We would like to thank all referees for their appreciation of our results and the useful feedback. Below is our reply.

Reviewer 3: Thank you for your encouragement to demonstrate the significance of our findings, in response to which we have run further numerical results (Table 1 below). We will include these results in the final version of the manuscript.

Reviewer 4: Please find our responses to your comments below.

(1) Optimistic probabilities may not add up to 1. We apologize for the confusion. In problem (3), x is a single new observation that is used for our inference (we extend our setting to multiple observations in Appendix B.4). The optimization problem (3) is over probability measures $\nu \in \mathbb{B}_\theta(\widehat{\nu}_\theta)$, and as such its maximizer is by construction a probability measure. Please note that we do *not* solve problem (3) for all values of x – we solve it once for a single observation x (or once for a batch of observations x, as discussed in Appendix B.4). We will clarify our intentions in the revised version of the manuscript; thank you!

(2) One area I found a bit confusing in the paper was how this could be actually used to solve the problem of Bayesian inference referenced at the beginning of the paper. Thank you. We do indeed assume that we can sample from the conditional distribution $p(\cdot|\theta)$, which is a common assumption in Bayesian statistics and stochastic approximations. We apologize for the lack of clarity in the definition of $\widehat{\nu}_i$. In the numerical section, we construct $\widehat{\nu}_i$ as described in Assumption 5.2 to ensure the convergence guarantees of the ELBO problem. We will elaborate more in the final version.

(3) Experimental results are not convincing. We have run further numerical results (see Table 1 below), which we will include in the manuscript. We hope that this gives further support to our methods. Thank you for this suggestion!

You have also provided some more detailed suggestions. We will implement all of these in the revision; thank you!

Reviewer 6: Please find our responses to your comments below.

(1) The distribution $p(\cdot|\theta)$. We apologize for using informal notation. The unknown true distribution $p(\cdot|\theta)$ can be any measure; it does *not* need to be discrete or continuous. The maximizer $\nu^\star$ in our optimistic likelihood problem (3), in contrast, is discrete for all our ambiguity sets. The hope—which is supported by our convergence analysis as well as our numerical results—is that our discrete approximations are close to $p(\cdot|\theta)$. We will clarify this point in the revision.

(2) Optimistic likelihood estimation vs. optimism in the face of uncertainty. Our reasons for relating our paper to the optimism in the face of uncertainty literature are twofold. Firstly, we wanted to highlight that an optimistic treatment of ambiguity, which may be counterintuitive to the (distributionally) robust optimization community, has been successfully applied in a different discipline. Secondly, we did not want to claim that we are the first to exercise optimism in the face of ambiguity and give due credit to the existing literature. We will clarify this in the revision, thank you!

(3) Theorem 2.4 is missing the definition of e. e denotes the vector of all ones. We will add this, thank you!

(4) Likelihood of a set S of observations. This is a very interesting question. In our understanding, maximizing the likelihood of a set S of observations amounts to a multi-objective optimization problem, where we need to assign a 'priority' (e.g., a weight or a lexicographic ordering) to the probability assigned to each data point $x \in S$. To us, the most natural choice is to maximize the sum of the log likelihoods as done in Appendix B.4, as it maximizes the likelihood of observing all data points $x \in S$ as a batch. Moreover, $\nu^\star$ that solves problem (B.4) in the appendix is also a probability measure, which can be used to evaluate $\nu^\star(x)$ separately for any $x \in S$. We will elaborate on our intentions in Appendix B.4 in the final version of the manuscript.

(5) Experimental results are not convincing. We have run further numerical results (see Table 1 below), which we will include in the manuscript. We hope that this gives further support to our methods. Thank you for this suggestion!

Additional numerical experiments: Table 1 extends the classification results for real-life datasets in Section 6.1. Our Python source code and experimental data will be published on Github to ensure reproducibility of our results.

Table 1: Average area under the precision-recall curve for various datasets. Bold numbers highlight the best results.

	Kernel	Moment	Wasserstein		Kernel	Moment	Wasserstein
Banknote	99.05	99.99	**100.00**	Housing	80.75	81.89	**83.02**
Blood Transfusion	66.44	**71.28**	69.71	ILPD	71.75	**72.95**	70.12
Breast Cancer	98.03	**99.26**	97.35	Ionosphere	91.15	97.05	**98.96**
Climate Model	**93.82**	81.94	93.72	Mammographic	84.11	86.53	**88.28**
Cylinder	77.37	75.00	**86.59**	Pima	80.90	**82.37**	80.81
German Credit	67.84	**75.50**	75.47	QSAR	84.49	**90.85**	90.55
Haberman	72.88	70.20	**73.26**	Sonar	87.18	83.49	**94.45**
Heart	79.46	**86.87**	77.07	Thoracic	58.73	**64.73**	59.89