.g., line 192). **5)** *"If regularizations are not added, it is very likely to overfit the data distribution."* — In this context, regularization was introduced to the Lagrangian to stabilize optimization (line 136), not to address overfitting. **6)** *"The only difference is that the author uses the augmented Lagrangian method"* — We are *not* using an augmented Lagrangian method, which would lead to a double sampling problem as explained in Sec 3.2. We have had to therefore develop to an alternative approach. **7)** *"Only comparing with the method of this article"* and *"not comparing with other state-of-the-art methods"* — The recent DICE estimators are considered state-of-the-art in OPE, and we compared to all such methods recoverable from the regularized Lagrangian. It is unfortunate no particular work was pointed out to support such an assertion.

Reviewer #4. Theorem 1 as OPE starting point. The constraints in the theorem characterize the dual and primal quantities (d^π and Q^π), which can be used to estimate policy value, either alone or combined (lines 171-173, with a change-of-variable $\zeta = d^\pi/d^{\mathcal{D}}$). It is thus a natural starting point for OPE, which we will make explicit in the final version. The variables $d(s,a)$ play the roles of both Lagrangian multipliers and the visitation distribution: The Lagrangian of the primal LP is $\mathcal{L} = (1-\gamma)\mu_0^\top Q + d^\top(R + \gamma PQ - Q)$, with multipliers d. By taking the gradient of $\mathcal{L}$ with respect to Q and setting the gradient to 0, we get $d = (1-\gamma)\mu_0 + \gamma P^\top d$, which is exactly what the stationary state-action visitation satisfies (Eq. (4) in the paper). **Proposal of a new method.** We did propose a new method, *BestDICE*, which outperforms others (Fig 1 below). We consider it possible to develop new meta-algorithms for model selection that can work better than BestDICE. **Experiment presentation.** We present the estimates produced during training to highlight the optimization behavior, as our major empirical contributions is to systematically apply regularizations to solve the challenging minimax optimization problem present in previous DICE algorithms.

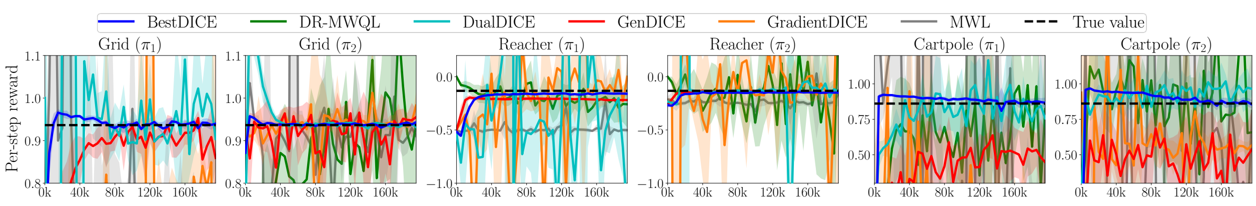

Figure 1: Comparison of BestDICE to other state-of-the-art OPE methods. Note that variants of this figure appeared in the original submission under different legends (e.g., rather than using GenDICE as the legend, we used "Dual est. + Primal reg. + Positivity + Normalization").