We thank the reviewers for their constructive feedback. All the reviewers agree that the proposed method, multi-label contrastive predictive coding (ML-CPC), leads to mutual information estimators with lower bias at negligible computational cost, and its value is demonstrated empirically on mutual information estimation, knowledge distillation and unsupervised representation learning. Here, we answer some major questions raised by the reviewers.

[R1] Did you tune for each case separately in the knowledge distillation experiments? No. We use the default hyperparameter setting of $a = 0.0$ (which is without KD loss) and $b = 0.8$ (which is the weight of distillation loss against classification loss) for all CPC and ML-CPC experiments. This is the default setting given by the CRD implementation here[1], and we did not tune that specifically for our case. We will clarify this in the final version.

[R1] Please include variances for KD results. Thanks for the suggestion. Our results are averaged over 3 seeds, but unfortunately we were not able to include variances due to width limitations, which is around 0.2 for each case. We will include variances in the supplementary material and add a reference to this in the main paper in the final version.

[R1] The expectation symbol seems redundant. We believe this is necessary since our lower bound arguments would rely on taking the expectation; the objective value for some mini-batch is not necessarily a lower bound to the mutual information of the entire distribution.

[R1] Joint and separable critic? Our setup follows that of [Poole et al. 2019][2]. "Joint" means a single neural network $f(x, y)$ is used, while "separable" means that the inner product of two separate neural networks $g(x)^\top h(y)$ is used. "Joint" is more flexible, but requires $O(NM)$ network evaluations for each batch, while "separable" is less flexible but only requires $O(N + M)$ evaluations. We will explain this more clearly in the final version.

[R2] Why is ML-CPC expected to have higher variance than CPC? It is indeed difficult to provably compare the variance *between ML-CPC and CPC* since under finite n and m, it is hard to obtain the optimal critic in closed form, except for some special cases (although for the same α, the variances seem similar). However, we believe that for $\alpha_1 < \alpha_2$, the variance of α_1-ML-CPC tends to be larger than that of α_2-ML-CPC (and similar for CPC). Apart from the usual bias-variance trade-off argument that you mentioned, our intuition for this is as follows. The target of the (ML)-CPC objectives is to estimate density ratio $r(x, y) = p(x, y)/p(x)p(y)$ with

$$\frac{g(x,y)}{\frac{\alpha}{m} g(x,y) + \sum g(x,\bar{y})}$$

where smaller α indicates that the estimates are more flexible (i.e. being able to take larger values), so this estimate becomes larger for large $r(x, y)$ and relatively the same for small $r(x, y)$, which increases variance. We will add a discussion about this in the final version.

[R3] ImageNet results. We tried modifying the knowledge distillation code for ImageNet (since the code for ImageNet is not provided out-of-the-box), but were unable to fully reproduce the results in the CRD paper, so the results here remain inconclusive. As of now, we are still running the representation learning experiments, which takes days with 4 GPUs; we will update ImageNet results in the final version, where we would have more time to run the experiments.

[R3] Why do ML-CPC estimates surpass MI without smoothing for $\alpha = 0.001, 0.0001$? First, while our results guarantee that the expectation of batch-based estimates to be a lower bound, and we optimize an unbiased estimate to the lower bound via SGD, it is still possible for estimates of particular batches to exceed MI. Second, our theory guarantees the lower bound property for $\alpha \approx 1/128$ (since batch size is 128), so for $\alpha = 0.001, 0.0001$ we do not have guarantees. Nevertheless, we wished to demonstrate that these choices of α can still be used empirically; we will clarify this in the final version.

[R4] "In Sec 2, . . . should it be "minimize $L(g)$"?

It is indeed a maximization. Our goal is to "maximize" mutual information through unbiased estimates of a lower bound to it; the setup is analogous to maximizing the "evidence lower bound" in variational inference.

[R4] "... why $L_\alpha(g)$ is upper bounded by $\log(m/\alpha)$"? Because our critic g outputs only positive values, the division in the expectation satisfies:

$$\frac{g(x,y)}{\frac{\alpha}{m} g(x,y) + \sum g(x,\bar{y})} \leq \frac{g(x,y)}{\frac{\alpha}{m} g(x,y)} = \frac{\alpha}{m}$$

Taking the expectation over values that are not greater than $\log(m/\alpha)$ then proves the claim. We will add a lemma to support this claim in the final version.

[R1] [R2] [R3] Typos and Figures Thank you for the suggestion; we will fix these issues in the final version.

[1] https://github.com/HobbitLong/RepDistiller/blob/master/scripts/run_cifar_distill.sh#L29

[2] On Variational Lower Bounds of Mutual Information, ICML 2019