SDP Methods for Sensitivity-Constrained Privacy Funnel and Information Bottleneck Problems

Yuheng Bu Tony Wang Gregory W. Wornell
Massachusetts Institute of Technology, Cambridge, MA 02139
Email: buyuheng@mit.edu, twang6@mit.edu, gww@mit.edu

***Abstract*—We generalize the information bottleneck (IB) and privacy funnel (PF) problems by introducing the notion of a sensitive attribute, which arises in a growing number of applications. In this generalization, we seek to construct representations of observations that are maximally (or minimally) informative about a target variable, while also satisfying constraints with respect to a variable corresponding to the sensitive attribute. In the Gaussian and discrete settings, we show that by suitably approximating the Kullback-Liebler (KL) divergence defining traditional Shannon mutual information, the generalized IB and PF problems can be formulated as semi-definite programs (SDPs), and thus efficiently solved, which is important in applications of high-dimensional inference. We validate our algorithms on synthetic data and demonstrate their use in imposing fairness in machine learning on real data as an illustrative application.**

I. Introduction

The Information Bottleneck (IB) problem introduced in [1] seeks to construct a representation U from observation X that is maximally informative about a target Y but minimally informative about X; specifically,

$$\max_{P_{U|X}} I(Y;U) \qquad \text{s.t.} \qquad I(X;U) \leq \epsilon. \tag{1}$$

A dual to the IB problem is the privacy funnel (PF) problem introduced in [2], which takes the form

$$\min_{P_{U|X}} I(Y;U) \qquad \text{s.t.} \qquad I(X;U) \geq R. \tag{2}$$

In this work, we generalize the IB and PF problems by introducing a sensitive attribute E, such that we are optimizing $P_{U|X}$ in the undirected graphical model depicted in Fig. 1 rather than in the (simpler) Markov chain $Y \leftrightarrow X \leftrightarrow U$.

Such generalizations capture a wide variety of privacy and fairness constraints that arise in existing and emerging machine learning applications [3]. For example, in the criminal justice system, predictions about the chance of recidivism of a convicted criminal (Y) given information such as the nature of the crimes and the number of prior arrests (X) may be overly correlated with race (E), which may violate privacy and fairness criteria [4], [5]. To avoid such issues, we could first construct a representation U from X that satisfies information-theoretic privacy/fairness constraints. Any inferences based on U would then automatically satisfy those constraints.

This work was supported, in part, by NSF under Grant CCF-1717610 and by the MIT-IBM Watson AI Lab.

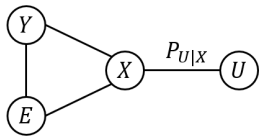

Fig. 1. Graphical model for the generalized IB and PF problems.

The generalized IB problem also arises naturally when seeking to generate domain-invariant features in transfer learning [6] and domain generalization [7]. In these settings, the goal is to construct new features U from the observation X that are informative about the label Y while being invariant across different domains E. As shown in [8], by imposing suitable independence constraints between U and E, inferences based on U will generalize well across the different domains.

Despite their practical importance, such IB/PF problems are inherently non-convex in general, making solutions difficult to obtain in general. Indeed, as shown in [1], [9], a closed-form solution to the original IB problem is only available in the binary or Gaussian case. For general distributions, variational methods based on Lagrangian functions of the IB problem have also been developed in [10], [11]. For the original PF problem, the greedy algorithm proposed by [2] and the submodularity-based clustering algorithm in [12] can only construct a deterministic transition $P_{U|X}$ by merging the alphabet of X.

In this paper, we use reparameterization techniques developed in [13] to show that by suitably approximating the Kullback-Liebler (KL) divergence defining mutual information, the generalized IB and PF problems can be formulated as semi-definite programs (SDPs). In both the Gaussian and discrete settings, we provide efficient SDP-based algorithms for solving the generalized IB and PF problems. We validate our algorithms on synthetic data and show how they can be used to address machine learning fairness issues involving real data.

A. Notation

Bold capital letters represent matrices (e.g. $\boldsymbol{A}$), and bold lower case letters represent vectors (e.g. $\boldsymbol{a}$). Capital calligraphic letters denote sets, and $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$. If P_X is a discrete probability mass function, $\mathbf{P}_X \in \mathbb{R}^{|\mathcal{X}|}$ is its column vector representation. Likewise $\mathbf{P}_{X,Y} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ is the matrix representation of $P_{X,Y}$.

Finally, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and $\|\cdot\|_{\mathrm{s}}$ denotes the spectral norm.

II. The generalized IB and PF problems

Let the observation X, target variable Y, and sensitive attribute E be random variables defined on alphabets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{E}$, respectively, with a known joint distribution $P_{X,Y,E}$. We want to construct a stochastic mapping $P_{U|X}(u|x) : \mathcal{X} \to \mathcal{U}$ from the observation X to a representation U satisfying certain requirements.

We first consider the generalized IB problem, where the goal is to construct U so that it is *maximally* informative about Y under (one of) the following constraints

$$\max_{P_{U|X}} I(Y;U) \quad \text{s.t.} \quad \begin{aligned} &\text{(IB-1).} \quad I(X;U) \le \epsilon; \\ &\text{(IB-2).} \quad I(E;U) \le \epsilon; \\ &\text{(IB-3).} \quad I(E;U|Y) \le \epsilon. \end{aligned}$$

Among these constraints, (IB-1) corresponds to the original IB problem, (IB-2) constrains how informative U can be about the sensitive attribute E, and (IB-3) promotes conditional independence between U and E given Y.

With respect to the existing literature, we note that [14] considers an inherently different generalized IB problem in which the standard IB problem is replaced with one using f-divergences instead of KL divergence. In fact, our generalized IB problem is more closely related to existing work on "state masking" ([15], [16]), where the goal is also to minimize the information leak about the channel state (E). However the model in state-masking formulation is much simpler than that of Fig. 1, as the transmitted symbols Y are independent of E in the former.

The PF counterparts of the generalized IB problem are as follows. In this case, the goal is to construct U so that it is *minimally* informative about Y under (one of) the following constraints

$$\min_{P_{U|X}} I(Y;U) \quad \text{s.t.} \quad \begin{aligned} &\text{(PF-1).} \quad I(X;U) \ge R; \\ &\text{(PF-2).} \quad I(E;U) \ge R. \end{aligned}$$

Note that $I(Y;U)$ is convex with respect to $P_{U|Y}$, and $P_{U|Y}$ is linear in $P_{U|X}$. This means the objective $I(Y;U)$ is convex in $P_{U|X}$. Thus, the IB problem is not convex, since it maximizes a convex function. Likewise, the PF problem is not convex due to the non-convex constraint $I(X;U) \ge R$.

While the generalized IB and PF problems are nonconvex, we show that there exist closely related problems that are convex. Specifically, in the Gaussian and discrete settings, we show that by approximating the standard KL divergence via suitable second-order Taylor series, the generalized IB and PF problems can be formulated as SDPs which can be solved efficiently.

III. Gaussian case

In this section, we assume the observation $X \in \mathbb{R}^{d_X}$, target $Y \in \mathbb{R}^{d_Y}$, and sensitive attribute $E \in \mathbb{R}^{d_E}$ are jointly Gaussian random variables. We first introduce a new divergence measure $\bar{D}$ from [13], along with its corresponding $\bar{I}$-information measure. We show that under $\bar{I}$-information, the IB and PF problem can be solved efficiently via SDPs.

A. Local Gaussian information geometry

The following $\bar{D}$ divergence was introduced in [13] as a second-order approximation for the KL divergence between two Gaussian distributions.

Definition 1. *The $\bar{D}$-divergence between Gaussian distributions $P = \mathcal{N}(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ and $Q = \mathcal{N}(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$ is*

$$\begin{aligned} \bar{D}(P\|Q) \triangleq\ & (\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_Q^{-1} (\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q) \\ & + \frac{1}{2} \left\| \boldsymbol{\Sigma}_Q^{-1/2} (\boldsymbol{\Sigma}_P - \boldsymbol{\Sigma}_Q) \boldsymbol{\Sigma}_Q^{-1/2} \right\|_{\mathrm{F}}^2. \end{aligned} \tag{3}$$

In turn, for jointly Gaussian random variables X, Y, their $\bar{I}$-information is $\bar{I}(X;Y) \triangleq \bar{D}(P_{X,Y}\|P_X P_Y)$. Similarly, for jointly Gaussian X, Y, Z, the conditional $\bar{I}$-information is $\bar{I}(X;Y|Z) \triangleq \mathbb{E}_{P_Z}\left[\bar{D}(P_{X,Y|Z}\|P_{X|Z}P_{Y|Z})\right]$.

For simplicity of exposition, we restrict our attention to zero-mean Gaussian random variables, since the means of Gaussian distributions affects neither their mutual information nor $\bar{I}$-information. So in particular X, Y, E are assumed to have zero mean.

B. Canonical correlation matrices

When analyzing $\bar{I}$-information, it is convenient to their equivalent representation via canonical correlation matrices (CCMs) instead of covariance matrices. We begin with the definition.

Definition 2. *The canonical correlation matrix (CCM) $\widetilde{\mathbf{B}}_{X,Y} \in \mathbb{R}^{d_X \times d_Y}$ between jointly Gaussian variables X and Y is given by*

$$\widetilde{\mathbf{B}}_{X,Y} \triangleq \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_Y^{-1/2}, \tag{4}$$

where $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ are the covariance matrices of X and Y, and $\boldsymbol{\Sigma}_{XY} = \mathbb{E}[XY^T]$.

The value of CCMs in representing $\bar{I}$-information is expressed via the following lemma from [13].

Lemma 1. *[13, Lemma 68] For jointly Gaussian random variables X and Y,*

$$\bar{I}(X;Y) = \left\|\widetilde{\mathbf{B}}_{X,Y}\right\|_{\mathrm{F}}^2. \tag{5}$$

In addition, CCMs also satisfy a simple chain rule:

Lemma 2. *[13, Fact 79] Suppose $X \leftrightarrow Y \leftrightarrow Z$ is a Markov chain of jointly Gaussian variables, then*

$$\widetilde{\mathbf{B}}_{X,Z} = \widetilde{\mathbf{B}}_{X,Y} \widetilde{\mathbf{B}}_{Y,Z}. \tag{6}$$

Finally the correspondence between joint Gaussian distributions and CCMs is given by the following lemma:

Lemma 3. *[13, Fact 60] Let $\boldsymbol{\Sigma}_X \in \mathbb{R}^{d_X \times d_X}$ and $\boldsymbol{\Sigma}_Y \in \mathbb{R}^{d_Y \times d_Y}$ be two positive-definite matrices, and let $\mathbf{M} \in \mathbb{R}^{d_X \times d_Y}$. There exist jointly Gaussian random variables X and Y with covariances $\boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y$ and canonical correlation matrix $\widetilde{\mathbf{B}}_{X,Y} = \mathbf{M}$ if and only if $\|\mathbf{M}\|_{\mathrm{s}} \le 1$.*

C. CCM-parameterizations of the IB and PF problems

From Fig. 1, we see that $Y \leftrightarrow X \leftrightarrow U$ is a Markov chain. By Lemma 1 and Lemma 2, this means

$$\bar{I}(Y;U) = \left\|\widetilde{\mathbf{B}}_{Y,X}\widetilde{\mathbf{B}}_{X,U}\right\|_{\mathrm{F}}^2. \tag{7}$$

Thus, the $\bar{I}$-information IB and PF problems can be written as optimization problems with respect to $\widetilde{\mathbf{B}}_{X,U}$ as

$$\begin{aligned} &\max_{\widetilde{\mathbf{B}}_{X,U}} \|\widetilde{\mathbf{B}}_{Y,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2, \\ \text{s.t.}\quad &\text{(IB-1).}\quad \|\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 \le \epsilon; \\ &\text{(IB-2).}\quad \|\widetilde{\mathbf{B}}_{E,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 \le \epsilon, \end{aligned} \tag{8}$$

and

$$\begin{aligned} &\min_{\widetilde{\mathbf{B}}_{X,U}} \|\widetilde{\mathbf{B}}_{Y,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2, \\ \text{s.t.}\quad &\text{(PF-1).}\quad \|\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 \ge R; \\ &\text{(PF-2).}\quad \|\widetilde{\mathbf{B}}_{E,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 \ge R. \end{aligned} \tag{9}$$

The (IB-3) constraint is more challenging to accommodate, as it cannot in general be written in terms of $\widetilde{\mathbf{B}}_{X,U}$ in a way that is amenable to SDP. However, in certain weak-dependence regimes, there exist SDP-friendly inequalities that accurately approximate the (IB-3) constraint. We develop this approach below.

Lemma 4. *[13, Corollary 73] Let X and Y be δ-dependent Gaussian random variables, i.e., $\bar{I}(X;Y) \le \delta$, then*

$$I(X;Y) = \frac{1}{2}\bar{I}(X;Y) + o(\delta^2). \tag{10}$$

Suppose that $\bar{I}(X;U) \le \delta$. Then $\bar{I}(E,Y;U) \le \delta$ due to the data processing inequality, which means

$$\begin{aligned} I(E;U|Y) &= I(E,Y;U) - I(Y;U) \\ &\approx \frac{1}{2}\bar{I}(E,Y;U) - \frac{1}{2}\bar{I}(Y;U) \\ &= \frac{1}{2}\|\widetilde{\mathbf{B}}_{EY,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 - \frac{1}{2}\|\widetilde{\mathbf{B}}_{Y,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2, \end{aligned} \tag{11}$$

where the first equality follows by the chain rule of mutual information, and $\widetilde{\mathbf{B}}_{EY,X} \in \mathbb{R}^{(d_E+d_Y)\times d_X}$ denotes the CCM between $\left[\begin{smallmatrix}E\\Y\end{smallmatrix}\right]$ and X.

Thus, when X and U are weakly dependent, the $\bar{I}$-information IB problem with an (IB-3) constraint can be approximated via

$$\begin{aligned} &\max_{\widetilde{\mathbf{B}}_{X,U}} \|\widetilde{\mathbf{B}}_{Y,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2, \\ \text{s.t.}\quad &\text{(IB-3).}\quad \begin{cases} \|\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 \le \delta, \\ \|\widetilde{\mathbf{B}}_{EY,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 - \|\widetilde{\mathbf{B}}_{Y,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2 \le \epsilon. \end{cases} \end{aligned} \tag{12}$$

D. SDPs for the Gaussian IB and PF problems

We are now ready to solve the CCM-parameterized optimization problems given in the previous section using semidefinite programming. The optimization problems in (8), (9) and (12) are all equivalent to SDPs. Due to space constraints, we show this only for (IB-2), as a representative example.

Theorem 1. *The optimization problem in (8) with the (IB-2) constraint is equivalent to the following SDP:*

$$\begin{aligned} \max_{\boldsymbol{A}\in\mathbb{S}^{d_X}} \ &\mathrm{tr}\left(\widetilde{\mathbf{B}}_{Y,X}^\top\widetilde{\mathbf{B}}_{Y,X}\boldsymbol{A}\right) \quad \text{s.t.} \quad \mathrm{tr}\left(\widetilde{\mathbf{B}}_{E,X}^\top\widetilde{\mathbf{B}}_{E,X}\boldsymbol{A}\right) \le \epsilon, \\ &\qquad\qquad\qquad\qquad\quad \mathbf{0} \preceq \boldsymbol{A} \preceq \boldsymbol{I}, \end{aligned} \tag{13}$$

where $\mathbb{S}^{d_X}$ is the space of $d_X \times d_X$ symmetric matrices.

Proof: First, via standard SDP theory ([17]), the problem in (13) can be easily verified to be a valid SDP. It remains to show that (13) is equivalent to (8) with the (IB-2) constraint. The key idea is to let $\boldsymbol{A} = \widetilde{\mathbf{B}}_{X,U}\widetilde{\mathbf{B}}_{X,U}^\top$. Then

$$\mathrm{tr}(\widetilde{\mathbf{B}}_{Y,X}^\top\widetilde{\mathbf{B}}_{Y,X}\boldsymbol{A}) = \|\widetilde{\mathbf{B}}_{Y,X}\widetilde{\mathbf{B}}_{X,U}\|_{\mathrm{F}}^2, \tag{14}$$

so we are maximizing the same objective as in (8). Similarly, we can show the (IB-2) constraint is equivalent to $\mathrm{tr}(\widetilde{\mathbf{B}}_{E,X}^\top\widetilde{\mathbf{B}}_{E,X}\boldsymbol{A}) \le \epsilon$.

By Lemma 3, $\widetilde{\mathbf{B}}_{Y,X}$ is a valid CCM precisely when $\|\widetilde{\mathbf{B}}_{Y,X}\|_s \le 1$. So $\boldsymbol{A}$ is the Gramian matrix of a CCM when $\mathbf{0} \preceq \boldsymbol{A} \preceq \boldsymbol{I}$, which means a solution to (13) will correspond to the Gramian matrix of a solution to (8) with the (IB-2) constraint.

Once we obtain an optimal solution $\boldsymbol{A}^*$ to (13), we convert $\boldsymbol{A}^*$ to an optimal CCM $\widetilde{\mathbf{B}}_{X,U}^*$ via the following procedure:

1) Compute an eigen-decomposition $\boldsymbol{A}^* = \sum_{i=1}^{d_U}\lambda_i\boldsymbol{v}_i\boldsymbol{v}_i^\top$, where λ_i are eigenvalues, $\boldsymbol{v}_i$ are eigenvectors, and d_U denotes the number of nonzero eigenvalues.
2) Construct an optimal CCM as $\widetilde{\mathbf{B}}_{X,U}^* = \sum_{i=1}^{d_U}\sqrt{\lambda_i}\boldsymbol{v}_i\boldsymbol{u}_i^\top$, where the $\boldsymbol{u}_i$ are any orthonormal basis of $\mathbb{R}^{d_U}$.

■

Note that we can also easily convert an optimal CCM to an optimal covariance matrix if desired. Via Lemma 3, the covariance matrix $\boldsymbol{\Sigma}_U$ can be set arbitrarily, so we let $\boldsymbol{\Sigma}_U = \boldsymbol{I}$ and get $\boldsymbol{\Sigma}_{X,U}^* = \boldsymbol{\Sigma}_X^{1/2}\widetilde{\mathbf{B}}_{X,U}^*$.

The proof of Theorem 1 can easily be extended to show that all our $\bar{I}$-information IB and PF problems are special cases of the general SDP

$$\begin{aligned} \min_{\boldsymbol{A}\in\mathbb{S}^{d_X}} \ &\mathrm{tr}(\boldsymbol{C}_0\boldsymbol{A}) \qquad \text{s.t.} \quad \mathrm{tr}(\boldsymbol{C}_1\boldsymbol{A}) \le \epsilon, \\ &\qquad\qquad\qquad\quad 0 \preceq \boldsymbol{A} \preceq \boldsymbol{I}, \end{aligned}$$

where $\boldsymbol{C}_0 = -\widetilde{\mathbf{B}}_{Y,X}^\top\widetilde{\mathbf{B}}_{Y,X}$ for IB problems, $\boldsymbol{C}_0 = \widetilde{\mathbf{B}}_{Y,X}^\top\widetilde{\mathbf{B}}_{Y,X}$ for PF problems, and

$$\begin{aligned} &\text{(IB-1).}\quad \boldsymbol{C}_1 = \boldsymbol{I}, \\ &\text{(IB-2).}\quad \boldsymbol{C}_1 = \widetilde{\mathbf{B}}_{E,X}^\top\widetilde{\mathbf{B}}_{E,X}, \\ &\text{(IB-3).}\quad \boldsymbol{C}_1 = \widetilde{\mathbf{B}}_{EY,X}^\top\widetilde{\mathbf{B}}_{EY,X} - \widetilde{\mathbf{B}}_{E,X}^\top\widetilde{\mathbf{B}}_{E,X}, \\ &\text{(PF-1).}\quad \boldsymbol{C}_1 = -\boldsymbol{I}, \\ &\text{(PF-2).}\quad \boldsymbol{C}_1 = -\widetilde{\mathbf{B}}_{E,X}^\top\widetilde{\mathbf{B}}_{E,X}. \end{aligned}$$

For (IB-3), since the approximation is only valid in weakly-dependent regime, the constraint $\mathrm{tr}(\boldsymbol{A}) \le \delta$ is also needed.

IV. Discrete Case

In this section, we assume the observation X, target variable Y, and sensitive attribute E are discrete random variables defined on alphabets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{E}$ respectively. Without loss of generality, we assume the marginals P_X, P_Y and P_E are strictly positive, since otherwise symbols with zero probability mass may be removed from their respective alphabets.

Similar to the Gaussian case, we show that when χ^2-information is used to approximate mutual information, the IB and PF problems have efficient SDP-based solutions. The SDP-based solutions are constructed from convenient matrix representations of discrete distributions.

A. χ^2-Information

Analogous to the way $\bar{I}$-information is a second-order approximation to mutual information in the Gaussian case, χ^2-information is a second-order approximation to mutual information in the discrete case.

Definition 3. *The χ^2-divergence between two discrete distributions P_X and Q_X is given by*

$$D_{\chi^2}(P_X\|Q_X) \triangleq \sum_{x\in\mathcal{X}} \frac{(Q_X(x)-P_X(x))^2}{Q_X(x)}. \tag{15}$$

We can define the χ^2-information between X and Y as

$$I_{\chi^2}(X;Y) \triangleq D_{\chi^2}(P_{X,Y}\|P_XP_Y), \tag{16}$$

and conditional and the conditional χ^2-information as,

$$I_{\chi^2}(X;Y|Z) \triangleq \mathbb{E}_{P_Z}[D_{\chi^2}(P_{X,Y|Z}\|P_{X|Z}P_{Y|Z})]. \tag{17}$$

B. Divergence transfer and canonical dependence matrices

Analogous to the way the CCMs are convenient representations when analyzing $\bar{I}$-information, divergence transfer matrices (DTMs) and canonical dependence matrices (CDMs) are convenient parameterizations for working χ^2-information. We describe these matrices below.

Definition 4. *The divergence transfer matrix (DTM) $\mathbf{B}_{X,Y} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ associated with joint distribution $P_{X,Y}$ is given by*

$$\mathbf{B}_{X,Y}(x,y) \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}. \tag{18}$$

While DTMs will prove to be very useful, when working with χ^2-information the more fundamental object is the canonical dependence matrix.

Definition 5. *The canonical dependence matrix (CDM) $\widetilde{\mathbf{B}}_{X,Y} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ associated with joint distribution $P_{X,Y}$ is given by*

$$\widetilde{\mathbf{B}}_{X,Y}(x,y) = \mathbf{B}_{X,Y}(x,y) - \sqrt{P_X(x)}\sqrt{P_Y(x)}. \tag{19}$$

The CDM captures the joint distribution between two discrete random variables while ignoring their marginals. It serves for discrete random variables the role that the CCM serves for Gaussian random variables. Similar to Lemma 1, the following lemma shows that the χ^2-information can be fully determined by the CDM.

Lemma 5. *For discrete random variables X and Y,*

$$I_{\chi^2}(X;Y) = \left\|\widetilde{\mathbf{B}}_{X,Y}\right\|_{\mathrm{F}}^2. \tag{20}$$

The correspondence between CDMs and joint distributions is characterized by the following lemma derived from [13].

Lemma 6. *[13, Proposition 97] Let P_X and P_Y be strictly positive marginals, and let $\mathbf{M} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$. There exists a strictly positive joint distribution $P_{X,Y}$ with marginals P_X, P_Y and CDM $\widetilde{\mathbf{B}}_{X,Y} = \mathbf{M}$ if and only if*

$$\|\mathbf{M}\|_{\mathrm{s}} \le 1, \quad \sqrt{\mathbf{P}_X}^{\top}\mathbf{M} = \mathbf{0}, \quad \mathbf{M}\sqrt{\mathbf{P}_Y} = \mathbf{0}, \tag{21}$$

$$\text{and} \quad \mathbf{M}(x,y) + \sqrt{P_X(x)}\sqrt{P_Y(y)} > 0. \tag{22}$$

Above, $\sqrt{\mathbf{P}_X}$ and $\sqrt{\mathbf{P}_Y}$ are the coordinate-wise square roots of $\mathbf{P}_X$ and $\mathbf{P}_Y$. This association between strictly positive joint distribution and CDMs is bijective and continuous.

Finally, we show that the following chain rule holds for DTMs and CDMs analogous to Lemma 2.

Lemma 7. *Let $X \leftrightarrow Y \leftrightarrow Z$ be a Markov chain of discrete random variables, then*

$$\widetilde{\mathbf{B}}_{X,Z} = \mathbf{B}_{X,Y}\widetilde{\mathbf{B}}_{Y,Z} = \widetilde{\mathbf{B}}_{X,Y}\mathbf{B}_{Y,Z}. \tag{23}$$

C. SDP formulation

In the discrete case, the χ^2-information IB and PF problems can be written as optimization problems with respect to $\widetilde{\mathbf{B}}_{X,U}$ by Lemma 5 and Lemma 7. If we further denote $\boldsymbol{A} = \widetilde{\mathbf{B}}_{X,U}\widetilde{\mathbf{B}}_{X,U}^{\top}$, we can also reformulate these optimization problems as a SDP of the form

$$\min_{\boldsymbol{A}} \operatorname{tr}(\boldsymbol{C}_0\boldsymbol{A}) \quad \text{s.t.} \quad \begin{aligned} &\operatorname{tr}(\boldsymbol{C}_1\boldsymbol{A}) \le \epsilon, \\ &\operatorname{tr}\left(\sqrt{\mathbf{P}_X}\sqrt{\mathbf{P}_X}^{\top}\boldsymbol{A}\right) = 0, \\ &0 \preceq \boldsymbol{A} \preceq \boldsymbol{I}, \end{aligned} \tag{24}$$

where $\boldsymbol{C}_0 = -\mathbf{B}_{Y,X}^{\top}\mathbf{B}_{Y,X}$ for IB problems, $\boldsymbol{C}_0 = \mathbf{B}_{Y,X}^{\top}\mathbf{B}_{Y,X}$ for PF problems, and

$$\begin{aligned} &\text{(IB-1).} \quad \boldsymbol{C}_1 = \boldsymbol{I}, \\ &\text{(IB-2).} \quad \boldsymbol{C}_1 = \mathbf{B}_{E,X}^{\top}\mathbf{B}_{E,X}, \\ &\text{(IB-3).} \quad \boldsymbol{C}_1 = \mathbf{B}_{E\otimes Y,X}^{\top}\mathbf{B}_{E\otimes Y,X} - \mathbf{B}_{E,X}^{\top}\mathbf{B}_{E,X}, \\ &\text{(PF-1).} \quad \boldsymbol{C}_1 = -\boldsymbol{I}, \\ &\text{(PF-2).} \quad \boldsymbol{C}_1 = -\mathbf{B}_{E,X}^{\top}\mathbf{B}_{E,X}. \end{aligned} \tag{25}$$

In (25), $E \otimes Y$ is the Cartesian product of E and Y. Additionally, like the Gaussian case, (25) is just an approximation of the original (IB-3) constraint and only holds when U is weakly-dependent on X. Thus we also need the constraint $\operatorname{tr}(\boldsymbol{A}) \le \delta$ for (25).

Now strictly speaking, the SDP formulation in (24) is a *relaxation* of the χ^2-information IB and PF problems. This is because the solution of (24) does not meet all the conditions of Lemma 6. In particular, while the conditions in (21) are captured by the last two constraints of (24), the positive matrix constraint in (22) is not enforced.

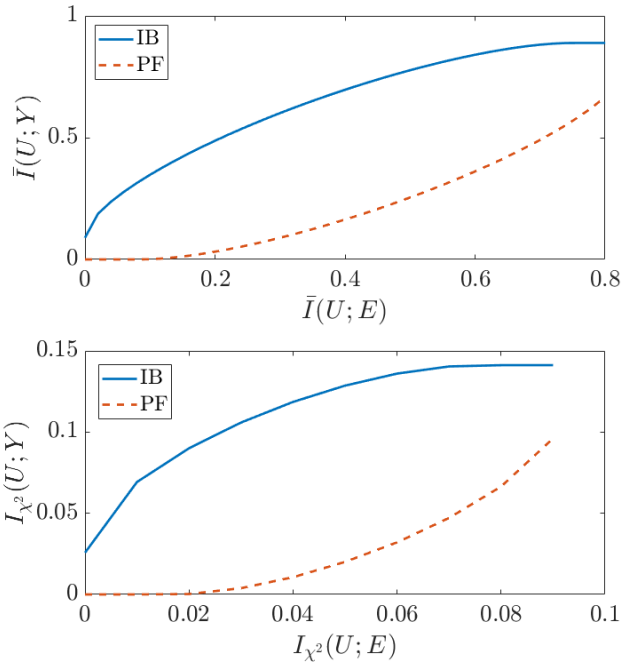

Fig. 2. The top plot (Gaussian setting) shows $\bar{I}(U;Y)$ plotted against $\bar{I}(U;E)$ with $d_X = 3$ and $d_Y = d_E = 1$. The bottom plot (discrete setting) shows $I_{\chi^2}(U;Y)$ plotted against $I_{\chi^2}(U;E)$ with $|\mathcal{X}| = 18$ and $|\mathcal{Y}| = |\mathcal{E}| = 3$.

The effect of this relaxation is that converting from $\boldsymbol{A}$ to $P_{U|X}$ may result in a conditional distribution with negative entries. However, we show in Section V that clipping negative entries to zero does not have a large impact on solution quality.

Another way to resolve the negative issue is to limit the norm of $\widetilde{\mathbf{B}}_{X,U}$ enough so that (21) is always satisfied. This is equivalent to adding a constraint of the form $\mathrm{tr}(\boldsymbol{A}) \leq \delta$.

Once we obtain an optimal solution $\boldsymbol{A}^*$ to the SDP in (24), we can convert $\boldsymbol{A}^*$ to $P^*_{U|X}$ via the following procedure:

1) Compute an eigen-decomposition $\boldsymbol{A}^* = \sum_{i=1}^{k-1} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$, where λ_i are non-zero eigenvalues, $\boldsymbol{v}_i$ are eigenvectors, and k denotes the number of non-zero eigenvalues.
2) The alphabet and marginal distribution of U can be set semi-arbitrarily, and here we let $|\mathcal{U}| = k$ and let $P_U(u) = 1/k$ be uniform.
3) Construct the CDM $\widetilde{\mathbf{B}}^*_{X,U} = \sum_{i=1}^{k-1} \sqrt{\lambda_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top$, where $\{\sqrt{\mathbf{P}_U}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}\}$ form a orthonormal basis of $\mathbb{R}^k$.
4) Construct the DTM $\mathbf{B}^*_{X,U} = \widetilde{\mathbf{B}}^*_{X,U} + \sqrt{\mathbf{P}_X}\sqrt{\mathbf{P}_U}^\top$.
5) Construct the conditional distribution

$$P^*_{U|X}(u|x) = P_X(x)^{-1/2} \cdot \mathbf{B}^*_{X,U}(x,u) \cdot P_U(u)^{1/2}.$$

V. Numerical Results

A. Synthetic data

We test our SDP formulations of (IB-2) and (PF-2) in both the Gaussian and discrete cases with a randomly generated joint distribution $P_{X,Y,E}$. We use the default SDP solver in CVX [18], [19]. By changing the value of ϵ and R in the constraints, we can obtain the maximum and minimum of $\bar{I}(U;Y)$ and $I_{\chi^2}(U;Y)$ for a given $\bar{I}(U;E)$ and $I_{\chi^2}(U;E)$.

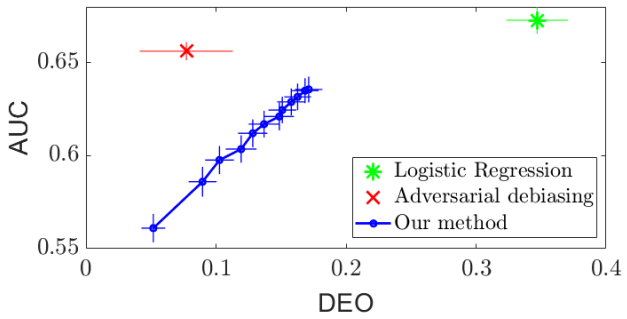

Fig. 3. Results of COMPAS with AUC plotted against DEO. Error bars are estimated from 200 Monte Carlo trials.

We plot our results in Fig. 2. The top right corner of each plot can be achieved by letting $U = X$, and $\bar{I}(U;E) \leq \bar{I}(X;E)$, $\bar{I}(U;Y) \leq \bar{I}(X;Y)$ due to data processing inequality. We also note that the achievable region of $(\bar{I}(U;E), \bar{I}(U;Y))$ is a convex set. This is a consequence of our SDP formulation, and this convexity can be proved formally.

B. COMPAS data

We also apply our SDP formulation in (24) for (IB-3) to a fairness task. We use ProPublica's COMPAS recidivism dataset [4], which contains categorical features and has been used in prior works [5], [20]. The goal of this dataset is to predict whether an individual recidivated (re-offended) (Y) using the severity of charge, number of prior crimes, and age category as the observation variables (X). As discussed in [21], COMPAS scores are biased against African-Americans, so race is set to be the sensitive attribute (E) and filtered to contain only Caucasian and African-American individuals.

We randomly generate a 80-20 train/test split, and estimate the DTMs $\hat{\mathbf{B}}_{E,X}$ and $\hat{\mathbf{B}}_{Y,X}$ using the empirical distributions of the training set. We then run our SDP algorithm to construct $\hat{P}_{U|X}$. Given a test observation x, we sample u from $\hat{P}_{U|X}(\cdot|x)$ and predict $\hat{y}$ using the maximum a posteriori (MAP) rule: $\hat{y} = \arg\max_{y \in \mathcal{Y}} \hat{P}_{Y|U}(y|u)$.

We compare the performance of the proposed (IB-3) algorithm with other two baselines: naïve logistic regression, and the adversarial debiasing method in [22] (implementation given in [23]). In Fig. 3, we plot the area under ROC curve (AUC) of $\hat{Y}$ generated different algorithms against the difference in equalized opportunities (DEO)

$$\mathrm{DEO} = \mathbb{P}(\hat{Y}{=}1|E{=}1, Y{=}1) - \mathbb{P}(\hat{Y}{=}1|E{=}0, Y{=}1),$$

which is a standard fairness measure used commonly in the literature [3]. A small DEO is equivalent to the conditional independence constraint $I_{\chi^2}(E;U|Y) \leq \epsilon$ in (IB-3). Although it might appear from Fig. 3 that adversarial debiasing is a superior approach, this is due to the fact that adversarial debiasing uses all available decision variables, while our algorithm only uses the decision variables that are discrete. In addition, our algorithm provides a smooth trade-off curve between performance and DEO, so that a desired level of fairness can be achieved by setting ϵ in practice.

REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[2] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. Information Theory Workshop (ITW)*. IEEE, 2014, pp. 501–505.

[3] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, 2016. [Online]. Available: https://github.com/propublica/compas-analysis

[5] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[6] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[7] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Proc. International Conference on Machine Learning (ICML)*. PMLR, 2013, pp. 10–18.

[8] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[9] G. Chechik, A. Globerson, N. Tishby, Y. Weiss, and P. Dayan, "Information bottleneck for Gaussian variables." *J. Mach. Learn. Res.*, vol. 6, no. 1, 2005.

[10] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[11] A. Zaidi, I. Estella-Aguerri *et al.*, "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 2, p. 151, 2020.

[12] N. Ding and P. Sadeghi, "A submodularity-based clustering algorithm for the information bottleneck and privacy funnel," in *Proc. Information Theory Workshop (ITW)*. IEEE, 2019, pp. 1–5.

[13] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," Preprint, 2019, http://allegro.mit.edu/~gww/unifeatures.

[14] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*. IEEE, 2018, pp. 531–535.

[15] N. Merhav and S. Shamai, "Information rates subject to state masking," *IEEE Trans. Inform. Theory*, vol. 53, no. 6, pp. 2254–2261, 2007.

[16] L. Wang and G. W. Wornell, "Communication subject to state obfuscation," in *Proc. Int. Zurich Seminar on Information and Communication (IZS)*. ETH Zurich, 2020, pp. 78–82.

[17] R. M. Freund, "Introduction to semidefinite programming (SDP)," *Massachusetts Institute of Technology*, pp. 8–11, 2004.

[18] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[19] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[20] J. Lee, Y. Bu, P. Sattigeri, R. Panda, G. Wornell, L. Karlinsky, and R. Feris, "A maximal correlation approach to imposing fairness in machine learning," *arXiv preprint arXiv:2012.15259*, 2020.

[21] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 3992–4001.

[22] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[23] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.