We would like to thank the reviewers for thoughtful feedback. We are glad to see that all 4 reviewers vote for acceptance, with the reviewers recognizing the paper to be "insightful", "novel and interesting", and to have "very important take-home-messages". Overall, the reviewers appreciated the unified view framework (R1, R3, R4, R5), our theory's relevance (R1, R3, R5), and our extensive experiments (R1, R3, R5). Below, we respond to each reviewer's concerns:

Reply to R1: Thank you for the positive assessment and for highlighting the importance of our theoretical results.

"analyze even white box predictors": This is an interesting suggestion. Perhaps by leveraging the structure of a predictor, one might be able to design more efficient estimators.

"The title feels a little broad compared to the work." Thanks for this feedback. We will review all claims, including the title, for the camera-ready version to make sure that our contributions are presented accurately.

"In the code, there are some hardcoded paths..." Sorry for this inconvenience. For the camera-ready, we will fix the hardcoded paths and provide detailed instructions for execution with expected run time.

Reply to R3: Thank you for the positive feedback and for championing our paper. We are glad that you found our message to be very important and unknown to the community. We will fix Figure 2 and typos in the final.

"how the calibration error can be determined in practice." Procedure for estimating canonical calibration error from samples do not yet exist. However, surrogate measures like Expected Calibration Error work well in practice. Estimating canonical calibration error and efficiently calibrating models remain important open problems.

Reply to R4: Thanks for the detailed review and positive assessment. We are glad that you appreciate the novelty of our unified framework and consider the label shift problem relevant to the community.

"The proposed method is not very novel." The novelty of our contribution does not include the introduction of new estimation methods. Instead, our work contributes the theoretical underpinnings for understanding popular methods. We believe that this form of novelty is equally deserving of publication as the introduction as new methods.

"The analysis of the miscalibration error is not very informative. Our theory shows two-fold benefits of calibration: (i) canonical calibration (Def 1) and an invertible confusion matrix (as BBSE requires) are necessary & sufficient for MLLS's consistency (Th 1-2, Cor 1). (ii) We bound one term in the finite sample error by miscalibration error (Lem 4).

"The miscalibration error can be large if the source and target data differ." Our bounds hold with calibration error *on the source distribution only*. While the RHS in the first part of Lemma 4 is unobservable, it is upper-bound by the calibration error on the source domain (the second part of Lemma 4). The multiplicative constant in the bounding step depends on $\max_y p_t(y)/p_s(y)(> 0)$ which also appears in existing blackbox estimation guarantees (e.g. BBSE, RLLS).

"analysis does not support that the calibration is the main reason that MLLS outperforms blackbox methods." Our finite-sample error bound hint at the importance of calibration (Th 3). Informally, the bound highlights dependency on 2 factors: (i) calibration error on source data; (ii) minimum eigenvalue of the Hessian of the likelihood (which increases with the granularity of calibration). The first factor explains the superior performance of calibrated MLLS over uncalibrated MLLS and is consistent with existing empirical observations (L 177-180). The latter factor elucidates the efficacy of BCTS-calibrated MLLS over BBSE. When MLLS is calibrated with BCTS on source data (Lem 5, L 295-298), granular calibration in practice tightens the bound (L 299-307).

"A good label shift estimation does not mean it would perform well in down-stream tasks." Existing bounds on downstream classification error (under label shift) depend on the MSE of the label distribution estimates (Th 1 Azizzadenesheli 2019[2])—thus improving these estimates translates into better downstream guarantees (Lines 83-85).

"but there is still a gap in this question, which is when and why (calibration is necessary)." Calibration on the source distribution is sufficient (Th 2). Weaker notions of calibration (e.g. marginal calibration) is insufficient to guarantee consistency (L 228-230). We agree that we only proved sufficiency and haven't fully characterize the necessity of calibration. Necessity is shown for only a restrict class of classifiers (e.g. the thresholding classifiers in Ex 1).

"BCTS method is dominating in the experiments. This mismatch is not validated in the experiments." Inspired from Alexendari 2020[1], we use BCTS as a surrogate for canonical calibration (no known method is guaranteed to achieve canonical calibration). However, in our synthetic GMM setting, we can perfectly calibrate the classifier (without BCTS) and show clearly how more granular calibration improves MLLS estimates (Figure 2, Lines 337-344).

Reply to R5: We thank the reviewer for the thoughtful review and positive feedback. We are glad that you consider our work as insightful to the ML community with interesting take-home-messages.

"Label shift is a strong assumption which in practice may not hold... how this can be plugged into more general domain adaptation." We agree that the label shift assumption is strong & unlikely to hold exactly in practice. However, we believe that rigorously understanding these idealized settings is a fundamental building block towards more complex settings. Moreover, in many important problems (e.g. medical diagnosis during an epidemic), label shift is nevertheless a useful model because prevalence is likely to change faster than conditional probabilities of symptoms given disease.