Trade-offs and Guarantees of Adversarial Representation Learning for Information Obfuscation

Jianfeng Chi[§*] Han Zhao[†*], Yuan Tian[§], Geoffrey J. Gordon[†]

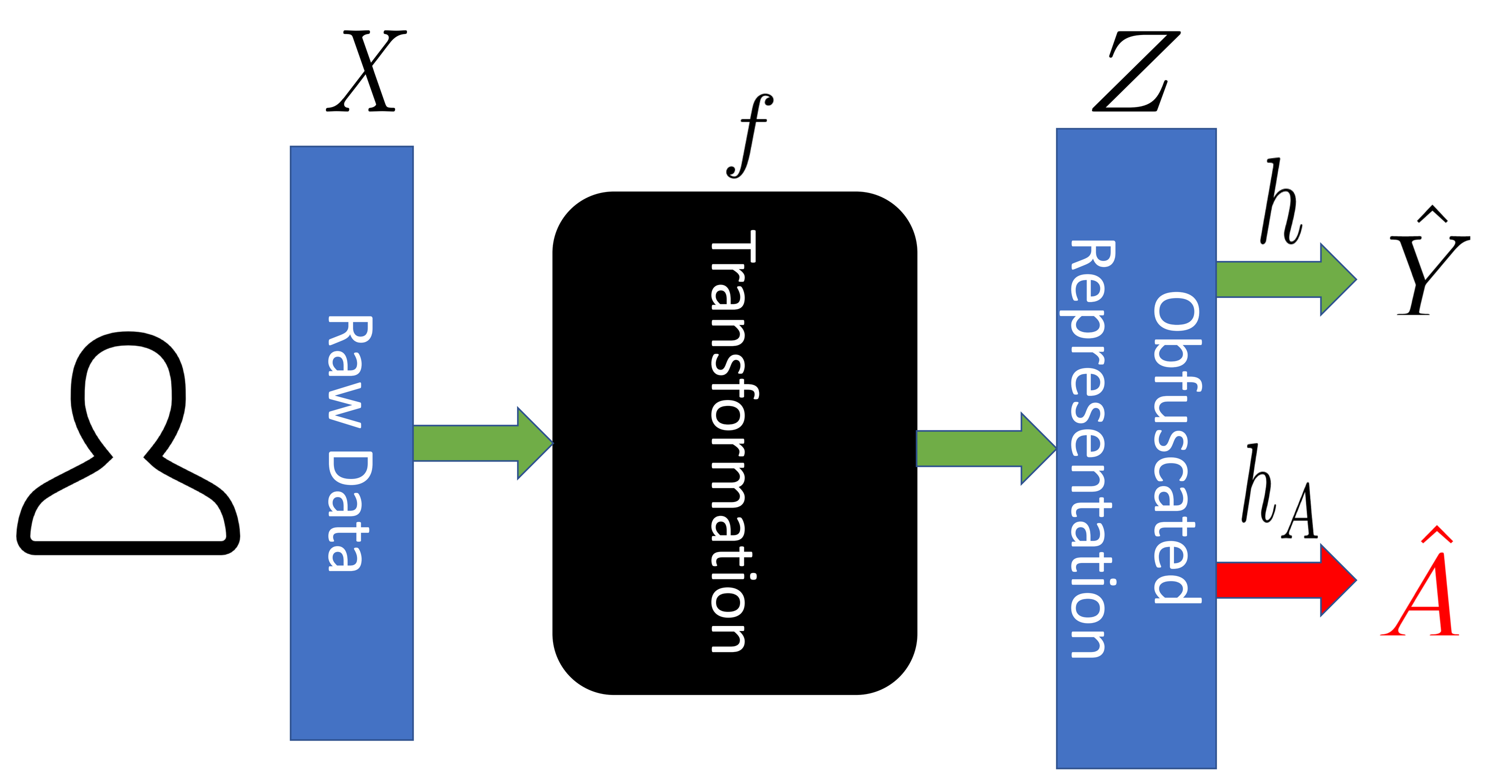
[§]University of Virginia, [†]Carnegie Mellon University

{jc6ub,yuant}@virginia.edu, {han.zhao,ggordon}@cs.cmu.edu

[*] equal contribution

Overview

Learning Representations that Obfuscate Sensitive Attributes:



Question:

Can we prevent the information leakage of the sensitive attribute while still maximizing the task accuracy? Furthermore, what is the fundamental trade-off between attribute obfuscation and accuracy maximization in the minimax problem?

Preliminaries

Utility:

$$\mathrm{Acc}(h) := 1 - \mathbb{E}_{\mathcal{D}}\left[|Y - h(X)|\right]$$

Attribute Inference Advantage:

$$\textsc{Adv}(\mathcal{H}_A) := \max_{h_A \in \mathcal{H}_A} \left| \Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 1) - \Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 0) \right|$$

- $\textsc{Adv}_A(h) = 0$ iff $I(h(X); A) = 0$ and $\textsc{Adv}_A(h) = 1$ iff $h(X) = A$ almost surely or $h(X) = 1 - A$
- $\textsc{Adv}(\mathcal{H}_A) + \min_{h_A \in \mathcal{H}_A} \Pr(h_A(X) = 0 \mid A = 1) + \Pr(h_A(X) = 1 \mid A = 0) = 1$ if $\mathcal{H}_A$ is symmetric: the larger the attribute inference advantage of $\mathcal{H}_A$, the smaller the minimum sum of Type-I and Type-II error under attacks from $\mathcal{H}_A$.

Theoretical Analysis

Formal Guarantees against Attribute Inference

$$\min_{h \in \mathcal{H}, f} \max_{h_A \in \mathcal{H}_A} \widehat{\mathrm{Err}}(h \circ f) - \lambda\Big(\Pr_{S}(h_A(f(X)) = 0 \mid A = 1) + \Pr_{S}(h_A(f(X)) = 1 \mid A = 0)\Big) \quad (1)$$

In practice, we have:

$$\min_{h \in \mathcal{H}, f} \max_{h_A \in \mathcal{H}_A} \mathrm{CE}_Y(h \circ f) - \lambda \cdot \mathrm{CE}_A(h_A \circ f) \quad (2)$$

Theorem:
Let f^* be the optimal feature map such that $f^* = \arg\min H(Y \mid Z = f(X)) - \lambda H(A \mid Z = f(X))$ and define $H^* := H(A \mid Z = f^*(X))$. Then for any adversary $\widehat{A}$ such that $I(\widehat{A}; A \mid Z) = 0$, we have

$$\Pr_{\mathcal{D}^{f^*}}(\widehat{A} \neq A) \geq H^*/2\lg(6/H^*).$$

Implication: If the obfuscated representation Z contains little information on A, then the inference error made by any adversary has to be large.

Inherent trade-off between Accuracy Maximization and Attribute Obfuscation

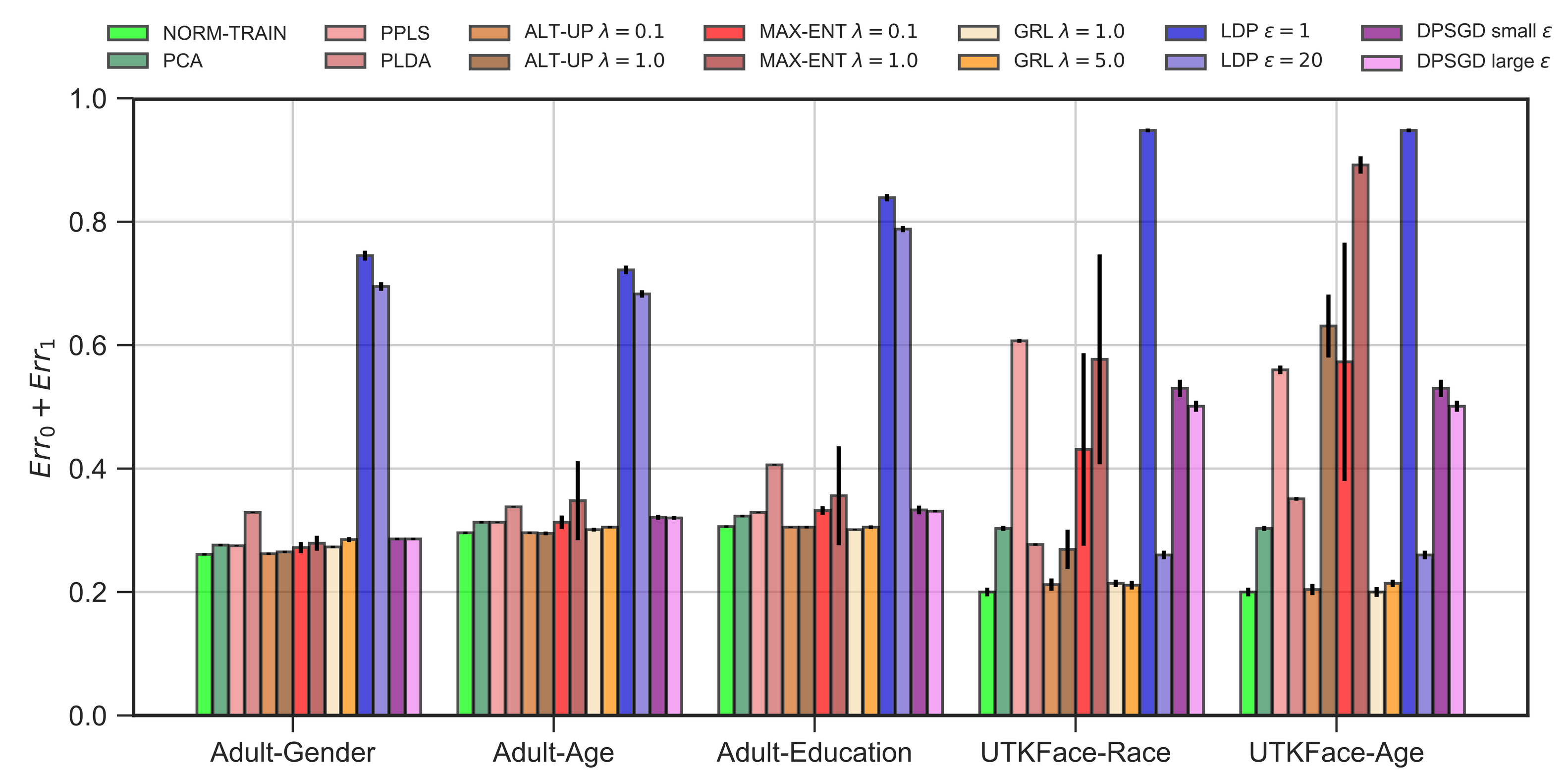
Theorem: Let $\mathcal{H} \subseteq 2^{\mathcal{Z}}$ contains all the measurable functions from $\mathcal{Z}$ to $\{0, 1\}$ and $\mathcal{D}_0^Y, \mathcal{D}_1^Y$ be two distributions over $\mathcal{Y}$ conditioned on $A = 0$ and $A = 1$ respectively. Assume the Markov chain $X \xrightarrow{f} Z \xrightarrow{h} \widehat{Y}$ holds, If $\textsc{Adv}(\mathcal{H}_A \circ f) \leq D_{\mathrm{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y)$, then $\forall h \in \mathcal{H}$, we have

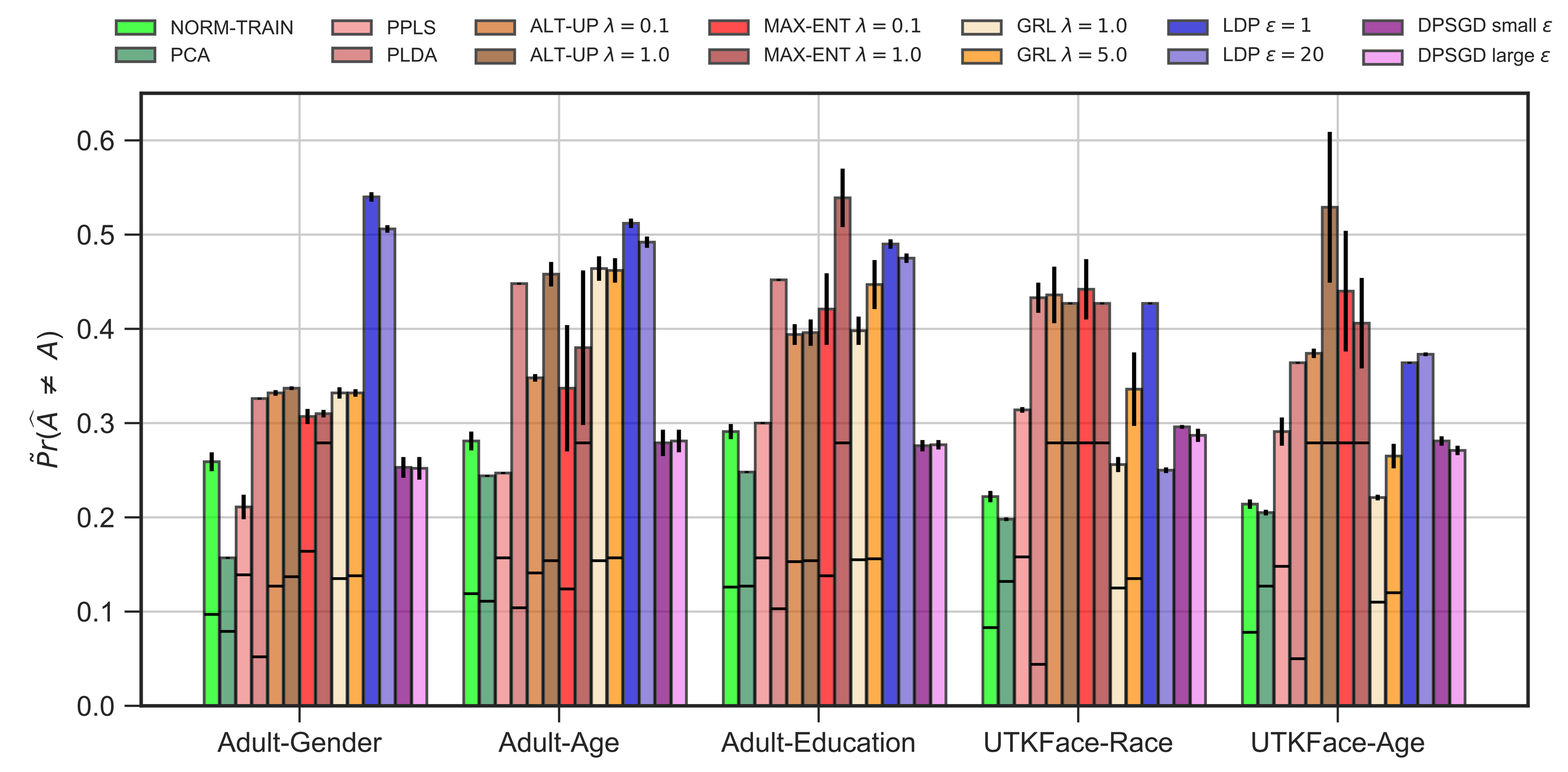
$$\mathrm{Err}_0(h \circ f) + \mathrm{Err}_1(h \circ f) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) - \sqrt{\textsc{Adv}(\mathcal{H}_A \circ f)}\right)^2.$$

Implication: If the label and the sensitive attribute are highly correlated, we cannot obfuscate the sensitive attribute while still maximizing the task accuracy simultaneously.

Empirical Results

(1) Income prediction on the UCI Adult dataset with sensitive attributes: gender, age, and education; (2) Gender estimation on UTKFace dataset with sensitive attributes: age and race.





- The formal guarantees hold for all representation learning based approaches;
- Inherent trade-offs between accuracy maximization and attribute obfuscation exist for all methods;
- Compared to DP-related methods, adversarial representation learning based approaches leads to better trade-offs;

Conclusion: The adversarial representation learning approaches achieve the best trade-offs in terms of attribute obfuscation and accuracy maximization.