Response for “Information Theoretic Counterfactual Learning from MNAR Feedback”

We thank the reviewers for their time and for their valuable advice to our work. Due to the space restriction, we will try our best to address their major concerns but we assure that minor comments are also addressed. We would like to stress that this work proposed an information-theoretic method for dealing with MNAR feedback motivated by balancing information of representations. To the best of our knowledge, this is the first time information bottleneck (IB) is adopted in recommendation, and is adapted for **MAR-free** counterfactual learning. We provide an easy-to-implement solution to intractable IB, and verify its effectiveness empirically. It is promising since we prove that without MAR data, using CVIB can reach comparable even superior results to MAR-based methods, since MAR data is really expensive or even impossible to collect in practice.

To Reviewer #1. (1). We thank the reviewer for advice to perform evaluation using other metrics used in the literature. We agree with the reviewer and we have evaluated the methods with nDCG using 10 runs: our CVIB shows **more significant gain** over the baselines than on AUC. We will include these results in the final version of this paper. **(2).** Liang et al. (2015) considered recommendation from positive feedback alone (implicit data); however, we consider learning from both explicit data and unobserved events; Liang et al. (2016) adopted IPS, requiring MAR data as well.

COAT	MF	IPS	SNIPS	DR	DRJL	CVIB	YAHOO	MF	IPS	SNIPS	DR	DRJL	CVIB
nDCG@5	0.589	0.633	0.603	0.622	0.608	**0.663**	nDCG@5	0.633	0.636	0.635	0.659	0.652	**0.734**
nDCG@10	0.667	0.689	0.676	0.693	0.679	**0.721**	nDCG@10	0.762	0.760	0.762	0.774	0.770	**0.820**

To Reviewer #2. (1). When embedding size is 16, MF, IPS, SNIPS, DR and CVIB get AUC 0.67, 0.663, 0.679, 0.667, $\mathbf{0.703}$ on COAT; and 0.661, 0.648, 0.674, 0.650, $\mathbf{0.717}$ on YAHOO. Further deterioration appears when embedding size gets larger. It means increasing embedding size **impairs** counterfactual learning due to overfitting, while CVIB still maintains its superiority. **(2).** About MSE. We postulate that the CVIB’s weakness in MSE lies in the absence of MAR data in learning. Apart from CVIB, all other baselines utilize MAR data to *regulate the average prediction* to approaching the mean MAR outcomes, by reweighting the loss function. **(3).** About result significance & robustness. We believe that randomness in experiments does exist. For resolving this concern, we try 10 runs of MF, IPS, SNIPS, DR with same learning rate and batch size as MF-CVIB, which yield mean AUC 0.687, 0.704, 0.705, 0.706 on COAT, and 0.678, 0.681, 0.682, 0.684 on YAHOO. On the other, Fig.3 in paper shows the 10 runs of MF-CVIB where it indeed yields better result **on average**: 0.738 on COAT and 0.716 on YAHOO. Interestingly on another ranking metric, e.g., nDCG, CVIB demonstrates more significant advantage. Please refer to the response to **Reviewer #1**.

To Reviewer #3. (1). $y \leftrightarrow x \leftrightarrow z$ is a Markov chain in IB by Tishby in 1999 ($\rightarrow$ and $\leftrightarrow$ are just two equivalent marks here). It turns out to obtain compressed z that is predictive to y. It is reasonable to assume x contains information of y, otherwise by no means we can predict outcome of any event. As mutual information $I(x;y) = I(y;x)$ by definition, y should contain same quantity of information of x. **(2).** Since MAR data is absent, there is no way to access the true $p(y|z)$ but only $p(y|z^+)$. This yields $H_{p,q}(y|z^+)$ as the first term in Eq.(15). In fact, we mention the problem in footnote #3 on p.5. Because we use $H_{p,q}(y|z^+)$ as proxy of $I(z^+;y)$ for the sufficiency term, these two terms of $H_{p,q}(y|z^+)$ will *cancel out* in the final objective function (only second term in Eq.(15) plus minimality term in Eq.(18) left). Same result appears if we let $I(y;z^+) - I(y;z^-) = H(y|z^-) - H(y|z^+)$. Overall, using $q(y|z^+)$ instead of $p(y|z^+)$ as approximation is a compromise for ensuring tractability without the loss of empirical performance. **(3).** In Eq.(18), the balancing term is cross entropy between model output $q(y|z^+)$ and $q(y|z^-)$, and the penalty term is entropy of $q(y|z^+)$. They are optimized plus the cross entropy loss between $p(y|z^+)$ and $q(y|z^+)$ together by SGD. **(4).** We factorize $I(z;y) = H(y) - H(y|z)$ hence maximizing $I(z;y)$ is equivalent to minimizing $H(y|z)$, then uses $H_{p,q}(y|z)$ as a proxy. We agree that it would be better to rephrase it as suggested. **(5).** We believe the involved baselines are comprehensive enough, e.g., the strongest baseline in DRJL [26] (ICML2019) is IPS. Stephen Bonner’s work is mentioned by reference [10], and please refer to **(2)** to **Reviewer #1** about Liang’s works. **(6).** Please refer to **(3)** to **Reviewer #2** about repeat experiments. **(7).** In this work, z^- is embedding of counterfactual event, i.e., x^-. And z^+ is factual embedding. Given the logged feedback, one event could only belong to either factual or counterfactual set, never both. In this view, minimizing $H_{q,q}(y|z^+, y|z^-)$ amounts to balancing information between factual and counterfactual embeddings on average. Therefore, what we need to do is to sample separately from factual and counterfactual event sets, then optimize on it on average. **(8).** We are sorry that using log softmax is a typo and exaggerates $H_{q,q}(y|z^+, y|z^-)$ thus α in the original experiments should be small. Nonetheless, fixing it then leveling up α’s value can result in same performance. To verify this claim, we perform 10 runs on MF-CVIB with $\alpha = 1.0$, same batch size and learning rate. On COAT and YAHOO, the mean and std of AUC is 0.733 (0.007) and 0.719 (0.001), which are even better than the results shown in Table 2. **(9).** Please refer to **(1)** to **Reviewer #2** about the embedding size.

To Reviewer #4. (1). We argue that the proposed CVIB is proved robust in this paper, please refer to Fig.3 and **(3)** to **Reviewer #2**. We will show mean/std of other baselines in the final version. **(2).** The fourth term in Eq.(18) is simply ℓ_2-norm penalty on embeddings, so we add weight decay in ADAM optimizer and find the optimal via grid search. **(3).** Binarizing ratings is commonly used for recommendation. Please refer to section 4.1 in *Causal Inference for Recommendation*. **(4).** Please refer to **(1)** to **Reviewer #2** about the embedding size. **(5).** Although there are many methods in rating prediction, notably few of them are for **counterfactual learning**, i.e., learning from MNAR data and testing on MAR data. Besides, in our experiments, all the selected baselines just use MF & NCF as **backbones**, the same to CVIB for a fair comparison. That means, they are applicable to many backbones, e.g., FM and DeepFM.