Bottlenecks CLUB: Unifying Information-Theoretic Trade-offs Among Complexity, Leakage, and Utility

Behrooz Razeghi, *Member, IEEE,* Flavio P. Calmon, *Member, IEEE,* Deniz Gündüz, *Fellow, IEEE,* and Slava Voloshynovskiy, *Senior Member, IEEE*

***Abstract*—Bottleneck problems are an important class of optimization problems that have recently gained increasing attention in the domain of machine learning and information theory. They are widely used in generative models, fair machine learning algorithms, design of privacy-assuring mechanisms, and appear as information-theoretic performance bounds in various multi-user communication problems. In this work, we propose a general family of optimization problems, termed as *complexity-leakage-utility bottleneck (CLUB)* model, which (i) provides a unified theoretical framework that generalizes most of the state-of-the-art literature for the information-theoretic privacy models, (ii) establishes a new interpretation of the popular generative and discriminative models, (iii) constructs new insights for the generative compression models, and (iv) can be used to obtain fair generative models. We first formulate the CLUB model as a complexity-constrained privacy-utility optimization problem. We then connect it with the closely related bottleneck problems, namely information bottleneck (IB), privacy funnel (PF), deterministic IB (DIB), conditional entropy bottleneck (CEB), and conditional PF (CPF). We show that the CLUB model generalizes all these problems as well as most other information-theoretic privacy models. Then, we construct the deep variational CLUB (DVCLUB) models by employing neural networks to parameterize variational approximations of the associated information quantities. Building upon these information quantities, we present unified objectives of the *supervised* and *unsupervised* DVCLUB models. Leveraging the DVCLUB model in an unsupervised setup, we then connect it with state-of-the-art generative models, such as variational auto-encoders (VAEs), generative adversarial networks (GANs), as well as the Wasserstein GAN (WGAN), Wasserstein auto-encoder (WAE), and adversarial auto-encoder (AAE) models through the optimal transport (OT) problem. We then show that the DVCLUB model can also be used in fair representation learning problems, where the goal is to mitigate the undesired bias during the training phase of a machine learning model. We conduct extensive quantitative experiments on colored-MNIST and CelebA datasets.**

Manuscript received July 19, 2022; revised February 04, 2023; accepted March 03, 2023. This research was partly funded by the Swiss NSF (No. 200021_182063), U.S. NSF (CIF 1900750 & CAREER 1845852), and UK EPSRC (CHIST-ERA-18-SDCDN-001 & EP/T023600/1). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Valeria Loscri. (Corresponding author: Behrooz Razeghi.)

B. Razeghi and S. Voloshynovskiy are with the Department of Computer Sciecne, University of Geneva, Switzerland (e-mail: {behrooz.razeghi, svolos}@unige.ch).

F. P. Calmon is with the School of Engineering and Applied Sciences, Harvard University, US (e-mail: flavio@seas.harvard.edu).

D. Gündüz is with the Department of Electrical and Electronic Engineering, Imperial College London, UK (e-mail: d.gunduz@imperial.ac.uk).

Implementation codes available at: https://github.com/BehroozRazeghi
Expanded version available at: https://arxiv.org/abs/2207.04895
Digital Object Identifier 10.1109/TIFS.2022.XXXXXXX

***Index Terms*—Information-theoretic privacy, statistical inference, information bottleneck, obfuscation, generative models.**

1. Introduction

RELEASING an *'optimal' representation* of data for a given task while simultaneously assuring *privacy* of the individuals' identity and their associated data is an important challenge in today's highly connected and data-driven world, and has been widely studied in the information theory, signal processing, data mining and machine learning communities. An optimal representation is the most useful (sufficient), compressed (compact), and least privacy-breaching (minimal) representation of data. Indeed, an optimal representation of data can be obtained subject to constraints on the target task and its computational and storage complexities.

We investigate the problem of privacy-preserving data representation for a specific *utility task*, e.g., classification, identification, or reconstruction. Treating utility and privacy in a statistical framework [1] with mutual information as both the utility and obfuscation measures, we generalize the privacy funnel (PF) [2] and information bottleneck (IB) [3] models, and introduce a new and more general model called *complexity-leakage-utility bottleneck (CLUB)*. Consider two parties, a data owner and a utility service provider. The data owner observes a random variable $\mathbf{X}$ and acquires some utility from the service provider based on the information he discloses. Simultaneously, the data owner wishes to limit the amount of information revealed about a sensitive random variable $\mathbf{S}$ that depends on $\mathbf{X}$. Therefore, instead of revealing $\mathbf{X}$ directly to the service provider, the data owner releases a new representation, denoted by $\mathbf{Z}$. The amount of information leaked to the service provider (public domain) about the sensitive variable $\mathbf{S}$ is measured by the mutual information $\mathrm{I}(\mathbf{S};\mathbf{Z})$. Moreover, the data owner is subjected to a constraint on information complexity of representation that is revealed to the service provider. This imposed information complexity is measured by $\mathrm{I}(\mathbf{X};\mathbf{Z})$. Moreover, in general, the acquired utility depends on a utility random variable $\mathbf{U}$ that is dependent on $\mathbf{X}$ and may also be correlated with $\mathbf{S}$. The amount of useful information revealed to the service provider is measured by $\mathrm{I}(\mathbf{U};\mathbf{Z})$. Therefore, considering a Markov chain $(\mathbf{U},\mathbf{S}) \multimap \mathbf{X} \multimap \mathbf{Z}$, our aim is to share a *privacy-preserving (sanitized) representation* $\mathbf{Z}$ of *observed data* $\mathbf{X}$, through a stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$, while preserving information about

utility attribute $\mathbf{U}$ and obfuscating information about *sensitive attribute* $\mathbf{S}$. The stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$ is called *complexity-constrained obfuscation-utility-assuring mapping*. The general diagram of our setup is depicted in Fig. 1.

1.1. Contributions

- We propose the CLUB model, which provides a characterization of the complexity-leakage-utility trade-off in a data release mechanism. This model provides a statistical inference framework that generalizes most of the state-of-the-art literature in the information-theoretic privacy models.
- We provide new insights into the representation learning problems by bridging information-theoretic privacy with the generative models through the bottleneck principle. We start from a purely information-theoretic framework that has roots in the classical Shannon rate-distortion theory. Next we demonstrate the connection of our model to several recent research trends in generative models and representation learning. In particular, we show that the CLUB model has connections to the variational fair auto-encoder model [4] to learn representations for a prediction problem while removing potential biases against some variable (e.g., gender, ethnicity) from the results of a learned model [4]–[7].
- We generalize the CLUB perspective to comprise both *supervised* and *unsupervised* setups. In the unsupervised setup, the deep variational CLUB (DVCLUB) model is shown to have interesting connections with the several generative models, such as variational auto-encoder (VAE) [8], β-VAE [9], InfoVAE [10], generative adversarial network (GAN) [11], Wasserstein GAN (WGAN) [12], Wasserstein auto-encoders [13], adversarial auto-encoder (AAE) [14], VAE-GAN [15], and BIB-AE [16].
- We conduct extensive qualitative and quantitative experiments on several real-world datasets to evaluate and validate the effectiveness of the proposed CLUB model.

Please, see [17] for an extended version of this paper.

1.2. State-of-the-Art

In order to protect privacy, various mechanisms have been developed that aim to prevent certain statistical inferences about data. These mechanisms can add noise to the output or randomize data to conceal private information before it is shared with a third party. The effectiveness of these mechanisms is evaluated using different privacy metrics, which consider the type of adversary, and the data sources available to them. There are two main types of privacy-preserving mechanisms: *prior-independent* and *prior-dependent*. Prior-independent mechanisms make minimal assumptions about the data distribution and the information held by an adversary, and are designed to protect privacy regardless of the specific characteristics of the data being protected or the motivations and capabilities of any potential adversaries. Prior-dependent mechanisms, on the other hand, make use of knowledge about the probability distribution of private data and the abilities of adversaries in order to design privacy-preserving mechanisms.

Addressing data anonymization [18], various well-known statistical formulations and schemes were proposed, such as k-anonymity [19], ℓ-diversity [20], t-closeness [21], differential privacy (DP) [22], and pufferfish privacy [23], which are based on some form of data perturbation. These mechanisms mainly focus on the querying data, inference algorithms and transporting. DP is the most popular *context-free* (prior-independent) notion of privacy, which is characterized in terms of the distinguishability of 'neighboring' databases. However, DP does not provide any guarantee on the average or maximum information leakage [1].

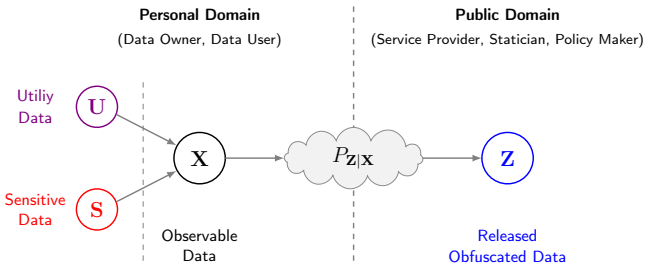


Fig. 1: The general CLUB framework.

Information-theoretic (IT) privacy is the study of designing mechanisms and metrics that preserve privacy when the statistical properties or probability distribution of data can be estimated or partially known. IT privacy approaches [1], [24]–[38] model and analyze the trade-off between privacy and utility using IT metrics, which quantify how much information an adversary can gain about private features from disclosed data. These metrics are often formulated in terms of divergences between probability distributions, such as f-divergences and Renyi divergence. IT privacy metrics can be operationalized in terms of an adversary's ability to infer sensitive data and can be used to balance the trade-off between allowing useful information to be drawn from disclosed data and preserving privacy. By using prior knowledge about the statistical properties of data and assumptions about the adversary's inference capabilities, IT privacy can help to understand the fundamental limits of privacy and how to balance privacy and utility. The IT privacy framework is inspired by Shannon's information-theoretic notion of secrecy [39], where security is measured through the equivocation rate at the eavesdropper, and by Reed [24] and Yamamoto's [25] treatment of security and privacy from a lossy source coding standpoint.

Data-driven privacy mechanisms, such as those inspired by generative adversarial networks (GANs) [11], aim to balance the protection of private information with the usefulness of released data. These mechanisms model the trade-off between privacy and utility as a game between a *defender (privatizer)* and an *adversary* [34], [40]–[42]. The privatizer encodes the dataset to minimize inference leakage on private/sensitive variables, while the adversary tries to infer these variables from the released data. Adversarial training algorithms, which are used to optimize privacy-preserving mechanisms, can be deterministic or incorporate randomness. Our model subsumes these models. We establish a more precise connection with the state-of-the-art after introducing the CLUB model.

Our model is inspired by [43], where the authors established the rate region of the extended Gray-Wyner system for two discrete memoryless sources, which include Wyner's common

information, Gács-Körner common information, IB, Körner graph entropy, necessary conditional entropy, and the PF, as extreme points. We extend and unify most of the previously proposed objectives in the literature based on IT privacy models. Our research is also closely related to [37], [44]–[46]. Considering the Markov chain $(\mathbf{U},\mathbf{S})-\circ-\mathbf{X}-\circ-\mathbf{Z}$, the authors in [44] addressed the problem of privacy-preserving representation learning in a scenario where the goal is to share a sanitized representation $\mathbf{Z}$ of high-dimensional data $\mathbf{X}$ while preserving information about utility attribute $\mathbf{U}$ and obfuscate information about private (sensitive) attribute $\mathbf{S}$. Inspired by GANs, the framework is formulated as a distribution matching problem. Compared with [44], our formulation is more general as it addresses a key missing component in their formulation, i.e., the rate (description length, information complexity) constraint. Another fundamental related work to ours is [37], which studied the rate-constrained privacy-utility trade-off problem and considered a similar model as [44], independently. The proposed framework is restricted to discrete alphabets and studied the necessary and sufficient conditions for the existence of positive utility, i.e., $\mathrm{I}(U;Z)>0$, under a perfect obfuscation regime, i.e., $\mathrm{I}(S;Z)=0$. Analogous to [37], considering the Markov chain $(U,S)-\circ-X-\circ-Z$, the authors in [45] adopted a local information geometry analysis to construct the modal decomposition of the joint distributions, divergence transfer matrices, and mutual information. Next, they obtained the locally sufficient statistics for inferences about the utility attribute, while satisfying the perfect obfuscation constraint. Furthermore, they developed the notion of perfect obfuscation based on χ^2-divergence and Kullback–Leibler divergence in the Euclidean information space. Considering the Markov chain $(\mathbf{U},\mathbf{S})-\circ-\mathbf{X}-\circ-\mathbf{Z}$, the role of information complexity in privacy leakage about an attribute of an adversary's interest is studied in [46]. In contrast to the PF and generative adversarial privacy models, they considered the setup in which the adversary's interest is *not known a priori* to the data owner. More detailed connections between the CLUB model and the state-of-the-art bottleneck models, fair machine learning models, generative models, and modern data compression models are presented in Sec. 6.

1.3. Outline

In Sec. 2, we present the general problem statement of the CLUB model, and next we briefly review and introduce a few preliminary concepts that are necessary to understand the problem formulation. We then present the CLUB model in Sec. 3. The DVCLUB model is presented in Sec. 4. Experimental results are provided in Sec. 5. The detailed connections between the CLUB model and the state-of-the-art bottleneck models, fair machine learning models, generative models, and modern data compression models are presented in Sec. 6. Finally, conclusions are drawn in Sec. 7.

1.4. Notations

Throughout this paper, random variables are denoted by capital letters (e.g., X, Y), deterministic values are denoted by small letters (e.g., x, y), random vectors are denoted by capital bold letter (e.g., $\mathbf{X}$, $\mathbf{Y}$), deterministic vectors are denoted by small bold letters (e.g., $\mathbf{x},\mathbf{y}$), alphabets (sets) are denoted by calligraphic fonts (e.g., $\mathcal{X},\mathcal{Y}$), and for specific quantities/values we use sans serif font (e.g., x, y, C, D, P, Ω). Superscript $(.)^T$ stands for the transpose. Also, we use the notation $[N]$ for the set $\{1,2,...,N\}$. $\mathrm{H}\left(P_{\mathbf{X}}\right) \coloneqq \mathbb{E}_{P_{\mathbf{X}}}\left[-\log P_{\mathbf{X}}\right]$ denotes the Shannon entropy; $\mathrm{H}\left(P_{\mathbf{X}}\|Q_{\mathbf{X}}\right) \coloneqq \mathbb{E}_{P_{\mathbf{X}}}\left[-\log Q_{\mathbf{X}}\right]$ denotes the cross-entropy of the distribution $P_{\mathbf{X}}$ relative to a distribution $Q_{\mathbf{X}}$; and $\mathrm{H}\left(P_{\mathbf{Z}|\mathbf{X}}\|Q_{\mathbf{Z}|\mathbf{X}} \mid P_{\mathbf{X}}\right) \coloneqq \mathbb{E}_{P_{\mathbf{X}}}\mathbb{E}_{P_{\mathbf{Z}|\mathbf{X}}}\left[-\log Q_{\mathbf{Z}|\mathbf{X}}\right]$ denotes the cross-entropy loss for $Q_{\mathbf{Z}|\mathbf{X}}$. The relative entropy is defined as $\mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{X}}\|Q_{\mathbf{X}}\right) \coloneqq \mathbb{E}_{P_{\mathbf{X}}}\left[\log \frac{P_{\mathbf{X}}}{Q_{\mathbf{X}}}\right]$. The conditional relative entropy is defined by $\mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Z}|\mathbf{X}}\|Q_{\mathbf{Z}|\mathbf{X}} \mid P_{\mathbf{X}}\right) \coloneqq \mathbb{E}_{P_{\mathbf{X}}}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}\|Q_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}\right)\right]$ and the mutual information is defined by $\mathrm{I}\left(P_{\mathbf{X}};P_{\mathbf{Z}|\mathbf{X}}\right) \coloneqq \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Z}|\mathbf{X}}\|P_{\mathbf{Z}} \mid P_{\mathbf{X}}\right)$. We abuse notation to write $\mathrm{H}\left(\mathbf{X}\right) = \mathrm{H}\left(P_{\mathbf{X}}\right)$ and $\mathrm{I}\left(\mathbf{X};\mathbf{Z}\right) = \mathrm{I}\left(P_{\mathbf{X}};P_{\mathbf{Z}|\mathbf{X}}\right)$ for random objects $\mathbf{X}\sim P_{\mathbf{X}}$ and $\mathbf{Z}\sim P_{\mathbf{Z}}$. We use the same notation for the probability distributions and the associated densities.

2. Problem Statement and Preliminaries

In this section, first we present the general problem statement of the CLUB model, and next we introduce the main concepts necessary to understand the fundamentals of our model. In particular, we concretely express our inference threat model and explain why we consider mutual information as the obfuscation and utility measures. Next, we briefly address the concepts of relevant information and minimal sufficient statistics.

2.1. General Problem Statement

Consider the problem of releasing a sanitized representation $\mathbf{Z}$ of observed data $\mathbf{X}$, through a stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$, while preserving information about a utility attribute $\mathbf{U}$ and obfuscating a sensitive attribute $\mathbf{S}$. In this scenario, $(\mathbf{U},\mathbf{S})-\circ-\mathbf{X}-\circ-\mathbf{Z}$ form a Markov chain. In general, one can consider a well-defined generic obfuscation measure as a functional of the joint distribution $P_{\mathbf{S},\mathbf{Z}}$ that captures the amount of leakage about $\mathbf{S}$ by releasing $\mathbf{Z}$. Let $\mathcal{C}_{\mathsf{L}} : \mathcal{P}\left(\mathcal{S}\times\mathcal{Z}\right) \to \mathbb{R}^+\cup\{0\}$ denote this generic leakage (obfuscation) measure. On the other hand, one can also consider an application-specific generic utility measure as a functional of the joint distribution $P_{\mathbf{U},\mathbf{Z}}$ that captures the utility that can be acquired about $\mathbf{U}$ by releasing $\mathbf{Z}$, instead of the original data, $\mathbf{X}$. Let $\mathcal{C}_{\mathsf{U}} : \mathcal{P}\left(\mathcal{U}\times\mathcal{Z}\right) \to \mathbb{R}^+\cup\{0\}$ denote a generic utility *loss* measure. Finally, let $\mathcal{C}_{\mathsf{C}} : \mathcal{P}\left(\mathcal{X}\times\mathcal{Z}\right) \to \mathbb{R}^+\cup\{0\}$ denote a generic measure of *complexity* of the distribution $P_{\mathbf{Z}|\mathbf{X}}$ for the data distribution $P_{\mathbf{X}}$. Then, one can consider the generic CLUB functional as:

$$\mathsf{CLUB}\left(P_{\mathbf{U},\mathbf{S},\mathbf{X}}\right) \coloneqq \inf_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ (\mathbf{U},\mathbf{S})-\circ-\mathbf{X}-\circ-\mathbf{Z}}} \mathcal{C}_{\mathsf{C}}\left(P_{\mathbf{X}},P_{\mathbf{Z}|\mathbf{X}}\right) + \mathcal{C}_{\mathsf{L}}\left(P_{\mathbf{S}},P_{\mathbf{Z}|\mathbf{X}}\right) + \mathcal{C}_{\mathsf{U}}\left(P_{\mathbf{U}},P_{\mathbf{Z}|\mathbf{X}}\right). \quad (1)$$

2.2. Obfuscation and Utility Measures under Logarithmic Loss

We consider obfuscation-utility trade-off model where both utility and obfuscation are measured under *logarithmic loss* (also often referred to as the self-information loss). The logarithmic loss function has been widely used in learning theory [47], image processing [48], IB [49], multi-terminal

source coding [50], as well as PF [2]. In this case, both the obfuscation and utility measures can be modelled by the mutual information. By minimizing the obfuscation measure under the logarithmic loss, one actually minimizes an upper bound on any bounded loss function [2].

Consider the *inference threat model* introduced in [1], which models a broad class of adversaries that perform statistical inference attacks on the sensitive data. Consider an inference cost function $\mathsf{C}: \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathbb{R}^+ \cup \{0\}$. Prior to observing $\mathbf{Z}$, the adversary chooses a belief distribution Q from the set $\mathcal{P}(\mathcal{S})$ of all possible distributions over $\mathcal{S}$, that minimizes the expected inference cost function $\mathsf{C}(\mathbf{S}, Q)$. Therefore:

$$Q^* = \arg \min_{Q \in \mathcal{P}(\mathcal{S})} \mathbb{E}_{P_{\mathbf{S}}} \left[\mathsf{C}(\mathbf{S}, Q)\right]. \tag{2}$$

Let $c_0^* = \mathbb{E}_{P_{\mathbf{S}}} \left[\mathsf{C}(\mathbf{S}, Q^*)\right]$ denote the corresponding minimum average cost. After observing $\mathbf{Z} = \mathbf{z}$, the adversary revises his belief distribution as:

$$Q_{\mathbf{z}}^* = \arg \min_{Q \in \mathcal{P}(\mathcal{S})} \mathbb{E}_{P_{\mathbf{S}|\mathbf{z}}} \left[\mathsf{C}(\mathbf{S}, Q) \mid \mathbf{Z} = \mathbf{z}\right]. \tag{3}$$

Let $c_{\mathbf{z}}^* = \mathbb{E}_{P_{\mathbf{S}|\mathbf{z}}} \left[\mathsf{C}(\mathbf{S}, Q_{\mathbf{z}}^*)\right]$ denote the corresponding minimum average cost of inferring $\mathbf{S}$ after observing $\mathbf{Z} = \mathbf{z}$. Therefore, the adversary obtains an average gain of $\Delta\mathsf{C} = c_0^* - \mathbb{E}_{P_{\mathbf{Z}}} \left[c_{\mathbf{Z}}^*\right]$ in inference cost. This cost gain measures the improvement in the quality of the inference of sensitive data $\mathbf{S}$ due to observation of $\mathbf{Z}$. Under the self-information loss cost function $\mathsf{C}(\mathbf{s}, Q) = -\log Q(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}$, the information leakage $\Delta\mathsf{C}$ can be measured by the Shannon mutual information $\mathrm{I}(\mathbf{S}; \mathbf{Z})$.

On the other hand, the stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$ should maintain the utility of desired data $\mathbf{U}$. Under the self-information loss, the utility measure is defined as $d(\mathbf{u}, \mathbf{z}) = -\log P_{\mathbf{U}|\mathbf{Z}}(\mathbf{u} \mid \mathbf{z})$, which is a function of $\mathbf{u}$ and $\mathbf{z}$ as well as stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$. Hence, the average utility loss is $\mathbb{E}_{P_{\mathbf{U},\mathbf{Z}}} \left[-\log P_{\mathbf{U}|\mathbf{Z}}\right] = \mathrm{H}(\mathbf{U} \mid \mathbf{Z})$ that can be minimized by designing the stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$. Consider some utility level $E^{\mathrm{u}} \geq 0$, such that we constrain to have $\mathrm{H}(\mathbf{U} \mid \mathbf{Z}) \leq E^{\mathrm{u}}$. Given $P_{\mathbf{U}}$, and therefore $\mathrm{H}(\mathbf{U})$, and assuming that $R^{\mathrm{u}} = \mathrm{H}(\mathbf{U}) - E^{\mathrm{u}} \geq 0$, the utility constraint can be recast as $\mathrm{I}(\mathbf{U}; \mathbf{Z}) \geq R^{\mathrm{u}}$.

In Sec. 3, built upon this general threat model and considering the self-information loss as both utility and obfuscation measures, we present a unified formulation for the complexity-leakage-utility trade-off that encompasses and extends upon the state-of-the-art.

2.3. Relevant Information

The relevant information is a common concept in information theory and statistics which captures the information that a random object $\mathbf{X}$ contains about another random object $\mathbf{U}$. Below, we present statistical and information-theoretical formulations proposed for measuring relevant information.

Definition 1 (Sufficient Statistic)**.** Let $\mathbf{U} \in \mathcal{U}$ be an unknown parameter and $\mathbf{X} \in \mathcal{X}$ be a random variable with conditional probability distribution function $P_{\mathbf{X}|\mathbf{U}}$. Given a function $f: \mathcal{X} \to \mathcal{Z}$, the random variable $\mathbf{Z} = f(\mathbf{X})$ is called a sufficient statistics for $\mathbf{U}$ if and only if $P_{\mathbf{X}|\mathbf{U},\mathbf{Z}}(\mathbf{x} \mid \mathbf{u}, \mathbf{z}) = P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x} \mid \mathbf{z}), \ \ \forall (\mathbf{x}, \mathbf{u}) \in \mathcal{X} \times \mathcal{U}$ [51], [52].

Equivalently we have $\mathrm{I}(\mathbf{U}; \mathbf{Z}) = \mathrm{I}(\mathbf{U}; \mathbf{X})$. Note that $\mathbf{U} - \circ - \mathbf{X} - \circ - f(\mathbf{X})$ forms a Markov chain, and by the data processing inequality (DPI), for a general statistic $f(\mathbf{X})$, we have $\mathrm{I}(\mathbf{U}; f(\mathbf{X})) \leq \mathrm{I}(f(\mathbf{X}); \mathbf{X})$. If equality holds, a sufficient statistic $\mathbf{Z}$ captures all the information in $\mathbf{X}$ about $\mathbf{U}$.

Definition 2 (Minimal Sufficient Statistic)**.** A sufficient statistics $\mathbf{Z}$ is said to be minimal if it is a function of all other sufficient statistics, that is, for all sufficient statistics $\mathbf{Z}'$, there exists f such that $\mathbf{Z} = f(\mathbf{Z}')$.

It means that $\mathbf{Z}$ induces the coarsest sufficient partition. In other words, $\mathbf{Z}$ achieves the maximum data reduction, while assuring $\mathrm{I}(\mathbf{U}; \mathbf{Z}) = \mathrm{I}(\mathbf{U}; \mathbf{X})$. Suppose that nature chooses a parameter $\mathbf{U}$ at random, after which the sample $\mathbf{X}$ is drawn from the distribution $P_{\mathbf{X}|\mathbf{U}}$. One can show that the statistic $\mathbf{Z}$ is a minimal sufficient statistic for $\mathbf{U}$, iff it is a solution of one of the two equivalent optimization problems:

$$\mathbf{Z} = \arg \min_{\substack{\mathbf{Z}' \text{: sufficient statistic} \\ \text{for } \mathbf{U}}} \mathrm{I}(\mathbf{X}; \mathbf{Z}') \equiv \arg \min_{\mathbf{Z}' \text{: } \mathrm{I}(\mathbf{U};\mathbf{Z}') = \mathrm{I}(\mathbf{U};\mathbf{X})} \mathrm{I}(\mathbf{X}; \mathbf{Z}').$$

It means that minimal sufficient statistic $\mathbf{Z}$ is the best compression of $\mathbf{X}$, with the zero information loss about parameter $\mathbf{U}$.

3. CLUB Model

3.1. General Setup

The CLUB model (Fig. 2) is a generalization of the sufficient statistic methods that formulate the problem of extracting the relevant information from a random object (data) $\mathbf{X}$ about the random object $\mathbf{U}$ that is of interest, while limiting statistical inference about a sensitive random object $\mathbf{S}$ that depends on $\mathbf{X}$ and is possibly depended on $\mathbf{U}$. Consider a scenario in which $P_{\mathbf{U},\mathbf{S},\mathbf{X}}$ is fixed and known by both the defender and the adversary, where $\mathbf{X} \in \mathcal{X}$ represents the data observed by the defender (e.g., high dimensional facial image), $\mathbf{U} \in \mathcal{U}$ denotes the utility attribute of our interest for a utility service provider (e.g., person identity), and $\mathbf{S} \in \mathcal{S}$ denotes the sensitive attribute (e.g., gender) that we wish to restrict its statistical inference. Intuitively, built upon our introduction of the concept of relevant information above, we intend to find a stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$ such that the posterior distribution of the utility attribute $\mathbf{U}$ is similar given the released representation $\mathbf{Z}$ and the original data $\mathbf{X}$, i.e., $P_{\mathbf{U}|\mathbf{Z}} \approx P_{\mathbf{U}|\mathbf{X}}$, while the posterior of private attribute $\mathbf{S}$ given released representation $\mathbf{Z}$ is as close as possible to its prior, i.e., $P_{\mathbf{S}|\mathbf{Z}} \approx P_{\mathbf{S}}$. In the rest of the paper, we assume our attributes are supported on a finite space.

3.2. Threat Model

We consider the inference threat model described in Sec. 2.2. In particular, we have the following assumptions:

- We assume the adversary is interested in an attribute $\mathbf{S}$ of data $\mathbf{X}$. The attribute $\mathbf{S}$ can be any (possibly randomized) function of $\mathbf{X}$. We restrict attribute $\mathbf{S}$ to be discrete, which captures the most scenarios of interest, e.g., a facial attribute, an identity, etc.
- The adversary observes released representation $\mathbf{Z}$ and the Markov chain $\mathbf{S} - \circ - \mathbf{X} - \circ - \mathbf{Z}$ holds.
- We assume that the adversary knows the *complexity-constrained obfuscation-utility-assuring mapping* $P_{\mathbf{Z}|\mathbf{X}}$ designed by the data owner (defender), i.e., the defence mechanism is public.

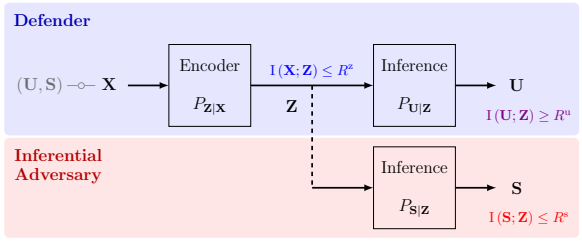


Fig. 2: The CLUB model.

3.3. Problem Formulation

Given three dependent (correlated) random variables $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{X}$ with joint distribution $P_{\mathbf{U},\mathbf{S},\mathbf{X}}$, the goal of the CLUB model is to find a representation $\mathbf{Z}$ of $\mathbf{X}$ using a stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$ such that: (i) $(\mathbf{U},\mathbf{S}) \multimap\!\!-\mathbf{X}\!-\!\!\circ\!\!-\mathbf{Z}$, and (ii) representation $\mathbf{Z}$ is maximally informative about $\mathbf{U}$ (maximizing $\mathrm{I}(\mathbf{U};\mathbf{Z})$) while being minimally informative about $\mathbf{X}$ (minimizing $\mathrm{I}(\mathbf{X};\mathbf{Z})$) and minimally informative about $\mathbf{S}$ (minimizing $\mathrm{I}(\mathbf{S};\mathbf{Z})$). We can formulate this three-dimensional trade-off by imposing constraints on the two of them. That is, for a given information complexity and information leakage constraints, $R^z \geq 0$ and $R^s \geq 0$, respectively, this trade-off can be formulated by a CLUB functional[1]:

$$\mathsf{CLUB}\left(R^z, R^s, P_{\mathbf{U},\mathbf{S},\mathbf{X}}\right) \coloneqq \sup_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ (\mathbf{U},\mathbf{S}) \multimap\!\!-\mathbf{X}\!-\!\!\circ\!\!-\mathbf{Z}}} \mathrm{I}(\mathbf{U};\mathbf{Z}) \quad \text{s.t. } \mathrm{I}(\mathbf{X};\mathbf{Z}) \leq R^z,\ \mathrm{I}(\mathbf{S};\mathbf{Z}) \leq R^s. \quad (4)$$

The constraint $\mathrm{I}(\mathbf{S};\mathbf{Z}) \leq R^s$ ensures that $\mathrm{H}(\mathbf{S} \mid \mathbf{Z}) \geq \mathrm{H}(\mathbf{S}) - R^s = E^a$, where E^a quantifies the amount of uncertainty for adversary. On the other hand, note that $\mathrm{I}(\mathbf{S};\mathbf{Z}) = \mathbb{E}_{P_\mathbf{Z}}\left[\mathrm{D_{KL}}\left(P_{\mathbf{S}|\mathbf{Z}} \| P_\mathbf{S}\right)\right]$. This means that for small values of R^s, the posterior of private attribute $\mathbf{S}$ given released representation $\mathbf{Z}$, i.e., $P_{\mathbf{S}|\mathbf{Z}}$, is as close as possible to the prior $P_\mathbf{S}$. The constraint $\mathrm{I}(\mathbf{X};\mathbf{Z}) \leq R^z$ controls *information complexity*[2] (aka *compactness*) of representations and goes beyond a simple regularization term. Indeed, it is related to the notion of *encoder capacity* [56], which is a measure of distinguishability among input data samples from their released representations. Moreover, the information complexity $\mathrm{I}(\mathbf{X};\mathbf{Z})$ establishes a fundamental relation to the generalization capability of the stochastic encoder model $P_{\mathbf{Z}|\mathbf{X}}$. The trade-off in (4) was studied in [37], where the constraint on $\mathrm{I}(\mathbf{U};\mathbf{Z})$ is motivated as a rate-constraint, and the trade-off is studied under perfect privacy regime (i.e., $\mathrm{I}(\mathbf{S};\mathbf{Z}) = 0$). Note that by controlling the rate R^z, the defender (data owner) exploits the imposed distortion at the utility service provider (authorized decoder) to control the uncertainty for the adversary (non-authorized decoder). The values $\mathsf{CLUB}\left(R^z, R^s, P_{\mathbf{U},\mathbf{S},\mathbf{X}}\right)$ for different R^z and R^s specify the CLUB curve.

In order to explore the CLUB curve, one must find optimal bottleneck representation $\mathbf{Z}$ for different values of R^z and R^s. In practice, the CLUB curve is explored by maximizing its associated Lagrangian functional. Therefore, equivalently, with the introduction of a Lagrange multipliers $\beta, \alpha \in [0,1]$[3], we can formulate the CLUB problem by the associated Lagrangian functional as follows:

$$\mathcal{L}_{\mathrm{CLUB}}\left(P_{\mathbf{Z}|\mathbf{X}}, \beta, \alpha\right) \coloneqq \mathrm{I}(\mathbf{U};\mathbf{Z}) - \beta\,\mathrm{I}(\mathbf{X};\mathbf{Z}) - \alpha\,\mathrm{I}(\mathbf{S};\mathbf{Z}). \quad (5)$$

Consider the set of 3-dimensional mutual information region for $(\mathbf{U},\mathbf{S},\mathbf{X}) \sim P_{\mathbf{U},\mathbf{S},\mathbf{X}}$, defined as $\mathcal{I}(P_{\mathbf{U},\mathbf{S},\mathbf{X}}) \coloneqq \bigcup_{P_{\mathbf{Z}|\mathbf{X}}} \left\{\mathrm{I}(\mathbf{U};\mathbf{Z}), \mathrm{I}(\mathbf{S};\mathbf{Z}), \mathrm{I}(\mathbf{X};\mathbf{Z}) : (\mathbf{U},\mathbf{S}) \multimap\!\!-\mathbf{X}\!-\!\!\circ\!\!-\mathbf{Z}\right\} \subseteq \mathbb{R}^3$. Also, consider the constrained information regions $\mathcal{R}(P_{\mathbf{U},\mathbf{S},\mathbf{X}})$, defined as $\mathcal{R}(P_{\mathbf{U},\mathbf{S},\mathbf{X}}) \coloneqq \left\{(R^u, R^s, R^z) \in \mathcal{I}_{\mathbf{U},\mathbf{S},\mathbf{X}} : \mathrm{I}(\mathbf{U};\mathbf{Z}) \geq R^u, \mathrm{I}(\mathbf{S};\mathbf{Z}) \leq R^s, \mathrm{I}(\mathbf{X};\mathbf{Z}) \leq R^z\right\}$. In Sec. 6, we the CLUB model's superiority by subsuming previous information-theoretic privacy models. The mutual information region $\mathcal{I}(P_{\mathbf{U},\mathbf{S},\mathbf{X}})$ encompasses the models proposed in [2], [3], [44], [57]–[59], highlighting the CLUB model's generality.

4. Deep Variational CLUB (DVCLUB)

Direct optimization of (5) to obtain the optimal stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$ is generally challenging. Instead, a tight variational bound can be optimized. Let $Q_{\mathbf{U}|\mathbf{Z}} : \mathcal{Z} \to \mathcal{P}(\mathcal{U})$, $Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}} : \mathcal{S} \times \mathcal{Z} \to \mathcal{P}(\mathcal{X})$ and $Q_\mathbf{Z} : \mathcal{Z} \to \mathcal{P}(\mathcal{Z})$ be variational approximations of the optimal *utility* decoder distribution $P_{\mathbf{U}|\mathbf{Z}}$, *uncertainty* decoder distribution $P_{\mathbf{X}|\mathbf{S},\mathbf{Z}}$, and latent space distribution $P_\mathbf{Z}$, respectively. In the sequel, we will obtain a variational bound $\mathcal{L}_{\mathrm{VCLUB}}\left(P_{\mathbf{Z}|\mathbf{X}}, Q_{\mathbf{U}|\mathbf{Z}}, Q_\mathbf{Z}, Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}, \beta, \alpha\right)$, such that for any valid mapping $P_{\mathbf{Z}|\mathbf{X}}$ satisfying the Markov chain condition $(\mathbf{U},\mathbf{S}) \multimap\!\!-\mathbf{X}\!-\!\!\circ\!\!-\mathbf{Z}$, we have:

$$\mathcal{L}_{\mathrm{CLUB}}\left(P_{\mathbf{Z}|\mathbf{X}}, \beta, \alpha\right) \geq \mathcal{L}_{\mathrm{VCLUB}}\left(P_{\mathbf{Z}|\mathbf{X}}, Q_{\mathbf{U}|\mathbf{Z}}, Q_\mathbf{Z}, Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}, \beta, \alpha\right). \quad (6)$$

The inequality holds with equality, if the variational approximations match the true distributions, i.e., when $Q_{\mathbf{U}|\mathbf{Z}} = P_{\mathbf{U}|\mathbf{Z}}$, $Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}} = P_{\mathbf{X}|\mathbf{S},\mathbf{Z}}$, and $Q_\mathbf{Z} = P_\mathbf{Z}$. Since (6) holds for any $P_{\mathbf{Z}|\mathbf{X}}$, instead of maximizing $\mathcal{L}_{\mathrm{CLUB}}\left(P_{\mathbf{Z}|\mathbf{X}}, \beta, \alpha\right)$, we will maximize $\mathcal{L}_{\mathrm{VCLUB}}\left(P_{\mathbf{Z}|\mathbf{X}}, Q_{\mathbf{U}|\mathbf{Z}}, Q_\mathbf{Z}, Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}, \beta, \alpha\right)$ over the variational distributions. We refer the reader to the expanded version of this manuscript for the derivations and interpretations of the information quantities.

The common approach is to use neural networks to parameterize variational inference bounds. To this goal, we parameterize the encoding distribution $P_{\mathbf{Z}|\mathbf{X}}$, utility decoding distribution $Q_{\mathbf{U}|\mathbf{Z}}$, uncertainty decoding distribution $Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}$, and prior $Q_\mathbf{Z}$. Let $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X})$ denote the family of encoding probability distributions $P_{\mathbf{Z}|\mathbf{X}}$ over $\mathcal{Z}$ for each element of space $\mathcal{X}$, parameterized by the output of a deep neural network $f_{\boldsymbol{\phi}}$ with parameters $\boldsymbol{\phi}$. In the context of inference problems, $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X})$ is called the amortized variational inference (AVI)

[1] Note that one can generalize this formulation and consider Arimoto's mutual information [53] and/or f-information [54] in the CLUB model (4). For instance, $\mathrm{I}_f(\mathbf{X};\mathbf{Z}) = \mathrm{D}_f\left(P_{\mathbf{X},\mathbf{Z}} \| P_\mathbf{X} P_\mathbf{Z}\right)$ and $D_f(\cdot\|\cdot)$ is f-divergence [52]. For the sake of brevity, we consider Kullback–Leibler divergence, a special case of the family of f-divergences between probability distributions associated with $f(t) = t \log t$. Moreover, note that although all $\mathrm{D}_f(\cdot\|\cdot)$ quantify the dissimilarity between a pair of distributions, their operational meanings are different.

[2] The notion of 'information complexity' inspired by the concept of *stochastic complexity* of the data relative to a model [55], which can be interpreted as the shortest code-length of the data given a model. We will elaborate on this notion in Sec. 4

[3] Note that by the DPI [52], $\mathrm{I}(\mathbf{U};\mathbf{Z}) \leq \mathrm{I}(\mathbf{X};\mathbf{Z})$ and $\mathrm{I}(\mathbf{S};\mathbf{Z}) \leq \mathrm{I}(\mathbf{X};\mathbf{Z})$. Hence, for $\beta, \alpha > 1$, based on the considered Lagrangian functional the model may learn a trivial representation, independent of $\mathbf{X}$.

distribution or variational posterior distribution. Analogously, let $P_{\boldsymbol{\theta}}(\mathbf{U} \mid \mathbf{Z})$ and $P_{\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})$ denote the corresponding family of decoding probability distributions $Q_{\mathbf{U}|\mathbf{Z}}$ and $Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}$, respectively, parameterized by the output of the deep neural networks $g_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\varphi}}$. Moreover, for the prior distribution $Q_{\mathbf{Z}}$ we consider the family of distributions $Q_{\boldsymbol{\psi}}(\mathbf{Z})$, which can be interpreted as the target (proposal) distribution in the latent space. The choice for these distributions is considered by trading off computational complexity with model expressiveness. Let $P_{\mathsf{D}}(\mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N}\delta(\mathbf{x}-\mathbf{x}_n)$, $\mathbf{x}_n \in \mathcal{X}$, denote the empirical data distribution. In this case, $P_{\boldsymbol{\phi}}(\mathbf{X},\mathbf{Z}) = P_{\mathsf{D}}(\mathbf{X})P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})$ denotes our joint inference data distribution, and $P_{\boldsymbol{\phi}}(\mathbf{Z}) = \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})\right]$ denotes the learned *aggregated* posterior distribution over latent space $\mathcal{Z}$.

Parameterized Information Complexity: The parameterized variational approximation of *information complexity* can be defined as:

$$\mathrm{I}_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) \coloneqq \mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}) \mid P_{\mathsf{D}}(\mathbf{X})) - \mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}))\,. \quad (7)$$

Indeed, the information complexity $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$ measures the amount of Shannon's mutual information between the parameters of the model and the dataset D, given a prior $Q_{\boldsymbol{\psi}}(\mathbf{Z})$ and stochastic map $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X}) : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$. Note that the posterior distribution $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X})$ depends on the choice of the optimization algorithm, therefore, the information complexity implicitly depends on this choice. The relative entropy $\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}))$ is usually ignored in the literature. A critical challenge is to guarantee that the learned aggregated posterior distribution $P_{\boldsymbol{\phi}}(\mathbf{Z})$ conforms well to the proposed prior $Q_{\boldsymbol{\psi}}(\mathbf{Z})$ [60]–[62]. We can tackle this issue by employing a more *expressive* form for $Q_{\boldsymbol{\psi}}(\mathbf{Z})$, which would allow us to provide a good fit for an arbitrary space $\mathcal{Z}$, at the expense of additional *computational complexity*. Note that by imposing a constraint on the information complexity $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$, we are imposing a constraint on the entropy of the AVI distribution $P_{\boldsymbol{\phi}}(\mathbf{Z})$.

Parameterized Information Utility:

We can decompose the mutual information between the released representation $\mathbf{Z}$ and the *utility attribute* $\mathbf{U}$ in two analytically equivalent ways:

$$\mathrm{I}(\mathbf{U};\mathbf{Z}) = \underbrace{\mathrm{H}(\mathbf{U}) - \mathrm{H}(\mathbf{U} \mid \mathbf{Z})}_{\text{Discriminative View}} = \underbrace{\mathrm{H}(\mathbf{Z}) - \mathrm{H}(\mathbf{Z} \mid \mathbf{U})}_{\text{Generative View}}. \quad (8)$$

Therefore, maximizing $\mathrm{I}(\mathbf{U};\mathbf{Z})$ can have two different interpretations. The discriminative view expresses that (i) the utility attribute needs to be distributed as uniformly as possible in the data space[4], and (ii) the utility attribute should be confidently inferred from the released representation $\mathbf{Z}$. On the other hand, the generative view expresses that (i) the released representations should be spread as much as possible in the latent space (i.e., high entropy $\mathrm{H}(\mathbf{Z})$), and (ii) the released representation corresponding to the same utility attribute should be close together (i.e., minimizing conditional entropy $\mathrm{H}(\mathbf{Z} \mid \mathbf{U})$). Since our goal is to identify the utility attributes based on the revealed representation, the discriminative view of this decomposition is more aligned with our model.

[4]We assumed our attributes are supported on a finite space.

The parameterized variational approximation associated to the information utility bound can be defined as:

$$\begin{aligned}
&\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{U};\mathbf{Z}) \\
&\coloneqq \mathbb{E}_{P_{\mathbf{U},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log \frac{P_{\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z})}{P_{\mathbf{U}}} \cdot \frac{P_{\boldsymbol{\theta}}(\mathbf{U})}{P_{\boldsymbol{\theta}}(\mathbf{U})}\right]\right] && \text{(9a)}\\
&= \mathbb{E}_{P_{\mathbf{U},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z})\right]\right] - \mathbb{E}_{P_{\mathbf{U}}}\left[\log \frac{P_{\mathbf{U}}}{P_{\boldsymbol{\theta}}(\mathbf{U})}\right] \\
&\qquad - \mathbb{E}_{P_{\mathbf{U}}}\left[\log P_{\boldsymbol{\theta}}(\mathbf{U})\right] && \text{(9b)}\\
&= -\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z}) - \mathrm{D}_{\mathrm{KL}}(P_{\mathbf{U}}\|P_{\boldsymbol{\theta}}(\mathbf{U})) + \mathrm{H}(P_{\mathbf{U}}\|P_{\boldsymbol{\theta}}(\mathbf{U})) && \text{(9c)}\\
&\geq \underbrace{-\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z})}_{\text{Prediction Fidelity}} - \underbrace{\mathrm{D}_{\mathrm{KL}}(P_{\mathbf{U}} \,\|\, P_{\boldsymbol{\theta}}(\mathbf{U}))}_{\text{Distribution Discrepancy Loss}} =: \mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{U};\mathbf{Z}), && \text{(9d)}
\end{aligned}$$

where $\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z}) \coloneqq \mathrm{H}\left(P_{\mathbf{U}|\mathbf{Z}} \,\|\, P_{\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z}) \mid P_{\boldsymbol{\phi}}(\mathbf{Z})\right)$, and the inequality follows by noticing that $\mathrm{H}(P_{\mathbf{U}}\|P_{\boldsymbol{\theta}}(\mathbf{U})) \geq 0$. This decomposition will lead us to a *unified* formulation for both the *supervised* and *unsupervised* DVCLUB objectives.

Let us consider a scenario in which the data owner wishes to release the original domain data $\mathbf{X}$ (e.g., facial images) as accurately as possible (i.e., $\mathbf{U} \equiv \mathbf{X}$), without revealing a specific sensitive attribute $\mathbf{S}$ (e.g., gender, emotion, etc.). We call this setup as the *unsupervised* CLUB model and the formerly described general case, where $\mathbf{U}$ is a generic attribute of data $\mathbf{X}$ as the *supervised* CLUB model. When the utility task is to *reconstruct* (generate) the original data $\mathbf{X}$, i.e., $P_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z}) = P_{\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z})$, let us denote the generated data distribution as $P_{\boldsymbol{\theta}}(\mathbf{X}) \coloneqq \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[P_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\right]$. The parameterized variational approximation associated to the information utility $\mathrm{I}(\mathbf{U};\mathbf{Z}) = \mathrm{I}(\mathbf{X};\mathbf{Z})$ can be defined as:

$$\begin{aligned}
&\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z}) \\
&\coloneqq -\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z}) - \mathrm{D}_{\mathrm{KL}}(P_{\mathsf{D}}(\mathbf{X})\|P_{\boldsymbol{\theta}}(\mathbf{X})) + \mathrm{H}(P_{\mathsf{D}}(\mathbf{X})\|P_{\boldsymbol{\theta}}(\mathbf{X})) && \text{(10a)}\\
&\geq \underbrace{-\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})}_{\text{Reconstruction Fidelity}} \underbrace{-\mathrm{D}_{\mathrm{KL}}(P_{\mathsf{D}}(\mathbf{X}) \,\|\, P_{\boldsymbol{\theta}}(\mathbf{X}))}_{\text{Distribution Discrepancy Loss}} =: \mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z}), && \text{(10b)}
\end{aligned}$$

where $\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z}) \coloneqq \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\right]\right] = \mathrm{H}\left(P_{\mathbf{X}|\mathbf{Z}} \,\|\, P_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z}) \mid P_{\boldsymbol{\phi}}(\mathbf{Z})\right)$.

Parameterized Information Leakage: Let $P_{\boldsymbol{\xi}}(\mathbf{S} \mid \mathbf{Z})$ denote the corresponding family of decoding probability distribution $Q_{\mathbf{S}|\mathbf{Z}}$, where $Q_{\mathbf{S}|\mathbf{Z}} : \mathcal{Z} \to \mathcal{P}(\mathcal{S})$ is a variational approximation of optimal decoder distribution $P_{\mathbf{S}|\mathbf{Z}}$. Similarly to (9), one can recast the parameterized variational approximation associated with the information leakage:

$$\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z}) \geq -\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S}|\mathbf{Z}) - \mathrm{D}_{\mathrm{KL}}(P_{\mathbf{S}}\|P_{\boldsymbol{\xi}}(\mathbf{S})) =: \mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z}). \quad (11)$$

However, we want to minimize $\mathrm{I}(\mathbf{S};\mathbf{Z})$; therefore, we instead need a variational upper bound. Alternatively, we can express $\mathrm{I}(\mathbf{S};\mathbf{Z})$ as:

$$\begin{aligned}
\mathrm{I}(\mathbf{S};\mathbf{Z}) &= \mathrm{I}(\mathbf{X};\mathbf{Z}) - \mathrm{I}(\mathbf{X};\mathbf{Z} \mid \mathbf{S}) \\
&= \mathrm{I}(\mathbf{X};\mathbf{Z}) - \mathrm{H}(\mathbf{X} \mid \mathbf{S}) + \mathrm{H}(\mathbf{X} \mid \mathbf{S},\mathbf{Z}). \quad (12)
\end{aligned}$$

Eqn. (12) depicts the critical role of information complexity $\mathrm{I}(\mathbf{X};\mathbf{Z})$ in controlling the information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$. The less the information complexity, the less distinguishable the revealed representations, and

hence the less the information leakage. Considering $\mathrm{D_{KL}}\left(P_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\|Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\right) \geq 0$, we can obtain an upper bound on $\mathrm{I}(\mathbf{S};\mathbf{Z})$ as $\mathrm{I}(\mathbf{S};\mathbf{Z}) \leq \mathrm{D_{KL}}\left(P_{\mathbf{Z}|\mathbf{X}}\|Q_{\mathbf{Z}} \mid P_{\mathbf{X}}\right) + \mathrm{H}(\mathbf{X} \mid \mathbf{S}) + \mathrm{H}\left(P_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\|Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}} \mid P_{\mathbf{S},\mathbf{Z}}\right)$, where we have $\mathrm{H}\left(P_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\|Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}} \mid P_{\mathbf{S},\mathbf{Z}}\right) := \mathbb{E}_{P_{\mathbf{S},\mathbf{X},\mathbf{Z}}}\left[\log Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\right]$. The term $\mathrm{H}\left(P_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\|Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}} \mid P_{\mathbf{S},\mathbf{Z}}\right)$ can be interpreted as follows. Consider the inference threat model presented in Sec. 2.2. The inferential adversary has access to the revealed representation $\mathbf{Z}$ and is interested in inferring the sensitive attribute $\mathbf{S}$. Note that the adversary's inference is through reconstructing the original data $\mathbf{X}$. Therefore, after observing $\mathbf{Z}$, the adversary chooses a belief distribution over $\mathcal{S}$ and then tries to reconstruct the original data $\mathbf{X}$ to revise his belief. If the released representation is statistically independent of the sensitive attribute $\mathbf{S}$, then the adversary cannot revise his belief by injecting various $\mathbf{S}$ to his inferential model. Hence, by maximizing the average log-likelihood $\log Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}$ over $P_{\mathbf{Z}|\mathbf{X}}$, the defender minimizes the average adversarial inference about $\mathbf{S}$.

The parameterized variational approximation of conditional entropy $\mathrm{H}(\mathbf{X}|\mathbf{S},\mathbf{Z})$ can be defined as:

$$\begin{aligned}\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z}) :=& -\mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})\right]\right] \\ &- \mathrm{D_{KL}}\left(P_{\mathbf{X}|\mathbf{S},\mathbf{Z}} \,\|\, P_{\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})\right) \quad (13a)\\ \leq& -\mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})\right]\right] \quad (13b)\\ =:& \ \mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z}). \quad (13c)\end{aligned}$$

Using (12) and (13), the upper bound on $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z})$ is given as:

$$\begin{aligned}\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z}) \ &\leq \ \mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z}) + \mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z}) + \mathrm{c} \quad (14a)\\ &=: \ \mathrm{I}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\varphi}}(\mathbf{S};\mathbf{Z}) + \mathrm{c}, \quad (14b)\end{aligned}$$

where c is a constant term, independent of the neural network parameters. The upper bound in (14) encourages the model to directly minimize the information complexity $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$ as well as the information uncertainty $\mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})$. By minimizing the information uncertainty $\mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})$ the model forces to forget the sensitive attribute $\mathbf{S}$ at the expense of reducing the uncertainty about the original data $\mathbf{X}$, i.e., encourages the model to reconstruct the original data $\mathbf{X}$.

Considering (6), and using the addressed parameterized approximations, we have:

$$\begin{aligned}\max_{P_{\mathbf{Z}|\mathbf{X}}} \max_{Q_{\mathbf{U}|\mathbf{Z}},Q_{\mathbf{Z}},Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}} &\mathcal{L}^{\mathsf{S/U}}_{\mathrm{VCLUB}}\left(P_{\mathbf{Z}|\mathbf{X}},Q_{\mathbf{U}|\mathbf{Z}},Q_{\mathbf{Z}},Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}},\beta,\alpha\right)\\ &\geq \max_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\varphi}} \mathcal{L}^{\mathsf{S/U}}_{\mathrm{DVCLUB}}\left(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\varphi},\beta,\alpha\right), \quad (15)\end{aligned}$$

where $\mathcal{L}^{\mathsf{S/U}}_{\mathrm{DVCLUB}}(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\varphi},\beta,\alpha)$ denotes the associated *Deep Variational CLUB (DVCLUB)* Lagrangian functional in the '*supervised*' or '*unsupervised*' scenarios, which are given in (16) and (17), respectively. Fig. 3 demonstrates the training architecture associated with problems (P1:S) and (P1:U). We will discuss the alternative objectives in the expanded version of this manuscript. In practice, we need to train the model using alternating block coordinate descent algorithms. In the following, we gradually build the fundamentals of the learning model.

4.1. Learning Algorithm

Given a collection of i.i.d. training samples $\{(\mathbf{u}_n,\mathbf{s}_n,\mathbf{x}_n)\}_{n=1}^{N} \subseteq \mathcal{U}\times\mathcal{S}\times\mathcal{X}$, and using the stochastic gradient descent (SGD)-type algorithms, the deep neural networks $f_{\boldsymbol{\phi}}$, $g_{\boldsymbol{\theta}}$, and $g_{\boldsymbol{\varphi}}$ (or $g_{\boldsymbol{\xi}}$) can be trained jointly to maximize a Monte-Carlo approximation of the DVCLUB functionals over parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\xi}$. In order to have a stable gradient with respect to the encoder, the reparameterization trick [8] is used to sample from the *learned posterior* distribution $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X})$. To do this, we need to *explicitly* consider $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X})$ to belong to a tractable parametric family of distributions (e.g., Gaussian distributions) such that we are able to sample from $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X})$ by: (i) sampling a random vector $\boldsymbol{\mathcal{E}}$ with distribution $P_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\varepsilon})$, $\boldsymbol{\varepsilon} \in \mathcal{E}$, which does not depend on $\boldsymbol{\phi}$[5]; (ii) transforming the samples using some parametric function $f_{\boldsymbol{\phi}} : \mathcal{X}\times\mathcal{E} \to \mathcal{Z}$, such that $\mathbf{Z} = f_{\boldsymbol{\phi}}(\mathbf{x},\boldsymbol{\mathcal{E}}) \sim P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X}=\mathbf{x})$. One can consider multivariate Gaussian parametric encoders of mean $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$, and co-variance $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\mathbf{x})$, i.e., $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}),\boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\mathbf{x}))$, where $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\mathbf{x})$

[5]Hence, does not impact differentiation of the network.

$$\begin{aligned}(\mathrm{P1{:}S}):\ \mathcal{L}^{\mathsf{S}}_{\mathrm{DVCLUB}}(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\varphi},\beta,\alpha) :=& \overbrace{\mathbb{E}_{P_{\mathbf{U},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{U}|\mathbf{Z})\right]\right] - \mathrm{D_{KL}}\left(P_{\mathbf{U}} \,\|\, P_{\boldsymbol{\theta}}(\mathbf{U})\right)}^{\text{Information Utility: } \mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{U};\mathbf{Z})}\\ &- (\beta+\alpha)\underbrace{\Big(\mathrm{D_{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}) \mid P_{\mathsf{D}}(\mathbf{X})) - \mathrm{D_{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}))\Big)}_{\text{Information Complexity: } \mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})}\\ &- \alpha\underbrace{\Big(-\mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})\right]\right]\Big)}_{\text{Information Uncertainty: } \mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})}. \quad (16)\end{aligned}$$

$$\begin{aligned}(\mathrm{P1{:}U}):\ \mathcal{L}^{\mathsf{U}}_{\mathrm{DVCLUB}}(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\varphi},\beta,\alpha) :=& \overbrace{\mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\right]\right] - \mathrm{D_{KL}}\left(P_{\mathsf{D}}(\mathbf{X}) \,\|\, P_{\boldsymbol{\theta}}(\mathbf{X})\right)}^{\text{Information Utility: } \mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z})}\\ &- (\beta+\alpha)\underbrace{\Big(\mathrm{D_{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}) \mid P_{\mathsf{D}}(\mathbf{X})) - \mathrm{D_{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}))\Big)}_{\text{Information Complexity: } \mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})}\\ &- \alpha\underbrace{\Big(-\mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})\right]\right]\Big)}_{\text{Information Uncertainty: } \mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X}|\mathbf{S},\mathbf{Z})}. \quad (17)\end{aligned}$$

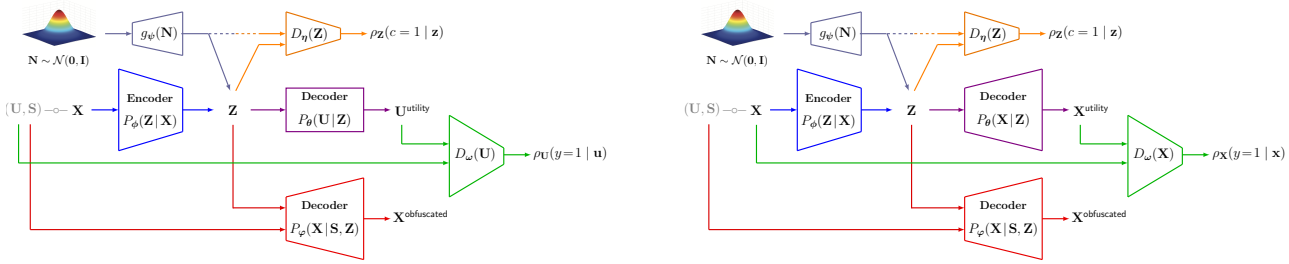


Fig. 3: DVCLUB training architecture associated with (P1): (a) supervised setup; (b) unsupervised setup.

are the two output vectors of the network $f_{\boldsymbol{\phi}}$ for the given input sample $\mathbf{x}$. The inferred posterior distribution is typically a multi-variate Gaussian with diagonal co-variance, i.e., $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{x}) = \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}))\big)$. Suppose $\mathcal{Z} = \mathbb{R}^{d_{\mathbf{z}}}$, therefore, we first sample a random variable $\boldsymbol{\mathcal{E}}$ i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\mathbf{z}}})$, then given a data sample $\mathbf{x} \in \mathcal{X}$, we generate the sample $\mathbf{z} = \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) + \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}) \odot \boldsymbol{\varepsilon}$, where $\odot$ is the element-wise (Hadamard) product. Noting that $f_{\boldsymbol{\phi}}$ is a deterministic mapping, the stochasticity of encoder $P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})$ is relegated to the random variable $\boldsymbol{\mathcal{E}}$. The decoder $P_{\boldsymbol{\theta}}(\mathbf{U}\,|\,\mathbf{Z})$ (likewise $P_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})$) utilizes a suitable distribution for the data and task under consideration.

The latent space prior distribution is typically considered as a *fixed* d-dimensional standard isotropic multi-variate Gaussian, i.e., $Q_{\boldsymbol{\psi}}(\mathbf{Z}) = Q_{\mathbf{Z}} = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. For this simple *explicit* choice, the information complexity upper bound $\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{X},\mathbf{Z})}\big[\log \frac{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}{Q_{\mathbf{Z}}}\big] = \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}) \,\|Q_{\mathbf{Z}})\right]$ has a closed-form expression for a given sample $\mathbf{x}$, which reads as $2\,\mathrm{D}_{\mathrm{KL}}\left(P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}=\mathbf{x}) \,\|Q_{\mathbf{Z}}\right) = \|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})\|_2^2 + d + \sum_{i=1}^{d}(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})_i - \log \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})_i)$. However, simple prior can lead to under-fitting and, as a consequence, poor representations. On the other hand, having a complete match, i.e., $Q_{\mathbf{Z}} = \frac{1}{N}\sum_{n=1}^{N} P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{x}_n)$, may potentially lead to over-fitting. Moreover, it is a computationally expensive task due to the summation over all training samples. One possible solution to overcome this issue is to explicitly consider a mixture of diagonal Gaussian distributions as the proposal prior, i.e., $Q_{\boldsymbol{\psi}}(\mathbf{Z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{\psi}_k}, \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\psi}_k})\big)$, $\sum_{k=1}^{K}\pi_k = 1$, where $K \ll N$, and learn the mixture weights. In [61], the authors considered the proposal prior as a mixture of equi-probable variational posteriors, such that $Q_{\boldsymbol{\psi}}(\mathbf{Z}) = \frac{1}{K}\sum_{k=1}^{K} P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \widetilde{\mathbf{x}}_k)$, where $\{\widetilde{\mathbf{x}}_k\}_{k=1}^{K}$, $K \ll N$, are referred to as the pseudo-inputs, which are learned through back-propagation. In the same line of research, in [62], the authors constructed the proposal prior by multiplying a simple prior with a learned acceptance probability function, which re-weights the considered simple prior. Alternatively, one can adversarially learn the prior distribution $Q_{\boldsymbol{\psi}}(\mathbf{Z})$ through a generator model $g_{\boldsymbol{\psi}}(\mathbf{N})$, where $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This choice gives us an *implicit* prior distribution $Q_{\boldsymbol{\psi}}(\mathbf{Z})$. Implicit distributions are probability distributions that are learned via passing noise through a deterministic function, which is parameterized by a neural network. This allows us to easily sample from them and take derivatives of samples with respect to the model parameters.

Divergence Estimation: We can estimate the KL-divergences in (16) and (17) using the *density-ratio trick* [63], [64], utilized in the GAN framework to directly match the data distribution $P_{\mathbf{X}}$ and the marginal model distribution $P_{\boldsymbol{\theta}}(\mathbf{X})$. The trick is to express two distributions as conditional distributions, conditioned on a label $C \in \{0, 1\}$, and reduce the task to binary classification. The key point is that we can estimate the KL-divergence, and indeed all the well-defined f-divergences, by estimating the *ratio* of two distributions without modeling each distribution explicitly. Consider $\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z})) = \mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z})}\Big[\log \frac{P_{\boldsymbol{\phi}}(\mathbf{Z})}{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\Big]$. Define $\rho_{\mathbf{Z}}(\mathbf{z} \mid c)$ as follows:

$$\rho_{\mathbf{Z}}(\mathbf{z}\,|\,c) = \begin{cases} P_{\boldsymbol{\phi}}(\mathbf{Z}) & , \text{ if } \quad c = 1 \\ Q_{\boldsymbol{\psi}}(\mathbf{Z}) & , \text{ if } \quad c = 0 \end{cases} . \tag{18}$$

Suppose that a perfect binary classifier (discriminator) $D_{\boldsymbol{\eta}}(\mathbf{z})$, with parameters $\boldsymbol{\eta}$, is trained to associate label $c = 1$ to samples from distribution $P_{\boldsymbol{\phi}}(\mathbf{Z})$ and label $c = 0$ to samples from $Q_{\boldsymbol{\psi}}(\mathbf{Z})$. Using the Bayes' rule and assuming that the marginal class probabilities are equal, i.e., $\rho(c = 1) = \rho(c = 0)$, the density ratio can be expressed as:

$$\frac{P_{\boldsymbol{\phi}}(\mathbf{Z} = \mathbf{z})}{Q_{\boldsymbol{\psi}}(\mathbf{Z} = \mathbf{z})} = \frac{\rho_{\mathbf{Z}}(\mathbf{z} \mid c = 1)}{\rho_{\mathbf{Z}}(\mathbf{z} \mid c = 0)} = \frac{\rho_{\mathbf{Z}}(c = 1 \mid \mathbf{z})}{\rho_{\mathbf{Z}}(c = 0 \mid \mathbf{z})} \approx \frac{D_{\boldsymbol{\eta}}(\mathbf{z})}{1 - D_{\boldsymbol{\eta}}(\mathbf{z})}.$$

Therefore, given a trained discriminator $D_{\boldsymbol{\eta}}(\mathbf{z})$ and M i.i.d. samples $\{\mathbf{z}_m\}_{m=1}^{M}$ from $P_{\boldsymbol{\phi}}(\mathbf{Z})$, one can estimate the divergence $\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}))$ as:

$$\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z})) \approx \frac{1}{M}\sum_{m=1}^{M} \log \frac{D_{\boldsymbol{\eta}}(\mathbf{z}_m)}{1 - D_{\boldsymbol{\eta}}(\mathbf{z}_m)}. \tag{19}$$

The density ratio trick opens the door to *implicit* prior and posterior distributions. The implicit generative models provide likelihood-free inference models. Interestingly, this trick allows us to *learn* the parameterized prior distribution $Q_{\boldsymbol{\psi}}(\mathbf{Z})$ through a generator model.

Given the discriminator (parameterized scoring function) $D_{\boldsymbol{\eta}}(\mathbf{z}) \approx \rho_{\mathbf{Z}}(c = 1 \mid \mathbf{z}) = P_{\boldsymbol{\phi}}(\mathbf{Z} = \mathbf{z})$, we now need to specify a proper scoring rule for binary discrimination to allow parameter learning. Binary cross-entropy loss is typically considered to this end. In this case, the *latent space discriminator* $D_{\boldsymbol{\eta}}(\mathbf{z})$ minimizes the following loss function:

$$\begin{aligned} \mathcal{L}^{\mathbf{z}}_{\mathrm{disc}}(\boldsymbol{\eta}, \boldsymbol{\phi}) &:= \mathbb{E}_{\rho_{\mathbf{Z}}(\mathbf{z}|c)\rho(c)}\left[-c\,\log D_{\boldsymbol{\eta}}(\mathbf{Z})\right. \\ &\qquad \left. -(1-c)\,\log(1 - D_{\boldsymbol{\eta}}(\mathbf{Z}))\right], \quad \text{(20a)} \\ &= \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log D_{\boldsymbol{\eta}}(\mathbf{Z})\right]\right] \\ &\qquad + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log(1 - D_{\boldsymbol{\eta}}(\mathbf{Z}))\right]. \quad \text{(20b)} \end{aligned}$$

Analogously, we can estimate the KL-divergence $\mathrm{D}_{\mathrm{KL}}\left(P_{\mathrm{D}}(\mathbf{X}) \,\|\, P_{\boldsymbol{\theta}}(\mathbf{X})\right)$, the discrepancy measure in the visible (perceptual) space in the unsupervised DVCLUB

functional (17). Let us introduce a random variable y, and assign a label $y = 1$ to samples drawn from $P_{\mathsf{D}}(\mathbf{X})$ and $y = 0$ to samples drawn from $P_{\boldsymbol{\theta}}(\mathbf{X})$. Also, consider the discriminator $D_{\boldsymbol{\omega}}(\mathbf{x}) \approx \rho_{\mathbf{X}}(y = 1 \mid \mathbf{x}) = P_{\mathsf{D}}(\mathbf{X} = \mathbf{x})$, with parameters $\boldsymbol{\omega}$. Given a trained discriminator $D_{\boldsymbol{\omega}}(\mathbf{x})$ and M i.i.d. samples $\{\mathbf{x}_m\}_{m=1}^{M}$ from $P_{\mathsf{D}}(\mathbf{X})$, we have:

$$\mathrm{D}_{\mathrm{KL}}\left(P_{\mathsf{D}}(\mathbf{X}) \,\|\, P_{\boldsymbol{\theta}}(\mathbf{X})\right) \approx \frac{1}{M}\sum_{m=1}^{M} \log \frac{D_{\boldsymbol{\omega}}(\mathbf{x}_m)}{1 - D_{\boldsymbol{\omega}}(\mathbf{x}_m)}. \quad (21)$$

In this case, the *visible space discriminator* $D_{\boldsymbol{\omega}}(\mathbf{x})$ minimizes the following loss function:

$$\begin{aligned}
&\mathcal{L}^{\mathsf{x}}_{\mathrm{disc}}(\boldsymbol{\omega}, \boldsymbol{\theta}) \quad (22)\\
&\coloneqq \mathbb{E}_{\rho_{\mathbf{X}}(\mathbf{x}|y)\rho(y)}\left[-y \log D_{\boldsymbol{\omega}}(\mathbf{X}) \; -(1-y)\log(1 - D_{\boldsymbol{\omega}}(\mathbf{X}))\right]\\
&= \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[-\log D_{\boldsymbol{\omega}}(\mathbf{X})\right] + \mathbb{E}_{P_{\boldsymbol{\theta}}(\mathbf{X})}\left[-\log(1 - D_{\boldsymbol{\omega}}(\mathbf{X}))\right]\\
&= \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[-\log D_{\boldsymbol{\omega}}(\mathbf{X})\right] + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log\left(1 - D_{\boldsymbol{\omega}}(g_{\boldsymbol{\theta}}(\mathbf{Z}))\right)\right],
\end{aligned}$$

or equivalently, maximizes $\mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\log D_{\boldsymbol{\omega}}(\mathbf{X})\right] + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[\log\left(1 - D_{\boldsymbol{\omega}}(g_{\boldsymbol{\theta}}(\mathbf{Z}))\right)\right]$.

Learning Procedure: The DVCLUB models (P1: S) and (P1: U) are trained using alternating block coordinate descent across five steps:

(1) Train the Encoder, Utility Decoder and Uncertainty Decoder.

- Supervised Setup:

$$\begin{aligned}
\max_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\varphi}} \quad & \mathbb{E}_{P_{\mathbf{U},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[P_{\boldsymbol{\theta}}(\mathbf{U}\,|\,\mathbf{Z})\right]\right]\\
& - (\beta + \alpha)\;\; \mathrm{D}_{\mathrm{KL}}\left(P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X}) \| Q_{\boldsymbol{\psi}}(\mathbf{Z}) \,|\, P_{\mathsf{D}}(\mathbf{X})\right)\\
& - \alpha \;\; \mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log P_{\boldsymbol{\varphi}}(\mathbf{X}\,|\,\mathbf{S},\mathbf{Z})\right]\right]. \quad (23)
\end{aligned}$$

- Unsupervised Setup:

$$\begin{aligned}
\max_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\varphi}} \quad & \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right]\\
& - (\beta + \alpha)\;\; \mathrm{D}_{\mathrm{KL}}\left(P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X}) \| Q_{\boldsymbol{\psi}}(\mathbf{Z}) \,|\, P_{\mathsf{D}}(\mathbf{X})\right)\\
& - \alpha \;\; \mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log P_{\boldsymbol{\varphi}}(\mathbf{X}\,|\,\mathbf{S},\mathbf{Z})\right]\right]. \quad (24)
\end{aligned}$$

(2) Train the Latent Space Discriminator.

$$\begin{aligned}
\min_{\boldsymbol{\eta}} \quad & \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log D_{\boldsymbol{\eta}}(\mathbf{Z})\right]\right]\\
& + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log(1 - D_{\boldsymbol{\eta}}(\mathbf{Z}))\right]. \quad (25)
\end{aligned}$$

(3) Train the Encoder and Prior Distribution Generator Adversarially.

$$\begin{aligned}
\max_{\boldsymbol{\phi},\boldsymbol{\psi}} \quad & \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log D_{\boldsymbol{\eta}}(\mathbf{Z})\right]\right]\\
& + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log(1 - D_{\boldsymbol{\eta}}(\mathbf{Z}))\right]. \quad (26)
\end{aligned}$$

(4) Train the Output Space Discriminator.

- Supervised Scenario: the *Attribute Class Discriminator* $D_{\omega}(\mathbf{U})$ is updated as:

$$\begin{aligned}
\min_{\boldsymbol{\omega}} \quad & \mathbb{E}_{P_{\mathbf{U}}}\left[-\log D_{\boldsymbol{\omega}}(\mathbf{U})\right]\\
& + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log\left(1 - D_{\boldsymbol{\omega}}(g_{\boldsymbol{\theta}}(\mathbf{Z}))\right)\right]. \quad (27)
\end{aligned}$$

- Unsupervised Scenario: the *Visible Space Discriminator* $D_{\omega}(\mathbf{X})$ is updated as:

$$\begin{aligned}
\min_{\boldsymbol{\omega}} \quad & \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[-\log D_{\boldsymbol{\omega}}(\mathbf{X})\right]\\
& + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log\left(1 - D_{\boldsymbol{\omega}}(g_{\boldsymbol{\theta}}(\mathbf{Z}))\right)\right]. \quad (28)
\end{aligned}$$

(5) Train the Prior Distribution Generator and Utility Decoder Adversarially.

$$\max_{\boldsymbol{\psi},\boldsymbol{\theta}} \quad \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log\left(1 - D_{\boldsymbol{\omega}}(g_{\boldsymbol{\theta}}(\mathbf{Z}))\right)\right]. \quad (29)$$

The complete training algorithm of the supervised DVCLUB model is shown in the Algorithm 1. The iterative alternating block coordinate descent algorithm associated with the unsupervised DVCLUB, the alternative DVCLUB algorithms, as well as network architectures are provided in the expanded version of this manuscript.

5. Experiments

We conduct experiments on the large-scale colored-MNIST and CelebA datasets. The Colored-MNIST is *our modified* version of the MNIST [65] data-set, which is a collection of $70,000$ 'colored' digits of size 28×28. The digits are randomly colored into Red, Green, and Blue based on the *uniform* and *non-uniform (biased)* distributions. The CelebA [66] dataset contains around $200,000$ colored images of size 128×128. We used TensorFlow 2.3 [67] with Integrated Keras API to implement and train the proposed DVCLUB models.

5.1. Colored-MNIST Experiments

The experiments on colored-MNIST dataset are depicted in Fig. 4 and Fig. 5. In these experiments the digits are randomly colored with the probabilities $P_S(\mathsf{Red}) = \frac{1}{2}$, $P_S(\mathsf{Green}) = \frac{1}{6}$, $P_S(\mathsf{Blue}) = \frac{1}{3}$, and set $Q_{\boldsymbol{\psi}}(\mathbf{Z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The complete results on utility attribute accuracy, estimated information utility $\mathrm{I}(\mathbf{U};\mathbf{Z})$, estimated information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$, MSE of reconstructed $\mathbf{U} \equiv \mathbf{X}$, as well as various qualitative evaluation setups for sample quality are provided in the expanded version of this manuscript.

Qualitative evaluation for sample quality of colored-MNIST at the inferential adversary for two CLUB models, for different trade-offs between β and α, are depicted in Fig. 4. Five scenarios are considered to investigate the adversary's beliefs about sensitive attribute $\mathbf{S}$. $\mathbf{S} = \emptyset$ corresponds to the case in which the adversary recovers $\mathbf{X}$ without any assumption on the digit color; $\mathbf{S} = [1, 1, 1]$ corresponds to a possible, probably meaningless, adversary's belief about sensitive attribute to infer $\mathbf{S}$ from reconstructed $\mathbf{X}$; $\mathbf{S} = [1, 0, 0]$, $\mathbf{S} = [0, 1, 0]$, and $\mathbf{S} = [0, 0, 1]$ correspond to different beliefs about the sensitive attribute, i.e., the digit color, associated with Red, Green and Blue colors, respectively. The results show that the adversary cannot revise his belief about $\mathbf{S}$. It means that the DVCLUB model learned representation $\mathbf{Z}$ independent to $\mathbf{S}$.

Fig. 5 depicts the information utility and information leakage curves for different trade-offs between β and α, which are estimated using MINE [68]. We observe that as β increases, i.e., the information complexity is reduced, both the information utility and information leakage starts to reduce, especially for small values of α. This behavior is consistent with our expectations, as α contributes both to the information complexity $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$ and information uncertainty $\mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})$. We also see that the information leakage is further reduced when the dimension of the released representation $\mathbf{Z}$, i.e., d_{z}, is reduced. This forces the data owner to obtain a more succinct representation of the utility variable, removing any extra information. Note that the digit color is independent of the digit number. Therefore, as expected, we have an almost flat curve for the information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$, invariant with respect to β, when the sensitive attribute $\mathbf{S}$ is the digit color. That is, the

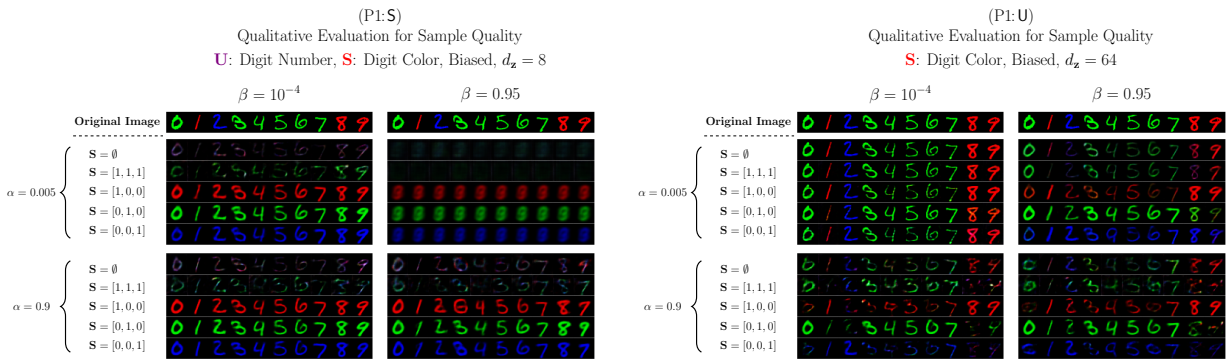


Fig. 4: Qualitative evaluation for sample quality at the inferential adversary for CLUB models (P1: S) (Left Panel) and (P1: U) (Right Panel) on Colored-MNIST, for different trade-offs between β and α.

model aims to learn representation $\mathbf{Z}$ that captures relevant information about $\mathbf{U}$ (digit number), as the information about digit color is independent of $\mathbf{U}$.

5.2. *CelebA Experiments*

The experiments on CelebA dataset are depicted in Fig. 6, Fig. 7, and Fig. 8. We provide utility accuracy curves of the test set of CelebA for (P1: S) and (P1: U) in Fig. 6, where in the supervised scenario it corresponds to recognition accuracy curve of the utility attribute $\mathbf{U}$, while in the unsupervised scenario it corresponds to MSE curve of the reconstructed utility data, i.e., $\mathbf{U} = \mathbf{X}$. The results show that by increasing the information complexity weight β, i.e., by reducing information complexity $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$, the recognition accuracy decreases and MSE increases. We see similar behavior by increasing the information leakage weight α.

Fig. 7 depicts the information leakage for different trade-offs between β and α. In the supervised scenario, we observe that increasing the information leakage weight α may not significantly reduce the information leakage. This behavior can be interpreted by possible correlation between the utility and sensitive attributes.

Fig. 8 depicts the qualitative evaluation for sample quality of CelebA at the inferential adversary for two CLUB models. Four scenarios are considered to investigate the adversary beliefs about sensitive attribute $\mathbf{S}$. $\mathbf{S} = \mathbf{S}$ corresponds to the case in which the adversary recovers $\mathbf{X}$ when his belief coincides with the original $\mathbf{S}$ (i.e., $\mathbf{S} = [1, 0]$ for man, and $\mathbf{S} = [0, 1]$ for woman); $\mathbf{S} = 1 - \mathbf{S}$ corresponds to the case in which the adversary recovers $\mathbf{X}$ when his belief coincides with the inverse of original $\mathbf{S}$; $\mathbf{S} = \emptyset$ corresponds to the case in which the adversary recovers $\mathbf{X}$ without any assumption on the sensitive attribute $\mathbf{S}$; $\mathbf{S} = [1, 1]$ corresponds to adversary's possible, probably meaningless, belief about sensitive attribute, to infer the sensitive attribute $\mathbf{S}$ from reconstructed $\mathbf{X}$. Gender and emotion probabilities are computed and depicted under each image. These probabilities show how accurately an adversary can revise his belief about $\mathbf{S}$, at different information leakage weights α and information complexity weights β.

The training details and network architectures are provided in the expanded version of this manuscript.

Algorithm 1 Supervised DVCLUB training algorithm associated with (P1: S)

Input: Training Dataset: $\{(\mathbf{u}_n, \mathbf{s}_n, \mathbf{x}_n)\}_{n=1}^N$;
Hyper-Parameters: α, β
$\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\varphi}, \boldsymbol{\eta}, \boldsymbol{\omega}$ ← Initialize Network Parameters
repeat

(1) **Train the Encoder, Utility Decoder, Uncertainty Decoder** $(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\varphi})$

Sample a mini-batch $\{\mathbf{u}_m, \mathbf{s}_m, \mathbf{x}_m\}_{m=1}^M \sim P_{\mathrm{D}}(\mathbf{X}) P_{\mathbf{U},\mathbf{S}|\mathbf{X}}$
Compute $\mathbf{z}_m \sim f_{\boldsymbol{\phi}}(\mathbf{x}_m), \forall m \in [M]$
Sample $\{\widehat{\mathbf{z}}_m\}_{m=1}^M \sim Q_{\boldsymbol{\psi}}(\mathbf{Z})$
Compute $\widehat{\mathbf{x}}_m = g_{\boldsymbol{\varphi}}(\widehat{\mathbf{z}}_m, \mathbf{s}_m), \forall m \in [M]$
Back-propagate loss:

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\varphi}) = -\frac{1}{M}\sum_{m=1}^{M}\big(\log P_{\boldsymbol{\theta}}(\mathbf{u}_m|\mathbf{z}_m) - (\beta+\alpha)\mathrm{D}_{\mathrm{KL}}\big(P_{\boldsymbol{\phi}}(\mathbf{z}_m|\mathbf{x}_m)\|Q_{\boldsymbol{\psi}}(\mathbf{z}_m)\big) + \alpha \log P_{\boldsymbol{\varphi}}(\widehat{\mathbf{x}}_m|\mathbf{s}_m, \mathbf{z}_m)\big)$$

(2) **Train the Latent Space Discriminator** $\boldsymbol{\eta}$

Sample $\{\mathbf{x}_m\}_{m=1}^M \sim P_{\mathrm{D}}(\mathbf{X})$
Sample $\{\mathbf{n}_m\}_{m=1}^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Compute $\mathbf{z}_m \sim f_{\boldsymbol{\phi}}(\mathbf{x}_m), \forall m \in [M]$
Compute $\widetilde{\mathbf{z}}_m \sim g_{\boldsymbol{\psi}}(\mathbf{n}_m), \forall m \in [M]$
Back-propagate loss:

$$\mathcal{L}(\boldsymbol{\eta}) = \frac{-(\beta+\alpha)}{M}\sum_{m=1}^{M}\Big(\log D_{\boldsymbol{\eta}}(\mathbf{z}_m) + \log\big(1 - D_{\boldsymbol{\eta}}(\widetilde{\mathbf{z}}_m)\big)\Big)$$

(3) **Train the Encoder and Prior Distribution Generator** $(\boldsymbol{\phi}, \boldsymbol{\psi})$ **Adversarially**

Sample $\{\mathbf{x}_m\}_{m=1}^M \sim P_{\mathrm{D}}(\mathbf{X})$
Sample $\{\mathbf{n}_m\}_{m=1}^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Compute $\mathbf{z}_m \sim f_{\boldsymbol{\phi}}(\mathbf{x}_m), \forall m \in [M]$
Compute $\widetilde{\mathbf{z}}_m \sim g_{\boldsymbol{\psi}}(\mathbf{n}_m), \forall m \in [M]$
Back-propagate loss:

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\psi}) = \frac{(\beta+\alpha)}{M}\sum_{m=1}^{M}\Big(\log D_{\boldsymbol{\eta}}(\mathbf{z}_m) + \log\big(1 - D_{\boldsymbol{\eta}}(\widetilde{\mathbf{z}}_m)\big)\Big)$$

(4) **Train the Attribute Class Discriminator** $\boldsymbol{\omega}$

Sample $\{\mathbf{u}_m\}_{m=1}^M \sim P_{\mathbf{U}}$
Sample $\{\mathbf{n}_m\}_{m=1}^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Compute $\widetilde{\mathbf{u}}_m \sim g_{\boldsymbol{\theta}}\left(g_{\boldsymbol{\psi}}(\mathbf{n}_m)\right), \forall m \in [M]$
Back-propagate loss:

$$\mathcal{L}(\boldsymbol{\omega}) = -\frac{1}{M}\sum_{m=1}^{M}\Big(\log D_{\boldsymbol{\omega}}(\mathbf{u}_m) + \log\big(1 - D_{\boldsymbol{\omega}}(\widetilde{\mathbf{u}}_m)\big)\Big)$$

(5) **Train the Prior Distribution Generator and Utility Decoder** $(\boldsymbol{\psi}, \boldsymbol{\theta})$ **Adversarially**

Sample $\{\mathbf{n}_m\}_{m=1}^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Compute $\widetilde{\mathbf{u}}_m \sim g_{\boldsymbol{\theta}}\left(g_{\boldsymbol{\psi}}(\mathbf{n}_m)\right), \forall m \in [M]$
Back-propagate loss:

$$\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M}\log\big(1 - D_{\boldsymbol{\omega}}(\widetilde{\mathbf{u}}_m)\big)$$

until Convergence
return $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\varphi}, \boldsymbol{\eta}, \boldsymbol{\omega}$

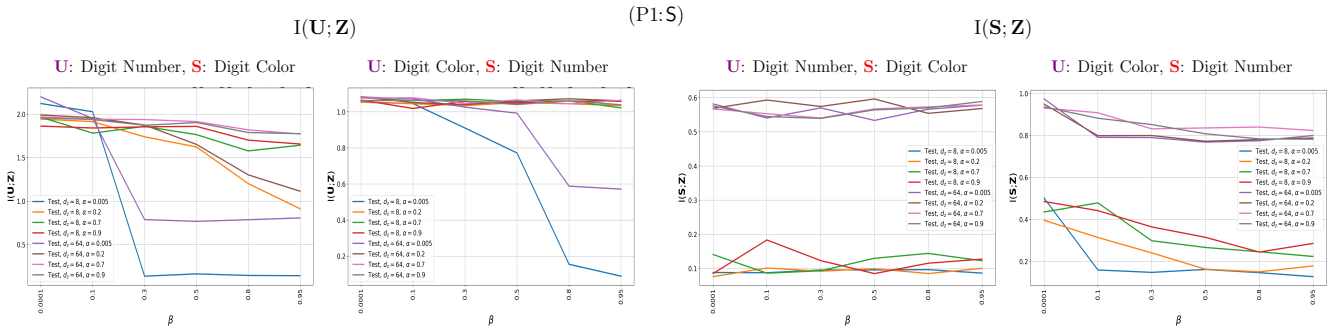


Fig. 5: Estimated information utility $\mathrm{I}(\mathbf{U};\mathbf{Z})$ (Left Panel) and information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$ (Right Panel) on Colored-MNIST dataset using MINE, considering supervised scenario (P1:S), for $d_z \in \{8, 64\}$, setting $P(\mathsf{Red}) = \frac{1}{2}$, $P(\mathsf{Green}) = \frac{1}{6}$, $P(\mathsf{Blue}) = \frac{1}{3}$.

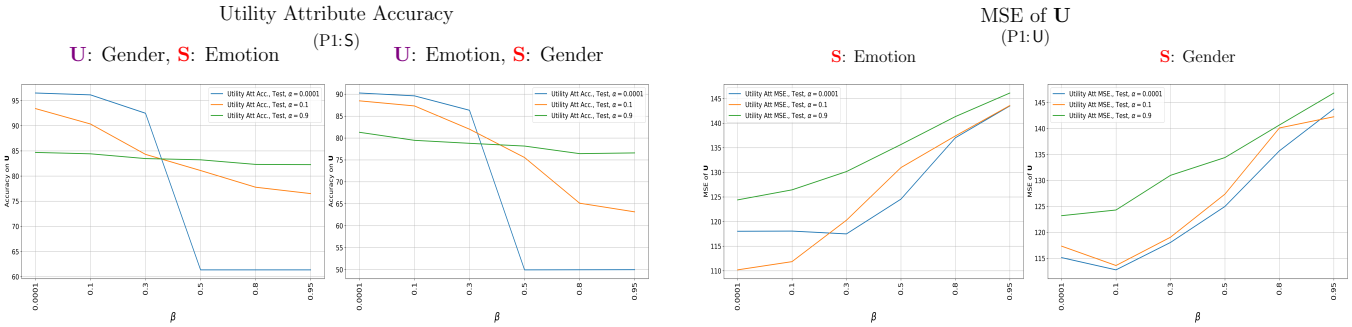


Fig. 6: Recognition accuracy of the utility attribute $\mathbf{U}$ for supervised CLUB model (P1:S) (Left Panel) and Mean Square Error (MSE) of reconstructed $\mathbf{U} \equiv \mathbf{X}$ for unsupervised CLUB model (P1:U) (Right Panel) on 'Test' dataset of CelebA, considering $d_{\mathbf{z}} = 64$, setting $Q_{\boldsymbol{\psi}}(\mathbf{Z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\mathbf{z}}})$.

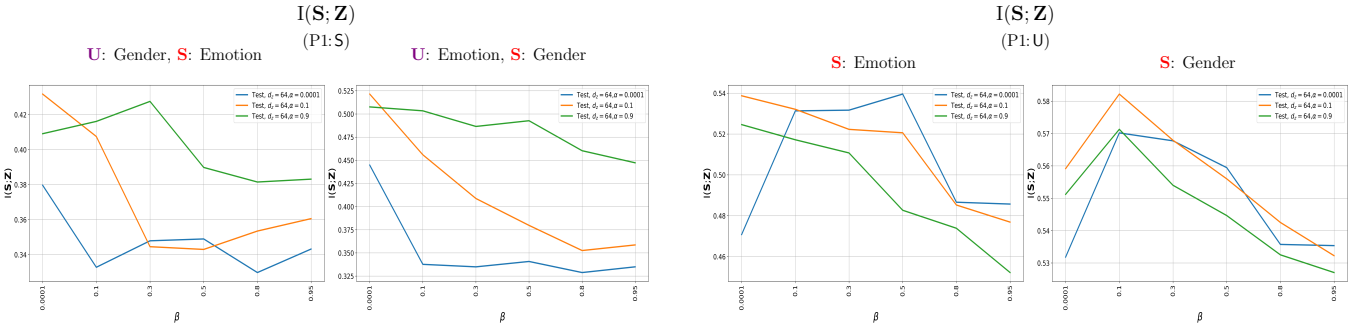


Fig. 7: Estimated information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$ for CLUB models (P1:S) (Left Panel) and (P1:U) (Right Panel) on 'Test' dataset of CelebA using MINE, considering $d_{\mathbf{z}} = 64$, setting $Q_{\boldsymbol{\psi}}(\mathbf{Z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\mathbf{z}}})$.

Fig. 8: Qualitative evaluation for sample quality at the inferential adversary for CLUB models (P1:S) (Left Panel) and (P1:U) (Right Panel) on CelebA, for different trade-offs between β and α.

6. Connections with other Problems

6.1. Connection with State-of-the-Art Bottleneck Models

In this subsection, we present the connection of CLUB model with the most relative state-of-the-art literature. We show that CLUB model generalizes most of the previously proposed models. To help interpretation, we use I-Diagram [69], an analogy between information theory and set theory, to visualize a clear connection between the different considered objectives. By constructing a unique measure, called I-Measure, consistent with Shannon's information measure, we can geometrically represent the relationship among the Shannon's information measures. The Markov chain $(\mathbf{U},\mathbf{S})$ –○–$\mathbf{X}$–○–$\mathbf{Z}$ implies $\mathrm{I}(\mathbf{U};\mathbf{S};\mathbf{Z}\mid\mathbf{X})+\mathrm{I}(\mathbf{U};\mathbf{Z}\mid\mathbf{S},\mathbf{X})+\mathrm{I}(\mathbf{S};\mathbf{Z}\mid\mathbf{U},\mathbf{X})=\mathrm{I}(\mathbf{U};\mathbf{Z}\mid\mathbf{X})+\mathrm{I}(\mathbf{S};\mathbf{Z}\mid\mathbf{U},\mathbf{X})=\mathrm{I}(\mathbf{U},\mathbf{S};\mathbf{Z}\mid\mathbf{X})=0$. Considering the Markov chain $(\mathbf{U},\mathbf{S})$ –○–$\mathbf{X}$–○–$\mathbf{Z}$, the corresponding I-Diagram is depicted in Fig. 9, where various information quantities are also highlighted.

Information Bottleneck (IB). The IB principle [3] formulates the problem of extracting the relevant information from the random variable $\mathbf{X}$ about the random variable $\mathbf{U}$ that is of interest. Given two correlated random variables $\mathbf{U}$ and $\mathbf{X}$ with joint distribution $P_{\mathbf{U},\mathbf{X}}$, the goal of *original* IB is to find a representation $\mathbf{Z}$ of $\mathbf{X}$ using a stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$ such that: (i) $\mathbf{U}$–○–$\mathbf{X}$–○–$\mathbf{Z}$ and (ii) representation $\mathbf{Z}$ is maximally informative about $\mathbf{U}$ (maximizing $\mathrm{I}(\mathbf{U};\mathbf{Z})$) while being minimally informative about $\mathbf{X}$ (minimizing $\mathrm{I}(\mathbf{X};\mathbf{Z})$). This trade-off can be formulated by bottleneck functional:

$$\mathsf{IB}(R,P_{\mathbf{U},\mathbf{X}}) \coloneqq \sup_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ \mathbf{U}-\circ-\mathbf{X}-\circ-\mathbf{Z}}} \mathrm{I}(\mathbf{U};\mathbf{Z}) \;\; \text{s.t.} \;\; \mathrm{I}(\mathbf{X};\mathbf{Z}) \leq R. \quad (30)$$

In the IB model, $\mathrm{I}(\mathbf{U};\mathbf{Z})$ is referred to as the relevance of $\mathbf{Z}$ and $\mathrm{I}(\mathbf{X};\mathbf{Z})$ is referred to as the complexity of $\mathbf{Z}$. The values $\mathsf{IB}(R,P_{\mathbf{U},\mathbf{X}})$ for different R specify the IB curve. Analogously, with the introduction of a Lagrange multiplier $\beta\in[0,1]$ we can quantify the IB problem by the associated Lagrangian functional $\mathcal{L}_{\mathrm{IB}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta\right)\coloneqq\mathrm{I}(\mathbf{U};\mathbf{Z})-\beta\,\mathrm{I}(\mathbf{X};\mathbf{Z})$. Clearly, IB is a specific case of the CLUB model (4), when the information leakage is disregarded, or equivalently, when $R^{\mathrm{s}}\geq\mathrm{H}(\mathbf{S})$.

Privacy Funnel (PF). In contrast to the IB principle, which seeks to obtain a representation $\mathbf{Z}$ that is maximally expressive (information preserving) about $\mathbf{U}$ while maximally compressive about $\mathbf{X}$, considering Markov chain $\mathbf{S}$–○–$\mathbf{X}$–○–$\mathbf{Z}$, the goal of PF [2] is to determine a representation $\mathbf{Z}$ that minimizes the information leakage between the private (sensitive) data $\mathbf{S}$ and the disclosed representation $\mathbf{Z}$, i.e., $\mathrm{I}(\mathbf{S};\mathbf{Z})$, while maximizing the amount of information between non-private (useful) data $\mathbf{X}$ and $\mathbf{Z}$, i.e., $\mathrm{I}(\mathbf{X};\mathbf{Z})$. Therefore, the PF method addresses the trade-off between the information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$ and the revealed useful information $\mathrm{I}(\mathbf{X};\mathbf{Z})$. Analogously, this trade-off can be formulated by the PF functional:

$$\mathsf{PF}(R,P_{\mathbf{S},\mathbf{X}}) \coloneqq \inf_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ \mathbf{S}-\circ-\mathbf{X}-\circ-\mathbf{Z}}} \mathrm{I}(\mathbf{S};\mathbf{Z}) \;\; \text{s.t.} \;\; \mathrm{I}(\mathbf{X};\mathbf{Z}) \geq R. \quad (31)$$

The values $\mathsf{PF}(R,P_{\mathbf{S},\mathbf{X}})$ for different R specify the PF curve. By introducing a Lagrange multiplier $\beta\in[0,1]$ we can quantify the PF problem by the associated Lagrangian functional $\mathcal{L}_{\mathrm{PF}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta\right)\coloneqq\mathrm{I}(\mathbf{S};\mathbf{Z})-\beta\,\mathrm{I}(\mathbf{X};\mathbf{Z})$. Setting $\mathbf{U}\equiv\mathbf{X}$ and $R^{\mathrm{z}}\geq\mathrm{H}(P_{\mathbf{X}})$ in the CLUB objective (4), the CLUB model reduces to the PF model. This corresponds to a scenario, for instance, in which the goal is to release facial images, without revealing a specific sensitive attribute.

In [44], the authors obtained a map $P_{\mathbf{Z}|\mathbf{X}}$ that minimizes $\mathrm{I}(\mathbf{U};\mathbf{X}\mid\mathbf{Z})$ under an information leakage constraint $\mathrm{I}(\mathbf{S};\mathbf{Z})$. Considering a similar setting to ours, the following objective is considered:

$$\inf_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ (\mathbf{U},\mathbf{S})-\circ-\mathbf{X}-\circ-\mathbf{Z}}} \mathrm{I}(\mathbf{U};\mathbf{X}\mid\mathbf{Z}) \quad \text{s.t.} \quad \mathrm{I}(\mathbf{S};\mathbf{Z})\leq R^{\mathrm{s}}. \quad (32)$$

One can interpret $\mathrm{I}(\mathbf{U};\mathbf{X}\mid\mathbf{Z})$ as the amount of information about $\mathbf{U}$ we lose by observing the released representation $\mathbf{Z}$ instead of the original data $\mathbf{X}$. To see its connection to CLUB, note that, under the Markov chain $\mathbf{U}$–○–$\mathbf{X}$–○–$\mathbf{Z}$, we have:

$$\mathrm{I}(\mathbf{U};\mathbf{Z})=\mathrm{I}(\mathbf{U};\mathbf{X})-\mathrm{I}(\mathbf{U};\mathbf{X}\mid\mathbf{Z}). \quad (33)$$

Therefore, maximizing $\mathrm{I}(\mathbf{U};\mathbf{Z})$ is equivalent to minimizing $\mathrm{I}(\mathbf{U};\mathbf{X}\mid\mathbf{Z})$. Clearly, the objective (32) can be considered as a special case of the CLUB model in (4). In particular, we highlight that our model addresses the key missed component in the considered formulation, i.e., the information complexity constraint. Note that $P_{\mathbf{U},\mathbf{X}|\mathbf{Z}}=P_{\mathbf{U}|\mathbf{X},\mathbf{Z}}P_{\mathbf{X}|\mathbf{Z}}=P_{\mathbf{U}|\mathbf{X}}P_{\mathbf{X}|\mathbf{Z}}$, hence, $\mathrm{I}(\mathbf{U};\mathbf{X}\mid\mathbf{Z})=\mathbb{E}_{P_{\mathbf{X},\mathbf{Z}}}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{U}|\mathbf{X}}\|P_{\mathbf{U}|\mathbf{Z}}\right)\right]$. This gives us another interpretation of the CLUB objective as a distribution matching problem, which aims at obtaining a stochastic map $P_{\mathbf{Z}|\mathbf{X}}$ such that $P_{\mathbf{U}|\mathbf{Z}}\approx P_{\mathbf{U}|\mathbf{X}}$. This means that the posterior distributions of the utility attribute $\mathbf{U}$ are similar conditioned on the released representation $\mathbf{Z}$ or the original data $\mathbf{X}$. Finally, we remark that Eq. (33) is also related to the source coding problem with a common reconstruction constraint studied in [70], [71].

Deterministic IB (DIB). In the information complexity decomposition $\mathrm{I}(\mathbf{X};\mathbf{Z})=\mathrm{H}(\mathbf{Z})-\mathrm{H}(\mathbf{Z}\mid\mathbf{X})$, the conditional entropy $\mathrm{H}(\mathbf{Z}\mid\mathbf{X})$ measures the *encoding uncertainty* (see Fig. 9c). The deterministic IB model [57] is based on a deterministic encoder. In this case, we have $\mathrm{H}(\mathbf{Z}\mid\mathbf{X})=0$. Considering the Markov chain $\mathbf{U}$–○–$\mathbf{X}$–○–$\mathbf{Z}$, the following objective is considered:

$$\mathsf{DIB}(R,P_{\mathbf{U},\mathbf{X}}) \coloneqq \inf_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ \mathbf{U}-\circ-\mathbf{X}-\circ-\mathbf{Z}}} \mathrm{H}(\mathbf{Z}) \;\; \text{s.t.}\; \mathrm{I}(\mathbf{U};\mathbf{Z})\geq R. \quad (34)$$

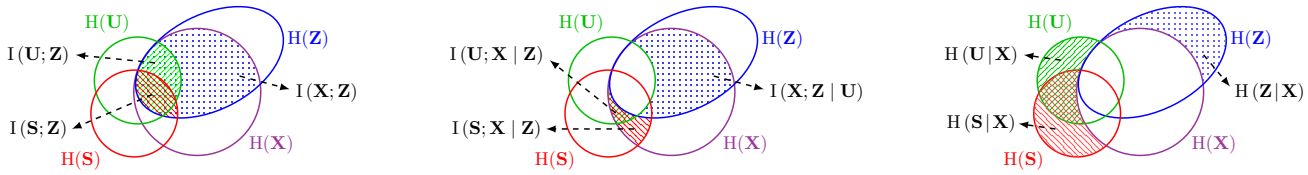
In this case, the associated Lagrangian functional is given by $\mathcal{L}_{\mathrm{DIB}}(P_{\mathbf{Z}|\mathbf{X}},\beta)\coloneqq\mathrm{H}(\mathbf{Z})-\beta\mathrm{I}(\mathbf{U};\mathbf{Z})$. Clearly, DIB is a specific case of the IB model; and hence, the CLUB model.

Conditional Entropy Bottleneck (CEB). The conditional entropy bottleneck (CEB) [58] is motivated by the principle of minimum necessary information, and considers the following objective:

$$\mathsf{CEB}(R,P_{\mathbf{U},\mathbf{X}}) \coloneqq \sup_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ \mathbf{U}-\circ-\mathbf{X}-\circ-\mathbf{Z}}} \mathrm{I}(\mathbf{U};\mathbf{Z}) \;\; \text{s.t.}\; \mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{U})\leq R. \quad (35)$$

One can interpret $\mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{U})$ as the *redundant* information in the released representation $\mathbf{Z}$ about the utility attribute $\mathbf{U}$. To see its connection to CLUB, let us consider the Markov chain $\mathbf{U}$–○–$\mathbf{X}$–○–$\mathbf{Z}$, we have:

$$\mathrm{I}(\mathbf{X};\mathbf{Z})=\mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{U})+\mathrm{I}(\mathbf{U};\mathbf{Z}). \quad (36)$$



(a) (b) (c)

Fig. 9: Information diagrams for $(\mathbf{U},\mathbf{S})$ —○—$\mathbf{X}$—○—$\mathbf{Z}$. (a) mutual information $\mathrm{I}(\mathbf{U};\mathbf{Z})$, $\mathrm{I}(\mathbf{X};\mathbf{Z})$ and $\mathrm{I}(\mathbf{S};\mathbf{Z})$; (b) redundant information $\mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{U})$, residual information $\mathrm{I}(\mathbf{U};\mathbf{X}\mid\mathbf{Z})$ and residual information $\mathrm{I}(\mathbf{S};\mathbf{X}\mid\mathbf{Z})$; (c) utility attribute uncertainty $\mathrm{H}(\mathbf{U}\,|\,\mathbf{X})$, private attribute uncertainty $\mathrm{H}(\mathbf{S}\,|\,\mathbf{X})$ and encoding uncertainty $\mathrm{H}(\mathbf{Z}\,|\,\mathbf{X})$.

Therefore, using (36), the associated CEB Lagrangian $\mathcal{L}_{\mathrm{CEB}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta'\right) := \mathrm{I}(\mathbf{U};\mathbf{Z}) - \beta'\,\mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{U})$ can recast as:

$$\begin{aligned}\mathcal{L}_{\mathrm{CEB}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta'\right) &= (\beta'+1)\,\mathrm{I}(\mathbf{U};\mathbf{Z}) - \beta'\mathrm{I}(\mathbf{X};\mathbf{Z})\\ &= (\beta'+1)\left(\mathrm{I}(\mathbf{U};\mathbf{Z}) - \frac{\beta'}{\beta'+1}\mathrm{I}(\mathbf{X};\mathbf{Z})\right)\\ &= (\beta'+1)\,\mathcal{L}_{\mathrm{IB}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta\right), \qquad (37)\end{aligned}$$

where $\beta = \beta'/(\beta'+1)$. Hence, maximizing $\mathcal{L}_{\mathrm{CEB}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta'\right)$ is equivalent to maximizing $\mathcal{L}_{\mathrm{IB}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta\right)$, which is a specific case of CLUB model (4).

Conditional PF (CPF). Inspired by the conditional entropy bottleneck [58] and noting that the PF (31) is dual of the Information Bottleneck (30), the recent work of [59], addressed the CPF as:

$$\mathrm{CPF}\,(R,P_{\mathbf{S},\mathbf{X}}) := \inf_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ \mathbf{S}-\!\circ\!-\mathbf{X}-\!\circ\!-\mathbf{Z}}} \mathrm{I}(\mathbf{S};\mathbf{Z}) \text{ s.t. } \mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{S}) \geq R^{\mathrm{c|s}}. \quad (38)$$

To see its connection to CLUB, we use the Markov chain $\mathbf{S}-\!\circ\!-\mathbf{X}-\!\circ\!-\mathbf{Z}$ to obtain:

$$\mathrm{I}(\mathbf{X};\mathbf{Z}) = \mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{S}) + \mathrm{I}(\mathbf{S};\mathbf{Z}). \qquad (39)$$

Hence, CPF ignores the shared information in the released representation $\mathbf{Z}$ about the private attribute $\mathbf{S}$, and imposes the constraint on the *residual* information in the released representation $\mathbf{Z}$ about the useful data $\mathbf{X}$[6]. The associated CPF Lagrangian functional is given by $\mathcal{L}_{\mathrm{CPF}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta'\right) := \mathrm{I}(\mathbf{S};\mathbf{Z}) - \beta'\mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{S})$. Analogous to (37), and using (39), we have:

$$\mathcal{L}_{\mathrm{CPF}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta'\right) = (\beta'+1)\,\mathcal{L}_{\mathrm{PF}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta\right), \qquad (40)$$

where $\beta = \beta'/(\beta'+1)$. Hence, minimizing $\mathcal{L}_{\mathrm{CPF}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta'\right)$ is equivalent to minimizing $\mathcal{L}_{\mathrm{PF}}\left(P_{\mathbf{Z}|\mathbf{X}},\beta\right)$, which is again a specific case of the CLUB model in (4).

6.2. Connection with State-of-the-Art Generative Models

We now compare the the DVCLUB functionals (16) and (17) to other generative modeling approaches in the literature. Note that the latent variable generative models, like the VAE and GAN families, aim at capturing data distributions by minimizing specific discrepancy measures $\mathrm{dist}(P_{\mathbf{X}},P_{\boldsymbol{\theta}}(\mathbf{X}))$ between the true (but unknown) data distribution $P_{\mathbf{X}}$ and the generated model distribution $P_{\boldsymbol{\theta}}(\mathbf{X})$. Furthermore, note that maximizing the objective (17) is equivalent to maximizing the information utility $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z})$, which is equivalent to minimizing the KL-divergence discrepancy measure $\mathrm{D}_{\mathrm{KL}}(P_{\mathrm{D}}(\mathbf{X})\,\|\,P_{\boldsymbol{\theta}}(\mathbf{X}))$. Finally note that although the CLUB model is formulated using Shannon's mutual information, which leads us to self-consistent equations, one can use other information measures with different operational meanings.

In β-VAE [9] the Lagrangian functional is defined as:

$$\begin{aligned}\mathcal{L}_{\beta\text{-VAE}}\,(\boldsymbol{\phi},\boldsymbol{\theta},\beta) := \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right]\\ -\beta\,\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}\,|\,\mathbf{X})\,\|\,Q_{\boldsymbol{\psi}}(\mathbf{Z})\mid P_{\mathrm{D}}(\mathbf{X}))\,. \qquad (41)\end{aligned}$$

In the original VAE [8] framework the parameter β associated with the information complexity is set to 1. Hence, VAE and β-VAE force $P_{\boldsymbol{\phi}}(\mathbf{Z}\,|\,\mathbf{X}{=}\mathbf{x})$ to match the proposal prior $Q_{\boldsymbol{\psi}}(\mathbf{Z})$ for all input samples $\mathbf{x}\in\mathcal{X}$ drawn from $P_{\mathrm{D}}(\mathbf{X})$. However, there is no constraint to penalize the discrepancy between the learned aggregated posterior $P_{\boldsymbol{\phi}}(\mathbf{Z})$ and the proposal prior distribution $Q_{\boldsymbol{\psi}}(\mathbf{Z})$. Clearly, β-VAE functional (41) is a specific case of DVCLUB functional (17).

The InfoVAE [10], considered the following Lagrangian functional:

$$\begin{aligned}\mathcal{L}_{\text{InfoVAE}}\,(\boldsymbol{\phi},\boldsymbol{\theta},\beta) := \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right]\\ -\beta\,\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}\,|\,\mathbf{X})\,\|\,Q_{\boldsymbol{\psi}}(\mathbf{Z})\mid P_{\mathrm{D}}(\mathbf{X}))\\ +(\beta-\tau)\,\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z})\,\|\,Q_{\boldsymbol{\psi}}(\mathbf{Z}))\,, \qquad (42)\end{aligned}$$

where $\tau > 0$. When $\beta = \tau$, we get β-VAE. Let $\tau = 0$ and consider DVCLUB Lagrangian functional (17), therefore, the first term of InfoVAE functional corresponds to the first term of information utility lower bound $\mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{U};\mathbf{Z})$, while the second and third terms correspond to the information complexity $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$. Hence, there is no term to capture a discrepancy between the observed distribution $P_{\mathrm{D}}(\mathbf{X})$ and the generated model distribution $P_{\boldsymbol{\theta}}(\mathbf{X})$.

GANs compute the optimal generator $g^*_{\boldsymbol{\theta}}$ by minimizing a distance between the observed distribution $P_{\mathrm{D}}(\mathbf{X})$ and the generated distribution $P_{\boldsymbol{\theta}}(\mathbf{X})$, without considering an explicit probability model for the observed data. The original GAN [11] problem, also called the vanilla GAN, considers the following objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}\,(\boldsymbol{\theta},\boldsymbol{\omega}) := \;& \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\log D_{\boldsymbol{\omega}}(\mathbf{X})\right]\\ &+\mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[\log\left(1-D_{\boldsymbol{\omega}}(\,g_{\boldsymbol{\theta}}(\mathbf{Z})\,)\right)\right], \qquad (43)\end{aligned}$$

where first a random code $\mathbf{Z}\in\mathcal{Z}$ is sampled from a fixed distribution $Q_{\psi}(\mathbf{Z})$, next $\mathbf{Z}$ is mapped to the $\mathbf{X}\in\mathcal{X}$ using a deterministic generator (decoder) $g_{\boldsymbol{\theta}}:\mathcal{Z}\to\mathcal{X}$. The generator $g_{\boldsymbol{\theta}}$ and the visual space discriminator $D_{\boldsymbol{\omega}}$ neural networks are trained adversarially:

$$\min_{\boldsymbol{\theta}}\max_{\boldsymbol{\omega}}\;\; \mathcal{L}_{\mathrm{GAN}}\,(\boldsymbol{\theta},\boldsymbol{\omega})\,. \qquad (44)$$

To see its relation to our DVCULB model, consider the visible space discriminator training step (28) and utility decoder adversarial training step (29).

[6]Note that in the PF model in (31), $\mathrm{I}(\mathbf{X};\mathbf{Z})$ measures the 'useful' information, which is of the designer's interest. Hence, $\mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{S})$ in PF quantifies the residual information, while $\mathrm{I}(\mathbf{X};\mathbf{Z}\mid\mathbf{U})$ in IB quantifies the redundant information.

Remark 1. In the Euclidean space we can represent any convex function as the point-wise supremum of a family of affine functions, and vice versa. Noting that f-divergence is a convex function of probability measures, we can represent it as a point-wise supremum of affine functions. Let $f : \mathcal{X} \to \mathbb{R}$ be a convex and lower semi-continuous function. The convex conjugate (also known as the Fenchel dual or Legendre transform) $f^* : \mathcal{X} \to \mathbb{R}$ of f is defined by $f^*(\mathbf{v}) := \sup_{\mathbf{c} \in \mathsf{dom}(f)} \{\langle \mathbf{c}, \mathbf{v} \rangle - f(\mathbf{c})\}$, where $\mathsf{dom}(f)$ denotes the effective domain of f which is given by $\mathsf{dom}(f) = \{\mathbf{c} \mid f(\mathbf{c}) < \infty\}$ [63], [72]. Hence, for any $\mathbf{v} \in \mathsf{dom}(f^*)$ and $\mathbf{c} \in \mathsf{dom}(f)$ we have $f(\mathbf{c}) \geq \langle \mathbf{c}, \mathbf{v} \rangle - f^*(\mathbf{v})$ (Fenchel-Young inequality). Using the notion of convex conjugate one can obtain a variational representation of a *well-defined* f-divergence $\mathrm{D}_f(P\|Q) = \int_{\mathcal{X}} Q(\mathbf{x}) f\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) \mathrm{d}\mathbf{x}$ in terms of the convex conjugate of f:

$$\mathrm{D}_f(P\|Q) = \int_{\mathcal{X}} Q(\mathbf{x}) \sup_{\mathbf{c} \in \mathsf{dom}(f^*)} \left\{ \mathbf{c}\, \frac{P(\mathbf{x})}{Q(\mathbf{x})} - f^*(\mathbf{c}) \right\} \mathrm{d}\mathbf{x} \tag{45a}$$

$$\geq \sup_{C : \mathcal{X} \to \mathbb{R}} \left(\int_{\mathcal{X}} P(\mathbf{x})\, C(\mathbf{x})\, \mathrm{d}\mathbf{x} - \int_{\mathcal{X}} Q(\mathbf{x})\, f^*(C(\mathbf{x}))\, \mathrm{d}\mathbf{x} \right) \tag{45b}$$

$$= \sup_{C : \mathcal{X} \to \mathbb{R}} \left(\mathbb{E}_P\left[C(\mathbf{X}) \right] + \mathbb{E}_Q\left[-f^*(C(\mathbf{X})) \right] \right), \tag{45c}$$

where the supremum is taken over all measurable functions $C : \mathcal{X} \to \mathbb{R}$. Under mild conditions on f [63], the bound is tight for $C^*(x) = f'\left(\frac{p(x)}{q(x)}\right)$, where f' denotes the first derivative of f.

Using the above remark, we now suppose $C(\mathbf{x})$ is a variational function parameterized by $\boldsymbol{\omega}$, and represent it as $C_{\boldsymbol{\omega}}(\mathbf{x}) = h\left(V_{\boldsymbol{\omega}}(\mathbf{x})\right)$, where $V_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{R}$, and $h : \mathbb{R} \to \mathsf{dom}(f^*)$ is an output activation function. We call $C_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{R}$ as the *critic* function. Choosing $f(\mathbf{c}) = \mathbf{c} \log \mathbf{c} - (\mathbf{c} + 1) \log(\frac{\mathbf{c}+1}{2})$ gives us the Jensen-Shannon's divergence (JSD) and we have $f^*(\mathbf{v}) = -\log\left(1 - e^{\mathbf{v}}\right)$. Using the JSD and choosing the output activation function as $h(t) = -\log(1 + e^{-t})$, the variational lower bound (45c) reduces to the original GAN objective (43), where $C_{\boldsymbol{\omega}}(\mathbf{x}) = h(V_{\boldsymbol{\omega}}(\mathbf{x})) = \log D_{\boldsymbol{\omega}}(\mathbf{x})$, and $-f^*\left(\log D_{\boldsymbol{\omega}}(\mathbf{x})\right) = \log(1 - D_{\boldsymbol{\omega}}(\mathbf{x}))$.

Remark 2 (Optimal Transport Problem)**.** A recent body of work study generative models from an optimal transport (OT) point of view. Let $\mathcal{X}$ and $\mathcal{Y}$ be two measurable spaces, and let $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ be the sets of all positive Radon probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively. For any measurable non-negative cost function $\mathsf{c} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, the optimal transport problem (Kantorovich problem) between distributions $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$ is defined as [73], [74]:

$$\mathsf{OT}_{\mathsf{c}}\left(P, Q\right) := \inf_{\boldsymbol{\pi} \in \Pi(P,Q)} \mathbb{E}_{\boldsymbol{\pi}}\left[\mathsf{c}(\mathbf{X}, \mathbf{Y}) \right], \tag{46}$$

where $\Pi(P, Q)$ denotes the set of joint distributions (couplings) over the product space $\mathcal{X} \times \mathcal{Y}$ with marginals P and Q, respectively. That is, for all measurable sets $\mathcal{A} \subset \mathcal{X}$ and $\mathcal{B} \subset \mathcal{Y}$, we have $\Pi(P, Q) := \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \pi(\mathcal{A} \times \mathcal{Y}) = P(\mathcal{A}), \pi(\mathcal{X} \times \mathcal{B}) = Q(\mathcal{B})\}$. The joint measures $\boldsymbol{\pi} \in \Pi(P, Q)$ are called the *transport plans*. The cost function c represents the cost to move a unit of mass from $\mathbf{x}$ to $\mathbf{y}$.

Remark 3 (Dual Formulation)**.** The Kantorovich problem (46) defines a constrained linear program, and hence admits an equivalent dual formulation:

$$\mathsf{OT}_{\mathsf{c}}\left(P, Q\right) := \sup_{(h_1, h_2) \in \mathcal{R}(\mathsf{c})} \mathbb{E}_P\left[h_1(\mathbf{X}) \right] + \mathbb{E}_Q\left[h_2(\mathbf{Y}) \right], \tag{47a}$$

where for any $\mathsf{c} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$, the set of admissible *dual potentials* (also known as Kantorovich potentials) is $\mathcal{R}(\mathsf{c}) := \{ (h_1, h_2) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) : h_1(\mathbf{x}) + h_2(\mathbf{y}) \leq \mathsf{c}(\mathbf{x}, \mathbf{y}), \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \}$, where $\mathcal{C}(\mathcal{X})$ and $\mathcal{C}(\mathcal{Y})$ are the space of continuous (real-valued) functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively.

For any $h_2 \in \mathcal{C}(\mathcal{Y})$, let us define its c-transform[7] $h_1^{\mathsf{c}} \in \mathcal{C}(\mathcal{Y})$ of $h_1 \in \mathcal{C}(\mathcal{X})$ as $h_1^{\mathsf{c}}(\mathbf{y}) := \inf_{\mathbf{x} \in \mathcal{X}} \mathsf{c}(\mathbf{x}, \mathbf{y}) - h_1(\mathbf{x}), \quad \forall \mathbf{y} \in \mathcal{Y}$. Given a candidate potential h_1 for the first variable, h_1^{c} is the best possible potential that can be paired with h_1. A function h_1^{c} in the form above is called a c-concave function. Note that Kantorovich potentials satisfy $h_2 = h_1^{\mathsf{c}}$. Denoting $h_1 = h$ we can rewrite the dual formulation in (47) as:

$$\mathsf{OT}_{\mathsf{c}}\left(P, Q\right) = \sup_{h \in \mathcal{C}(\mathcal{X})} \mathbb{E}_P\left[h(\mathbf{X}) \right] + \mathbb{E}_Q\left[h^{\mathsf{c}}(\mathbf{Y}) \right]. \tag{48}$$

Remark 4 (Wasserstein Distance)**.** In the Kantorovich problem (46), assume $\mathcal{X} = \mathcal{Y}$, let $(\mathcal{X}, \mathsf{d})$ be a metric space, and consider $\mathsf{c}(\mathbf{x}, \mathbf{y}) = \mathsf{d}^p(\mathbf{x}, \mathbf{y})$ for some $p \geq 1$. In this case, the p-Wasserstein distance[8] on $\mathcal{X}$ is defined as:

$$\mathcal{W}_p(P, Q) := \mathsf{OT}_{\mathsf{d}^p}\left(P, Q\right)^{1/p}. \tag{49}$$

The cases $p = 1$ and $p = 2$ are particularly interesting. The 1-Wasserstein distance is more flexible and easier to bound, moreover, the Kantorovich-Rubinstein duality holds for the 1-Wasserstein distance. The 2-Wasserstein distance is more appropriate to reflect the geometric features, moreover, it scales better with the dimension. Let $\mathrm{L}_p(h) = \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} \{ \frac{|h(\mathbf{x}) - h(\mathbf{y})|}{\mathsf{d}^p(\mathbf{x}, \mathbf{y})} : \mathbf{x} \neq \mathbf{y} \}$ denotes the Lipschitz constant of a function $h \in \mathcal{C}(\mathcal{X})$ with respect to cost $\mathsf{c}(\mathbf{x}, \mathbf{y}) = \mathsf{d}^p(\mathbf{x}, \mathbf{y})$. One can show that if $\mathrm{L}_p(h) \leq 1$, then $h^{\mathsf{c}} = -h$. Let $\mathcal{H}_{p,k} = \{h \in \mathcal{C}(\mathcal{X}) : \mathrm{L}_p(h) \leq k\}$ is the set of all bounded $\mathrm{L}_p(h)$-Lipschitz functions on $(\mathcal{X}, \mathsf{d})$ with $\mathsf{c}(\mathbf{x}, \mathbf{y}) = \mathsf{d}^p(\mathbf{x}, \mathbf{y})$ such that $\mathrm{L}_p(h) \leq k$. The $\mathcal{W}_1(P, Q)$ can be rewritten as:

$$\mathcal{W}_1(P, Q) = \sup_{h \in \mathcal{H}_{1,1}} \mathbb{E}_P\left[h(\mathbf{X}) \right] - \mathbb{E}_Q\left[h(\mathbf{Y}) \right]. \tag{50}$$

Note that, as stated before, the latent variable generative models aim at capturing data distributions by minimizing specific discrepancy measures between the true (but unknown) data distribution $P_{\mathbf{X}}$ and the generated model distribution $P_{\boldsymbol{\theta}}(\mathbf{X})$. Moreover, in practice, only an empirical data distribution $P_{\mathsf{D}}(\mathbf{X})$ is available. In the Optimal Transport problem, one can factor the mapping from $\mathbf{X} \in \mathcal{X}$ to $\mathbf{Y} \in \mathcal{X}$, i.e., the couplings $\Pi(P, Q)$, through a latent code $\mathbf{Z} \in \mathcal{Z}$. This shed light on the connections between the latent variable generative models and Optimal Transport problem.

Let $\mathbf{Y} = g_{\boldsymbol{\theta}}(\mathbf{Z}) \in \mathcal{X}$ denotes the generated samples by a generator $g_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathcal{X}$ and consider the Kantorovich-Rubinstein duality formulation (50). By restricting the dual potential h to have a parametric form, i.e., $h = D_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{R}$, the Wasserstein GAN (WGAN) [12] considers the following objective:

[7] The c-transform is a generalization of the Legendre transform from convex analysis. If $\mathsf{c}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$ on $\mathbb{R}^n \times \mathbb{R}^n$, the c-transform coincides with the Legendre transform.

[8] The p-Wasserstein satisfies the three metric axioms, hence it defines an actual distance between P and Q. Also, note that $\mathcal{W}_p$ depends on d.

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\omega}} \ \mathcal{L}_{\mathrm{WGAN}}(\boldsymbol{\theta},\boldsymbol{\omega}) \coloneqq \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[D_{\boldsymbol{\omega}}(\mathbf{X})\right] - \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[D_{\boldsymbol{\omega}}\left(g_{\boldsymbol{\theta}}(\mathbf{Z})\right)\right], \quad (51)$$

where the 1-Lipschitz constraint in (50) is satisfied by using a deep neural network with ReLu units.

Let $P_{\boldsymbol{\theta}}(\mathbf{Y}\,|\,\mathbf{Z}=\mathbf{z}) = \delta(\mathbf{y} - g_{\boldsymbol{\theta}}(\mathbf{z})), \forall \mathbf{z} \in \mathcal{Z}$, i.e., suppose $\mathbf{Y} \sim P_{\boldsymbol{\theta}}(\mathbf{X})$ is defined with a deterministic mapping, the parameterized Kantorovich problem associated with (46) can be expressed as follows:[9]:

$$\begin{aligned}\mathsf{OT}_{\mathsf{c}}\left(P_{\mathsf{D}}(\mathbf{X}), P_{\boldsymbol{\theta}}(\mathbf{X})\right) &= \inf_{\boldsymbol{\pi}\in\Pi(P_{\mathsf{D}}(\mathbf{X}),P_{\boldsymbol{\theta}}(\mathbf{X}))} \mathbb{E}_{\boldsymbol{\pi}}\left[\mathsf{c}(\mathbf{X},\mathbf{Y})\right] \\ &= \inf_{\substack{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}):\\ P_{\boldsymbol{\phi}}(\mathbf{Z})=Q_{\boldsymbol{\psi}}(\mathbf{Z})}} \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\mathsf{c}\left(\mathbf{X}, g_{\boldsymbol{\theta}}(\mathbf{Z})\right)\right]\right]. \quad (52)\end{aligned}$$

The Wasserstein Auto-Encoder (WAE) [13] is formulated as the relaxed unconstrained parameterized OT (52), and reads as:

$$\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \mathcal{L}_{\mathrm{WAE}}(\boldsymbol{\phi},\boldsymbol{\theta}) \coloneqq \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\mathsf{c}(\mathbf{X}, g_{\boldsymbol{\theta}}(\mathbf{Z}))\right]\right] + \lambda' \,\mathsf{dist}(P_{\boldsymbol{\phi}}(\mathbf{Z}), Q_{\boldsymbol{\psi}}(\mathbf{Z})), \quad (53)$$

where $\lambda' \geq 0$ is a regularization parameter, and $\mathsf{dist}(P_{\boldsymbol{\phi}}(\mathbf{Z}), Q_{\boldsymbol{\psi}}(\mathbf{Z}))$ is a discrepancy between $P_{\boldsymbol{\phi}}(\mathbf{Z})$ and $Q_{\boldsymbol{\psi}}(\mathbf{Z})$. For instance, one can consider $\mathsf{dist}(P_{\boldsymbol{\phi}}(\mathbf{Z}), Q_{\boldsymbol{\psi}}(\mathbf{Z})) = \mathrm{D}_f(P_{\boldsymbol{\phi}}(\mathbf{Z})\|Q_{\boldsymbol{\psi}}(\mathbf{Z}))$, or alternatively, one can use the Maximum Mean Discrepancy (MMD) for a characteristic positive-definite reproducing kernel [13]. Note that, in contrast to β-VAE (41), the WAE [13] directly captures discrepancy between aggregated posterior $P_{\boldsymbol{\phi}}(\mathbf{Z})$ and proposal prior $Q_{\boldsymbol{\psi}}(\mathbf{Z})$, and ignoring the conditional relative entropy $\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}\,|\,\mathbf{X}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}) \,|\, P_{\mathsf{D}}(\mathbf{X}))$. To see its connection with CLUB model, consider the DVCLUB objective (17). Note that maximizing the information utility $\mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z})$ is equivalent to minimizing the divergence measure between $P_{\mathsf{D}}(\mathbf{X})$ and $P_{\boldsymbol{\theta}}(\mathbf{X})$, as well as, minimizing the *negative* log-likelihood $\mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right]$.

In WAE objective (53), let $\mathsf{c}(\mathbf{x},\mathbf{y}) = \mathsf{d}^2(\mathbf{x},\mathbf{y})$. This will lead us to the Adversarial Auto-Encoders (AAEs) model [14]. This means that the AAEs minimizes 2-Wasserstein distance between $P_{\mathsf{D}}(\mathbf{X})$ and $P_{\boldsymbol{\theta}}(\mathbf{X})$. Hence, the AAE objective reads as:

$$\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \ \mathcal{L}_{\mathrm{AAE}}(\boldsymbol{\phi},\boldsymbol{\theta}) \coloneqq \mathbb{E}_{P_{\mathsf{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right] + \lambda' \,\mathsf{dist}(P_{\boldsymbol{\phi}}(\mathbf{Z}), Q_{\boldsymbol{\psi}}(\mathbf{Z})). \quad (54)$$

6.3. Connection with Modern Data Compression Models

The deterministic CLUB (DCLUB) model encourages to have a deterministic encoding function. Considering the Markov chain $(\mathbf{U},\mathbf{S})\multimap\!\circ\!-\mathbf{X}-\!\circ\!\multimap\mathbf{Z}$, the DCLUB functional can be expressed as:

$$\mathsf{DCLUB}(R^{\mathrm{u}}, R^{\mathrm{s}}, P_{\mathbf{U},\mathbf{S},\mathbf{X}}) \coloneqq \inf_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ (\mathbf{U},\mathbf{S})\multimap\!\circ\!-\mathbf{X}-\!\circ\!\multimap\mathbf{Z}}} \mathrm{H}(\mathbf{Z}) \quad \text{s.t. } \mathrm{I}(\mathbf{U};\mathbf{Z}) \geq R^{\mathrm{u}},\ \mathrm{I}(\mathbf{S};\mathbf{Z}) \leq R^{\mathrm{s}}. \quad (55)$$

Clearly, DIB model (34) is a specific case of the DCLUB (55); and hence, the CLUB model. For full details, refer to the expanded presentation available at arXiv:2207.04895.

[9] An almost similar formulation holds for the case in which $P_{\boldsymbol{\theta}}(\mathbf{Y}\,|\,\mathbf{Z})$ are not necessarily Dirac. We refer the reader to [13], [75] for more details.

6.4. Connection with Fair Machine Learning Models

In the field of algorithmic fairness, there is an effort to mitigate biases against certain variables (e.g. gender, ethnicity) in the results of models. Mitigation approaches are categorized into three groups: pre-processing, in-processing, and post-processing. The goal of in-processing approaches is to mitigate bias during training. Fair representation learning aims to learn a representation that is *invariant* to certain variables. The CLUB model can aid in fair representation learning while considering both information complexity and utility constraints.

7. Conclusion

We proposed a general family of optimization problems that unified and generalized most of the state-of-the-art information-theoretic privacy models. We first addressed obfuscation and utility measures under logarithmic loss, and then expressed the concept of relevant information from information-theoretic and statistical perspectives. We then introduced and characterized the CLUB optimization problem and established a variational lower bound to optimize the associated CLUB Lagrangian functional. We constructed the DVCLUB models by employing neural networks to parameterize variational approximations of the information complexity, information leakage, and information utility quantities. We also proposed alternating block coordinate descent training algorithms associated with the proposed DVCLUB models. The DVCLUB model sheds light on the connections between information theory and generative models, deep compression models, as well as, fair machine learning models. In addition, this unifying perspective allows us to relate the CLUB model to several information-theoretic coding problems. Constructing an I-measure, consistent with Shannon's information measures, we geometrically represent the relationship among Shannon's information measures in state-of-the-art bottleneck problems, interpreting the differences in their objectives. Although the CLUB model is formulated using Shannon's mutual information, which leads us to self-consistent equations, one can also use other information measures with different operational meanings.

Acknowledgement

Behrooz Razeghi would like to thank Amir A. Atashin for valuable discussions and contributions in implementing the proposed model. The authors would like to thank Dr. Shahab Asoodeh and Dr. Bernhard C. Geiger for insightful suggestions leading to improvements of this manuscript.

References

[1] F. P. Calmon and N. Fawaz, "Privacy against statistical inference," in *IEEE Allerton*, 2012.
[2] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *IEEE ITW*, 2014.
[3] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *IEEE Allerton*, 2000.
[4] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," in *ICLR*, 2016.
[5] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *NIPS*, 2017.
[6] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NIPS*, 2016.

[7] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. W. Michalak, S. Asoodeh, and F. P. Calmon, "Beyond adult and compas: Fairness in multi-class prediction," in *NeurIPS*, 2022.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[9] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[10] "Infovae: Information maximizing variational autoencoders," *arXiv preprint arXiv:1706.02262*, 2017.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair *et al.*, "Generative adversarial nets," in *NIPS*, 2014.

[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *ICML*, 2017.

[13] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf, "Wasserstein auto-encoders," in *ICLR*, 2018.

[14] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *ICLR Workshop*, 2016.

[15] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *ICML*, 2016.

[16] S. Voloshynovskiy, M. Kondah, S. Rezaeifar, O. Taran, T. Hotolyak, and D. J. Rezende, "Information bottleneck through variational glasses," in *NeurIPS Workshop on Bayesian Deep Learning*, 2019.

[17] B. Razeghi, F. P. Calmon, D. Gunduz, and S. Voloshynovskiy, "Bottlenecks CLUB: Unifying information-theoretic trade-offs among complexity, leakage, and utility," *arXiv preprint arXiv:2207.04895*, 2022.

[18] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, 2000.

[19] ——, "k-anonymity: A model for protecting privacy," *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[20] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *ICDE*, 2006.

[21] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE*, 2007.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, 2006.

[23] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *ACM Sym. on Prin. of Database Sys.*, 2012.

[24] I. S. Reed, "Information theory and privacy in data banks," in *ACM National Computer Conference and Exposition*, 1973.

[25] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.)," *IEEE Transactions on Information Theory*, 1983.

[26] F. P. Calmon, M. Varia, M. Médard, M. M. Christiansen, K. R. Duffy, and S. Tessaro, "Bounds on inference," in *IEEE Allerton*, 2013.

[27] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *IEEE ISIT*, 2015.

[28] Y. O. Basciftci, Y. Wang, and P. Ishwar, "On privacy-utility tradeoffs for constrained data release mechanisms," in *IEEE ITA*, 2016.

[29] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, p. 15, 2016.

[30] K. Kalantari, L. Sankar, and O. Kosut, "On information-theoretic privacy with general distortion cost functions," in *IEEE ISIT*.

[31] B. Rassouli, F. Rosas, and D. Gündüz, "Latent feature disclosure under perfect sample privacy," in *IEEE WIFS*, 2018.

[32] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Tran. on Inf. Theory*, vol. 65, 2018.

[33] S. A. Osia, A. Taheri, A. S. Shamsabadi, K. Katevas, H. Haddadi, and H. R. Rabiee, "Deep private-feature extraction," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[34] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," in *IEEE Allerton*, 2019.

[35] H. Hsu, S. Asoodeh, F. P. Calmon, and N. Fawaz, "Information-theoretic privacy watchdogs," in *IEEE ISIT*, 2019.

[36] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "Tunable measures for information leakage and applications to privacy-utility tradeoffs," *IEEE Tran. on Information Theory*, 2019.

[37] S. Sreekumar and D. Gündüz, "Optimal privacy-utility trade-off under a rate constraint," in *IEEE ISIT*, 2019.

[38] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE TIFS*, 2019.

[39] C. E. Shannon, "Communication theory of secrecy systems," *Bell system technical journal*, vol. 28, no. 4, pp. 656–715, 1949.

[40] H. Edwards and A. Storkey, "Censoring representations with an adversary," in *ICLR*, 2016.

[41] J. Hamm, "Enhancing utility and privacy with noisy minimax filters," in *IEEE ICASSP*, 2017.

[42] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, no. 12, 2017.

[43] C. T. Li and A. El Gamal, "Extended Gray–Wyner system with complementary causal side information," *IEEE Transactions on IT*, 2017.

[44] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, and G. Sapiro, "Adversarially learned representations for information obfuscation and inference," in *ICML*, 2019.

[45] B. Razeghi, F. P. Calmon, D. Gündüz, and S. Voloshynovskiy, "On perfect obfuscation: Local information geometry analysis," in *IEEE WIFS*, 2020.

[46] A. A. Atashin, B. Razeghi, D. Gündüz, and S. Voloshynovskiy, "Variational leakage: The role of information complexity in privacy leakage," in *ACM Workshop on Wireless Security and Machine Learning*, 2021.

[47] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[48] T. Andre, M. Antonini, M. Barlaud, and R. M. Gray, "Entropy-based distortion measure for image coding," in *IEEE ICIP*, 2006.

[49] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *IEEE ISIT*, 2007.

[50] T. A. Courtade and R. D. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *IEEE ISIT*, 2011.

[51] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[52] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes for ECE563 (UIUC) and*, vol. 6, 2014.

[53] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," *Topics in information theory*, 1977.

[54] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *studia scientiarum Mathematicarum Hungarica*, 1967.

[55] J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, 1986.

[56] M. Vera, P. Piantanida, and L. R. Vega, "The role of the information bottleneck in representation learning," in *IEEE ISIT*, 2018.

[57] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," *Neural computation*, 2017.

[58] I. Fischer, "The conditional entropy bottleneck," *arXiv preprint arXiv:2002.05379*, 2020.

[59] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, "A variational approach to privacy and fairness," *arXiv:2006.06332*, 2020.

[60] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *ICML*, 2015.

[61] J. Tomczak and M. Welling, "VAE with a VampPrior," in *AISTATS*, 2018.

[62] M. Bauer and A. Mnih, "Resampled priors for variational autoencoders," in *AISTATS*, 2019.

[63] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, 2010.

[64] M. Sugiyama, T. Suzuki, and T. Kanamori, "Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation," *Ann. of the Institute of Statistical Mathematics*, 2012.

[65] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[66] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.

[67] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis *et al.*, "Tensorflow: A system for large-scale machine learning," in *USENIX*, 2016.

[68] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio *et al.*, "Mutual information neural estimation," in *ICML*, 2018.

[69] R. W. Yeung, "A new outlook on shannon's information measures," *IEEE transactions on information theory*, 1991.

[70] Y. Steinberg, "Coding and common reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4995–5010, 2009.

[71] M. Benammar and A. Zaidi, "Rate-distortion function for a heegard-berger problem with two sources and degraded reconstruction sets," *IEEE Transactions on Information Theory*, 2016.

[72] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *NIPS*, 2016.

[73] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[74] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in ML*, 2019.

[75] O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf, "From optimal transport to generative modeling: the VEGAN cookbook," *arXiv:1705.07642*, 2017.