We thank all reviewers for the constructive comments! Following the suggestions, here is a list of new experiments.

(i) Added a baseline *Truth* for label shift where the importance weight is the validation/train ratio for each class:

Acc	ρ 100	Truth: 83.05 (0.58)	DIW3-L: 83.69 (1.21)	ρ 200	Truth: 79.92 (0.46)	DIW3-L: 81.38 (1.24)

Truth seems slightly worse since it is just roughly tuned; the difference is *statistically insignificant* by t-test.

(ii) Computed the ℓ_2 distances between the true weights (by Truth) and the estimated weights (by other methods):

Dist	$\rho = 200$	IW: 0.0324 (0.0010)	Reweight: 0.0321 (0.0010)	DIW3-L: 0.0166 (0.0003)

The weights by DIW3-L are *statistically significantly closer* to the true weights. It is a unit test possible under label shift, which is why we wanted the label shift experiments, but we didn't finish them before the deadline.

(iii) Tested Adam as the optimizer (we can see that the messages *do not change much* if changing SGD to Adam):

(CIFAR-10, 40% symmetric noise) Acc	IW: 44.51 (1.55)	Reweight: 72.96 (0.97)	DIW3-L: 79.80 (0.25)

(iv) TODO: add text-data experiments (based on *20Newsgroups*) and real-world experiments (based on *Clothing1M*). For 20Newsgroups, we need to choose the network architecture, which is not a trivial task. For Clothing1M, we are using the common practice ResNet-50, but the issue is that it goes in a speed of 2 days/epoch...

(v) TODO: add covariate shift experiments based on MNIST, Fashion-MNIST, or CIFAR-10. It is quite difficult to simulate covariate shifts based on benchmark datasets where we *cannot access/manipulate* $p(x)$ or $p(x|y)$, even though covariate shift should be extremely popular in the wild. Most of covariate shift papers made use of fully controlled toy data where *IW suffices* and there is *no circular dependency*.

We have two plans: first, we may try strongly regularized GANs (e.g., very early stopped); second, we may try mixup style data generation (mixing two images taken from two classes). We are thinking of ways to guarantee that the shift is precisely covariate shift and the shift is challenging to learning methods, i.e., $p(x)$ should change significantly, $p(x)$ must have a larger support in training, $p(y|x)$ cannot change at all. Any advice is welcome.

In the next version, we will definitely include all the aforementioned experiments!

Use of validation data and motivation of the problem setting (by R2 & R4 & R5) Our setting is the general IW, with the same data generation as *Learning to reweight* [38] and *MentorNet* [22] (but a different goal from them).

First of all, note that there is no *distributional assumption* about the *joint distribution shift*. Given that $p(x, y)$ can shift now, without any further information, the shift is *not identifiable*, or equivalently, the weights are *not estimable*. Hence, we should let the learning methods see some *clean validation data* coming from the test distribution. By careful algorithm design, the learning methods should be able to infer the shift by referring to the clean validation data.

On the other hand, what was done in more specific shifts (i.e., covariate shift, label/class-prior shift, and label noise) is that a specific distributional assumption is added instead of the clean validation data to the algorithm design:

- under covariate shift, only $p(x)$ can shift, and $p(y|x)$ must remain the same;
- under label shift, only $p(y)$ can shift, and $p(x|y)$ must remain the same;
- under label noise, only $p(y|x)$ can shift, and $p(x)$ must remain the same.

The learning methods should be able to infer *the parameters but not the type of* the shift, because the type is already given by the assumption. In this sense, *IW is more general* than these specific shifts without clean validation data; the *data-driven philosophy/methodology* can tackle more complex shifts where such an assumption is not obvious.

Last but not least, the small set of clean validation data can be obtained with the help of domain experts. Since it is small, it would not cost too much of the data-labeling budget. Some real-world label-noise datasets are designed in this manner, for example, Clothing1M whose clean training data can serve as our clean validation data.

Baselines are not latest for learning with noisy labels (by R3) The problem setting of the proposal is IW rather than label noise that is only a special case of IW. The latest label-noise methods *may not be applied* in the more general IW. Among the learning methods under the same problem setting (i.e., how the data look like), *[38] is already the latest*.

Class imbalance to label shift (by R4) Thanks for pointing the fact out! Yes, class imbalance refers to *cost-sensitive learning* under *no label shift*, but the goal is to maximize the AUC, which has been shown as same as to minimize the *balanced error* [Menon et al., ICML'15]. In our experiments, the test data is balanced so that the classification error is also the balanced error. That being said, our terminology following [5] is not as good as label shift. We will call it class-prior or *prior probability shift* following the book [37], since label shift may also be confused with label noise.

On training time and hyperparameter choice (by R4) Yes, we have used many tricks including regularizations and learning rate decays. In fact, not only weighted ERM with IW but also standard ERM relies on the training time and hyperparameter choice. Our contribution was to find that *complex data form is not necessarily the bottleneck of IW*.

No experiments for the hidden-layer-output transformation version in Algorithm 1 (by R5) In our experiments, *this version corresponds to "-F" methods*, and the other version corresponds to "-L" methods. Indeed, the experimental results you wanted are reported and discussed in Section 5.2 (more specifically, in Table 2 and Figure 5).