(a) Synthetic source data: blondness perfectly correlates with the male gender.

(b) Synthetic target data: each gender has a variety of hair colors.

(c) Samples corrected by self-training were mostly mistakenly predicted because of the spurious correlation.

We thank the reviewers for the detailed and insightful feedback. The reviewers noted that the paper "target[s] a timely and important problem" and "posits/analyzes a structured model... to study self-training" [R1], "the theoretical claims are well established" [R4], and "the experimental results support the theoretical implications" [R3]. We will address the major points and incorporate others in the next revision.

[R1]: "missing ... a bound on the target accuracy of the final classifier (in analogy to Theorem 3.2 which studies a simpler setting). Clarification on why this is not provided or difficult to provide ... would be useful."
• The strongest result obtained, as noted in Line 211-212, is that the final classifier is a min-entropy solution, or max unsupervised margin solution, **without** using spurious features. Proving this classifier also obtains good accuracy requires analyzing unsupervised margin maximization, which needs stronger data assumptions[1] and is beyond the scope of the paper. Unsupervised margin maximization was proposed 25 years ago (as transductive SVM [5]), but has little theoretical analysis (except [1]). Instead, we focus on removing the spurious features.
[R1]: "γ is not explained", "x_2 is assumed to be Gaussian... and x_1 is a mixture"
• γ/σ is indeed the SNR. The condition that w_S has $1-\rho$ accuracy is equivalent to requiring that $w_1\gamma$ is large (Lemma A.1 in Appendix A). x_2 should be x_1 (a typo) in line 141.
[R2]: "[theory not surprising because loss] would favor good features due to their correlation with the model", "unclear how... the theory can... provide guidance to empirical discoveries"
• We respectfully and strongly disagree. The spurious features are indeed independent with ground-truth target labels, but **no** target label is provided in self-training. The classifier is entirely self-trained on pseudolabels, which can in fact be correlated with the spurious features due to biases from the source-trained initial classifier.
• Our theory suggests that self-training on diverse unlabeled datasets can improve model robustness, which guides practitioners to collect larger and more diverse datasets even if labeling is impractical. Recent empirical results [6] are consistent with our theory but do not explain why self-training works. We also show entropy minimization can converge faster than pseudo-labeling (Appendix E.3) which may inform practitioners.
[R3]: "the [exponential] loss functions [is] not standard. I suggest analyzing log-loss"
• We would like to respectfully argue that the distinction between the two losses is unimportant in the context of our paper. In theoretical works, these losses have been regarded as equivalent in various contexts [3, 4], and analyzing exponential loss is convenient for the proof. In Figure 10, the two losses achieve equivalent empirical performance.
[R3]: "hard to identify which [assumptions] play the essential role in the result and which are just for technical issues", "why the additional source loss on labeled data in self-training is required in the experiment on CelebA"
• In the setting of Theorem 3.1, we make two main assumptions: 1. the signal x_1 is structured and 2. the source classifier is decent. Both assumptions are essential because they rule out failure cases of self-training in Section 3.1.
• The source loss constrains self-training to stay close to the source classifier, ensuring that the classifier correctly uses signal features while entropy minimization removes the spurious ones. Previous empirical works have found the need for explicit constraints to stay close to source classifier [2].
[R4]: "show that the weights for the spurious features are indeed minimized to zero", "what happens to ... accuracy if the classifier gets stuck ... Will self-training reject low confidence samples?", "any effect of regularization"
• Figure 8 shows spurious coordinates go to 0 in the Gaussian mixture experiment. For colored MNIST, Figures 3 and 4 show reduced reliance on x_2. For our CelebA experiment, visualization of a random sample of examples whose predictions are corrected by self-training shows that self-training removes the spurious feature (see attached figure).
• The local minima of the min-entropy objective that self-training converges to can have sub-optimal accuracy. We do not reject any examples in our experiments as our theory does not analyze rejection.
• Regularization is necessary; without it, scaling up the source predictions can make the loss arbitrarily small. Our theory analyzes l2 regularization by constraining the classifier norm to 1. We will clarify this point in the next revision.

References: [1] Derberko et al., Error bounds for transductive learning via compression and clustering, 2004. [2] Shu et al., A dirt-t approach to unsupervised domain adaptation, 2018. [3] Soudry et al., The implicit bias of gradient descent on separable data, 2018. [4] Telgarsky, Margins, shrinkage, and boosting, 2013. [5] Vapnik, The nature of statistical learning theory, 1995. [6] Xie et al., Self-training with noisy student improves imagenet classification, 2020.

[1] The authors suspect that a difficult situation is when all the data reside on the hypercube $\{-1,+1\}^d$, where a large unsupervised margin classifier may exist coincidentally due to the discrete nature. We need to make careful assumptions to rule out such cases.