Fundamental Limits of Perfect Concept Erasure

Somnath Basu Roy Chowdhury[1] **Avinava Dubey**[2] **Ahmad Beirami**[3] **Rahul Kidambi**[2]
Nicholas Monath[3] **Amr Ahmed**[2] **Snigdha Chaturvedi**[1]

[1]UNC Chapel Hill [2]Google Research [3]Google DeepMind
{somnath, snigdha}@cs.unc.edu, {avinavadubey, beirami, rahulkidambi, nmonath}@google.com

Abstract

Concept erasure is the task of erasing information about a concept (e.g., gender or race) from a representation set while retaining the maximum possible utility — information from original representations. Concept erasure is useful in several applications, such as removing sensitive concepts to achieve fairness and interpreting the impact of specific concepts on a model's performance. Previous concept erasure techniques have prioritized robustly erasing concepts over retaining the utility of the resultant representations. However, there seems to be an inherent tradeoff between erasure and retaining utility, making it unclear how to achieve perfect concept erasure while maintaining high utility. In this paper, we offer a fresh perspective toward solving this problem by quantifying the fundamental limits of concept erasure through an information-theoretic lens. Using these results, we investigate constraints on the data distribution and the erasure functions required to achieve the limits of perfect concept erasure. Empirically, we show that the derived erasure functions achieve the optimal theoretical bounds. Additionally, we show that our approach outperforms existing methods on a range of synthetic and real-world datasets using GPT-4 representations.

1 INTRODUCTION

Modern large-scale machine learning models (Achiam et al., 2023; Team et al., 2023) trained in a self-supervised manner have showcased an impressive array of capabilities across a range of applications (Ahn et al., 2022; Sun, 2023; Singhal et al., 2023). Owing to their success, in many applications practitioners rely on data representations from large models (Kenton and Toutanova, 2019; Dosovitskiy et al., 2021; Radford et al., 2021; Touvron et al., 2023) and fine-tune smaller networks on top of these representations. Such representations often encode a diverse set of *concepts* (Hinton et al., 1986) about the underlying data. For example, concepts associated with a video representation can be the geographic location or pose of a person. Concepts can also be more abstract, e.g., text representations may encode writing style (Patel et al., 2023), topics (Sarkar et al., 2023), or even sensitive information like gender or race of the author (Bolukbasi et al., 2016). Since practitioners cannot control the concepts encoded in representations, machine learning models may learn to rely on such sensitive concepts, thereby inadvertently impacting downstream applications. For example, while developing a resume screening system, an organization may want to ensure that the candidate's gender does not affect the hiring decisions. This calls for techniques that enable practitioners to erase certain concepts encoded in data representations.

To address these challenges, we focus on *concept erasure* (Ravfogel et al., 2020, 2022a; Chowdhury et al., 2024), i.e., removing information about a concept (e.g., gender) from a representation set. This is performed by transforming the original representations using an erasure function. The transformed representations should encode minimal information about the erased concept (*privacy*) while retaining the maximal information about the original representations (*utility*). This differs from conventional invariant representation learning (Zhao et al., 2019; Nguyen et al., 2021; Zhao et al., 2022; Sadeghi et al., 2022), which only retains information about a specific task (e.g., hiring). Since concept erasure is task-agnostic, its output representations can be used in a wide range of applications. This is useful when a data distributor wants to share data represen-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

tations with its customers while safeguarding sensitive concepts, such as the gender or race of the data producer (Zemel et al., 2013). Additionally, prior work has shown that concept erasure is useful in improving fairness (Ravfogel et al., 2020; Chowdhury and Chaturvedi, 2022; Belrose et al., 2024) and interpretability (Gonen et al., 2020; Elazar et al., 2021).

A long line of work (Bolukbasi et al., 2016; Ravfogel et al., 2020, 2022a,b; Chowdhury and Chaturvedi, 2022; Chowdhury et al., 2024) has focused on developing effective concept erasure techniques. Most of these approaches prioritize only erasing concept information. However, in concept erasure, we want to achieve perfect erasure while retaining maximum utility. Motivated by this challenge, we aim to answer the research question: *What is the maximum utility that can be retained while perfectly erasing a concept, and what functions achieve perfect erasure?*

In this work, we take a step towards answering the above question by deriving the functions that achieve perfect concept erasure. We formalize the concept erasure task using information-theoretic tools, introducing the definitions of perfect and relaxed concept erasure. We borrow tools from privacy and make a conceptual connection with trade-offs between utility and privacy (Sreekumar and Gündüz, 2019; Bertran et al., 2019; Razeghi et al., 2020; Atashin et al., 2021) and build on (Calmon et al., 2015, 2017) that characterize the fundamental limits of perfect privacy. In these data settings, we derive the erasure functions that achieve perfect concept erasure, which we call *perfect erasure functions* (`PEFs`). In controlled settings, we show that existing approaches either reveal a significant amount of concept information or entirely lose all utility information from the original representations.

To summarize, our primary contributions are:

- We formalize the task of concept erasure, introducing the definitions of perfect and relaxed concept erasure using information-theoretic tools.
- We hypothesize constraints on the underlying data required to achieve the information-theoretic outer bounds for perfect concept erasure.
- We derive the perfect erasure functions under different data conditions.
- We extensively evaluate `PEF` on synthetic and real-world datasets, including GPT-4 representations.

2 BACKGROUND

Concept Erasure. The idea of removing information from data representations was motivated by the problem of removing gender information from GloVe embeddings (Bolukbasi et al., 2016). This task was later termed as *concept erasure* (Ravfogel et al., 2020, 2022a), which involved removing information about a *concept* (a random variable) from a representation set. Concept erasure differs from conventional approaches to removing information from representations, such as adversarial learning (Zemel et al., 2013; Madras et al., 2018; Elazar and Goldberg, 2018; Sadeghi et al., 2022; Iwasawa et al., 2018; Chowdhury et al., 2021; Xie et al., 2017), which aim to learn invariant representations for a specific task. For concept erasure, the user does not have access to a specific task and should retain the maximum utility from original representations.

A simplified form of this task is *linear concept erasure* (Ravfogel et al., 2020; Dev et al., 2021; Ravfogel et al., 2022a), which removes concept information in a manner that prevents a linear adversary (e.g., a linear classifier) from extracting it. Initial approaches involve projecting representations onto the nullspace of linear concept classifiers, either once (Bolukbasi et al., 2016; Haghighatkhah et al., 2022) or in an iterative manner (Ravfogel et al., 2020). Subsequent work (Ravfogel et al., 2022a) generalized the iterative nullspace projection objective as a minimax game. Recently, Belrose et al. (2024) show that a concept is perfectly linearly erased only when different concept groups share the same centroid representation and provide an optimal algorithm to achieve this condition.

Though theoretically sound and effective in practice, linear concept erasure has limitations, as the erased concept can still be retrieved using a non-linear adversary. Recent works have focused on performing *general concept erasure*, which tries to protect the concept against an arbitrarily strong adversary. One approach, utilized by (Ravfogel et al., 2022b; Shao et al., 2023), involves projecting inputs onto a non-linear space and then applying linear concept erasure techniques. On the other hand, Chowdhury and Chaturvedi (2022); Chowdhury et al. (2024) learn a parameterized erasure function using rate-distortion based objectives. While empirically promising, these approaches cannot theoretically guarantee perfect erasure nor the retention of maximum possible utility.

General concept erasure techniques often showcase a significant information/utility loss (Chowdhury et al., 2024), which is expected as perfectly erasing a concept may remove other information. An extreme case occurs when a function outputs random representations, which perfectly erases a concept but results in complete information loss (Lowy et al., 2022). Therefore, we quantify the maximum possible information that can be retained while perfectly erasing a concept. In this work, we use mutual information to quantify the utility and privacy during concept erasure (Sankar et al., 2013). Mutual information minimization has been used to improve

privacy (Duchi et al., 2014; Bertran et al., 2019) and fairness (Mary et al., 2019; Lowy et al., 2022) in many applications. However, we do not optimize the mutual information terms directly; instead, we utilize their outer bounds. This problem can be studied using other divergence measures as well. We use mutual information as its outer bounds enable us to derive the analytic form of the erasure functions.

A long line of work in information theory (Calmon et al., 2017; Sreekumar and Gündüz, 2019; Bertran et al., 2019; Atashin et al., 2021) has focused on achieving an optimal tradeoff between privacy, utility, and function complexity by optimizing the associated mutual information terms. Bottleneck CLUB (Razeghi et al., 2023) provides a general framework to unify these approaches. Concept erasure is a special case of this framework, where we focus only on utility and privacy (also called the privacy funnel (Makhdoumi et al., 2014)). Calmon et al. (2015) was the first to study perfect privacy (or perfect erasure), deriving the PIC conditions necessary to achieve it within the privacy funnel framework. Calmon et al. (2017) derived the mutual information bounds for the privacy funnel setting, while Hsu et al. (2018) generalized the privacy funnel problem to a broader class of f-divergences. Building on this, Wang and Calmon (2017); Wang et al. (2019) focused on erasing functions of the data such that those cannot be reconstructed (using MMSE) and derived sharp privacy-utility tradeoff bounds using χ^2-divergence. More recent works (Razeghi et al., 2024; Huang and El Gamal, 2024), leveraged variational approximation using deep networks to optimize the privacy funnel. In contrast to these approaches, we do not optimize the mutual information terms. Instead, we build on the theoretical framework presented by Calmon et al. (2017). Using the mutual information outer bounds from (Calmon et al., 2017), we derive the data constraints and erasure functions necessary for achieving perfect concept erasure. Next, we discuss the minimum entropy coupling, which we incorporate into the erasure functions.

Minimum Entropy Coupling (MEC). Minimum entropy coupling Γ is a joint distribution over m input probability distributions $(p_1, \dots, p_m)$, with the minimum entropy. In our work, we are only interested in finding a coupling between two distributions, $\Gamma(p, q)$.

$$\Gamma_{\min}(p,q) = \arg\min_{\Gamma} H_{\Gamma}(p,q) = -\Gamma_{ij} \log \Gamma_{ij},$$

where $\sum_j \Gamma_{ij} = p_i, \sum_i \Gamma_{ij} = q_j$. It is easy to see that the minimum entropy, $H_{\min}(p,q)$, achieved by MEC, $\Gamma_{\min}(p,q)$, also maximizes the mutual information: $I(p;q) = H(p) + H(q) - H_{\min}(p,q)$, between two distributions p and q. However, computing the minimum entropy coupling is NP-Hard (Kovačević et al., 2015). Later, Kocaoglu et al. (2017a) introduced a greedy algorithm to efficiently approximate the MEC. Subsequent works (Kocaoglu et al., 2017b; Rossi, 2019; Compton et al., 2022) showed that the approximation produced by the greedy algorithm is tight. Specifically, while computing the MEC between two distributions, the greedy approach produces approximations within 0.53 bits (Compton et al., 2023). We use the greedy algorithm for efficiently computing the MEC to derive the perfect erasure functions. In the following sections, we will formalize concept erasure, quantify the outer bounds, and derive perfect erasure functions.

3 CONCEPT ERASURE

3.1 Problem Statement

Concept erasure tries to erase the effect of a concept, A (a random variable with support $\mathcal{A}$), from a representation set, X (random variable with support $\mathcal{X}$). This is usually done by learning a transformation of the original representations, $Z = f(X)$ (random variable with support $\mathcal{Z}$). We will refer to the transformation map, f, as the *erasure function*. In our setup, the erasure function f *only* has access to X and not A. This allows f to generalize to examples where the concept A is not available. Note that although prior works have utilized various erasure techniques, they can all be generalized using an erasure function, f. For example, linear concept erasure approaches (Bolukbasi et al., 2016; Ravfogel et al., 2020) use a product of projection matrices as $f(x) = \prod_j \mathbf{P}_j x$, kernelized linear erasure (Ravfogel et al., 2022b) use $f(x) = \prod_j \mathbf{P}_j \Phi(x)$ (where $\Phi(\cdot)$ is non-linear), Chowdhury and Chaturvedi (2022); Chowdhury et al. (2024); Huang and El Gamal (2024) explicitly parameterize f using neural networks.

Assumptions. We make the following assumptions in our concept erasure setup:

(A1) The Markov property: $A \longrightarrow X \xrightarrow{f} Z$, which implies $I(Z; A|X) = 0$.

(A2) The support of the input and output representation sets, $\mathcal{X}$ and $\mathcal{Z}$, are finite.

(A3) The support of the concept set is finite, $\mathcal{A} = \{a_1, a_2, \dots, a_{|\mathcal{A}|}\}$.

(A4) The representations for each concept group are sampled from the distribution, $\forall i \in \{1, \dots, |\mathcal{A}|\}$, $P_i = P(X|A = a_i)$. For all pairs of distributions (P_i, P_j) with $i \neq j$, their supports (defined as $\mathcal{X}_i = \text{supp}(P_i)$) are disjoint, i.e., $\mathcal{X}_i \cap \mathcal{X}_j = \phi$. Similarly, the output representations for each concept group is $P(Z|A = a_i)$.

(A5) The size of the input representation support is larger than the size of concept support, $|\mathcal{X}| > |\mathcal{A}|$.

An example of such a setup is erasing gender information from English word embeddings. The representation set is finite since the number of English words is finite. Therefore, even though data representations are typically continuous, assuming a categorical distribution with a large support set is reasonable.

Objective. The primary objective of concept erasure is to construct an erasure function, f, such that these two properties: (a) effectively erase a concept, A, achieving $I(Z;A) \approx 0$ and (b) retaining information about the original representations, X, achieving high $I(Z;X)$. The resultant objective is (for a small constant $\epsilon > 0$):

$$\max_{f} I(Z;X) \text{ subject to } I(Z;A) \leq \epsilon, \tag{1}$$

where $Z = f(X)$. The concept erasure objective shown above is reminiscent of the information bottleneck objective (Tishby et al., 2000), which focuses on achieving a good tradeoff between $I(Z;A)$ and $I(Z;X)$. In concept erasure, the objective is to robustly erase concept A, resulting in near-zero $I(Z;A) \approx 0$, even if it is at the expense of a reduced utility, $I(Z;X)$.

3.2 Types of Concept Erasure

We note that optimizing the objective in Eq. 1 is difficult, as estimating mutual information using a finite number of samples is challenging (Song and Ermon, 2019; McAllester and Stratos, 2020). Rather than optimizing the mutual information terms, we use their outer bounds to derive the erasure functions. Next, we introduce several definitions involving concept erasure.

Definition 3.1 (Perfect Concept Erasure)**.** *An erasure function, f, achieves perfect concept erasure if $I(Z;A) = 0$, where $Z = f(X)$.*

Perfect concept erasure ensures that no information about the concept is present in the resultant representations, $I(Z;A) = 0$. However, achieving perfect erasure may not always be feasible. Our next result provides the conditions under which perfect concept erasure is feasible. For deriving the conditions of perfect erasure, we rely on the notion of principal inertia components (PICs) (Calmon et al., 2017). Intuitively, PICs capture the correlation between two random variables X and A, and can be viewed as a decomposition of the joint distribution, $p_{A,X}$ (see details in Appendix A.1). Note that all of the results in this paper are derived under Assumptions (A1)-(A5) unless otherwise stated.

Lemma 3.2 (Perfect Concept Erasure Achievability)**.** *Under the assumptions (A1)-(A4), for a joint distribution $p_{A,X}$ with finite support defined over $\mathcal{A}\times\mathcal{X}$, perfect concept erasure, $I(Z;A) = 0$, is achievable if either (a) smallest principal inertia component, $\lambda_d(A,X) = 0$ or (b) $|\mathcal{X}| > |\mathcal{A}|$.*

This lemma states that perfect erasure is feasible when either the smallest PIC is zero, $\lambda_d(A,X) = 0$, or $|\mathcal{X}| > |\mathcal{A}|$ (proof in Appendix A.1). Throughout the paper, we assume (A5): $|\mathcal{X}| > |\mathcal{A}|$, which means perfect erasure is always feasible irrespective of $\lambda_d(A,X)$. This is a reasonable assumption in practice as the size of the representation space is typically much larger than the number of concepts.

Note that Definition 3.1 for perfect concept erasure only requires the concept to be erased from learned representations, $I(Z;A) = 0$. It does not impose any constraints on the utility of the learned representations, $I(Z;X)$. Since concept erasure also involves retaining maximum utility, it is natural to question the maximum achievable utility $I(Z;X)$ during perfect erasure.

Lemma 3.3 (Perfect Erasure Bound)**.** *Under the assumptions (A1)-(A5), if an erasure function $f : \mathcal{X} \to \mathcal{Z}$ achieves perfect concept erasure $I(Z;A) = 0$, then the utility $I(Z;X)$ is bounded as:*

$$I(Z;X) = I(Z;X|A) \leq H(X|A),$$

where the equality is satisfied when $H(X|Z,A) = 0$.

The above result presents the outer bound for the utility, $I(Z;X)$, during perfect concept erasure. The complete proof is in Appendix A.2. Next, we focus on a slightly different setting of concept erasure where the privacy leakage about the concept can be non-zero, $I(Z;A) = \epsilon$ ($\epsilon > 0$). Specifically, we are interested in understanding the best achievable privacy (minimum $I(Z;A)$) for any utility u, such that $I(Z;X) \geq u$. To address this question, we present the notion of an *erasure funnel*[1] introduced in information theory literature.

Definition 3.4 (Erasure Funnel (Calmon et al., 2017))**.** *For $0 \leq u \leq H(X)$, representations $Z = f(X)$, and a fixed joint distribution over $\mathcal{A}\times\mathcal{X}$, we define the erasure funnel function as:*

$$\epsilon(u) = \inf_{f} \{I(Z;A) | I(Z;X) \geq u, A \to X \xrightarrow{f} Z\}. \tag{2}$$

We use the erasure funnel to define the setting where concept leakage is non-zero, $\epsilon(u) > 0$. We term this setting *relaxed concept erasure*.

Definition 3.5 (Relaxed Concept Erasure)**.** *An erasure function f performs relaxed concept erasure for $\epsilon(u) > 0$, if the utility is $I(Z;X) \geq u$ and $I(Z;A) = \epsilon(u)$, where $\epsilon(u) = \inf_f \{I(Z;A) | I(Z;X) \geq u, A \to X \xrightarrow{f} Z\}$.*

In Definition 3.5, $\epsilon(u)$ is the minimum achievable privacy leakage, $I(Z;A)$, such that the utility is at least

[1] This is often termed as privacy funnel, but we will refer to it as erasure funnel in the context of concept erasure.

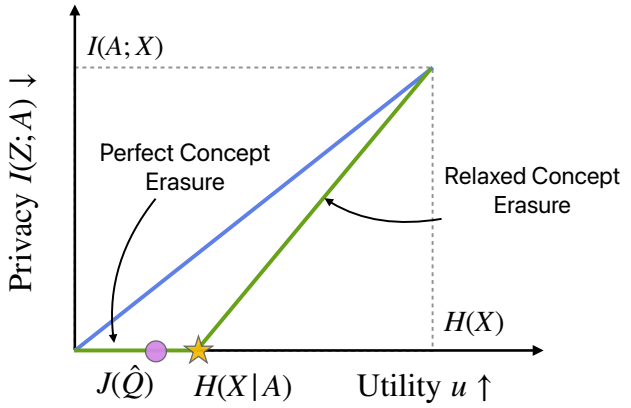

Figure 1: Schematic diagram of the privacy-utility tradeoff and illustration of the bounds of erasure funnel function, $\epsilon(u)$ (Calmon et al., 2017). The region shown by the **green** line is desirable for concept erasure as it is the minimum $I(Z;A)$ for a given utility $I(Z;X) \geq u$.

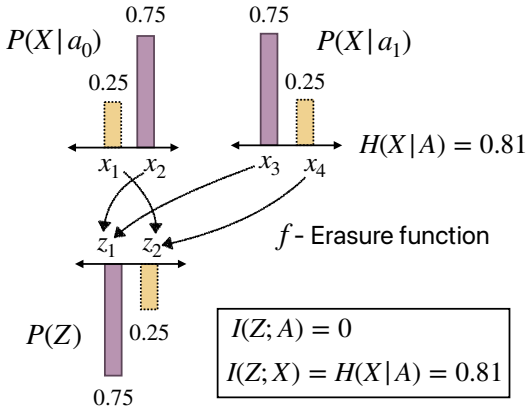

Figure 2: An illustration of a perfect erasure function. The input representations of each group are sampled from 2D distributions. The erasure function f in the figure achieves the outer bounds for concept erasure.

$I(Z;X) \geq u$. This definition considers the utility, $I(Z;X)$, and focuses on finding an erasure function f that does not erase all information. This is useful in applications such as fairness, where it is not necessary to completely erase concepts to achieve fair outcomes.

3.3 Outer Bounds For Concept Erasure

We build on the definitions introduced in the previous section and present the outer bounds for privacy. Ideally, we want to *minimize* privacy leakage, $I(Z;A)$ while *maximizing* utility, $I(Z;X)$. Here, we utilize the erasure funnel (Def. 3.4) to answer these questions. The erasure funnel extends beyond relaxed erasure and can offer insights into perfect erasure ($\epsilon(u) = 0$). We will use $\epsilon(u)$ to answer the following research questions:

(**RQ1**) What is the maximum utility u during perfect concept erasure (when $\epsilon = 0$)?

(**RQ2**) What is the outer bound of privacy $\epsilon(u)$ (minimum $\epsilon(u)$) during relaxed concept erasure?

As obtaining the precise analytic expression for $\epsilon(u)$ is difficult, we present its bounds (Calmon et al., 2017) to answer the above research questions.

Lemma 3.6 (Outer Bounds for $\epsilon(u)$ (Calmon et al., 2017))**.** *For $0 \leq u \leq H(X)$, erasure funnel function, $\epsilon(u)$, is bounded as:*

$$\max\{0, u - H(X|A)\} \leq \epsilon(u) \leq \frac{uI(A;X)}{H(X)}.$$

In Figure 1, we plot the result of Lemma 3.6, where the y-axis indicates the privacy $I(Z;A)$, and the x-axis indicates the utility, u. We observe that $\epsilon(u)$ lies within a funnel-like structure with upper bound shown in **blue** and lower bound in **green**. In this plot, perfect concept erasure is the flat portion of the **green** line where $\epsilon(u) = 0$. During perfect erasure, we observe that the maximum possible utility is $H(X|A)$ shown by ★ . This is the same result we derived in Lemma 3.3 and provides the solution for (**RQ1**). During relaxed concept erasure $\epsilon(u) > 0$, the minimum mutual information is $\epsilon(u) = \frac{uI(A;X)}{H(X)}$, which is a linear function of the utility, u. This provides the solution for (**RQ2**). We will discuss the importance of ● in Section 4.

We would like to emphasize that these outer bounds might not be achievable for any random variables X and A. Therefore, our next focus will be on understanding the data constraints under which the outer bounds can be attained and identifying the erasure functions that achieve them. In this work, we only consider perfect concept erasure and its corresponding erasure functions. We defer the derivation of erasure functions that achieve relaxed concept erasure to future works.

4 PERFECT ERASURE FUNCTIONS (PEF)

In this section, we present an algorithm for concept erasure, which is inspired by achieving perfect concept erasure, $I(Z;A) = 0$, along with the utility outer bound, $I(Z;X) = H(X|A)$ (★ in Fig. 1). We will refer to these methods as *perfect erasure functions*.

To motivate our method, we consider a simple example in Fig. 2. We consider a scenario with two concept groups, $\mathcal{A} = \{a_0, a_1\}$ with equal prior $p(a_0) = p(a_1) = 0.5$. We consider the data representation space $\mathcal{X}$ to be $\{x_1, x_2, x_3, x_4\}$. Each group has a two dimensional representation – $P(X|A = a_0)$ has support of $\{x_1, x_2\}$ and $P(X|A = a_1)$ has a support of $\{x_3, x_4\}$.

In this example, there is an erasure function, f, which

maps the data X to the erased representations, Z, which makes use of a helpful observation – $P(X|A=a_0)$ and $P(X|A=a_1)$ are permutations of one another. This erasure function meets the definition of a perfect erasure function. We hypothesize that the fact that the conditional distributions are permutations is essential to the achievement of perfect erasure. This permutation is essential as we can have a bijective map that transforms X into Z without revealing A. We formalize this in the following class of data.

Definition 4.1 (Data Constraints)**.** *We consider the class of data for which the distributions of different groups are permutations of each other:* $\forall(i,j)$, *(a)* $|\mathcal{X}_i| = |\mathcal{X}_j|$, *(b)* $\forall x \in \mathcal{X}_i,\ P_i(x) = P_j(\sigma_{ij}(x))$, *where* $P_i = P(X|A=a_i)$*, and for which there exists an erasure function* f *that achieves* $I(Z;A)=0$ *and* $I(X;Z) = H(X|A)$.

The above definition states that the distributions must be *equal up to permutation* and have equal support sizes ($\mathcal{X}_i = \text{supp}(P_i)$). We provide the justification behind the choice of these data constraints in Appendix A.3. Next, we turn to answer the question of how to find such an erasure function f to achieve this definition.

Equal Distributions (*up to permutations*). In the setting, where the distributions P_i's are equal up to permutation, we present the erasure functions. Intuitively, the erasure function is a one-to-one map from each distribution, P_i, to a common distribution, $Q = P(Z)$ (as shown in Figure 2). The common distribution Q can be any permutation of the P_i's. We formalize this in the following method.

Method 4.2 (Perfect Erasure Function)**.** *If the data constraints in Definition 4.1 hold, then define erasure function* f *as:*

$$f(x) = \{\sigma_i(x) |\ x \in \mathcal{X}_i,\ \sigma_i \in \Sigma_i\}, \tag{3}$$

where σ_i *is any bijective function that transforms* P_i *into* Q *defined by the set,* Σ_i *below:*

$$\Sigma_i = \{\sigma : \mathcal{X}_i \to \mathcal{Z} | \forall x \in \mathcal{X}_i,\ P_i(x) = Q(\sigma(x))\}\,. \tag{4}$$

The above method shows that a piecewise bijective mapping different concept groups, $\mathcal{X}_i$, to a common output support, $\mathcal{Z}$, achieves the outer bounds for perfect erasure (Definition 4.1). The justification is presented in Appendix A.4. Since the choice of Q and σ_i's are not unique, there can be multiple formulations of the erasure function. In the simple scenario, where all P_i's are uniform distributions, then f can be any random bijective map between $\mathcal{X}_i$ and $\mathcal{Z}$.

Unequal distributions. We consider the scenario where the constraints of Definition 4.1 are violated, i.e., the distributions, $\{P_i\}$'s, are not equal (even after permutation). In this setting, achieving the outer bounds

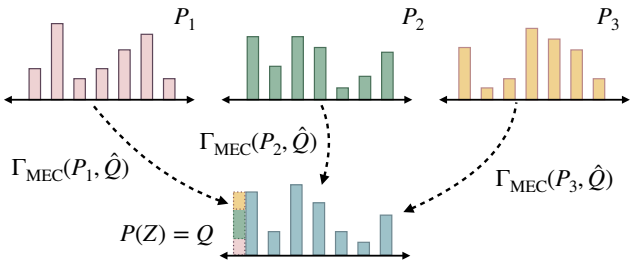

Figure 3: Schematic diagram of the proposed concept erasure process for unequal distributions. Given distributions from different concept groups, $\{P_1, P_2, P_3\}$, we map these representations to a common distribution, Q. Each distribution P_i is mapped to Q using the minimal entropy coupling map, $\Gamma_{\text{MEC}}(P_i, Q)$.

mentioned in Definition 4.1 is not possible. However, it is still possible to achieve perfect concept erasure, $I(Z;A)=0$. Operating with $P(Z) = P(Z|A=a_i)$, we want to find the piecewise map $f_i : \mathcal{X}_i \to \mathcal{Z}$ such that the distribution of the common support is fixed and equal to the conditional distribution, $Q = P(Z) = P(Z|A=a_i)$. Collectively the piecewise components f_i define the erasure function f. Specifically, we seek mapping functions f_i's that maximize the utility, $I(Z;X)$, as shown below:

$$\begin{aligned}
\max I(Z;X) &= \max I(Z;X|A) \qquad (5)\\
&= \max \sum_i p(a_i) I(Z;X|A=a_i)\\
&= \max \sum_i p(a_i)\left[H(Q) + H(P_i) - H_{\Gamma_i}(Q,P_i)\right]\\
&= \max \left[H(Q) - \sum_i p(a_i) H_{\Gamma_i}(Q,P_i)\right]. \qquad (6)
\end{aligned}$$

The optimization of the above objective depends on two key sets of parameters: (i) the distribution of the common support, Q, and (ii) Γ_i's – coupling joint distributions between pairwise distributions (P_i, Q). The second term in the optimization (Eq. 6) is equivalent to estimating the minimum entropy coupling, $H_{\Gamma_i}(P_i, Q)$, which can be obtained efficiently using the greedy algorithm (Section 3) if we know Q. Since Q is unknown, we can simplify the optimization problem as:

$$\begin{aligned}
Q^* &= \arg\max_Q J(Q)\\
&= \arg\max_Q \left[H(Q) - \sum_i p(a_i) H_{\min}(P_i, Q)\right]. \qquad (7)
\end{aligned}$$

Solving the above objective is difficult and cannot be optimized using standard techniques like gradient descent as we do not know the analytical form of the minimum entropy coupling function, $H_{\min}(\cdot, P_i)$. Even evaluation of the $H_{\min}(\cdot, P_i)$ function is expensive as

MEC estimation is NP-Hard and approximating it still requires running the greedy algorithm. Therefore, we resort to Bayesian Optimization (BO) (Frazier, 2018), which performs optimization by making a small number of queries to the objective $J(Q)$ (Eq. 7). Bayesian Optimization utilizes a surrogate model (usually a Gaussian process (Seeger, 2004)) to fit the observed query points. Using an acquisition function, BO explores different areas of the input space to maximize the objective. We will refer to the solution returned by Bayesian Optimization as $\hat{Q}_{\text{BO}}$; however, it is not guaranteed to be the optimal solution.

We hypothesize that the set of input distributions $\bar{P}$ are a good candidate for output distribution Q as they are stationary points for the objective $J(Q)$. The best achievable utility, $I(Z;X)$, when $Q \in \overline{P}$ is shown by the ● point in Figure 1. The set $\bar{P}$ only represents the set of stationary points and it may be still possible for the global maxima, Q^* to improve upon them. However, when distributions are unequal analytically deriving the global optima $Q^* \notin \bar{P}$ is non-trivial and remains an open question.[2]

To summarize, when underlying distributions P_i's are unequal, we select either by using the stationary points or the solution of Bayesian optimization: $\hat{Q} = \arg\max_{Q \in \bar{P} \cup \hat{Q}_{\text{BO}}} J(Q)$. In practice, we observe that the stationary points $\bar{P}$ often achieve better performance than the BO solution, $\hat{Q}_{\text{BO}}$. Once, we have selected the support distribution $\hat{Q}$, the erasure function is a stochastic map, $f(x) \sim P(Z|X=x)$, that samples the output z from the conditional distribution:

$$P(Z|X=x) = \left\{ \Gamma_i(x) |\ x \in \mathcal{X}_i, \Gamma_i = \arg\min_{\Gamma} H_{\Gamma}(P_i, \hat{Q}) \right\},$$

where Γ_i is the coupling map corresponding to the minimum entropy coupling between distribution pairs $(P_i, \hat{Q})$. An illustration of this erasure function is shown in Figure 3. Using the result from Lemma 3.3, it is easy to see that for unequal distributions the erasure function doesn't achieve the utility outer bound as $H(X|Z,A) > 0$ (as different $x \in \mathcal{X}_i$ can map to the same $z \in \mathcal{Z}$). Nonetheless, since the output distribution Q is fixed, and MEC yields a valid joint distribution, it still ensures perfect concept erasure, $I(Z;X) = 0$.

We summarize the overall concept erasure algorithm in Algorithm 1. The algorithm is provided with the entire representation set $\mathcal{S} = \{x_1, x_2, \ldots\}$ (where $x_i \in \mathcal{X}$) and the corresponding concept set, $\mathcal{G}$. First, it estimates the distribution of the representations from individual concept groups. Second, it checks whether the underlying distributions are equal up to permutation. Based on this step, the algorithm determines the optimal erasure function. Finally, the selected erasure function is used to generate the erased representations, $\mathcal{O}$.

[2] Further discussion of optimization in Appendix B.1.

Algorithm 1 Concept Erasure Using `PEF`

Input: Sample Set $\mathcal{S} = \{x_1, x_2, \ldots\}$, Concept Set $\mathcal{G} = \{a_1, a_2, \ldots\}$.
$\forall i, P_i = \text{EstimateDistribution}(\mathcal{S}_i)$ // ($\mathcal{S}_i = \{x | x \in \mathcal{X}_i\}$ is the sample set of the i-th concept group)
if $\forall (i,j), P_i = \sigma(P_j)$ **then**
    // Distributions are equal up to permutation
    $f(x) = \{\sigma_i(x) |\ x \in \mathcal{X}_i,\ \sigma_i \in \Sigma_i\}$ // (Σ_i is defined in Eq. 4)
    $\mathcal{O} = f(\mathcal{S})$
else
    // Unequal distributions
    $\hat{Q} = \arg\max_Q \left[H(Q) - \sum_i p(a_i) H_{\min}(P_i, Q) \right]$
    $P(Z|X=x) = \left\{ \Gamma_i(x) |\ x \in \mathcal{X}_i, \Gamma_i = \arg\min_{\Gamma} H_{\Gamma}(P_i, \hat{Q}) \right\}$
    $\mathcal{O} = \{z \sim P(Z|X=x) |\ x \in \mathcal{X}\}$
end if
return $\mathcal{O}$ // erased representation set

5 EXPERIMENTS

In this section, we discuss the experimental setup and evaluation results in detail. The implementation is available here: https://github.com/brcsomnath/PEF.

Baselines. We compare our approach, `PEF`, with several state-of-the-art baselines, including linear concept erasure techniques – INLP (Ravfogel et al., 2020), LEACE (Belrose et al., 2024) and general erasure techniques – FaRM (Chowdhury and Chaturvedi, 2022), KRaM (Chowdhury et al., 2024). We also compare with the state-of-the-art privacy funnel solver that directly optimizes mutual information terms using variational approximation, DCA (Huang and El Gamal, 2024).

Synthetic Data. We conduct experiments in controlled settings using synthetic data, which allows us to directly compare `PEF`'s performance with the theoretical bounds (shown in Figure 1). Otherwise, accurately estimating mutual information (MI) using high-dimensional data is challenging. Specifically, we consider a set of 3D representations sampled from a finite support set. The finite support set is also randomly sampled from a 3D uniform distribution. We use two underlying concepts and sample representations from distinct support for each group. We experiment in two settings where the groups have (i) *equal* and (ii) *unequal* support distributions to simulate the scenarios in Section 4. Finally, we report the MI estimates using $\mathcal{V}$-information (Xu et al., 2019), which uses a classifier to predict the concept (privacy) and original representation (utility) from the representations

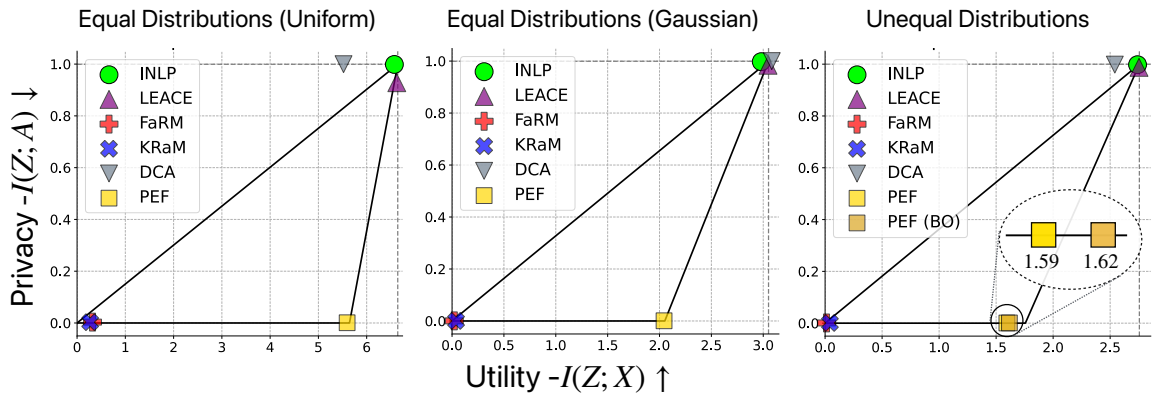

Figure 4: Privacy-Utility tradeoff plots for concept erasure using synthetic data when the underlying representations from different concept groups are sampled from equal uniform (*left*), equal Gaussian (*center*) and unequal (*right*) distributions. We observe that PEF retains significant utility (high $I(Z;X)$) with perfect erasure, $I(Z;A) = 0$.

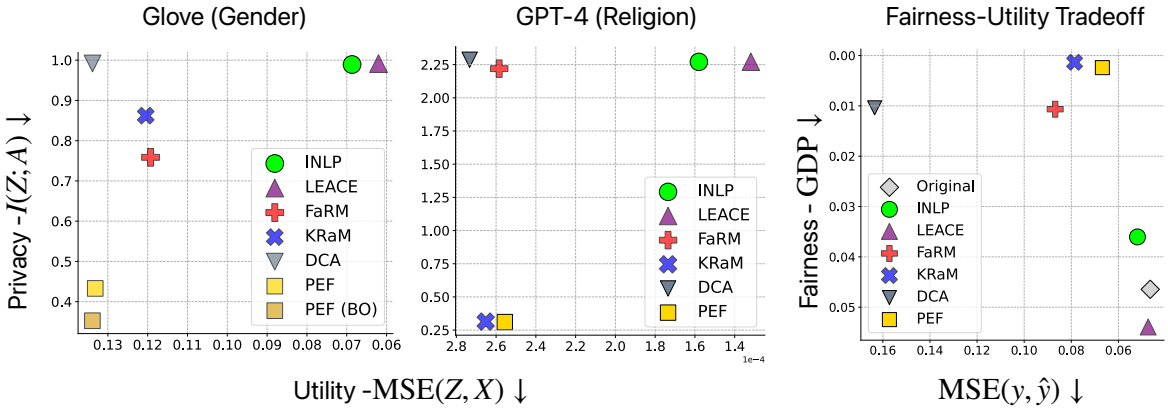

Figure 5: (*Left*) We show the privacy-utility tradeoff for erasing gender information from GloVe embeddings and religion information from GPT-4 embeddings. We observe that PEF achieves significant improvement in privacy compared to baseline approaches. (*Right*) We use the erased GPT-4 embeddings for toxicity classification and plot the fairness-utility tradeoff. We observe that PEF achieves the best fairness-utility tradeoff among the baselines.

post-concept erasure. Since the support of the input representation set is finite, we associate each sample representation with a categorical label. The classifier, used to compute the $\mathcal{V}$-information with the original representations, predicts the label given the erased representation. $\mathcal{V}$-information provides a strong lower bound for the MI estimate when the classifier is constrained by computational resources.

In Figure 4, we report the privacy, $I(Z;A)$, and utility, $I(Z;X)$, tradeoff during concept erasure. The lines describing the funnel structure indicate the theoretical erasure funnel bounds, $\epsilon(u)$, as described in Figure 1. We observe that linear concept erasure approaches, INLP and LEACE, retain significant utility, achieving high $I(Z;X)$ but at the cost of poor privacy, high $I(Z;A)$. DCA performs similar to linear erasure approaches. On the other hand, nonlinear erasure approaches such as FaRM and KRaM erase the concept robustly, but result in a significant drop in utility, low $I(Z;X)$. In contrast to these approaches, PEF achieves high utility while perfectly erasing the concept, $I(Z;A) = 0$. When representations are sampled from equal distributions in Figure 4 (left & center), we observe that PEF achieves the utility outer bound. When the distributions are unequal, in Figure 4 (right), PEF achieves perfect privacy but is unable to achieve the maximum utility, which is expected. In this setting, the Bayesian optimization solution, PEF (BO), slightly improves the utility while achieving perfect erasure.

Real-World Datasets. We evaluate our approach in real-world settings to erase gender from GloVe embeddings and religion from the Jigsaw (jig, 2019) dataset's GPT-4 text embeddings. As these representations are high-dimensional, computing the mutual information with the original representations is challenging. Therefore, we report the mean squared error (MSE) in predicting the original representations from the erased representations using a scikit-learn (Pedregosa et al., 2011) neural network regression model.

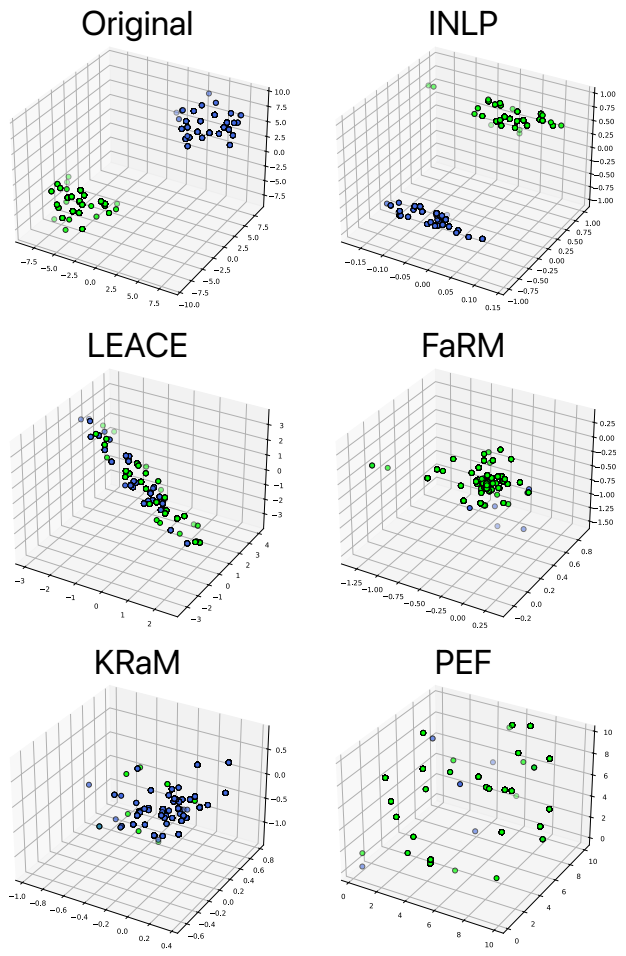

Figure 6: Visualization of the representations post concept erasure using different methods. We observe that existing techniques either leak concept information (INLP & LEACE) or lose utility as representations collapse (FaRM & KRaM). PEF can erase concepts properly while maintaining utility.

For Glove embeddings shown in Figure 5 (left), we observe that PEF achieves significantly better privacy (low $I(Z;A)$) compared to all baselines while achieving similar utility to methods like FaRM, KRaM, and DCA. We also observe that PEF (BO) slightly outperforms PEF in terms of privacy. PEF assumes access to the original probability distributions. In practice, we have finite high-dimensional samples and use kernel density estimation (Chen, 2017) to estimate the distributions. We hypothesize that the slight drop in utility is due to the unavailability of these original distributions.

In the second setting, we obtain GPT-4 (`text-embedding-3-large`) representations for online comments in the Jigsaw dataset. The objective of this experiment is to erase religion information and utilize the resultant representations for toxicity classification. This ensures that religion does not affect the predicted toxicity scores. We report the results in Figure 5 (center). Similar to the previous setting, we observe significant privacy improvements with similar utility to non-linear erasure techniques (FaRM, KRaM & DCA). In this setting, we found that the local minima $\{P_i\}$' provided a better solution than $\hat{Q}_{\text{BO}}$, therefore, the performance PEF and PEF (BO) is the same. This is expected as Bayesian optimization is not known to be effective in high dimensions (3072).

After concept erasure, we use the resultant representations to perform toxicity classification. We report the fairness-utility tradeoff, where fairness is computed using generalized demographic parity (Jiang et al., 2021) (for continuous-valued toxicity scores) and utility is the MSE loss of the predictions. In Figure 5 (right), we observe that PEF significantly improves fairness GDP scores while having minimal impact on performance (minimal increase in MSE). In general, PEF achieves a good balance between utility and fairness compared to the baselines. We provide more details in Appendix C.

Visualization. In this section, we visualize the representations obtained from various methods, as shown in Figure 6. We use synthetic 3D representations, which encode a binary concept shown by representations in blue and green. Post concept erasure, we expect representations from both groups to be indistinguishable. For linear erasure techniques (INLP & LEACE), we observe that the underlying concept can be identified, while non-linear techniques (FaRM & KRaM) achieve better overlap but collapse into a small region of space, signifying a utility loss. In contrast to these approaches, PEF shows robust erasure while retaining utility (no collapse observed). We further illustrate the erasure process of LEACE and PEF with a simpler example in Appendix C.3.

6 CONCLUSION

In this paper, we study the fundamental limits of perfect concept erasure. Theoretically, we study the maximum amount of information that can be retained while perfectly erasing a concept. Furthermore, we explore the conditions that the underlying data distribution and the analytical form of the erasure function required to achieve perfect erasure. Empirically, we demonstrate the effectiveness of the proposed perfect erasure function (PEF) in various synthetic and real-world settings. Although PEF can guarantee perfect erasure, it assumes access to a finite support representation distribution, which may be difficult to estimate in low-resource settings. Future works can focus on effectively estimating the underlying distribution using a small number of samples to improve the performance of PEF.

Acknowledgment

The authors are thankful to Flavio du pin Calmon for providing helpful pointers to related work. This work was supported in part by NSF grant DRL-2112635.

References

Jigsaw unintended bias in toxicity classification. 2019.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691, 2022.

Amir Ahooye Atashin, Behrooz Razeghi, Deniz Gündüz, and Slava Voloshynovskiy. Variational leakage: The role of information complexity in privacy leakage. In Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning, pages 91–96, 2021.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems, 36, 2024.

Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In International Conference on Machine Learning, pages 614–623. PMLR, 2019.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29, 2016.

Andreas Buja. Remarks on functional canonical variates, alternating least squares methods and ace. The Annals of Statistics, pages 1032–1069, 1990.

Flavio Calmon, Ali Makhdoumi, Muriel Médard, Mayank Varia, Mark Christiansen, and Ken R Duffy. Principal inertia components and applications. IEEE Transactions on Information Theory, 63(8):5011–5038, 2017.

Flavio P Calmon, Ali Makhdoumi, and Muriel Médard. Fundamental limits of perfect privacy. In 2015 IEEE International Symposium on Information Theory (ISIT), pages 1796–1800. IEEE, 2015.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. Biostatistics & Epidemiology, 1(1):161–187, 2017.

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. Learning fair representations via rate-distortion maximization. Transactions of the Association for Computational Linguistics, 10: 1159–1174, 2022.

Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier Oliva, Shashank Srivastava, and Snigdha Chaturvedi. Adversarial scrubbing of demographic information for text classification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 550–562, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.43. URL https://aclanthology.org/2021.emnlp-main.43.

Somnath Basu Roy Chowdhury, Nicholas Monath, Kumar Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. Robust concept erasure via kernelized rate-distortion maximization. Advances in Neural Information Processing Systems, 36, 2024.

Spencer Compton, Kristjan Greenewald, Dmitriy A Katz, and Murat Kocaoglu. Entropic causal inference: Graph identifiability. In International Conference on Machine Learning, pages 4311–4343. PMLR, 2022.

Spencer Compton, Dmitriy Katz, Benjamin Qi, Kristjan Greenewald, and Murat Kocaoglu. Minimum-entropy coupling approximation guarantees beyond the majorization barrier. In International Conference on Artificial Intelligence and Statistics, pages 10445–10469. PMLR, 2023.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5034–5050, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.411. URL https://aclanthology.org/2021.emnlp-main.411.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. Internation Conference on Learning Representations, 2021.

Jingfei Du, Myle Ott, Haoran Li, Xing Zhou, and

Veselin Stoyanov. General purpose text embeddings from pre-trained language models for scalable inference. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3018–3030, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.271. URL https://aclanthology.org/2020.findings-emnlp.271.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Privacy aware learning. Journal of the ACM (JACM), 61(6):1–57, 2014.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 11–21, 2018.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics, 9: 160–175, 2021.

Peter I Frazier. A tutorial on bayesian optimization. arXiv preprint arXiv:1807.02811, 2018.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. Its not greek to mbert: Inducing word-level translations from multilingual bert. In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 45–56, 2020.

Michael Greenacre. Correspondence analysis in practice. chapman and hall/crc, 2017.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8395–8416, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.575. URL https://aclanthology.org/2022.emnlp-main.575.

Geoffrey E Hinton et al. Learning distributed representations of concepts. In Proceedings of the eighth annual conference of the cognitive science society, volume 1, page 12. Amherst, MA, 1986.

Hsiang Hsu, Shahab Asoodeh, Salman Salamatian, and Flavio P Calmon. Generalizing bottleneck problems. In 2018 IEEE International Symposium on Information Theory (ISIT), pages 531–535. IEEE, 2018.

Teng-Hui Huang and Hesham El Gamal. An efficient difference-of-convex solver for privacy funnel. In 2024 IEEE International Symposium on Information Theory Workshops (ISIT-W), pages 1–6. IEEE, 2024.

Yusuke Iwasawa, Kotaro Nakayama, and Yutaka Matsuo. Censoring representations with multiple-adversaries over random subspaces, 2018. URL https://openreview.net/forum?id=ByuP8yZRb.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In International Conference on Learning Representations, 2021.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186, 2019.

Murat Kocaoglu, Alexandros Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017a.

Murat Kocaoglu, Alexandros G Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causality and greedy minimum entropy coupling. In 2017 IEEE International Symposium on Information Theory (ISIT), pages 1465–1469. IEEE, 2017b.

Mladen Kovačević, Ivan Stanojević, and Vojin Šenk. On the entropy of couplings. Information and Computation, 242:369–382, 2015.

Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization. Transactions on Machine Learning Research, 2022.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 3384–3393. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/madras18a.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In 2014 IEEE Information Theory Workshop (ITW 2014), pages 501–505. IEEE, 2014.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In International Conference on Machine Learning, pages 4382–4391. PMLR, 2019.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In International Conference on Artificial Intelligence and Statistics, pages 875–884. PMLR, 2020.

A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. Advances in Neural Information Processing Systems, 34:5264–5275, 2021.

Fernando Nogueira. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014–. URL https://github.com/bayesian-optimization/BayesianOptimization.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.

Ajay Patel, Delip Rao, and Chris Callison-Burch. Learning interpretable style embeddings via prompting llms. arXiv preprint arXiv:2305.12696, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In International Conference on Machine Learning, pages 18400–18421. PMLR, 2022a.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6034–6055, 2022b.

Behrooz Razeghi, Flavio P Calmon, Deniz Gündüz, and Slava Voloshynovskiy. On perfect obfuscation: Local information geometry analysis. In 2020 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–6. IEEE, 2020.

Behrooz Razeghi, Flavio P Calmon, Deniz Gunduz, and Slava Voloshynovskiy. Bottlenecks club: Unifying information-theoretic trade-offs among complexity, leakage, and utility. IEEE Transactions on Information Forensics and Security, 18:2060–2075, 2023.

Behrooz Razeghi, Parsa Rahimi, and Sébastien Marcel. Deep variational privacy funnel: General modeling with applications in face recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4920–4924. IEEE, 2024.

Massimiliano Rossi. Greedy additive approximation algorithms for minimum-entropy coupling problem. In 2019 IEEE International Symposium on Information Theory (ISIT), pages 1127–1131. IEEE, 2019.

Bashir Sadeghi, Sepehr Dehdashtian, and Vishnu Boddeti. On characterizing the trade-off in invariant representation learning. Transactions on Machine Learning Research, 2022.

Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. IEEE Transactions on Information Forensics and Security, 8(6):838–852, 2013.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. Zero-shot multi-label topic inference with sentence encoders and llms. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16218–16233, 2023.

Matthias Seeger. Gaussian processes for machine learning. International journal of neural systems, 14(02): 69–106, 2004.

Shun Shao, Yftah Ziser, and Shay B Cohen. Gold doesnt always glitter: Spectral removal of linear

and nonlinear guarded attribute information. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1603–1614, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. Nature, 620(7973):E19–E19, 2023.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In International Conference on Learning Representations, 2019.

Sreejith Sreekumar and Deniz Gündüz. Optimal privacy-utility trade-off under a rate constraint. In 2019 IEEE International Symposium on Information Theory (ISIT), pages 2159–2163. IEEE, 2019.

Zhongxiang Sun. A short survey of viewing large language models in legal aspect. arXiv preprint arXiv:2303.09136, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. arXiv preprint physics/0004057, 2000.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

Hao Wang and Flavio P Calmon. An estimation-theoretic view of privacy. In 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 886–893. IEEE, 2017.

Hao Wang, Lisa Vo, Flavio P Calmon, Muriel Médard, Ken R Duffy, and Mayank Varia. Privacy with estimation guarantees. IEEE Transactions on Information Theory, 65(12):8025–8042, 2019.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. Advances in neural information processing systems, 30, 2017.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In International Conference on Learning Representations, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In International conference on machine learning, pages 325–333. PMLR, 2013.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In International conference on machine learning, pages 7523–7532. PMLR, 2019.

Han Zhao, Chen Dan, Bryon Aragam, Tommi S Jaakkola, Geoffrey J Gordon, and Pradeep Ravikumar. Fundamental limits and tradeoffs in invariant representation learning. Journal of Machine Learning Research, 23(340):1–49, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A MATHEMATICAL PROOFS

Contents

A.1 Proof of Lemma 3.2

In this section, we present the necessary and sufficient conditions required for achieving perfect concept erasure. First, we introduce the notion of principal inertia components, which will be instrumental in deriving the conditions for perfect erasure.

Principal Inertia Components (PICs). PICs can be viewed as a decomposition of the joint distribution, $p_{X,A}$, between two random variables X and A (a branch of study dedicated to this is called *correspondence analysis* (Greenacre, 2017)). Principal inertia components exhibit interesting properties and are often used to define common measures like maximum correlation, $\rho_m(X,A)$, and Chi-squared correlation, $\chi^2(X,A)$.

Definition A.1 (Principal Inertia Components (Calmon et al., 2017))**.** *Let random variables X and A have finite support sets $\mathcal{X}$ and $\mathcal{A}$ respectively, and joint distribution $p_{X,A}$. In addition, let $f_0 : X \to \mathbb{R}$ and $g_0 : Y \to \mathbb{R}$ be the constant functions $f_0(x) = 1$ and $g_0(a) = 1$. For $k \in \mathbb{Z}^+$, we (recursively) define:*

$$\lambda_k(X;A) = \max\Big\{\mathbb{E}[f(X)g(A)]^2 \,\Big|\, \mathbb{E}[f(X)f_j(X)] = 0, \\ \mathbb{E}[g(Y)g_j(A)] = 0,\ j \in \{0,\dots,k-1\}\Big\},$$

where $f \in \mathcal{L}_2(p(X))$, $g \in \mathcal{L}_2(p(Y))$ are square-integrable functions, and

$$(f_k, g_k) \coloneqq \arg\max\Big\{\mathbb{E}[f(X)g(Y)]^2 \,\Big|\, \mathbb{E}[f(X)f_j(X)] = 0, \\ \mathbb{E}[g(Y)g_j(A)] = 0,\ j \in \{0,\dots,k-1\}\Big\}.$$

$\lambda_k(X;A)$ values are known as the principal inertia components (PICs) of the joint distribution, $p(X,A)$. The functions f_k and g_k are the principal functions of X and A.

The above definition presents a characterization of PICs as the construction of zero-mean, unit variance functions $f(X)$ and $g(A)$, which maximizes the correlation $\mathbb{E}[f(X)g(A)]$ without using any additional information. Note that alternative characterizations of PICs exist that include estimation errors and geometric interpretations (Calmon et al., 2017).

From Definition A.1, we observe that PICs satisfy: $\lambda_{k+1}(X;A) \le \lambda_k(X;A) \le 1$. PICs also correspond to the singular values of the joint probability matrix, specifically $\sqrt{\lambda_k(X;A)}$ is the $(k+1)$-st singular value of $\mathbf{Q}$, where $\mathbf{Q} = \mathbf{D}_X^{-1/2}\mathbf{P}\mathbf{D}_A^{-1/2}$. $\mathbf{D}_X \in \mathbb{R}^{m\times m}$ and $\mathbf{D}_A \in \mathbb{R}^{n\times n}$ are diagonal matrices with entries $[\mathbf{D}_X]_{i,i} = p_X(i)$ and $[\mathbf{D}_A]_{j,j} = p_A(j)$ respectively, where $m = |\mathcal{X}|$ and $n = |\mathcal{A}|$. $\mathbf{P} \in \mathbb{R}^{m\times n}$ is a matrix with entries $[\mathbf{P}]_{i,j} = p_{X,A}(i,j)$.

k-correlation. Next, we introduce the notion of k-correlation and discuss how it relates to commonly known correlation metrics.

Definition A.2 (k-correlation)**.** *For random variables X, A with finite support sets $\mathcal{X}, \mathcal{A}$ respectively, k-correlation between X and A is defined as:*

$$\mathcal{J}_k(X,A) \coloneqq \sum_{i=1}^{k} \lambda_k(X,A). \tag{8}$$

The special cases of k-correlation correspond to well-known correlation metrics. For example, $\mathcal{J}_1(X,A) = \rho_m(X,A)^2$ is the Rényi maximum correlation (Buja, 1990). Similarly, if $d = \min\{|\mathcal{X}|,|\mathcal{A}|\} - 1$, then $\mathcal{J}_d(X,A) = \chi^2(X,A)$ is the chi-squared correlation.

Note that the smallest PIC is $\lambda_d(X;A)$ because $\mathbf{Q}$ has rank $(d+1)$. We will show how $\lambda_d(X,A)$ plays an important role in understanding the feasibility of perfect concept erasure. Next, we introduce the concept of erasure-utility coefficient for a given joint distribution, $p_{A,X}$.

Definition A.3 (Erasure-Utility Coefficient)**.** *The optimal erasure-utility coefficient for $p_{A,X}$ is:*

$$v^* \coloneqq \inf \frac{I(Z;A)}{I(Z;X)}.$$

We observe that the erasure-utility coefficient is closely related to concept erasure, where we want to reduce $I(Z;A)$ while maximizing $I(Z;X)$. Under perfect concept erasure, $v^* = 0$. Next, we restate the result from Du et al. (2020), which provides an upper bound for the coefficient, v^*.

Lemma A.4 (Optimal Erasure-Utility Coefficient (Calmon et al., 2017))**.** *Let $d \coloneqq \min\{|\mathcal{A}|,|\mathcal{X}|\}-1$, and $\lambda_d(A,X)$ be the smallest principal inertia component of any distribution $p(A,X)$ with finite support $\mathcal{A}\times\mathcal{X}$. Then,*

$$v^* \leq \begin{cases} \lambda_d(A,X), & \text{if} |\mathcal{X}| \leq |\mathcal{A}| \\ 0, & \text{otherwise} \end{cases}.$$

Lemma 3.2 (Perfect Concept Erasure Achievability)**.** *Under the assumptions (A1)-(A4), for a joint distribution $p_{A,X}$ with finite support defined over $\mathcal{A}\times\mathcal{X}$, perfect concept erasure, $I(Z;A)=0$, is achievable if either (a) smallest principal inertia component, $\lambda_d(A,X)=0$ or (b) $|\mathcal{X}|>|\mathcal{A}|$.*

Proof of Lemma 3.2. From Lemma A.4 and Definition A.3, we can say that $v^*=0$ when either $\lambda_d(A,X)=0$ or $|\mathcal{X}|>|\mathcal{A}|$. □

A.2 Proof of Lemma 3.3

Lemma 3.3 (Perfect Erasure Bound)**.** *Under the assumptions (A1)-(A5), if an erasure function $f:\mathcal{X}\to\mathcal{Z}$ achieves perfect concept erasure $I(Z;A)=0$, then the utility $I(Z;X)$ is bounded as:*

$$I(Z;X) = I(Z;X|A) \leq H(X|A),$$

where the equality is satisfied when $H(X|Z,A)=0$.

Proof. The proof relies on the following chain rule for mutual information:

$$I(Z;X) = I(Z;A) + I(Z;X|A) - I(Z;A|X) \tag{9}$$

For completeness, we prove it below:

$$\begin{aligned}
& I(Z;A|X) = H(Z|X) - H(Z|A,X) \\
\Rightarrow\ & I(Z;A|X) = H(Z) - H(Z) + H(Z|A) - H(Z|A) + H(Z|X) - H(Z|A,X) \\
\Rightarrow\ & I(Z;A|X) = [H(Z) - H(Z|A)] - [H(Z) - H(Z|X)] + [H(Z|A) - H(Z|A,X)] \\
\Rightarrow\ & I(Z;A|X) = I(Z;A) - I(Z;X) + I(Z;X|A) \\
\Rightarrow\ & I(Z;X) = I(Z;A) - I(Z;A|X) + I(Z;X|A).
\end{aligned}$$

In the above result, we plug-in $I(Z;A)=0$ (due to perfect erasure) and $I(Z;A|X)=0$ (due to Markov assumption (A1)). We obtain the following result:

$$I(Z;X) = I(Z;X|A). \tag{10}$$

Now, replacing the upper bound of $I(Z;X|A) \leq H(X|A)$, we get:

$$I(Z;X) \leq H(X|A).$$

The equality is satisfied when the following holds:

$$\begin{aligned} & I(Z;X|A) = H(X|A) \\ \Rightarrow\ & H(X|A) - H(X|Z,A) = H(X|A) \\ \Rightarrow\ & H(X|Z,A) = 0. \end{aligned}$$

This completes the proof. □

A.3 Justification for Data Constraints in Definition 4.1

In this section, we provide the justification behind the choice of data constraints necessary for optimal erasure, where the distribution of representations should be permutations. First, we present an additional Lemma that would be useful in developing the intuition behind the data constraints.

Lemma A.5 (Zero Conditional Entropy)**.** *If X and Y are random variables defined on finite supports $\mathcal{X}$ and $\mathcal{Y}$ respectively and $Y = f(X)$, then $H(X|Y) = 0$ if and only if f is a bijective map.*

Proof of Lemma A.5. First, we prove the forward direction, which says that f is bijective given $H(X|Y) = 0$.

$$\begin{aligned} & H(X|Y) = 0 \\ \Rightarrow\ & \sum_{y\in\mathcal{Y}} p(y) H(X|Y=y) = 0 \\ \overset{(a)}{\Longrightarrow}\ & \forall y \in \mathcal{Y}, H(X|Y=y) = 0 \end{aligned} \tag{11}$$

(a) holds because $\forall y \in \mathcal{Y}$, $p(y) > 0$. Eq. 11 implies that the probability distribution $\forall y, P(X|Y=y)$ is degenerate and there exists $x \in \mathcal{X}$ such that

$$\forall y, \exists x \text{ s.t. } P(X=x|Y=y) = 1, \forall x' \neq x, P(X=x'|Y=y) = 0. \tag{12}$$

Recall that we have $Y = f(X)$, and we want to show that f is invertible (and therefore bijective). Suppose not, then $\exists x, x' \in X$ that $f(x) = f(x')$. This contradicts Eq. 12 because $P(x|y) = 1$. Therefore, f must be invertible and bijective.

Next, we prove the backward direction, which shows that $H(X|Y) = 0$ if f is bijective.

When f is bijective, the following conditional distribution holds for a given y:

$$P(X = f^{-1}(y)|Y=y) = 1 \text{ and } P(X=x|Y=y) = 1, \forall x \neq f^{-1}(y). \tag{13}$$

For a fixed y, using Eq. 13, we can write the conditional entropy as:

$$\begin{aligned} H(X|Y=y) &= -\sum_{x\in\mathcal{X}} P(X=x|Y=y) \log P(X=x|Y=y) \\ &= \left[1 \cdot \log 1 + \sum_{x \neq f^{-1}(y)} 0 \cdot \log 0 \right] \\ &= 0. \end{aligned} \tag{14}$$

Using the above result, we can write the following:

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p(y) H(X|Y=y) = \sum_{y\in\mathcal{Y}} p(y) \cdot 0 = 0. \tag{15}$$

This completes the proof. □

With the above result, we are prepared to explain the data constraints in Definition 4.1.

Definition 4.1 (Data Constraints)**.** *We consider the class of data for which the distributions of different groups are permutations of each other:*

$$\forall (i,j),\ (a)\ |\mathcal{X}_i| = |\mathcal{X}_j|,\ (b)\ \forall x \in \mathcal{X}_i,\ P_i(x) = P_j(\sigma_{ij}(x)),$$

where $P_i = P(X|A = a_i)$*, and for which there exists an erasure function* f *that achieves* $I(Z;A) = 0$ *and* $I(X;Z) = H(X|A)$*.*

The intuition of this definition is that we want to find an f that makes use of the bijective maps between

First, we note that due to the disjoint support assumption (A4), we can write f as a piecewise function:

$$f(x) = \{f_i(x) |\ x \in \mathcal{X}_i\},\ \text{where } f_i : \mathcal{X}_i \to \mathcal{Z}. \tag{16}$$

From Lemma 3.3, we know that the utility outer bound holds when the following condition is true:

$$\begin{aligned} & H(X|Z,A) = 0 \\ \overset{(a)}{\implies} & H(X|Z, A = a_i) = 0, \forall i \in \{1,\dots,|\mathcal{A}|\} \end{aligned} \tag{17}$$

(a) holds because $\forall i \in \{1,\dots,|\mathcal{A}|\}, p(a_i) > 0$ by definition of the support. Applying Lemma A.5, we observe that Eq. 17 is satisfied only when all f_i's are bijective maps.

The bijective maps f_i's imply that the support size of the input and output must be the same $|\mathcal{X}_i| = |\mathcal{Z}|$ (the support of $\mathcal{X}$ and $\mathcal{Z}$ is finite as stated in Assumption (A2)). This implies for every pair of concept group $\forall (i,j), |\mathcal{X}_i| = |\mathcal{X}_j|$, which proves the first data constraint.

The bijective maps f_i's also imply that $\forall i$ the distributions $P(Z|A = a_i)$ and $P(X|A = a_i)$ are permutations of each other, as shown below:

$$\begin{aligned} \forall i, P[Z = z|A = a_i] &= \sum_{x \in \mathcal{X}_i} \mathbf{1}\left[f_i(x) = z\right] P\left[X = x|A = a_i\right] \\ &= P_i\left[X = f_i^{-1}(z)\right] \\ &= P_i\left[X = \sigma_i(x)\right], \end{aligned} \tag{18}$$

where $\sigma_i(\cdot)$ is a permutation function and $P_i(X) = P(X|A = a_i)$.

Next, we focus on the condition of perfect erasure:

$$\begin{aligned} & I(Z;A) = 0 \\ \Rightarrow\ & H(Z) - H(Z|A) = 0 \\ \Rightarrow\ & H(Z) = H(Z|A) \\ \Rightarrow\ & H(Z) = H(Z|A = a_i), \forall i \in \{1,\dots,|\mathcal{A}|\}. \end{aligned} \tag{19}$$

Next, we note that the output distribution can be written as: $P(Z) = \sum_i p(a_i) P(Z|A = a_i)$. We can write down the entropy of Z as the entropy of the underlying distributions as below:

$$\begin{aligned} H(Z) &= H\left(\sum_i p(a_i) P[Z|A = a_i]\right) \\ &\geq \sum_i p(a_i) H\left(Z|A = a_i\right). \end{aligned} \tag{20}$$

Eq. 20 holds due to the concavity of the entropy function. However, for the Eq. 19 to hold, the equality in Eq. 20 needs to be satisfied. The equality is satisfied only when all the underlying distributions are equal, as shown below:

$$\begin{aligned} & P[Z|A = a_i] = P[Z|A = a_j], \forall (i,j) \\ \Rightarrow\ & P_i[X = \sigma_i(x)] = P_j[X = \sigma_j(x)] && (21) \\ \Rightarrow\ & P_i[X = x] = P_j[X = \sigma_i^{-1}\sigma_j(x)] && (22) \\ \Rightarrow\ & P_i[X = x] = P_j[X = \sigma_{ij}(x)], \end{aligned}$$

where $\sigma_{ij} = \sigma_i \odot \sigma_j$. Eq. 21 holds due to the result in Eq. 18. The result in Eq. 22 implies that for all pairs of concept groups (i,j), the underlying distributions are permutations of each other. This explains the second data constraint.

A.4 Justification for Method 4.2

In this section, we will show that the perfect erasure function defined in Eq. 3 achieves the MI outer bounds.

Method 4.2 (Perfect Erasure Function)**.** *If the data constraints in Definition 4.1 hold, then define erasure function f as:*

$$f(x) = \{\sigma_i(x) |\ x \in \mathcal{X}_i,\ \sigma_i \in \Sigma_i\}, \tag{3}$$

where σ_i is any bijective function that transforms P_i into Q defined by the set, Σ_i below:

$$\Sigma_i = \{\sigma : \mathcal{X}_i \to \mathcal{Z} | \forall x \in \mathcal{X}_i,\ P_i(x) = Q(\sigma(x))\}\,. \tag{4}$$

First, we show that the perfect erasure function defined in Eq. 3 achieves the utility outer bound. Using Lemma 3.3, we get the following:

$$\begin{aligned} I(Z;X) &= I(Z;X|A) \\ &= \sum_i p(a_i) I(Z;X|A=a_i) \\ &= \sum_i p(a_i)[H(X|A=a_i) - H(X|Z,A=a_i)] \\ &= \sum_i p(a_i) H(X|A=a_i) \qquad (23) \\ &= H(X|A). \end{aligned}$$

Eq. 23 can hold since $H(X|Z, A=a_i) = 0$ if $f_i : \mathcal{X}_i \to \mathcal{Z}$ is a bijective map.

Next, we show that the erasure function achieves perfect concept erasure. We compute the conditional distribution:

$$\begin{aligned} \forall z \in \mathcal{Z},\ P(Z=z|A=a_i) &= \sum_{x \in \mathcal{X}_i} \mathbf{1}[f_i(x)=z] P(X=x|A=a_i) \qquad (24) \\ &= P_i(X = f_i^{-1}(z)) \\ &= Q(z). \qquad (25) \end{aligned}$$

Eq. 24 holds because f_i is a bijective map. Eq. 25 holds because of the erasure function definition in Eq. 4, where $P(X=x|A=a_i) = P_i(x) = Q(f_i(x))$. This implies the probability distribution over $\mathcal{Z}$ is:

$$\begin{aligned} \forall z \in \mathcal{Z},\ P(Z=z) &= \sum_i p(a_i) P[Z=z|A=a_i] \\ &= \sum_i p(a_i) Q(z) \\ &= Q(z). \qquad (26) \end{aligned}$$

Next, we use Eq. 25 and Eq. 26 to derive the privacy mutual information as shown below:

$$\begin{aligned} I(Z;A) &= H(Z) - H(Z|A) \\ &= H(Z) - \sum_i p(a_i) H(Z|A=a_i) \\ &= H(Q) - \sum_i p(a_i) H(Q) \qquad (27) \\ &= H(Q) - H(Q) \\ &= 0. \end{aligned}$$

Eq. 27 holds because $P(Z=z) = Q(z)$ and $\forall i \in \{1,\ldots,|\mathcal{A}|\},\ P(Z=z|A=a_i) = Q(z)$.

The idea of this is to find the erasure function defined in Method 4.2, which achieves the outer bounds $I(Z;X) = H(X|A)$ and $I(Z;A) = 0$.

B METHOD

B.1 Optimization Procedure

In this section, we present an alternative method to optimize the objective function presented in Eq. 7, when the input distributions of the concept groups are unequal. This optimization is dependent on two sets of variables: common support Q and coupling maps $\Gamma^{(i)} = \Gamma(P_i, Q)$. The key idea is to eliminate one set of variables and write the objective function only in terms of the coupling maps, since $Q = \sum_k \Gamma^{(i)}_{kj}$ is equal to the marginals of the coupling maps.

Therefore, we can write the objective as shown below:

$$\max_{\{\Gamma^{(i)}\}_{i=1}^{|\mathcal{A}|}} \frac{1}{|\mathcal{A}|} H\left(\sum_k \Gamma^{(i)}_{kj}\right) - \sum_i p(a_i) H_{\Gamma^{(i)}}\left(P_i, \sum_k \Gamma^{(i)}_{kj}\right), \tag{28}$$

$$\text{where } \forall (i_1, i_2) \in [1, |\mathcal{A}|],\ \sum_k \Gamma^{(i_1)}_{kj} - \sum_k \Gamma^{(i_2)}_{kj} = 0 \text{ and } \sum_k \Gamma^{(i)}_{kj} = P_i. \tag{29}$$

The above optimization can be solved using projected gradient descent (Madry et al., 2017), where we take unconstrainted gradient steps using the differentiable objective in Eq. 28 and then projecting the solutions obtained, $\{\Gamma_i\}$ to the linear constraints in Eq. 29. However, since Eq. 28 is non-convex, solving this optimization still doesn't guarantee convergence to the global optima.

We applied this optimization method in small-scale synthetic setups and found it to function well. However, this method is not scalable to probability distributions with a large support set. This is because the number of linear constraints in Eq. 29 blows up, and the projection step is extremely expensive in practice.

C EXPERIMENTS

Contents

C.1 Experimental Setup

Performing concept erasure using `PEF` does not require any parameter estimation except when the distributions are unequal. In this scenario, we use the Bayesian Optimization library (Nogueira, 2014–) and use the upper confidence bound (UCB) of exploitation and exploration as the acquisition function (with $\kappa = 2.5$). For evaluation of $\mathcal{V}$-information and MSE loss, we use the MLPClassifier and MLPRegressor functions from the scikit-learn (Pedregosa et al., 2011) library with its default settings. All the baseline models were trained on a 22GB NVIDIA Quadro RTX 6000 GPU and experiments were performed using the PyTorch (Paszke et al., 2019) framework. For baseline methods, we use the hyperparameters presented in the corresponding papers.

Next, we discuss the data generation process for the synthetic experiments reported in Section 5. We consider two concept groups and sample N 3D representations from a uniform distribution for each group. These representations form the support of the probability distributions we define next. We either define an equal (uniform or Gaussian) or unequal probability distribution using the support elements from each group. Using these distributions, we sample 10K representations from each group to form our final representation set. The concept label for each representation is the group from which it was sampled. For toxicity classification, we use the Jigsaw dataset, which contains online comments and the label is a fraction between 0 and 1 indicating the toxicity of the comment. We consider online comments associated with the following religion groups – Buddhist, Christian, Hindu, Jewish, Muslim, and others.

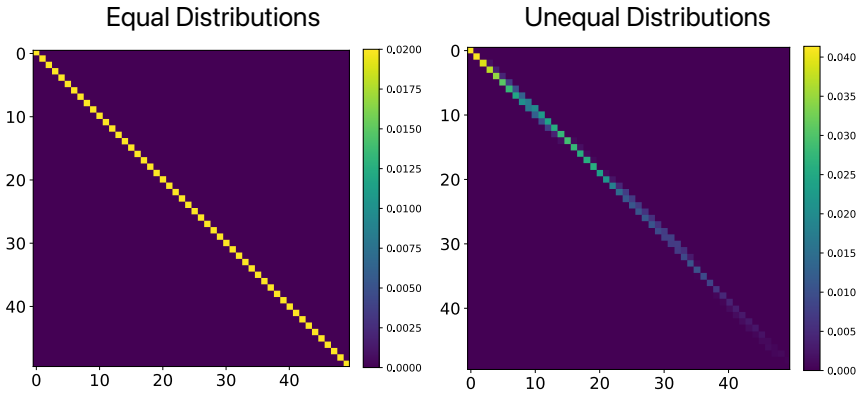

Figure 7: (*Left*) We show the minimum entropy coupling (joint distribution) between two equal distributions. We observe that the coupling map is a diagonal matrix indicating a 1-1 bijective map. (*Right*) We show that minimum entropy coupling between two unequal distributions is not a bijective map and individual support elements can be mapped to different elements.

C.2 Analysis Experiments

We conduct analysis experiments to inspect the erasure functions obtained using the formulation in Eq. 4. Specifically, we visualize the coupling maps (joint distributions) formed between two distributions $\Gamma(P_i, Q)$. We consider two scenarios: (i) equal distributions ($P_i = Q$) and (ii) unequal distributions ($P_i \neq Q$). In Figure 7, we visualize the minimum entropy coupling maps formed in these two settings. In Figure 7 (left), we observe that the coupling map is bijective (1-1 map), as predicted by Lemma 4.2. In Figure 7 (center), we observe that the coupling map is not bijective and therefore perfect erasure is not achieved.

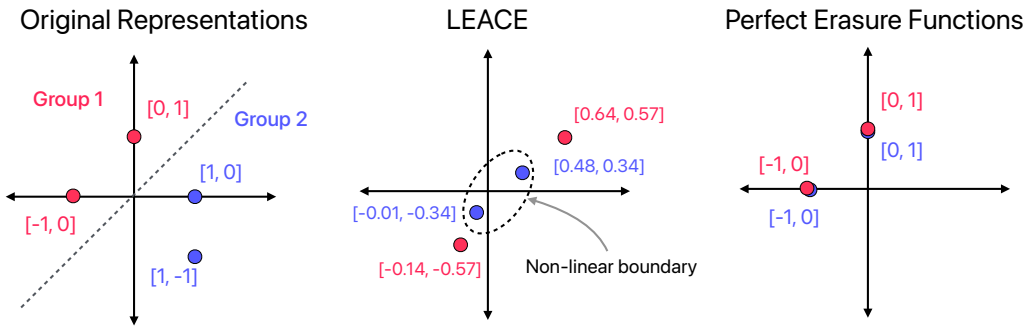

Figure 8: A toy example to illustrate the efficacy of erasure techniques LEACE and PEF. (*Left*) The original representation space with two linearly separable concept groups. (*Middle*) The erased representations obtained using LEACE, where there still exists a non-linear boundary separating the two groups. (*Right*) The erased representations from PEF where the concept has been perfectly erased.

C.3 Comparison with LEACE

In this section, we compare our proposed method PEF with LEACE (Belrose et al., 2024). LEACE is another perfect erasure technique that prevents linear adversaries from extracting concept information from erased representations. However, stronger non-linear adversaries may still be able to extract concept information. We illustrate this using a toy example in Figure 8. In this example, we consider a 2D original representation space (left plot in Figure 8) with two concept groups (each with only two elements). After erasure using LEACE (Figure 8 (middle)), we observe that there still exists a non-linear boundary that can separate the two groups and thereby reveal concept information. In contrast, PEF precisely maps elements between the two groups (Figure 8 (right)), making it impossible for any adversary to distinguish them. This example illustrates a scenario where LEACE can reveal concept information while PEF remains effective.

D BROADER IMPACT

Erasing sensitive concepts from data representations can reduce bias and enhance privacy, but it may also lead to the loss of valuable information, potentially diminishing the effectiveness of a machine learning model. The definition of sensitive concept attributes varies widely across cultural, ethical, and legal contexts. This work assumes that these attributes can be clearly defined and universally agreed upon, which is not always feasible. Therefore, developers should consider the societal impact carefully before implementing such erasure frameworks in practice.

`PEF` is intended to be used in applications where the developer is aware of the concept that needs to be erased. `PEF` can only erase concepts where the labels are annotated either as categorical attributes. One potential misuse of `PEF` would be to define relevant features for a task (e.g., educational background) as a concept to be erased. In such scenarios, the decision-making system can end up relying on arbitrary personal information to make hiring decisions. It is important to have oversight over such practices by using standard fairness measures like demographic parity.