Learning Robust and Privacy-Preserving Representations via Information Theory

Binghui Zhang[1], Sayedeh Leila Noorbakhsh[1], Yun Dong[2], Yuan Hong[3], Binghui Wang[1]

[1]Illinois Institute of Technology
[2]Milwaukee School of Engineering
[3]University of Connecticut
bzhang57@hawk.iit.edu, snoorbakhsh@hawk.iit.edu, dong@msoe.edu, yuan.hong@uconn.edu, bwang70@iit.edu

Abstract

Machine learning models are vulnerable to both security attacks (e.g., adversarial examples) and privacy attacks (e.g., private attribute inference). We take the first step to mitigate both the security and privacy attacks, and maintain task utility as well. Particularly, we propose an information-theoretic framework to achieve the goals through the lens of representation learning, i.e., learning representations that are robust to both adversarial examples and attribute inference adversaries. We also derive novel theoretical results under our framework, e.g., the inherent trade-off between adversarial robustness/utility and attribute privacy, and guaranteed attribute privacy leakage against attribute inference adversaries.

Code — https://github.com/ARPRL/ARPRL

Introduction

Machine learning (ML) has achieved breakthroughs in many areas such as computer vision and natural language processing. However, recent works show current ML design is vulnerable to both security and privacy attacks, e.g., adversarial examples and private attribute inference. Adversarial examples (Szegedy et al. 2013; Carlini and Wagner 2017) are typically generated by carefully adding imperceptible perturbations to natural data and remain a serious problem that prevents the deployment of modern ML models in safety-critical applications such as autonomous driving (Eykholt et al. 2018) and medical imaging (Bortsova et al. 2021). In addition, many real-world applications involve data that contain sensitive information, such as race, gender, and income. When applying ML to these applications, it poses a great challenge since private attribute can often be accurately inferred (Jia et al. 2017; Aono et al. 2017; Melis et al. 2019).

To mitigate adversarial examples and attribute inference attacks, many defenses are proposed but follow two separate lines with different techniques. For instance, state-of-the-art defenses against adversarial examples are based on adversarial training (Madry et al. 2018; Zhang et al. 2019; Wang et al. 2019), which solves a min-max optimization problem. In contrast, a representative defense against inference attacks is based on differential privacy (Abadi et al. 2016), which is a statistical method. Further, some works (Song, Shokri, and Mittal 2019b,a) show adversarially robust models can even leak more private information (also verified in our results).

In this paper, we focus on the research question: *1) Can we design a model that ensures adversarial robustness and attribute privacy protection while maintaining the utility of any (unknown) downstream tasks simultaneously? 2) Further, can we theoretically understand the relationships among adversarial robustness, utility, and attribute privacy?* To achieve the goal, we propose an information-theoretic defense framework through the lens of *representation learning*, termed **ARPRL**. Particularly, instead of training models from scratch, which requires huge computational resources and is time consuming, shared learnt representations ensures the community to save much time and costs for future use. Our ARPRL is partly inspired by two works (Zhu, Zhang, and Evans 2020; Zhou et al. 2022), which show adversarially robust representations based defenses outperform the *de facto* adversarial training based methods, while being *the first work* to contrivially generalize learning data representations that are robust to both adversarial examples and attribute inference adversaries. More specifically, we formulate learning representations via three mutual information (MI) objectives: one for adversarial robustness, our for attribute privacy protection, and one for utility preservation. We point out that our ARPRL is *task-agnostic*, meaning the learnt representations does not need to know the target task at hand and can be used for any downstream task. Further, based on our MI objectives, we can derive several theoretical results. For instance, we obtain an inherent tradeoff between adversarial robustness and attribute privacy, as well as between utility and attribute privacy. These tradeoffs are also verified through the experimental evaluations on multiple benchmark datasets. We also derive the guaranteed attribute privacy leakage. Our key contributions are as below:

- We propose the first information-theoretic framework to unify both adversarial robustness and privacy protection.
- We formulate learning adversarially robust and privacy-preserving representations via mutual information goals and train neural networks to approximate them.
- We provide novel theoretical results: the tradeoff between adversarial robustness/utility and attribute privacy, and guaranteed attribute privacy leakage.

Preliminaries and Problem Setup

Notations. We use s, $\mathbf{s}$, and $\mathcal{S}$ to denote (random) scalar, vector, and space, respectively. Given a data $\mathbf{x} \in \mathcal{X}$, we denote its label as $y \in \mathcal{Y}$ and private attribute as $u \in \mathcal{U}$, where $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{U}$ are input data space, label space, and attribute space, respectively. An l_p ball centered at a data $\mathbf{x}$ with radius ϵ is defined as $\mathcal{B}_p(\mathbf{x}, \epsilon) = \{\mathbf{x}' \in \mathcal{X} : \|\mathbf{x}' - \mathbf{x}\|_p \leq \epsilon\}$. The joint distribution of $\mathbf{x}$, y, and u is denoted as $\mathcal{D}$. We further denote $f : \mathcal{X} \to \mathcal{Z}$ as the representation learner that maps $\mathbf{x} \in \mathcal{X}$ to its representation $\mathbf{z} \in \mathcal{Z}$, where $\mathcal{Z}$ is the representation space. We let $C : \mathcal{Z} \to \mathcal{Y}$ be the *primary task classifier*, which predicts label y based on $\mathbf{z}$, and $A : \mathcal{Z} \to \mathcal{U}$ be the *attribute inference classifier*, which infers u based on the representation $\mathbf{z}$. The mutual information (MI) of two random variables $\mathbf{x}$ and $\mathbf{z}$ is denoted by $I(\mathbf{x}; \mathbf{z})$.

Adversarial example/perturbation, adversarial risk, and representation vulnerability (Zhu et al., 2020). Let ϵ be the l_p perturbation budget. For any classifier $C : \mathcal{X} \to \mathcal{Y}$, the *adversarial risk* of C with respect to ϵ is defined as:

$$\begin{aligned} \mathrm{AdvRisk}_\epsilon(C) &= \Pr[\exists \mathbf{x}' \in \mathcal{B}_p(\mathbf{x}, \epsilon), \text{ s.t. } C(\mathbf{x}') \neq y] \\ &= \sup\nolimits_{\mathbf{x}' \in \mathcal{B}_p(\mathbf{x}, \epsilon)} \Pr[C(\mathbf{x}') \neq y], \end{aligned} \tag{1}$$

where $\mathbf{x}'$ is called *adversarial example* and $\delta = \mathbf{x}' - \mathbf{x}$ is *adversarial perturbation* with an l_p budget ϵ, i.e., $\|\delta\|_p \leq \epsilon$. Formally, adversarial risk captures the vulnerability of a classifier to adversarial perturbations. When $\epsilon = 0$, adversarial risk reduces to the standard risk, i.e., $\mathrm{AdvRisk}_0(C) = \mathrm{Risk}(C) = \Pr(C(\mathbf{x}) \neq y)$. Motivated by the empirical and theoretical difficulties of robust learning with adversarial examples, Zhu, Zhang, and Evans (2020); Zhou et al. (2022) target learning adversarially robust representations based on mutual information. They introduced the term *representation vulnerability* as follow: Given a representation learner $f : \mathcal{X} \to \mathcal{Z}$ and an l_p perturbation budget ϵ, the representation vulnerability of f with respect to ϵ is defined as

$$\mathrm{RV}_\epsilon(f) = \max\nolimits_{\mathbf{x}' \in \mathcal{B}_p(\mathbf{x}, \epsilon)} [I(\mathbf{x}; \mathbf{z}) - I(\mathbf{x}'; \mathbf{z}')], \tag{2}$$

where $\mathbf{z} = f(\mathbf{x})$ and $\mathbf{z}' = f(\mathbf{x}')$ are the learnt representation for $\mathbf{x}$ and $\mathbf{x}'$, respectively. We note *higher/smaller* $\mathrm{RV}_\epsilon(f)$ *values imply the representation is less/more robust to adversarial perturbations*. Further, (Zhu, Zhang, and Evans 2020) linked the connection between adversarial robustness and representation vulnerability through the following theorem:

Theorem 1. *Consider all primary task classifiers as* $\mathcal{C} = \{C : \mathcal{Z} \to \mathcal{Y}\}$. *Given the perturbation budget* ϵ*, for any representation learner* $f : \mathcal{X} \to \mathcal{Z}$*,*

$$\inf_{C \in \mathcal{C}} \mathrm{AdvRisk}_\epsilon(C \circ f) \geq 1 - \frac{\big(I(\mathbf{x}; \mathbf{z}) - \mathrm{RV}_\epsilon(f) + \log 2\big)}{\log |\mathcal{Y}|}. \tag{3}$$

The theorem states that a smaller representation vulnerability implies a smaller lower bounded adversarial risk, which means better adversarial robustness, and vice versa. Finally, f is called (ϵ, τ)-robust if $\mathrm{RV}_\epsilon(f) \leq \tau$.

Attribute inference attacks and advantage. Following existing privacy analysis (Salem et al. 2023), we assume the private attribute space $\mathcal{U}$ is binary. Let $\mathcal{A}$ be the set of all binary attribute inference classifiers, i.e. $\mathcal{A} = \{A : \mathcal{Z} \to \mathcal{U} = \{0, 1\}\}$. Then, we formally define the *attribute inference advantage* of the worst-case attribute inference adversary with respect to the joint distribution $\mathcal{D} = \{\mathbf{x}, y, u\}$ as below:

$$\begin{aligned} \mathrm{Adv}_{\mathcal{D}}(\mathcal{A}) = \max_{A \in \mathcal{A}} & |\Pr_{\mathcal{D}}(A(\mathbf{z}) = a | u = a) \\ & - \Pr_{\mathcal{D}}(A(\mathbf{z}) = a | u = 1 - a)|, \ \forall a = \{0, 1\}. \end{aligned} \tag{4}$$

We can observe that: if $\mathrm{Adv}_{\mathcal{D}}(\mathcal{A}) = 1$, an adversary can *completely* infer the privacy attribute through the learnt representations. In contrast, if $\mathrm{Adv}_{\mathcal{D}}(\mathcal{A}) = 0$, an adversary obtains a *random guessing* inference performance. To protect the private attribute, we aim to obtain a small $\mathrm{Adv}_{\mathcal{D}}$.

Threat model and problem setup. We consider an attacker performing both attribute inference and adversarial example attacks. *We assume the attacker does not have the access to the representation learner (i.e., f), but can obtain the shared data representations.* Our goal is to learn task-agnostic representations that are adversarially robust, protect attribute privacy, and maintain the utility of any downstream task. Formally, given data $\{\mathbf{x}, y, u\}$ from an underlying distribution $\mathcal{D}$, and a perturbation budget ϵ, we aim to obtain the representation learner f such that the representation vulnerability $RV_\epsilon(f)$ is small, attribute inference advantage $\mathrm{Adv}_{\mathcal{D}}(\mathcal{A})$ is small, but the performance is high, i.e., $\mathrm{Risk}(C)$ is small.

Design of ARPRL

In this section, we will design our **a**dversarilly **r**obust and **p**rivacy-preserving **r**epresentation **l**earning method, termed **ARPRL**, inspired by information theory.

Formulating ARPRL via MI Objectives

Given a data $\mathbf{x}$ with private attribute u sampled from a distribution $\mathcal{D}$, and a perturbation budget ϵ, we aim to learn the representation $\mathbf{z} = f(\mathbf{x})$ for $\mathbf{x}$ that satisfies three goals:

- **Goal 1: Protect attribute privacy.** $\mathbf{z}$ contains as less information as possible about the private attribute u. Ideally, when $\mathbf{z}$ does not include information about u, i.e., $\mathbf{z} \perp u$, it is impossible to infer u from $\mathbf{z}$.
- **Goal 2: Preserve utility.** $\mathbf{z}$ should be be useful for many downstream tasks. Hence it should include as much information about $\mathbf{x}$ as possible, while excluding the private u. Ideally, when $\mathbf{z}$ retains the most information about $\mathbf{x}$, the model trained on $\mathbf{z}$ will have the same performance as the model trained on the raw $\mathbf{x}$ (though we do not know the downstream task), thus preserving utility.
- **Goal 3: Adversarially robust.** $\mathbf{z}$ should be not sensitive to adversarial perturbations on the data $\mathbf{x}$, indicating a small representation vulnerability.

We propose to formalize the above goals via MI. Formally, we quantify the goals as below:

Formalize Goal 1: $$\min_f I(\mathbf{z}; u); \tag{5}$$

Formalize Goal 2: $$\max_f I(\mathbf{x}; \mathbf{z} | u); \tag{6}$$

Formalize Goal 3: (7)

$$\min_f RV_\epsilon(f|u) = \min_f \max_{\mathbf{x}' \in \mathcal{B}_p(\mathbf{x}, \epsilon)} I(\mathbf{x}; \mathbf{z}|u) - I(\mathbf{x}'; \mathbf{z}'|u).$$

where 1) we minimize $I(\mathbf{z}; u)$ to maximally reduce the correlation between $\mathbf{z}$ and the private attribute u; 2) we maximize $I(\mathbf{x}; \mathbf{z}|u)$ to keep the raw information in $\mathbf{x}$ as much as possible in $\mathbf{z}$ while excluding information about the private u; 3) $RV_\epsilon(f|u)$ is the representation vulnerability of f conditional on u with respect to ϵ. Minimizing it learns adversarially robust representations that exclude the information about private u. Note that $I(\mathbf{x}; \mathbf{z}|u)$ in Equation (7) can be merged with that in Equation (6). Hence Equation (7) can be reduced to the below min-max optimization problem:

$$\max_f \min_{\mathbf{x}' \in \mathcal{B}_p(\mathbf{x}, \epsilon)} I(\mathbf{x}'; \mathbf{z}'|u). \tag{8}$$

Estimating MI via Variational Bounds

The key challenge of solving the above MI objectives is that calculating an MI between two arbitrary random variables is likely to be infeasible (Peng et al. 2019). To address it, we are inspired by the existing MI neural estimation methods (Alemi et al. 2017; Belghazi et al. 2018; Poole et al. 2019; Hjelm et al. 2019; Cheng et al. 2020), which convert the intractable exact MI calculations to the tractable variational MI bounds. Then, we parameterize each variational MI bound with a neural network, and train the neural networks to approximate the true MI. *We clarify that we do not design new MI neural estimators, but adopt existing ones to aid our customized MI terms for learning adversarially robust and privacy-preserving representations.*

Minimizing upper bound MI in Equation (5) for privacy protection. We adapt the variational upper bound CLUB proposed in (Cheng et al. 2020). Specifically, via some derivations (see Appendix B), we have

$$I(\mathbf{z}; u) \leq \min I_{vCLUB}(\mathbf{z}; u) \Longleftrightarrow \max \mathbb{E}_{p(\mathbf{z},u)}[\log q_\Psi(u|\mathbf{z})]$$

where $q_\Psi(u|\mathbf{z})$ is an auxiliary posterior distribution of $p(u|\mathbf{z})$. Then our **Goal 1** for privacy protection can be reformulated as solving the min-max objective function below:

$$\min_f \min_\Psi I_{vCLUB}(\mathbf{z}; u) \Longleftrightarrow \min_f \max_\Psi \mathbb{E}_{p(\mathbf{z},u)}[\log q_\Psi(u|\mathbf{z})].$$

Remark. This equation can be interpreted as an *adversarial game* between: (1) an adversary q_Ψ (i.e., attribute inference classifier) who aims to infer the private attribute u from the representation $\mathbf{z}$; and (2) a defender (i.e., the representation learner f) who aims to protect u from being inferred.

Maximizing lower bound MI in Equation (6) for utility preservation. We adopt the MI estimator proposed in (Nowozin, Cseke, and Tomioka 2016) to estimate the lower bound of the MI Equation (6). Via some derivations (see details in Appendix B) we have:

$$I(\mathbf{x}; \mathbf{z}|u) \geq H(\mathbf{x}|u) + \mathbb{E}_{p(\mathbf{x},\mathbf{z},u)}[\log q_\Omega(\mathbf{x}|\mathbf{z}, u)],$$

where q_Ω is an *arbitrary* auxiliary posterior distribution. Since $H(\mathbf{x}|u)$ is a constant, our **Goal 2** can be rewritten as the below max-max objective function:

$$\max_f I(\mathbf{x}; \mathbf{z}|u) \iff \max_{f,\Omega} \mathbb{E}_{p(\mathbf{x},\mathbf{z},u)} \left[\log q_\Omega(\mathbf{x}|\mathbf{z}, u)\right].$$

Remark. This equation can be interpreted as a *cooperative game* between the representation learner f and q_Ω who aim to preserve the utility collaboratively.

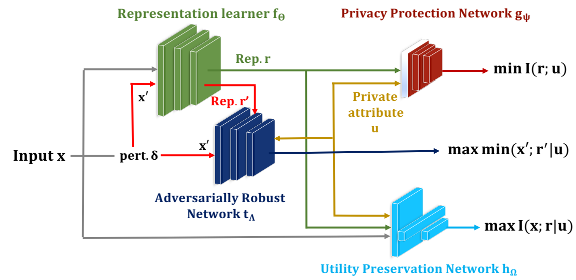


Figure 1: Overview of ARPRL.

Maximizing the worst-case MI in Equation (8) for adversarial robustness. To solve Equation (8), one needs to first find the perturbed data $\mathbf{x}' \in \mathcal{B}_p(\mathbf{x}, \epsilon)$ that minimizes MI $I(\mathbf{x}'; \mathbf{z}'|u)$, and then maximizes this MI by training the representation learner f. As claimed in (Zhu, Zhang, and Evans 2020; Zhou et al. 2022), minimizing the MI on the worst-case perturbed data is computational challenging. An approximate solution (Zhou et al. 2022) is first performing a strong white-box attack, e.g., the projected gradient descent (PGD) attack (Madry et al. 2018), to generate a set of adversarial examples, and then selecting the adversarial example that has the smallest MI. Assume the strongest adversarial example is $\mathbf{x}^a = \arg\min_{\mathbf{x}' \in \mathcal{B}_p(\mathbf{x}, \epsilon)} I(\mathbf{x}'; \mathbf{z}'|u)$. The next step is to maximize the MI $I(\mathbf{x}^a; \mathbf{z}^a|u)$. Zhu, Zhang, and Evans (2020) used the MI Neural Estimator (MINE) (Belghazi et al. 2018) to estimate this MI. Specifically,

$$I(\mathbf{x}^a; \mathbf{z}^a|u) \geq I_\Lambda(\mathbf{x}^a; \mathbf{z}^a|u) = \mathbb{E}_{p(\mathbf{x}^a,\mathbf{z}^a,u)}[t_\Lambda(\mathbf{x}^a, \mathbf{z}^a, u)] \\ - \log \mathbb{E}_{p(\mathbf{x}^a)p(\mathbf{z}^a)p(u)}[\exp(t_\Lambda(\mathbf{x}^a, \mathbf{z}^a, u))],$$

where $t_\Lambda : \mathcal{X} \times \mathcal{Z} \times \{0, 1\} \to \mathbb{R}$ can be any family of neural networks parameterized with Λ. More details about calculating the MI are deferred to the implementation section.

Objective function of ARPRL. By using the above MI bounds, our objective function of ARPRL is as follows:

$$\max_f \Big(\alpha \min_\Psi -\mathbb{E}_{p(\mathbf{x},u)} \left[\log q_\Psi(u|f(\mathbf{x}))\right] + \beta \max_\Lambda I_\Lambda(\mathbf{x}^a; \mathbf{z}^a|u) \\ + (1-\alpha-\beta)\big(\max_\Omega \mathbb{E}_{p(\mathbf{x},u)} \left[\log q_\Omega(\mathbf{x}|f(\mathbf{x}), u)\right] \\ + \lambda \max_\Delta \mathbb{E}_{p(\mathbf{x},y)} \left[\log q_\Delta(y|f(\mathbf{x}))\right]\big)\Big), \tag{9}$$

where $\alpha, \beta \in [0, 1]$ tradeoff between privacy and utility, and robustness and utility, respectively. That is, a larger/smaller α indicates a stronger/weaker attribute privacy protection and a larger/smaller β indicates a stronger/weaker robustness against adversarial perturbations.

Implementation in Practice via Training Parameterized Neural Networks

In practice, Equation (9) is solved via training four neural networks, i.e., the representation learner f_Θ (parameterized with Θ), privacy-protection network g_Ψ associated with the auxiliary distribution q_Ψ, robustness network t_Λ associated with the MINE estimator, and utility-preservation network h_Ω associated with the auxiliary distribution q_Ω, on a set of training data. Suppose we have collected a set of samples $\{(\mathbf{x}_j, y_j, u_j)\}$ from the dataset distribution $\mathcal{D}$. We can then approximate each term in Equation (9).

Specifically, we approximate the expectation associated with the privacy-protection network network g_Ψ as

$$\mathbb{E}_{p(u,\mathbf{x})} \log q_\Psi(u|f(\mathbf{x}))) \approx -\sum\nolimits_j CE(u_j, g_\Psi(f(\mathbf{x}_j)))$$

where $CE(\cdot)$ means the cross-entropy loss function.

Further, we approximate the expectation associated with the utility-preservation network h_Ω via the *Jensen-Shannon* (JS) MI estimator (Hjelm et al. 2019). That is,

$$\begin{aligned}&\mathop{\mathbb{E}}_{p(\mathbf{x},u)} \log q_\Omega(\mathbf{x}|f(\mathbf{x}), u) \approx I^{(JS)}_{\Theta,\Omega}(\mathbf{x}; f(\mathbf{x}), u)\\ &= \mathop{\mathbb{E}}_{p(\mathbf{x},u)}[-\text{sp}(-h_\Omega(\mathbf{x}, f(\mathbf{x}), u)] - \mathop{\mathbb{E}}_{p(\mathbf{x},u,\bar{\mathbf{x}})}[\text{sp}(h_\Omega(\bar{\mathbf{x}}, f(\mathbf{x}), u)]\end{aligned}$$

where $\bar{\mathbf{x}}$ is an independent random sample of the same distribution as $\mathbf{x}$, and expectation can be replaced by samples $\{\mathbf{x}^i_j, \bar{\mathbf{x}}^i_j, u^i_j\}$. $\text{sp}(z) = \log(1+\exp(z))$ is a softplus function.

Regarding the MI related to the robustness network t_Λ, we can adopt the methods proposed in (Zhu, Zhang, and Evans 2020; Zhou et al. 2022). For instance, (Zhu, Zhang, and Evans 2020) proposed to avoid searching the whole ball, and restrict the search space to be the set of empirical distributions with, e.g., m samples: $\mathcal{S}_m(\epsilon) = \{\frac{1}{m}\sum_{i=1}^m \delta_{\mathbf{x}'_i} : \mathbf{x}'_i \in \mathcal{B}_p(\mathbf{x}_i, \epsilon), \forall i \in [m]\}$. Then it estimates the MI $\min_{\mathbf{x}' \in \mathcal{B}_p(\mathbf{x},\epsilon)} I(\mathbf{x}'; f(\mathbf{x}')|u)$ as

$$\min_{\mathbf{x}'} I^{(m)}_\Lambda(\mathbf{x}'; f(\mathbf{x}')|u) \text{ s.t. } \mathbf{x}' \in \mathcal{S}_m(\epsilon), \tag{10}$$

where $I^{(m)}_\Lambda(\mathbf{x}'; f(\mathbf{x}')|u) = \frac{1}{m}\sum_{i=1}^m t_\Lambda(\mathbf{x}_i, f(\mathbf{x}_i), u_i) - \log[\frac{1}{m}\sum_{i=1}^m e^{t_\Lambda(\bar{\mathbf{x}}_i, f(\mathbf{x}_i), u_i)}]$, where $\{\bar{\mathbf{x}}_i\}$ are independent and random samples that have the same distribution as $\{\mathbf{x}_i\}$.

Zhu et al. 2020 propose an alternating minimization algorithm to solve Equation (10)—it alternatively performs gradient ascent on Λ to maximize $I^{(m)}_\Lambda(\mathbf{x}'; f(\mathbf{x}')|u)$ given $\mathcal{S}_m(\epsilon)$, and then searches for the set of worst-case perturbations on $\{\mathbf{x}'_i : i \in [m]\}$ given Λ based on, e.g., projected gradient descent. Figure 1 overviews our ARPRL. Algorithm 1 in Appendix details the training of ARPRL.

Theoretical Results [1]

Robustness vs. Representation Vulnerability. We first show the relationship between adversarial risk (or robustness) and representation vulnerability in ARPRL.

Theorem 2. *Let all binary task classifiers be $\mathcal{C} = \{C : \mathcal{Z} \to \mathcal{Y}\}$. Then for any representation learner $f : \mathcal{X} \to \mathcal{Z}$,*

$$\inf_{C \in \mathcal{C}} \text{AdvRisk}_\epsilon(C \circ f) \geq \frac{\text{RV}_\epsilon(f|u) - I(\mathbf{x}; \mathbf{z}|u)}{\log 2}. \tag{11}$$

Remark. Similar to Theorem 1, Theorem 2 shows a smaller representation vulnerability implies a smaller lower bounded adversarial risk for robustness. In addition, a larger MI $I(\mathbf{x}; \mathbf{z}|u)$ (**Goal 2** for utility preservation) produces a smaller adversarial risk, implying better robustness.

Utility vs. Privacy Tradeoff. The following theorem shows the tradeoff between utility and privacy:

Theorem 3. *Let $\mathbf{z} = f(\mathbf{x})$ be with a bounded norm R (i.e., $\max_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z}\| \leq R$), and $\mathcal{A}$ be the set of all binary inference classifiers taking $\mathbf{z}$ as an input. Assume the task classifier C is C_L-Lipschitz, i.e., $\|C\|_L \leq C_L$. Then, we have the below relationship between the standard risk and the advantage:*

$$\text{Risk}(C \circ f) \geq \Delta_{y|u} - 2R \cdot C_L \cdot Adv_{\mathcal{D}}(\mathcal{A}), \tag{12}$$

where $\Delta_{y|u} = |Pr_{\mathcal{D}}(y = 1|u = 0) - Pr_{\mathcal{D}}(y = 1|u = 1)|$ is a dataset-dependent constant.

Remark. Theorem 3 says that any task classifier using learnt representations leaks attribute privacy: the smaller the advantage $\text{Adv}_{\mathcal{D}}(\mathcal{A})$ (meaning less attribute privacy is leaked), the larger the lower bound risk, and vice versa. Note that the lower bound is independent of the adversary, meaning it covers the *worst-case* attribute inference adversary. Hence, Equation (12) reflects an inherent tradeoff between utility preservation and attribute privacy leakage.

Robustness vs. Privacy Tradeoff. Let $\mathcal{D}'$ be a joint distribution over the adversarially perturbed input $\mathbf{x}'$, sensitive attribute u, and label y. By assuming the representation space is bounded by R, the perturbed representations also satisfy $\max_{\mathbf{z}' \in \mathcal{Z}} \|\mathbf{z}'\| \leq R$, where $\mathbf{z}' = f(\mathbf{x}')$. Following Equation 4, we have an associated adversary *advantage* $\text{Adv}_{\mathcal{D}'}(\mathcal{A})$ with respect to the joint distribution $\mathcal{D}'$. Similarly, $\text{Adv}_{\mathcal{D}'}(\mathcal{A}) = 1$ means an adversary can *completely* infer the privacy attribute u through the learnt adversarially perturbed representations $\mathbf{z}'$, and $\text{Adv}_{\mathcal{D}'}(\mathcal{A}) = 0$ implies an adversary only obtains a *random guessing* inference performance. Then we have the following theorem:

Theorem 4. *Let $\mathbf{z}' = f(\mathbf{x}')$ be the learnt representation for $\mathbf{x}' \in \mathcal{B}(\mathbf{x}, \epsilon)$ with a bounded norm R (i.e., $\max_{\mathbf{z}' \in \mathcal{Z}} \|\mathbf{z}'\| \leq R$), and $\mathcal{A}$ be the set of all binary inference classifiers. Under a C_L-Lipschitz task classifier C, we have the below relationship between the adversarial risk and the advantage:*

$$\text{AdvRisk}_\epsilon(C \circ f) \geq \Delta_{y|u} - 2R \cdot C_L \cdot Adv_{\mathcal{D}'}(\mathcal{A}). \tag{13}$$

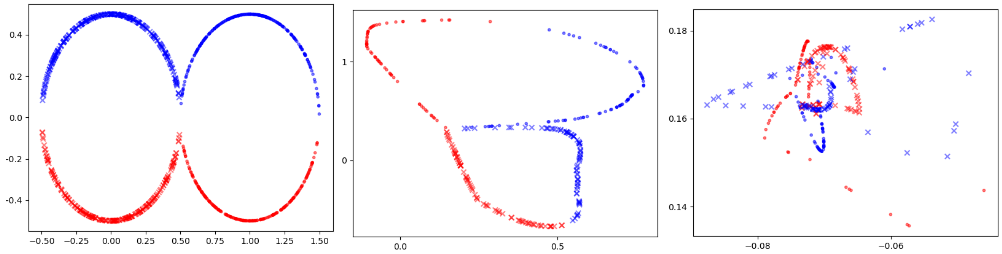
Remark. Similar to Theorem 3, Theorem 4 states that, any task classifier using adversarially learnt representations has to incur an adversarial risk on at least a private attribute value. Moreover, the lower bound covers the *worst-case* adversary. Equation (13) hence reflects an inherent trade-off between adversarial robustness and privacy.

Guaranteed Attribute Privacy Leakage. The attribute inference accuracy induced by the worst-case adversary is bounded in the following theorem:

Theorem 5. *Let $\mathbf{z}$ be the learnt representation by Equation (9). For any attribute inference adversary $\mathcal{A} = \{A : \mathcal{Z} \to \mathcal{U} = \{0, 1\}\}$, $Pr(A(\mathbf{z}) = u) \leq 1 - \frac{H(u|\mathbf{z})}{2\log_2(6/H(u|\mathbf{z}))}$.*

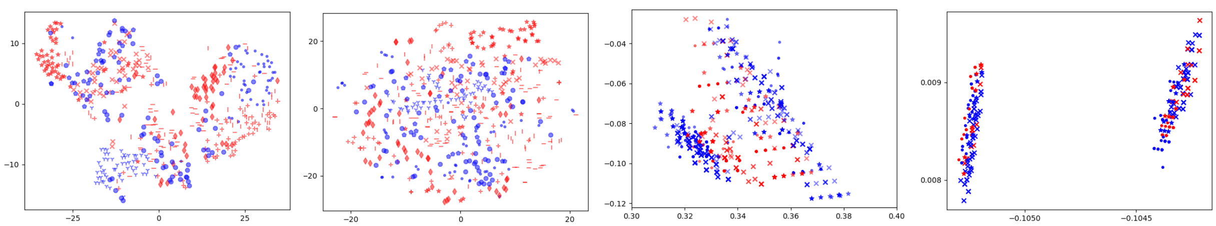
Remark. Theorem 5 shows when $H(u|\mathbf{z})$ is larger, the inference accuracy induced by any adversary is smaller, i.e., less attribute privacy leakage. From another perspective, as $H(u|\mathbf{z}) = H(u) - I(u; \mathbf{z})$, achieving the largest $H(u|\mathbf{z})$ implies minimizing $I(u; \mathbf{z})$ (note that $H(u)$ is a constant)—This is exactly our **Goal 1** aims to achieve.

[1](Zhao et al. 2020) also has theoretical results of privacy protection against attribute inference attacks. The differences between theirs and our theoretical results are discussed in the Appendix.

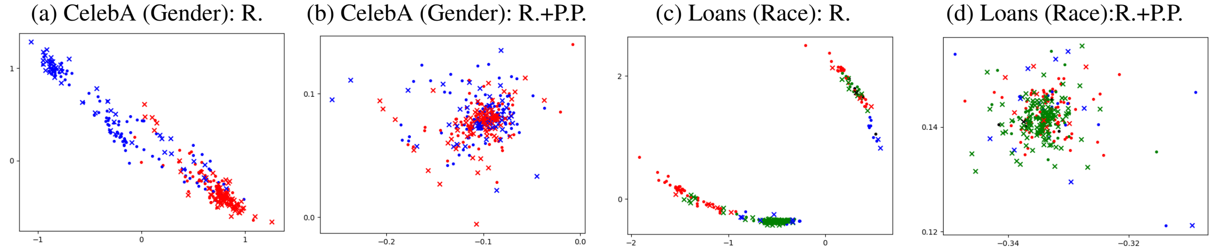


(a) Raw circle data (b) $\alpha = 0, \beta = 0.5$ (c) $\alpha = 0.1, \beta = 0.5$

Figure 2: 2D representations learnt by ARPRL. (a) Raw data; (b) only robust representations (privacy acc: 99%, robust acc: 88%, test acc: 99%); and (c) robust + privacy preserving representations (privacy acc: 55%, robust acc: 75%, test acc: 85%). red vs. blue: binary private attribute values; cross $\times$ vs. circle $\circ$: binary task labels.



(a) CelebA (Gender): R. (b) CelebA (Gender): R.+P.P. (c) Loans (Race): R. (d) Loans (Race):R.+P.P.



(e) Adult (Gender): R. (f) Adult (Gender): R.+P.P. (g) Adult (Marital): R. (h) Adult (Marital): R.+P.P.

Figure 3: 2D t-SNE representations learnt by AdvPPRL. *Left:* only robust representations; *Right:* robust + privacy preserving representations (under the best tradeoff in Table 1). Colors indicate attribute values, while point patterns mean labels.

Evaluations

We evaluate ARPRL on both synthetic and real-world datasets. The results on the synthetic dataset is for visualization and verifying the tradeoff purpose.

Experimental Setup

We train the neural networks via Stochastic Gradient Descent (SGD), where the batch size is 100 and we use 10 local epochs and 50 global epochs in all datasets. The learning rate in SGD is set to be $1e^{-3}$. The detailed network architecture is shown in Table 2 in Appendix D The hyperparameters used in the adversarially robust network are following (Zhu, Zhang, and Evans 2020). We also discuss how to choose the hyperparameters α and β in real-world datasets in Appendix D. W.l.o.g, we consider the most powerful l_∞ perturbation. Following (Zhu, Zhang, and Evans 2020), we use the PGD attack for both generating adversarial perturbations in the estimation of worst-case MI and evaluating model robustness[2]. We implement ARPRL in PyTorch and use the NSF Chameleon Cloud GPUs (Keahey et al. 2020) (CentOS7-CUDA 11 with Nvidia Rtx 6000) to train the model. We evaluate ARPRL on three metrics: utility preservation, adversarial robustness, and privacy protection.

Results on A Toy Example

We generate 2 2D circles with the center (0, 0) and (1, 0) respectively and radius 0.25, and data points are on the circumference. Each circle indicates a class and has 5,000 samples, where 80% of the samples are for training and the rest 20% for testing. We define the binary attribute value for each data as whether its y-value is above/below the x-axis. The network architecture is shown in Table D in Appendix. We use an l_∞ perturbation budget $\epsilon = 0.01$ and 10 PGD attack steps with step size 0.1. We visualize the learnt representations via 2D t-SNE (Van der Maaten and Hinton 2008) in Figure 2.

We can see that: by learning *only robust* representations, the 2-class data can be well separated, but their private

[2]Note our goal is not to design the best adversarial attack, i.e., generating the optimal adversarial perturbation. Hence, the achieved adversarial robustness might not the optimal. We also test CelebA against the CW attack (Carlini and Wagner 2017), and the robust accuracy is 85%, which close to 87% with the PGD attack.

attribute values can be also completely separated–almost 100% privacy leakage. In contrast, by learning both *robust and privacy-preserving* representations, the 2-class data can be separated, but their private attributes are mixed—only 55% inference accuracy. Note that the optimal random guessing inference accuracy is 50%. We also notice a tradeoff among robustness/utility and attribute privacy, as demonstrated in our theorems. That is, a more robust/accurate model leaks more attribute privacy, and vice versa.

Results on the Real-World Datasets

Datasets and setup. We use three real-world datasets from different applications, i.e., the widely-used CelebA (Liu et al. 2015) image dataset (150K training images and 50K for testing), the Loans (Hardt, Price, and Srebro 2016), and Adult Income (Dua and Graff 2017) datasets In CelebA, we treat binary 'gender' as the private attribute, and detect 'gray hair' as the primary (binary classification) task, following (Li et al. 2021; Osia et al. 2018). For the Loans dataset, the primary task is to predict the affordability of the person asking for the loan while protecting their race. For the Adult Income dataset, predicting whether the income of a person is above $50K or not is the primary task. The private attributes are the gender and marital status. For l_∞ perturbations, we set the budget ϵ = 0.01 on Loans and Adults, and 0.1 on CelebA. We use 10 PGD attack steps with step size 0.1.

Results. Tables 1 shows the results on the three datasets, where we report the robust accuracy (under the l_∞ attack), normal test accuracy, and attribute inference accuracy (as well as the gap to random guessing). We have the following observations: 1) When $\alpha = 0$, it means ARPRL only focuses on learning robust representation (similar to (Zhu, Zhang, and Evans 2020)) and obtains the best robust accuracy. However, the inference accuracy is rather high, indicating a serious privacy leakage. 2) Increasing α can progressively better protect the attribute privacy, i.e., the inference accuracy is gradually reduced and finally close to random guessing (note different datasets have different random guessing value). 3) α and β together act as the tradeoff among robustness, utility, and privacy. Particularly, a better privacy protection (i.e., larger α) implies a smaller test accuracy, indicating an utility and privacy tradeoff, as validated in Theorem 3. Similarly, a better privacy protection also implies a smaller robust accuracy, indicating a robustness and privacy tradeoff, as validated in Theorem 4.

Visualization. We further visualize the learnt representations via t-SNE in Figure 3. We can see that: When only focusing on learning robust representations, both the data with different labels and with different attribute values can be well separated. On the other hand, when learning both robust and privacy-preserving representations, the data with different labels can be separately, but they are mixed in term of the attribute values—meaning the privacy of attribute values is protected to some extent.

Runtime. We only show runtime on the largest CelebA (150K training images). In our used platform, it took about 5 mins each epoch (about 15 hours in total) to learn the robust and privacy-preserving representation for each hyperparameter setting. The computational bottleneck is mainly from training robust representations (where we adapt the source code from (Zhu, Zhang, and Evans 2020)), which occupies 60% of the training time (e.g., 3 mins out of 5 mins in each epoch). Training the other neural networks is much faster.

Comparing with the State-of-the-arts

Comparing with task-known privacy-protection baselines. We compare ARPRL with two recent task-known methods for attribute privacy protection on CelebA: **DPFE** (Osia et al. 2018) that also uses mutual information (but in different ways), and **Deepobfuscator** (Li et al. 2021)[3], an adversarial training based defense. We align three methods with the same test accuracy 0.88, and compare attribute inference accuracy. *For fair comparison, we do not consider adversarial robustness in our ARPRL.* The attribute inference accuracy of DPFE and Deepobfuscator are 0.79 and 0.70, respectively, and our ARPRL's is 0.71. First, DPFE performs much worse because it assumes the distribution of the learnt representation to be Gaussian (which could be inaccurate), while Deepobfuscator and ARPRL have no assumption on the distributions; Second, Deepobfuscator performs slightly better than ARPRL. This is because both ARPRL and Deepobfuscator involve adversarial training, Deepobfuscator uses task labels, but ARPRL is task-agnostic, hence slightly sacrificing privacy.

Comparing with task-known adversarial robustness baselines. We compare ARPRL with the state-of-the-art task-known adversarial training based TRADES (Zhang et al. 2019) and test on CelebA, under the same adversarial perturbation and without privacy-protection (i.e., $\alpha = 0$). For task-agnostic ARPRL, its robust accuracy is 0.87, which is slightly worse than TRADES's is 0.89. However, when ARPRL also includes task labels during training, its robust accuracy increases to 0.91—This again verifies that adversarially robust representations based defenses outperform the classic adversarial training based method.

Comparing with task-known TRADES + Deepobfuscator for both robustness and privacy protection. A natural solution to achieve both robustness and privacy protection is by combining SOTAs that are individually adversarially robust or privacy-preserving. We test TRADES + Deepobfuscator on CelebA. Tuning the tradeoff hyperparameters, we obtain the best utility, privacy, and robustness tradeoff at: (Robust Acc, Test Acc, Infer. Acc) = (0.79, 0.84, 0.65) while the best tradeoff of ARPRL in Table 1 is (Robust Acc, Test Acc, Inference Acc) = (0.79, 0.85, 0.62), which is slightly better than TRADES + Deepobfuscator, though they both know the task labels. The results imply that simply combining SOTA robust and privacy-preserving methods is not the best option. Instead, our ARPRL learns both robust and privacy-preserving representations under the same information-theoretic framework.

[3]We observe that a most recent work (Jeong et al. 2023) has similar performance as Deepobfuscator, but 2 orders of memory consumption. We do not include their results for conciseness.

Table 1: Test accuracy, robust accuracy, vs. inference accuracy (and gap w.r.t. the optimal random guessing) on the considered three datasets and private attributes. Note that some datasets are unbalanced, so the random guessing values are different. Larger α means more privacy protection, while larger β means more robust against adversarial perturbation. $\alpha = 0$ means no privacy protection and only focuses on robust representation learning, same as (Zhu, Zhang, and Evans 2020; Zhou et al. 2022).

CelebA — Private attr.: Gender (binary), budget $\epsilon = 0.1$

α	β	Rob. Acc	Test Acc	Infer. Acc (gap)
0	0.50	0.87	0.91	0.81 (0.31)
0.1	0.45	0.84	0.88	0.75 (0.25)
0.5	0.25	0.79	0.85	0.62 (0.12)
0.9	0.05	0.71	0.81	0.57 (0.07)

Loans — Private attr.: Race (binary), budget $\epsilon = 0.01$

α	β	Rob. Acc	Test Acc	Infer. Acc (gap)
0	0.50	0.45	0.74	0.92 (0.22)
0.05	0.475	0.42	0.69	0.75 (0.05)
0.10	0.45	0.40	0.68	0.72 (0.02)
0.15	0.425	0.39	0.66	0.71 (0.01)

Adult income — Private attr.: Gender (binary), budget $\epsilon = 0.01$

α	β	Rob. Acc	Test Acc	Infer. Acc (gap)
0	0.5	0.63	0.68	0.88 (0.33)
0.05	0.475	0.57	0.67	0.72 (0.17)
0.10	0.45	0.55	0.65	0.59 (0.04)
0.20	0.4	0.53	0.63	0.55 (0.00)

Adult income — Private attr.: Marital status (7 values), budget $\epsilon = 0.01$

α	β	Rob. Acc	Test Acc	Infer. Acc (gap)
0	0.5	0.56	0.71	0.70 (0.14)
0.001	0.495	0.55	0.65	0.60 (0.04)
0.005	0.49	0.52	0.60	0.59 (0.03)
0.01	0.45	0.47	0.59	0.57 (0.01)

Related Work

Defenses against adversarial examples. Many efforts have been made to improve the adversarial robustness of ML models against adversarial examples (Goodfellow, Shlens, and Szegedy 2015; Kurakin, Goodfellow, and Bengio 2017; Pang et al. 2019; Zhang et al. 2019; Wong and Kolter 2018; Mao et al. 2019; Cohen, Rosenfeld, and Kolter 2019; Wang et al. 2019; Dong et al. 2020; Lecuyer et al. 2019; Zhai et al. 2020; Wong, Rice, and Kolter 2020; Zhou et al. 2021). Among them, adversarial training based defenses (Madry et al. 2018; Dong et al. 2020) has become the state-of-the-art. At a high level, adversarial training augments training data with adversarial examples, e.g., via CW attack (Carlini and Wagner 2017), PGD attack (Madry et al. 2018), AutoAttack (Croce and Hein 2020)), and uses a min-max formulation to train the target ML model (Madry et al. 2018). However, as pointed out by (Zhu, Zhang, and Evans 2020; Zhou et al. 2022), the dependence between the output of the target model and the input/adversarial examples has not been well studied, making the ability of adversarial training not fully exploited. To improve it, they propose to learn adversarially-robust representations via mutual information, which is shown to outperform the state-of-the-art adversarial training based defenses. Our ARPRL is inspired by them while having a nontrivial generalization to learn both robust and privacy-preserving representations with guarantees.

Defenses against inference attacks. Existing defense methods against inference attacks can be roughly classified as *adversarial learning* (Oh, Fritz, and Schiele 2017; Wu et al. 2018; Pittaluga, Koppal, and Chakrabarti 2019; Liu et al. 2019; Tripathy, Wang, and Ishwar 2019; Jeong et al. 2023), *differential privacy* (Shokri and Shmatikov 2015; Abadi et al. 2016), and *information obfuscation* (Bertran et al. 2019; Hamm 2017; Osia et al. 2018; Roy and Boddeti 2019; Zhao et al. 2020; Azam et al. 2022). Adversarial learning methods are inspired by GAN (Goodfellow et al. 2014) and they learn features from training data so that their private information cannot be inferred from a learnt model. However, these methods need to know the primary task and lack of formal privacy guarantees. Differential privacy methods have formal privacy guarantees, but they have high utility losses. Information obfuscation methods aim to maximize the utility, under the constraint of bounding the information leakage, but almost all of them are empirical and task-dependent. The only exception is (Zhao et al. 2020), which has guaranteed information leakage. However, this works requires stronger assumptions (e.g., conditional independence assumption between variables). Our work can be seen as a combination of information obfuscation with adversarial learning to learn both robust and privacy-preserving representations. It also has a privacy leakage guarantee and inherent trade-offs between robustness/utility and privacy.

Information-theoretical representation learning against inference attacks. Wang et al. (2021) propose to use mutual information to learn privacy-preserving representation on graphs against node (or link) inference attacks, while keeping the primary link (or node) prediction performance. Arevalo et al. (2024) learns task-agnostic privacy-preserving representations for federated learning against attribute inference attacks with privacy guarantees. Further, Noorbakhsh et al. (2024) developed an information-theoretic framework (called Inf^{2}Guard) to defend against common inference attacks including membership inference, property inference, and data reconstruction, while offering privacy guarantees.

Conclusion

We develop machine learning models to be robust against adversarial examples and protect sensitive attributes in training data. We achieve the goal by proposing ARPRL, which learns adversarially robust, privacy preserving, and utility preservation representations formulated via mutual information. We also derive theoretical results that show the inherent tradeoff between robustness/utility and privacy and guarantees of attribute privacy against the worst-case adversary.

Acknowledgements

We sincerely thank the anonymous reviewers for their constructive feedback. This research was partially supported by the Cisco Research Award and the National Science Foundation under grant Nos. ECCS-2216926, CNS-2241713, CNS-2331302, CNS-2339686, CNS-2302689, and CNS-2308730, CNS-2319277, and CMMI-2326341.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *CCS*.

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep variational information bottleneck. In *ICLR*.

Aono, Y.; Hayashi, T.; Wang, L.; and Moriai, S. 2017. Privacy-preserving deep learning: Revisited and enhanced. In *ATIS*.

Arevalo, C. A.; Noorbakhsh, S. L.; Dong, Y.; Hong, Y.; and Wang, B. 2024. Task-Agnostic Privacy-Preserving Representation Learning for Federated Learning against Attribute Inference Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10909–10917.

Azam, S. S.; Kim, T.; Hosseinalipour, S.; Joe-Wong, C.; Bagchi, S.; and Brinton, C. 2022. Can we generalize and distribute private representation learning? In *International Conference on Artificial Intelligence and Statistics*, 11320–11340. PMLR.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *ICML*.

Bertran, M.; Martinez, N.; Papadaki, A.; Qiu, Q.; Rodrigues, M.; Reeves, G.; and Sapiro, G. 2019. Adversarially learned representations for information obfuscation and inference. In *ICML*.

Bortsova, G.; González-Gonzalo, C.; Wetstein, S. C.; Dubost, F.; Katramados, I.; Hogeweg, L.; Liefers, B.; van Ginneken, B.; Pluim, J. P.; Veta, M.; et al. 2021. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*.

Calabro, C. 2009. *The exponential complexity of satisfiability problems*. University of California, San Diego.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE S & P*.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *ICML*.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified adversarial robustness via randomized smoothing. In *ICML*.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*.

Dong, Y.; Deng, Z.; Pang, T.; Zhu, J.; and Su, H. 2020. Adversarial distributional training for robust deep learning. In *NeurIPS*.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *CVPR*.

Gibbs, A. L.; and Su, F. E. 2002. On choosing and bounding probability metrics. *International statistical review*, 70(3): 419–435.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

Hamm, J. 2017. Minimax filter: Learning to preserve privacy from inference attacks. *JMLR*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3323–3331. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Jeong, J.; Cho, M.; Benz, P.; and Kim, T.-h. 2023. Noisy adversarial representation learning for effective and efficient image obfuscation. In Evans, R. J.; and Shpitser, I., eds., *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, 953–962. PMLR.

Jia, J.; Wang, B.; Zhang, L.; and Gong, N. Z. 2017. AttriInfer: Inferring User Attributes in Online Social Networks Using Markov Random Fields. In *WWW*.

Keahey, K.; Anderson, J.; Zhen, Z.; Riteau, P.; Ruth, P.; Stanzione, D.; Cevik, M.; Colleran, J.; Gunawi, H. S.; Hammock, C.; Mambretti, J.; Barnes, A.; Halbach, F.; Rocha, A.; and Stubbs, J. 2020. Lessons Learned from the Chameleon Testbed. In *USENIX ATC*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial machine learning at scale. In *ICLR*.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, 656–672. IEEE.

Li, A.; Guo, J.; Yang, H.; and Chen, Y. 2021. Deepobfuscator: Adversarial training framework for privacy-preserving image classification. *arXiv*.

Liao, P.; Zhao, H.; Xu, K.; Jaakkola, T.; Gordon, G. J.; Jegelka, S.; and Salakhutdinov, R. 2021. Information obfuscation of graph neural networks. In *ICML*.

Liu, S.; Du, J.; Shrivastava, A.; and Zhong, L. 2019. Privacy Adversarial Network: Representation Learning for Mobile Data Privacy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4): 1–18.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Mao, C.; Zhong, Z.; Yang, J.; Vondrick, C.; and Ray, B. 2019. Metric learning for adversarial robustness. *Advances in Neural Information Processing Systems*, 32.

McAllester, D.; and Stratos, K. 2020. Formal Limitations on the Measurement of Mutual Information. arXiv:1811.04251.

Melis, L.; Song, C.; De Cristofaro, E.; and Shmatikov, V. 2019. Exploiting unintended feature leakage in collaborative learning. In *IEEE SP*.

Noorbakhsh, S. L.; Zhang, B.; Hong, Y.; and Wang, B. 2024. Inf2Guard: An Information-Theoretic Framework for Learning Privacy-Preserving Representations against Inference Attacks. In *USENIX Security*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*.

Oh, S. J.; Fritz, M.; and Schiele, B. 2017. Adversarial image perturbation for privacy protection a game theory perspective. In *ICCV*.

Osia, S. A.; Taheri, A.; Shamsabadi, A. S.; Katevas, K.; Haddadi, H.; and Rabiee, H. R. 2018. Deep private-feature extraction. *IEEE TKDE*.

Pang, T.; Xu, K.; Du, C.; Chen, N.; and Zhu, J. 2019. Improving adversarial robustness via promoting ensemble diversity. In *ICML*.

Peng, X. B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; and Levine, S. 2019. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *ICLR*.

Pittaluga, F.; Koppal, S.; and Chakrabarti, A. 2019. Learning privacy preserving encodings through adversarial training. In *WACV*.

Poole, B.; Ozair, S.; Oord, A. v. d.; Alemi, A. A.; and Tucker, G. 2019. On variational bounds of mutual information. In *ICML*.

Roy, P. C.; and Boddeti, V. N. 2019. Mitigating information leakage in image representations: A maximum entropy approach. In *CVPR*.

Salem, A.; Cherubin, G.; Evans, D.; Köpf, B.; Paverd, A.; Suri, A.; Tople, S.; and Zanella-Béguelin, S. 2023. SoK: Let the privacy games begin! A unified treatment of data inference privacy in machine learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, 327–345. IEEE.

Shokri, R.; and Shmatikov, V. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1310–1321. ACM.

Song, L.; Shokri, R.; and Mittal, P. 2019a. Membership inference attacks against adversarially robust deep learning models. In *SPW*.

Song, L.; Shokri, R.; and Mittal, P. 2019b. Privacy risks of securing machine learning models against adversarial examples. In *CCS*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv*.

Tripathy, A.; Wang, Y.; and Ishwar, P. 2019. Privacy-Preserving Adversarial Networks. 495–505. IEEE Press.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).

Wang, B.; Guo, J.; Li, A.; Chen, Y.; and Li, H. 2021. Privacy-preserving representation learning on graphs: A mutual information perspective. In *Proceedings of the 27th acm sigkdd conference on knowledge discovery & data mining*, 1667–1676.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*.

Wong, E.; and Kolter, Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. In *ICLR*.

Wu, Z.; Wang, Z.; Wang, Z.; and Jin, H. 2018. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *ECCV*.

Zhai, R.; Dan, C.; He, D.; Zhang, H.; Gong, B.; Ravikumar, P.; Hsieh, C.-J.; and Wang, L. 2020. MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius. In *ICLR*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *ICML*.

Zhao, H.; Chi, J.; Tian, Y.; and Gordon, G. J. 2020. Trade-offs and guarantees of adversarial representation learning for information obfuscation. In *NeurIPS*.

Zhou, D.; Liu, T.; Han, B.; Wang, N.; Peng, C.; and Gao, X. 2021. Towards defending against adversarial examples via attack-invariant features. In *ICML*.

Zhou, D.; Wang, N.; Gao, X.; Han, B.; Wang, X.; Zhan, Y.; and Liu, T. 2022. Improving Adversarial Robustness via Mutual Information Estimation. In *ICML*.

Zhu, S.; Zhang, X.; and Evans, D. 2020. Learning adversarially robust representations via worst-case mutual information maximization. In *ICML*.

Algorithm 1

Algorithm 1: Adversarially robust and privacy-preserving representation learning (**ARPRL**)

Input: A dataset $\mathcal{D} = \{\mathbf{x}_i, y_i, u_i\}$, perturbation budget ϵ, $\alpha, \beta \in [0, 1]$, $\lambda > 0$, learning rates $lr_1, lr_2, lr_3, lr_4, lr_5$; #global epochs I, #local gradient steps J;
Output: Representation learner parameters Θ.

Initialize $\Theta, \Psi, \Omega, \Lambda$ for the representation learner f_Θ, privacy protection network g_Φ, utility preservation network h_Ω, and adversarially robust network t_Λ;
for $t = 0; t < T; t++$ **do**
  $L_1 = \sum_i CE(u_i, g_\Psi(f(\mathbf{x}_i)))$;
  $L_2 = \frac{1}{|\mathcal{D}|} \sum_i t_\Lambda(\mathbf{x}_i, f_\Theta(\mathbf{x}_i), u_i) - \log[\frac{1}{|\mathcal{D}|} \sum_i e^{t_\Lambda(\bar{\mathbf{x}}_i, f_\Theta(\mathbf{x}_i), u_i)}]$;
  $L_3 = I^{(JS)}_{\Theta,\Omega}(\mathbf{x}; f(\mathbf{x}), u)$;
  for $i = 0; i < I; i++$ **do**
    for $j = 0; j < J; j++$ **do**
      $\Psi \leftarrow \Psi - lr_1 \cdot \frac{\partial L_1}{\partial \Psi}$;
      $\Lambda \leftarrow \Lambda + lr_2 \cdot \frac{\partial L_2}{\partial \Lambda}$;
      $\Omega \leftarrow \Omega + lr_3 \cdot \frac{\partial L_3}{\partial \Omega}$;
    end for
    $\Theta \leftarrow \Theta + lr_4 \cdot \frac{\partial(\alpha L_1 + \beta L_2 + (1-\alpha-\beta) L_3)}{\partial \Theta}$;
  end for
end for

Derivation Details in Section 3

Deriving a upper bound of $I(\mathbf{z}; u)$: By applying the vCLUB bound proposed in (Cheng et al. 2020), we have:

$$I(\mathbf{z}; u) \leq I_{vCLUB}(\mathbf{z}; u) = \mathbb{E}_{p(\mathbf{z},u)}[\log q_\Psi(u|\mathbf{z})] - \mathbb{E}_{p(\mathbf{z})p(u)}[\log q_\Psi(u|\mathbf{z})], \quad (14)$$

where $q_\Psi(u|\mathbf{z})$ is an auxiliary posterior distribution of $p(u|\mathbf{z})$. and it needs to satisfy the condition:

$$KL(p(\mathbf{z}, u)||q_\Psi(\mathbf{z}, u)) \leq KL(p(\mathbf{z})p(u)||q_\Psi(\mathbf{z}, u)).$$

To achieve this, we need to minimize:

$$\begin{aligned} &\min_\Psi KL(p(\mathbf{z}, u)||q_\Psi(\mathbf{z}, u)) \\ &= \min_\Psi KL(p(u|\mathbf{z})||q_\Psi(u|\mathbf{z})) \\ &= \min_\Psi \mathbb{E}_{p(\mathbf{z},u)}[\log p(u|\mathbf{z})] - \mathbb{E}_{p(\mathbf{z},u)}[\log q_\Psi(u|\mathbf{z}))] \\ &\Longleftrightarrow \max_\Psi \mathbb{E}_{p(\mathbf{z},u)}[\log q_\Psi(u|\mathbf{z})], \end{aligned} \quad (15)$$

where we use that $\mathbb{E}_{p(\mathbf{z},u)}[\log p(u|\mathbf{z})]$ is irrelevant to Ψ.

Hence, minimizing $I_{vCLUB}(\mathbf{z}; u)$ reduces to maximizing $\mathbb{E}_{p(\mathbf{z},u)}[\log q_\Psi(u|\mathbf{z})]$.

Deriving a lower bound of $I(\mathbf{x}; \mathbf{z}|u)$: From (Nowozin, Cseke, and Tomioka 2016), we know

$$\begin{aligned} I(\mathbf{x}; \mathbf{z}|u) &= H(\mathbf{x}|u) - H(\mathbf{x}|\mathbf{z}, u) \\ &= H(\mathbf{x}|u) + \mathbb{E}_{p(\mathbf{x},\mathbf{z},u)}[\log p(\mathbf{x}|\mathbf{z}, u))] \\ &= H(\mathbf{x}|u) + \mathbb{E}_{p(\mathbf{x},\mathbf{z},u)}[\log q_\Omega(\mathbf{x}|\mathbf{z}, u))] \\ &\quad + \mathbb{E}_{p(\mathbf{x},\mathbf{z},u)}[KL(p(\cdot|\mathbf{z}, u)||q_\Omega(\cdot|\mathbf{z}, u))] \\ &\geq H(\mathbf{x}|u) + \mathbb{E}_{p(\mathbf{x},\mathbf{z},u)}[\log q_\Omega(\mathbf{x}|\mathbf{z}, u)], \end{aligned} \quad (16)$$

where q_Ω is an *arbitrary* auxiliary posterior distribution that aims to maintain the information $\mathbf{x}$ in the representation $\mathbf{z}$ conditioned on the private u.

Proofs

Proof of Theorem 2

Theorem 2. *Let all binary task classifiers be $\mathcal{C} = \{C : \mathcal{Z} \to \mathcal{Y}\}$. Then for any representation learner $f : \mathcal{X} \to \mathcal{Z}$,*

$$\inf_{C \in \mathcal{C}} \mathrm{AdvRisk}_\epsilon(C \circ f) \geq \frac{\mathrm{RV}_\epsilon(f|u) - I(\mathbf{x}; \mathbf{z}|u)}{\log 2}. \quad (11)$$

Proof. Replacing $I(\mathbf{x}; \mathbf{z})$ and $\mathrm{RV}_\epsilon(f)$ in Theorem 1 with $I(\mathbf{x}; \mathbf{z}|u)$ and $\mathrm{RV}_\epsilon(f|u)$, and setting $|\mathcal{Y}| = 2$, we reach Theorem 2. □

Proof of Theorem 3

We first introduce the following definitions and lemmas that will be used to prove Theorem 3.

Definition 1 (Lipschitz function and Lipschitz norm)**.** *We say a function $f : A \to \mathbb{R}^m$ is L-Lipschitz continuous, if for any $a, b \in A$, $\|f(a) - f(b)\| \leq L \cdot \|a - b\|$. Lipschitz norm of f, i.e., $\|f\|_L$, is defined as $\|f\|_L = \max \frac{\|f(a)-f(b)\|_L}{\|a-b\|_L}$.*

Definition 2 (Total variance (TV) distance)**.** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two distributions over the same sample space Γ, the TV distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is defined as: $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) = \max_{E \subseteq \Gamma} |\mathcal{D}_1(E) - \mathcal{D}_2(E)|$.*

Definition 3 (1-Wasserstein distance)**.** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two distributions over the same sample space Γ, the 1-Wasserstein distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is defined as $W_1(\mathcal{D}_1, \mathcal{D}_2) = \max_{\|f\|_L \leq 1} |\int_\Gamma f d\mathcal{D}_1 - \int_\Gamma f d\mathcal{D}_2|$, where $\|\cdot\|_L$ is the Lipschitz norm of a real-valued function.*

Definition 4 (Pushforward distribution)**.** *Let $\mathcal{D}$ be a distribution over a sample space and g be a function of the same space. Then we call $g(\mathcal{D})$ the pushforward distribution of $\mathcal{D}$.*

Lemma 1 (Contraction of the 1-Wasserstein distance)**.** *Let g be a function defined on a space and L be constant such that $\|g\|_L \leq C_L$. For any distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over this space, $W_1(g(\mathcal{D}_1), g(\mathcal{D}_2)) \leq C_L \cdot W_1(\mathcal{D}_1, \mathcal{D}_2)$.*

Lemma 2 (1-Wasserstein distance on Bernoulli random variables)**.** *Let y_1 and y_2 be two Bernoulli random variables with distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Then, $W_1(\mathcal{D}_1, \mathcal{D}_2) = |Pr(y_1 = 1) - Pr(y_2 = 1)|$.*

Lemma 3 (Relationship between the 1-Wasserstein distance and the TV distance (Gibbs and Su 2002))**.** *Let g be a function defined on a norm-bounded space $\mathcal{Z}$, where $\max_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z}\| \leq R$, and $\mathcal{D}_1$ and $\mathcal{D}_1$ are two distributions over the space $\mathcal{Z}$. Then $W_1(g(\mathcal{D}_1), g(\mathcal{D}_2)) \leq 2R \cdot d_{TV}(g(\mathcal{D}_1), g(\mathcal{D}_2))$.*

Lemma 4 (Relationship between the TV distance and advantage (Liao et al. 2021))**.** *Given a binary attribute $u \in \{0, 1\}$. Let $\mathcal{D}_{u=a}$ be the conditional data distribution of $\mathcal{D}$ given $u = a$ over a sample space Γ. Let $Adv_\mathcal{D}(\mathcal{A})$ be the advantage of adversary. Then for any function f, we have $d_{TV}(f(\mathcal{D}_{u=0}), f(\mathcal{D}_{u=1})) = Adv_\mathcal{D}(\mathcal{A})$.*

We now prove Theorem 3, which is restated as below:

Theorem 3. *Let $\mathbf{z} = f(\mathbf{x})$ be with a bounded norm R (i.e., $\max_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{z}\| \leq R$), and $\mathcal{A}$ be the set of all binary inference classifiers taking $\mathbf{z}$ as an input. Assume the task classifier C is C_L-Lipschitz, i.e., $\|C\|_L \leq C_L$. Then, we have the below relationship between the standard risk and the advantage:*

$$\mathrm{Risk}(C \circ f) \geq \Delta_{y|u} - 2R \cdot C_L \cdot Adv_{\mathcal{D}}(\mathcal{A}), \tag{12}$$

where $\Delta_{y|u} = |Pr_{\mathcal{D}}(y = 1|u = 0) - Pr_{\mathcal{D}}(y = 1|u = 1)|$ is a dataset-dependent constant.

Proof. We denote $\mathcal{D}_{u=a}$ as the conditional data distribution of $\mathcal{D}$ given $u = a$, and $\mathcal{D}_{y|u}$ as the conditional distribution of label y given u. cf is denoted as the (binary) composition function of $c \circ f_\Theta$. As c is binary task classifier on the learnt representations, it follows that the pushforward $cf(\mathcal{D}_{u=0})$ and $cf(\mathcal{D}_{u=1})$ induce two distributions over the binary label space $\mathcal{Y} = \{0, 1\}$. By leveraging the triangle inequalities of the 1-Wasserstein distance, we have

$$\begin{aligned}
&W_1(\mathcal{D}_{y|u=0}, \mathcal{D}_{y|u=1}) \\
&\leq W_1(\mathcal{D}_{y|u=0}, cf(\mathcal{D}_{u=0})) + W_1(cf(\mathcal{D}_{u=0}), cf(\mathcal{D}_{u=1})) \\
&\quad + W_1(cf(\mathcal{D}_{u=1}), \mathcal{D}_{y|u=1})
\end{aligned} \tag{17}$$

Using Lemma 2 on Bernoulli random variables $y|u = a$:

$$\begin{aligned}
&W_1(\mathcal{D}_{y|u=0}, \mathcal{D}_{y|u=1}) \\
&= |\mathrm{Pr}_{\mathcal{D}}(y = 1|u = 0) - \mathrm{Pr}_{\mathcal{D}}(y = 1|u = 1)| \\
&= \Delta_{y|u}.
\end{aligned} \tag{18}$$

Using Lemma 1 on the contraction of the 1-Wasserstein distance and that $\|c\|_L \leq C_L$, we have

$$W_1(cf(\mathcal{D}_{u=0}), cf(\mathcal{D}_{u=1})) \leq C_L \cdot W_1(f(\mathcal{D}_{u=0}), f(\mathcal{D}_{u=1})). \tag{19}$$

Using Lemma 3 with $\max_{\mathbf{z}} \|\mathbf{z}\| \leq R$, we have

$$\begin{aligned}
&W_1(f(\mathcal{D}_{u=0}), f(\mathcal{D}_{u=1})) \\
&\leq 2R \cdot d_{TV}(f(\mathcal{D}_{u=0}), f(\mathcal{D}_{u=1})) \\
&= 2R \cdot \mathrm{Adv}_{\mathcal{D}}(\mathcal{A}),
\end{aligned} \tag{20}$$

where the last equation is based on Lemma 4.

Combing Equations 19 and 20, we have

$$W_1(cf(\mathcal{D}_{u=0}), cf(\mathcal{D}_{u=1})) \leq 2R \cdot C_L \cdot \mathrm{Adv}_{\mathcal{D}}(\mathcal{A}).$$

Furthermore, using Lemma 2 on Bernoulli random variables y and $cf(\mathbf{x})$, we have

$$\begin{aligned}
&W_1(\mathcal{D}_{y|u=a}, cf(\mathcal{D}_{u=a})) \\
&= |\mathrm{Pr}_{\mathcal{D}}(y = 1|u = a) - \mathrm{Pr}_{\mathcal{D}}(cf(\mathbf{x}) = 1|u = a))| \\
&= |\mathbb{E}_{\mathcal{D}}[y|u = a] - \mathbb{E}_{\mathcal{D}}[cf(\mathbf{x})|u = a]| \\
&\leq \mathbb{E}_{\mathcal{D}}[|y - cf(\mathbf{x})||u = a] \\
&= \mathrm{Pr}_{\mathcal{D}}(y \neq cf(\mathbf{x})|u = a) \\
&= \mathrm{Risk}_{u=a}(c \circ f).
\end{aligned} \tag{21}$$

Hence, $W_1(\mathcal{D}_{y|u=0}, cf(\mathcal{D}_{u=0})) + W_1(\mathcal{D}_{y|u=1}, cf(\mathcal{D}_{u=1})) \leq \mathrm{Risk}(c \circ f)$.

Finally, by combining Equations (17) - (21), we have:

$$\Delta_{y|u} \leq \mathrm{Risk}(c \circ f) + 2R \cdot C_L \cdot \mathrm{Adv}_{\mathcal{D}}(\mathcal{A}), \tag{22}$$

thus $\mathrm{Risk}(c \circ f) \geq \Delta_{y|u} - 2R \cdot C_L \cdot \mathrm{Adv}_{\mathcal{D}}(\mathcal{A})$, completing the proof. □

Proof of Theorem 4

We follow the way as proving Theorem 3. We first restate Theorem 4 as below:

Theorem 4. *Let $\mathbf{z}' = f(\mathbf{x}')$ be the learnt representation for $\mathbf{x}' \in \mathcal{B}(\mathbf{x}, \epsilon)$ with a bounded norm R (i.e., $\max_{\mathbf{z}'\in\mathcal{Z}} \|\mathbf{z}'\| \leq R$), and $\mathcal{A}$ be the set of all binary inference classifiers. Under a C_L-Lipschitz task classifier C, we have the below relationship between the adversarial risk and the advantage:*

$$\mathrm{AdvRisk}_\epsilon(C \circ f) \geq \Delta_{y|u} - 2R \cdot C_L \cdot Adv_{\mathcal{D}'}(\mathcal{A}). \tag{13}$$

Proof. Recall that D' is a joint distribution of the perturbed input $\mathbf{x}'$, the label y, and private attribute u. We denote $\mathcal{D}'_{u=a}$ as the conditional perturbed data distribution of $\mathcal{D}'$ given $u = a$, and $\mathcal{D}'_{y|u}$ as the conditional distribution of label y given u. Also, the pushforward $cf(\mathcal{D}'_{u=a})$ induces two distributions over the binary label space $\mathcal{Y} = \{0, 1\}$ with $a = \{0, 1\}$. Via the triangle inequalities of the 1-Wasserstein distance, we have

$$\begin{aligned}
&W_1(\mathcal{D}'_{y|u=0}, \mathcal{D}'_{y|u=1}) \\
&\leq W_1(\mathcal{D}'_{y|u=0}, cf(\mathcal{D}'_{u=0})) + W_1(cf(\mathcal{D}'_{u=0}), cf(\mathcal{D}'_{u=1})) \\
&\quad + W_1(cf(\mathcal{D}'_{u=1}), \mathcal{D}'_{y|u=1})
\end{aligned} \tag{23}$$

Using Lemma 2 on Bernoulli random variables $y|u = a$:

$$\begin{aligned}
W_1(\mathcal{D}'_{y|u=0}, \mathcal{D}'_{y|u=1}) &= |\mathrm{Pr}_{\mathcal{D}'}(y = 1|u = 0) - \mathrm{Pr}_{\mathcal{D}'}(y = 1|u = 1)| \\
&= |\mathrm{Pr}_{\mathcal{D}}(y = 1|u = 0) - \mathrm{Pr}_{\mathcal{D}}(y = 1|u = 1)| \\
&= \Delta_{y|u},
\end{aligned} \tag{24}$$

where we use that the perturbed data and clean data share the same label y condition on u.

Then following the proof of Theorem 3, we have:

$$W_1(cf(\mathcal{D}'_{u=0}), cf(\mathcal{D}'_{u=1})) \leq C_L \cdot W_1(f(\mathcal{D}'_{u=0}), f(\mathcal{D}'_{u=1})); \tag{25}$$

$$W_1(f(\mathcal{D}'_{u=0}), f(\mathcal{D}'_{u=1})) \leq 2R \cdot d_{TV}(f(\mathcal{D}'_{u=0}), f(\mathcal{D}'_{u=1})). \tag{26}$$

We further show $d_{TV}(f(\mathcal{D}'_{u=0}), f(\mathcal{D}'_{u=1})) = \mathrm{Adv}_{\mathcal{D}'}(\mathcal{A})$:

$$\begin{aligned}
&d_{TV}(f(\mathcal{D}'_{u=0}), f(\mathcal{D}'_{u=1})) \\
&= \max_E |\mathrm{Pr}_{f(\mathcal{D}'_{u=0})}(E) - \mathrm{Pr}_{f(\mathcal{D}'_{u=1})}(E)| \\
&= \max_{A_E\in\mathcal{A}} |\mathrm{Pr}_{\mathbf{z}'\sim f(\mathcal{D}'_{u=0})}(A_E(\mathbf{z}') = 1) - \mathrm{Pr}_{\mathbf{z}'\sim f(\mathcal{D}'_{u=1})}(A_E(\mathbf{z}) = 1)| \\
&= \max_{A_E\in\mathcal{A}} |\mathrm{Pr}(A_E(\mathbf{z}') = 1|u = 0) - \mathrm{Pr}(A_E(\mathbf{z}') = 1|u = 1)| \\
&= \mathrm{Adv}_{\mathcal{D}'}(\mathcal{A}),
\end{aligned} \tag{27}$$

where the first equation uses the definition of TV distance, and $A_E(\cdot)$ is the characteristic function of the event E in the second equation.

With Equations (25) - (27), we have

$$W_1(cf(\mathcal{D}'_{u=0}), cf(\mathcal{D}'_{u=1})) \leq 2R \cdot C_L \cdot \mathrm{Adv}_{\mathcal{D}'}(\mathcal{A}).$$

Furthermore, using Lemma 2 on Bernoulli random variables y and $cf(\mathbf{x})$, we have

$$\begin{aligned}
&W_1(\mathcal{D}'_{y|u=0}, cf(\mathcal{D}'_{u=0})) + W_1(\mathcal{D}'_{y|u=1}, cf(\mathcal{D}'_{u=1})) \\
&= |\mathrm{Pr}_{\mathcal{D}'}(y=1|u=0) - \mathrm{Pr}_{\mathcal{D}'}(cf(\mathbf{x}')=1|u=0))| \\
&\quad + |\mathrm{Pr}_{\mathcal{D}'}(y=1|u=1) - \mathrm{Pr}_{\mathcal{D}'}(cf(\mathbf{x}')=1|u=1))| \\
&= |\mathbb{E}_{\mathcal{D}'}[y|u=0] - \mathbb{E}_{\mathcal{D}'}[cf(\mathbf{x}')|u=0]| \\
&\quad + |\mathbb{E}_{\mathcal{D}'}[y|u=1] - \mathbb{E}_{\mathcal{D}'}[cf(\mathbf{x}')|u=1]| \\
&\leq \mathbb{E}_{\mathcal{D}'}[|y-cf(\mathbf{x}')||u=0] + \mathbb{E}_{\mathcal{D}'}[|y-cf(\mathbf{x}')||u=1] \\
&= \mathrm{Pr}_{\mathcal{D}'}(y \neq cf(\mathbf{x}')|u=0) + \mathrm{Pr}_{\mathcal{D}'}(y \neq cf(\mathbf{x}')|u=1) \\
&= \mathrm{Pr}_{\mathcal{D}'}(y \neq cf(\mathbf{x}')) \\
&= \mathrm{Pr}_{\mathcal{D}}[\exists \mathbf{x}' \in \mathcal{B}(\mathbf{x},\epsilon), \text{ s.t. } cf(\mathbf{x}') \neq y] \\
&= \mathrm{AdvRisk}_\epsilon(c \circ f). \qquad (28)
\end{aligned}$$

Finally, by combining Equations (23) - (28), we have:

$$\Delta_{y|u} \leq \mathrm{AdvRisk}_\epsilon(c \circ f) + 2R \cdot C_L \cdot \mathrm{Adv}_{\mathcal{D}'}(\mathcal{A})$$

Hence, $\mathrm{AdvRisk}_\epsilon(c \circ f) \geq \Delta_{y|u} - 2R \cdot C_L \cdot \mathrm{Adv}_{\mathcal{D}'}(\mathcal{A})$, completing the proof.

□

Proof of Theorem 5

We first point out that (Zhao et al. 2020) also provide the theoretical result in Theorem 3.1 against attribute inference attacks. However, there are two key differences between theirs and our Theorem 5: First, Theorem 3.1 requires an assumption $I(\hat{A}; A|Z) = 0$, while our Theorem 5 does not need extra assumption; 2) The proof for Theorem 3.1 decomposes an joint entropy $H(A, \hat{A}, E)$, while our proof decomposes a conditional entropy $H(s, u|A(z))$. We note that the main idea to prove both theorems is by introducing an indicator and decomposing an entropy in two different ways.

The following lemma about the inverse binary entropy will be useful in the proof of Theorem 5:

Lemma 5 ((Calabro 2009) Theorem 2.2)**.** *Let $H_2^{-1}(p)$ be the inverse binary entropy function for $p \in [0,1]$, then $H_2^{-1}(p) \geq \frac{p}{2\log_2(\frac{6}{p})}$.*

Lemma 6 (Data processing inequality)**.** *Given random variables X, Y, and Z that form a Markov chain in the order $X \rightarrow Y \rightarrow Z$, then the mutual information between X and Y is greater than or equal to the mutual information between X and Z. That is $I(X;Y) \geq I(X;Z)$.*

With the above lemma, we are ready to prove Theorem 5 restated as below.

Theorem 5. *Let $\mathbf{z}$ be the learnt representation by Equation (9). For any attribute inference adversary $\mathcal{A} = \{A : \mathcal{Z} \rightarrow \mathcal{U} = \{0,1\}\}$, $Pr(A(\mathbf{z}) = u) \leq 1 - \frac{H(u|\mathbf{z})}{2\log_2(6/H(u|\mathbf{z}))}$.*

Proof. Let s be an indicator that takes value 1 if and only if $\mathcal{A}(\mathbf{z}) \neq u$, and 0 otherwise, i.e., $s = 1[\mathcal{A}(\mathbf{z}) \neq u]$. Now consider the conditional entropy $H(s, u|\mathcal{A}(\mathbf{z}))$ associated with $\mathcal{A}(\mathbf{z})$, u, and s. By decomposing it in two different ways, we have

$$\begin{aligned}
H(s, u|\mathcal{A}(\mathbf{z})) &= H(u|\mathcal{A}(\mathbf{z})) + H(s|u, \mathcal{A}(\mathbf{z})) \\
&= H(s|\mathcal{A}(\mathbf{z})) + H(u|s, \mathcal{A}(\mathbf{z})). \qquad (29)
\end{aligned}$$

Note that $H(s|u, \mathcal{A}(\mathbf{z})) = 0$ as when u and $\mathcal{A}(\mathbf{z})$ are known, s is also known. Similarly,

$$\begin{aligned}
H(u|s, \mathcal{A}(\mathbf{z})) &= Pr(s=1)H(u|s=1, \mathcal{A}(\mathbf{z})) \\
&\quad + Pr(s=0)H(u|s=0, \mathcal{A}(\mathbf{z})) \\
&= 0 + 0 = 0, \qquad (30)
\end{aligned}$$

because when we know s's value and $\mathcal{A}(\mathbf{z})$, we also actually knows u.

Thus, Equation 29 reduces to $H(u|\mathcal{A}(\mathbf{z})) = H(s|\mathcal{A}(\mathbf{z}))$. As conditioning does not increase entropy, i.e., $H(s|\mathcal{A}(\mathbf{z})) \leq H(s)$, we further have

$$H(u|\mathcal{A}(\mathbf{z})) \leq H(s). \qquad (31)$$

On the other hand, using mutual information and entropy properties, we have $I(u; \mathcal{A}(\mathbf{z})) = H(u) - H(u|\mathcal{A}(\mathbf{z}))$ and $I(u; \mathbf{z}) = H(u) - H(u|\mathbf{z})$. Hence,

$$I(u; \mathcal{A}(\mathbf{z})) + H(u|\mathcal{A}(\mathbf{z})) = I(u; \mathbf{z}) + H(u|\mathbf{z}). \qquad (32)$$

Notice $\mathcal{A}(\mathbf{z})$ is a random variable such that $u \perp \mathcal{A}(\mathbf{z})|\mathbf{z}$. Hence, we have the Markov chain $u \rightarrow \mathbf{z} \rightarrow \mathcal{A}(\mathbf{z})$. Based on the data processing inequality in Lemma 6, we know $I(u; \mathcal{A}(\mathbf{z})) \leq I(u; \mathbf{z})$. Combining it with Equation 32, we have

$$H(u|\mathcal{A}(\mathbf{z})) \geq H(u|\mathbf{z}). \qquad (33)$$

Combing Equations (31) and (33), we have $H(s) = H_2(Pr(s=1)) \geq H(u|\mathbf{z})$, which implies

$$Pr(\mathcal{A}(\mathbf{z}) \neq u) = Pr(s=1) \geq H_2^{-1}(H(u|\mathbf{z})),$$

where $H_2(t) = -t\log_2 t - (1-t)\log_2(1-t)$.

Finally, by applying Lemma 5, we have

$$Pr(\mathcal{A}(\mathbf{z}) \neq u) \geq \frac{H(u|\mathbf{z})}{2\log_2(6/H(u|\mathbf{z}))}.$$

Hence the attribute privacy leakage is bounded by $Pr(\mathcal{A}(\mathbf{z}) = u) \leq 1 - \frac{H(u|\mathbf{z})}{2\log_2(6/H(u|\mathbf{z}))}$.

□

Discussion on the Tightness of the MI Bound

We leverage mutual information as a tool to learn robust and privacy-preserving representations as well as deriving theoretical results. However, computing the exact value of mutual information on high-dimensional feature spaces, as commonly encountered in deep learning, remains a computational challenge. Several works (Alemi et al. 2017; Belghazi et al. 2018; Poole et al. 2019; Hjelm et al. 2019; Cheng et al. 2020) have introduced various neural estimators to approximate mutual information by providing bounds. Unfortunately, as shown by (McAllester and Stratos 2020), these bounds are generally loose. At the moment, developing a method to achieve a tight mutual information bound, along with analyzing the associated approximation error, remains an open question in the field.

Datasets and network architectures

Detailed dataset descriptions

CelebA dataset (Liu et al. 2015). CelebA consists of more than 200K face images with size 32x32. Each face image is labeled with 40 binary facial attributes. In the experiments, we use 150K images for training and 50K images for testing. We treat binary 'gender' as the private attribute, and detect 'gray hair' as the primary (binary classification) task.

Loans dataset (Hardt, Price, and Srebro 2016). This dataset is originally extracted from the loan-level Public Use Databases. The Federal Housing Finance Agency publishes these databases yearly, containing information about the Enterprises' single family and multifamily mortgage acquisitions. Specifically, the database used in this project is a single-family dataset and has a variety of features related to the person asking for a mortgage loan. All the attributes in the dataset are numerical, so no preprocessing from this side was required. On the other hand, in order to create a balanced classification problem, some of the features were modified to have a similar number of observations belonging to all classes. We use 80% data for training and 20% for testing.

The utility under this scope was measured in the system accurately predicting the affordability category of the person asking for a loan. This attribute is named *Affordability*, and has three possible values: 0 if the person belongs to a mid-income family and asking for a loan in a low-income area, 1 if the person belongs to a low-income family and asking for a loan in a low-income area, and 2 if the person belongs to a low-income family and is asking for a loan not in a low-income area. The private attribute was set to be binary *Race*, being White (0) or Not White (1).

Adult Income dataset (Dua and Graff 2017). This is a well-known dataset available in the UCI Machine Learning Repository. The dataset contains 32,561 observations each with 15 features, some of them numerical, other strings. Those attributes are not numerical were converted into categorical using an encoder. Again, we use the 80%-20% train-test split.

The primary classification task is predicting if a person has an income above $50,000, labeled as 1, or below, which is labeled as 0. The private attributes to predict are the *Gender*, which is binary, and the *Marital Status*, which has seven possible labels: 0 if Divorced, 1 if AF-spouse, 2 if Civil-spouse, 3 if Spouse absent, 4 if Never married, 5 if Separated, and 6 if Widowed.

Network architectures

The used network architectures for the three neural networks are in Table 2.

How to Choose α and β

Assume we reach the required utility with a (relatively large) value $1-\alpha-\beta$ (e.g., 0.7, 0.8; note its regularization controls the utility). Then we have a principled way to efficiently tune α and β based on their meanings:

1) We will start with three sets of $(\alpha 1, \beta 1), (\alpha 2, \beta 2), (\alpha 3, \beta 3)$, where one is with $\alpha 1 = \beta 1$, one is with a larger $\alpha 2 > \alpha 1$ (i.e., better privacy), and one is with a larger $\beta 3 > \beta 1$ (better robustness), respectively.

2) Based on the three results, we know whether a larger α or β is needed to obtain a better privacy-robustness tradeoff and set their values via a binary search. For instance, if needing more privacy protection, we can set a larger $\alpha 4 = \frac{\alpha 1 + \alpha 2}{2}$; or needing more robustness, we can set a larger $\beta 4 = \frac{\beta 1 + \beta 3}{2}$.

Step 2) continues until finding the optimal tradeoff α and β.

Table 2: Network architectures for the used datasets. Note that utility preservation network is the same as robust network.

Rep. Learner	Robust Network	Privacy Network	Utility Network
CelebA			
conv1-64	conv3-64	linear-32 MaxPool	conv3-64
conv64-128	conv64-128	linear-#priv. attri. values	conv64-128
linear-1024	conv128-256		conv128-256
linear-64	conv2048-2048		conv2048-2048
Loans and Adult Income			
linear-12	linear-64	linear-16	linear-64
linear-3	linear-3	linear-#priv. attri. values	linear-3
Toy dataset			
linear-10	linear-64	linear-5	linear-64
linear-2	linear-2	linear-#priv. attri. values	linear-2