We thank all the reviewers for the helpful reviews. We respond to each reviewer's specific questions here.

To Reviewer #1:

1. *The proposed $\mathcal{L}_{DMI}$ is not limited to DNN. Why does the paper focus on DNN practice?* Yes, $\mathcal{L}_{\text{DMI}}$ is not limited to DNN and can be applied to general settings. Our theory is valid in the general settings. Our experimental setting focuses on DNN practice since the noisy-labels problem in data-driven deep learning is especially important as large high-quality data is crucial to data-driven deep learning but may be extremely hard to obtain (see intro), thus a series of works [21, 11, 6, 36] also focus on designing noise-robust functions for DNN. Applying our method to non-DNN settings can be a future direction but we believe our current results not only make a significant theoretical step in the general noisy-label learning problem but also make a significant empirical step in the DNN practice.

2. *The sufficiency of the paper's literature review and experimental comparison are kind of weak.* We have compared our method with some very recent baselines like **GCE** [36] in NIPS 2018, **LCCN** [33] in AAAI 2019 and also cited several works that focus on network structure or learning paradigm (see the second and third paragraphs in related work). We would appreciate if you could provide our missing works and will add them in our final version.

To Reviewer #2:

1. *Once the latent true variable Y is not identifiable, how to find the optimal classifier for Y?* Yes, there exists $Y' \neq Y$ such that conditioning on Y', X is also independent of $\tilde{Y}$. $Y' = \tilde{Y}$ is one example. However, first, training with Y or Y' or $\tilde{Y}$ will obtain the same optimal classifier based on our main theorem. Second, based on the information-monotonicity of $\mathcal{L}_{\text{DMI}}$, if Y is more informative about X than Y' (e.g. when $Y' = \tilde{Y}$), the optimal classifier we obtained will be more similar with Y than Y'. When we assume the ground truth is the most informative among all variables that satisfy the conditional independent assumption, the optimal classifier will be closest to the ground truth. We will add this clarification in our final version.

2. *Why it doesn't work well on small noise?* We have the clean data comparisons since $r = 0$ is the clean case. **DMI** outperforms all other counterparts except **GCE** [36] when $r = \{0.0, 0.1, 0.2\}$ in the Dogs vs. Cats dataset and when $r = 0.0$ in CIFAR-10. **GCE** does a training optimization on **MAE** [6] with some hyperparameters while sacrifices the robustness a little bit theoretically. Our next step may be to employ some training optimizations in our method.

3. *A Rate of ... papers.* Thanks for your information! We will cite them and make it clear in the final version.

To Reviewer #3:

1. *Why converting to two classes and other baselines are missing for Fashion-MNIST?* We use it as an explanation experiment to compare distance-based (we use **CE** as an example here) and information-theoretic (our method) loss functions (see the third paragraph in intro). For clean presentation, we convert to two highly imbalanced classes and only compare with **CE**. We will make it clear in our final version. We attach the comparisons with other methods here (see the following figures) and will include them in Appendix in our final version.

2. *Why are CIFAR-10 uniform, Dogs vs. Cats uniform and dog->cat missing?* Due to space limitation, we omit some under-representative experiments although we have done them. For the uniform noise, first we want to clarify that noise patterns are divided into two main cases, diagonally dominant and diagonally non-dominant and uniform noise is a special case of diagonally dominant noise. Thus, we did not emphasize the uniform noise results in the submission. We did not present the results for dog->cat since, unlike the highly imbalanced Fashion-MNIST (90% clothes, 10% bags), Dogs vs. Cats is a balanced dataset (50% dogs, 50% cats) and thus the dog->cat results are very similar to cat->dog results. We attach these results here (see the following figures) and will include them in Appendix in our final version.