We would like to thank the reviewers for their insightful feedback. In the following, we address their key concerns.

R1, R2, R4: Suggest more extensive analysis on Assumption 2 and the normalization step in Algorithm 1.

Following reviewers' suggestions, we will add more thorough analysis in the final paper.

1. A brief explanation of Assumption 2: a) There is an anchor set $\mathcal{A}$ so that $\mathbb{P}(y = +1|x \in \mathcal{A}) = 1$, where $\mathcal{A}$ has a positive probability under f_P. This is a strong variant of the widely used irreducibility assumption [22] (Sec. A.2 in Suppl. Material). b) As pointed out by **R2**, $\mathcal{A}$ could be a small set. But there are *almost surely* samples in $\mathcal{A}$ as data size tends to infinity according to the assumption even if $\mathcal{A}$ does not cover the whole positive data set $\mathcal{P}$. c) In practical cases where $\mathcal{A}$ is too small and $|\mathcal{P}|$ is finite, $\mathcal{A} \cap \mathcal{P}$ could be empty. So we analyzed the misclassification rate under a relaxation of the assumption (Condition (ii) of Assumption 6 and Theorem 7).

2. The normalization step comes from Assumption 2 as indicated by **R4**. Without the step, the variational loss $\mathcal{L}_{\rm var}$ can be minimized by $\Phi = c \cdot \Phi^*$ for all $c > 0$ (Remark 4) due to the fact that the Bayesian classifier $\Phi^* \propto f_p/f$ is only identifiable up to a multiplicative constant without Assumption 2.

R2, R3: VPU seems to heavily rely on Mixup. Its advantages and applications are then limited.

Mixup was introduced in VPU as a regularizer to solve the overfitting problem (Table 4 and Lines 100–105, 376–384). We will conduct extensive comparison experiments for analysis of Mixup. A part of the results (percent accuracy) on the FashionMNIST (FM) dataset is shown in the table below, where class priors are estimated by KM2 [28] for nnPU.

	VPU (Mixup)	nnPU (Mixup)	VPU (Large-margin)		VPU (Mixup)	nnPU (Mixup)	VPU (Large-margin)
FM^1	92.7 ± 0.3	91.0 ± 0.6	92.6 ± 0.4	FM^2	90.8 ± 0.6	90.5 ± 0.7	91.1 ± 0.2

1. In many advanced methods (e.g., nnPU), the class prior is considered as a given constant, and it is difficult for Mixup to reduce the error caused by the inaccurate class prior estimation. Specifically, nnPU already has an effective strategy to tackle overfitting, and its accuracy is not significantly improved by Mixup as shown in the table.

2. In the case mentioned by **R3**, where Mixup is not applicable, the overfitting of VPU can be solved by regularization techniques without data augmentation (e.g., large-margin regularization). This will be investigated in the final paper, and the feasibility can be partially demonstrated by columns 4 and 8 in the above table.

3. As indicated by **R2**, a small mini-batch $\mathcal{B}^{\mathcal{P}}$ yields a unbiased but high-variance estimate of $\mathbb{E}_{x\in\mathcal{P}}[\log \Phi(x)]$. The Mixup is necessary only if f_P cannot accurately characterized by $\mathcal{P}$; otherwise the variance can be reduced by simply scaling up $\mathcal{B}^{\mathcal{P}}$ since labeled and unlabeled mini-batches are independently drawn with the same size in Algorithm 1.

R1: The proposed approach still relies on the "selected completely at random" (SCAR) assumption.

For this problem, we demonstrated the performance of VPU by experiments without SCAR (Fig. 2), and performed the asymptotic analysis by assuming the selection bias is bounded (Condition (i) of Assumption 6 and Theorem 7).

R2: Theorem 7 loses its significance under the assumption that $M, N \to \infty$.

We are now working on the error analysis for VPU with finite samples similar to that for nnPU [12], but this is beyond the scope of this paper. In addition, our experiments indicated that VPU can achieve high accuracies with $M \ll N$ (Table 8), and the ablation study on data size was shown in Fig. 3.

R3: In Table 10, VPU performs better even when the true class prior is given to previous methods.

In fact, nnPU slightly outperforms VPU on a learning task of "Grid Stability" as shown in Table 10. Although the comparison between VPU and uPU/nnPU with exact class priors requires further investigation, our experimental experience shows that the Mixup does not play an indispensable role (see our response to the second comment).

R4: The idea of variational bounds and minimizing KL divergence is not new.

This comment is quite enlightening. Motivated by it, we found similar variational principles developed in other fields (e.g., Donsker-Varadhan representation of KL-divergence). We will cite the related references, and clearly state our contributions on problem formulation, asymptotic analysis and regularization for the variational PU learning.

R4: Only the accuracy metric is used, which is not the best when data sets are imbalanced.

The accuracy metric was used because it is popular in the literature of PU learning [12, 14] and also due to page limit. We will provide AUC of classifiers in the final paper.

R4: How to prevent the log terms from taking negative infinite values? Why is alpha chosen as 0.3?

1. In experiments, we modeled $\log \Phi$ instead of Φ by the NN, and estimated $\log \mathbb{E}_f[\Phi]$ by the log-sum-exp function.
2. The performance is not sensitive to $\alpha \in [0.1, 0.4]$ for our experiments, which is similar to the conclusion in [23].