Supplementary Material for ICLR Submission #5214 Discussion

1 Additional Experiments

1.1 Varying Model Architectures on the LWFA+ Image Dataset

To demonstrate that, as predicted by Theorem 2, Section 3, the strict trade-off between features' utility for a downstream prediction task and the LPP applies regardless of a model's architecture or the structure of the feature encoder $Z = f_E(X)$, we conduct an additional experiment on the LFWA+ image dataset using a different model architecture. We use the ResNet-18 architecture from He et al. (2015) implemented by PyTorch. Training batch size is 32, SGD learning rate is 0.01.

Fig. 1 compares the trade-off between utility and attribute leakage of a CNN256 (*top*) and a ResNet18 (*bottom*) models trained with standard SGD. The blue horizontal bars in Fig. 1 (*right*) show the model's utility for learning task Y measured as $\tilde{I}_\infty(Y, Z)$. The heatmaps in Fig. 1 (*left*) show the difference between the adversary's inference gain and the model's utility $\tilde{I}_\infty(S, Z) - \tilde{I}_\infty(Y, Z)$. Each row corresponds to a different learning task Y, each column represents a different sensitive attribute targeted by the adversary. We observe that regardless of the model architecture, for any learning task there always exists a sensitive attribute for which $\tilde{I}_\infty(S, Z) > \tilde{I}_\infty(Y, Z)$ and thus violates the LPP.

1.2 Experiments on an Additional Tabular Dataset

We ran an additional experiment to demonstrate that the strict trade-off between model utility and the LPP also holds on a very different type of dataset and model. As for tabular data, together with image data, sharing feature encodings instead of raw data is often suggested as a solution to limit harmful inferences, we choose the Texas Hospital dataset (Texas Department of State Health Services, Austin, Texas, 2013) and the TabNet model architecture (Arik & Pfister, 2021) for these experiments.

Data. The Texas Hospital Discharge dataset (Texas Department of State Health Services, Austin, Texas, 2013) is a large public use data file provided by the Texas Department of State Health Services. The dataset we use consists of 5,202,376 records uniformly sampled from a pre-processed data file that contains patient records from the year 2013. We retain 18 data attributes of which 11 are categorical and 7 continuous.

Experiment Setup. In each experiment, we select one attribute as the model's learning task Y and a second attribute as the sensitive attribute S targeted by the adversary. We repeat each experiment 5 times to capture randomness of our measurements for both the model and adversary, and show average results across all 5 repetitions. At the start of the experiment, we split the data into the three sets D_T, D_E, and D_A. We train a TabNet model on the train set D_T for the chosen learning task and then estimate the model's utility on the evaluation set D_E. We measure the model's utility by estimating the multiplicative gain $\tilde{I}_\infty(Y; Z) = \log {}^{\tilde{\Pr}(Y=\hat{Y}(Z))}/_{\tilde{\Pr}(Y=\hat{Y})}$, where $\hat{Y}(Z)$ denotes the trained model's prediction for a record's task label Y and $\hat{Y}$ without the argument the majority class baseline guess. After model training and evaluation, we train both the label-only and features adversary on the auxiliary data D_A. The features adversary is given access to a record's representation at the last encoding layer of the TabNet encoder (see Arik & Pfister (2021) for details of the model architecture). For a given sensitive attribute S, we estimate the adversary's gain as $\tilde{I}_\infty(S, Z \mid Y) = \log {}^{\tilde{\Pr}[S=\hat{S}(Z,Y)]}/_{\tilde{\Pr}[S=\hat{S}(Y)]}$.

As above, the bar chart in Fig. 2 (*right*) shows the model's utility for learning task Y indicated in each row measured as $\tilde{I}_\infty(Y, Z)$. The heatmaps in Fig. 2 (*left*) show the difference between the adversary's inference gain and the model's

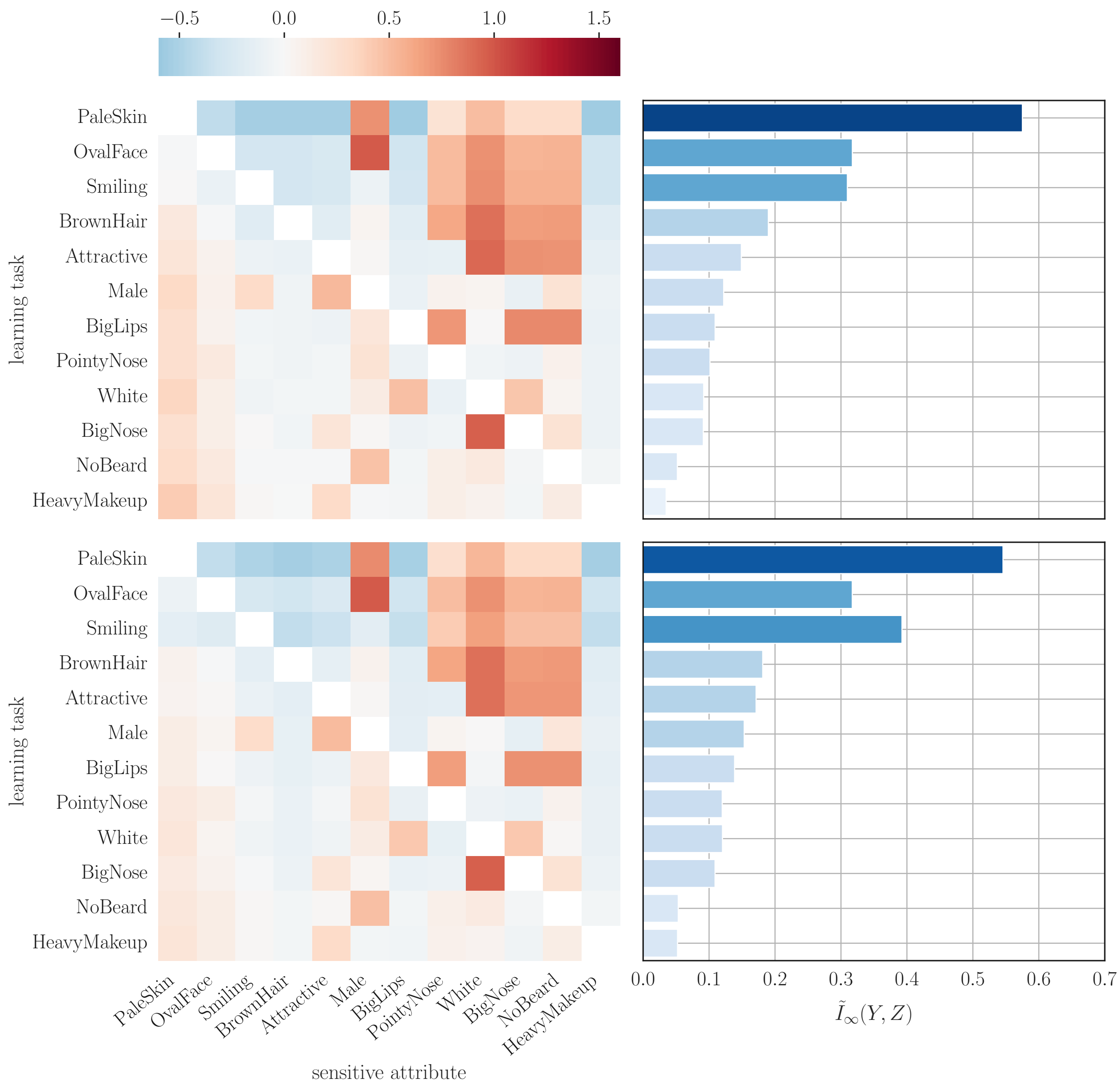

Figure 1: Attribute leakage (*left*) and model utility (*right*) for a CNN256 (*top*) and a ResNet18 model architecture (*bottom*) trained on the LFWA+ image dataset.

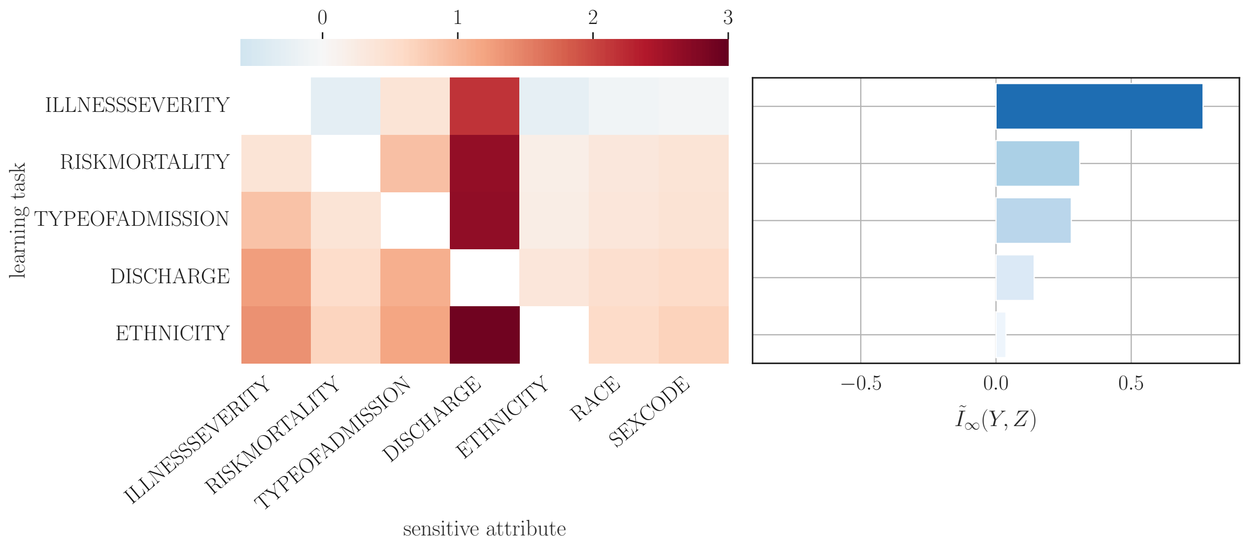

Figure 2: Attribute leakage (*left*) and model utility (*right*) for a TabNet model trained on the Texas Hospital dataset

utility $\tilde{I}_\infty(S, Z) - \tilde{I}_\infty(Y, Z)$. As on the LFWA+ dataset, for any learning task there always exists a sensitive attribute for which an adversary gains an advantage from observing a target record's feature representation.

References

Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv*, 2015.

Texas Department of State Health Services, Austin, Texas. Texas Hospital Inpatient Discharge Public Use Data File 2013 Q1-Q4. `https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm`, 2013. Accessed 2020-06-01.