We thank the reviewers for constructive feedback.

Multiple reviewers indicated the set of primal distributions being optimized over is not clearly articulated. Given the dataset, we optimize over the set of distributions which dominate the empirical distribution (i.e., place positive probability on each realized datum). Despite this being a broad class, the support of the optimum matches the empirical support plus one additional (w, r) tuple (cf. Sections 4, B, and C). We've adjusted the exposition to indicate this.

Detailed responses to concrete questions or comments follow. (numbers in brackets refer to references in the paper)

Reviewer 1: *it is presented as a specialization of . . . [Kallus and Uehara, 2019]*: This is not correct—our approach is an alternative. The estimators agree with an asymptotically large number of samples, but disagree with fewer samples as shown in Figure 2. Furthermore this approach enables our primary focus on CIs for both off-policy estimation and robust off-policy learning in contextual bandits. CIs are not addressed in [Kallus and Uehara, 2019].

what each experiment is investigating: We adjusted the beginning of section 6 to introduce the experiments. Experiments investigate the quality of estimation, CIs, and learning. In estimation we compare MSE of different estimators. In CIs we compare coverage and width of different CIs. In learning we compare accuracy of learned policies with training objective our CI lower bound (or estimator) against the algorithm used in the VW system, a mature software for contextual bandit learning.

the relationship to robust supervised learning: In Section 5 we refer the reader to [7] for how empirical likelihood applies to supervised learning. For contextual bandits one needs to account for the nature of the partial feedback and cannot simply use the formulation from [7]. The relationship then is that both [7] and our work propose learning a model under the worst distribution that is still plausible given the data.

Reviewer 2: *Does the support of the distribution over (w, r) need to be finite?* No. The asymptotic distribution of the dual likelihood statistic is due to a martingale CLT [21]; bounded moments are sufficient and finite support is not required.

Figure 2 using MLE instead of EL: fixed. Furthermore in the text we now consistently refer to equation (6) as "EL".

Introduce abbreviations when first used: fixed for IPS. EMP is the actual term used by [Kallus and Uehara, 2019].

behaves as a likelihood: Dual likelihood ratios are asymptotically distributed like parametric likelihood ratios in the well-specified case [21], providing a nonparametric analogue to Wilks' theorem. We've clarified the exposition.

unclear how eq(7) defines the value: We now discuss the "maximum possible dual likelihood value given the data" and refer the reader to section 3.2.

what is meant by lower (data) scale: adjusted to read "the amount of data required for success"

intuitively centered: dropped. *pleasing functional form*: "pleasing" dropped. all other comments: fixed.

Reviewer 3: *I do not find any theoretical guarantee for the estimation of confidence intervals*. The CIs have correct asymptotic coverage and coverage errors decay as $O(1/n)$ (cf. [23] section 2.6). We now indicate this explicitly.

When you . . . build an α-confidence interval, is the real error probability smaller than α theoretically?: This is an open question. Empirical evidence (Fig. 1 right) suggests we do.

whether this method can be extended to a more general setting: we are currently researching this. Unfortunately, due to space constraints, we had to drop discussion of follow-on research in section 7.

Reviewer 4: *it could be made more accessible for non-specialised audience*. We tried to include enough background, pointers to relevant work, and an example in Section 2.

give guarantees for the size of the confidence interval and also the mse of the estimator. Our estimator asymptotically coincides with [Kallus and Uehara 2019] and the asymptotic MSE was derived in that paper. For the CI size the results of [14] (also [23] Section 13.5) show that EL enjoys a kind of optimality similar to that of the likelihood ratio test for multinomial samples [10].

w_n can only have 3 possible values: the logging policy h is ϵ-greedy and the evaluated policy π is deterministic so the 3 possible values correspond to h and π disagree, or they agree and h explores or exploits. We now state this explicitly.

referring forward to equation (7). We now discuss the "maximum possible dual likelihood value given the data" and refer the reader to section 3.2.

all other comments: fixed.