On Information Theoretic Fairness With A Bounded Point-Wise Statistical Parity Constraint: An Information Geometric Approach

Amirreza Zamani†, Ayfer Özgür‡, Mikael Skoglund†
†Division of Information Science and Engineering, KTH Royal Institute of Technology
‡Department of Electrical Engineering, Stanford
Email: amizam@kth.se, aozgur@stanford.edu, skoglund@kth.se

Abstract

In this paper, we study an information-theoretic problem of designing a fair representation under a bounded point-wise statistical (demographic) parity constraint. More specifically, an agent uses some useful data (database) X to solve a task T. Since both X and T are correlated with some latent sensitive attribute or secret S, the agent designs a representation Y that satisfies a bounded point-wise statistical parity, that is, such that for all realizations of the representation $y \in \mathcal{Y}$, we have $\chi^2(P_{S|y}; P_S) \leq \epsilon$. In contrast to our previous work, here we use the point-wise measure instead of a bounded mutual information, and we assume that the agent has no direct access to S and T; hence, the Markov chains $S - X - Y$ and $T - X - Y$ hold. In this work, we design Y that maximizes the mutual information $I(Y;T)$ about the task while satisfying a bounded compression rate constraint, that is, ensuring that $I(Y;X) \leq r$. Finally, Y satisfies the point-wise bounded statistical parity constraint $\chi^2(P_{S|y}; P_S) \leq \epsilon$.

When ϵ is small, concepts from information geometry allow us to locally approximate the KL-divergence and mutual information. To design the representation Y, we utilize this approximation and show that the main complex fairness design problem can be rewritten as a quadratic optimization problem that has simple closed-form solution under certain constraints. For the cases where the closed-form solution is not obtained we obtain lower bounds with low computational complexity. Here, we provide simple fairness designs with low complexity which are based on finding the maximum singular value and singular vector of a matrix. Finally, in a numerical example we compare our obtained results with the optimal solution.

I. Introduction

As shown in Fig. 1, in this paper, an agent wants to use some observable useful data (database) X to draw inferences or make some decision about a task T. Here, we assume that both the useful data and the task are correlated with some latent sensitive attribute or secret S. For example, the data X could represent a person's resume, the task T could determine whether this person should be employed in a position or not, and the sensitive attribute S could correspond to the person's gender or ethnicity.

It is important to ensure that decisions are not unfair and that inferences do not lead to privacy breaches [1], [2]. To address this, one possible solution is to design a *private* or *fair* representation Y of the database X. This representation should preserve as much information as possible about the task T while satisfying fairness or privacy criteria with respect to the latent sensitive attribute or secret [1]–[21]. As outlined in [1], [2], [18], privacy and fairness problems are very similar, and their intersections can be utilized in broad areas such as classification algorithms. More specifically, in [1], the design of representations Y that contain no information about the sensitive attribute S yielding $I(Y;S) = 0$, has been considered. In the *information-theoretic privacy* literature, the independence of disclosed data Y and private attribute or secret S is referred to as *perfect privacy* or *perfect secrecy* [10], [13]. In the *algorithmic fairness* literature, the independence of an algorithm's output (representation of the database) Y from the sensitive attribute S is known as *perfect demographic (or statistical) parity* [1], [2], [4]–[6]. As discussed in [9], [11], [13], [18], [22], [23], there are many cases where perfect privacy (secrecy) or perfect demographic (statistical) parity is not attainable (for example, [22, Example 1]), hence, in [22], a bounded statistical parity criteria is used, that is $I(S;Y) \leq \epsilon$. In this work, instead of bounded mutual information, we use bounded point-wise statistical parity, as in [11] for a privacy design problem. More specifically, we use $\chi^2(P_{S|y}; P_S) \leq \epsilon$ as the fairness criterion which must hold for each $y \in \mathcal{Y}$. Using a point-wise measure can be motivated by the definition of bounded statistical (demographic) parity, which is based on the difference between the conditional distributions of the outcome given the sensitive information [18], [24], [25]. In contrast to [1] and [22], here we assume that the latent sensitive attribute or secret S and the task T are not directly accessible to the agent, as shown in Fig. 1. In other words, here, we assume that the Markov chains $S - X - Y$ and $T - X - Y$ hold.

Many information theory problems face challenges due to the lack of a geometric structure in the space of probability distributions. Assuming the distributions of interest are close, we can approximate the KL-divergence as well as the mutual information by a weighted squared Euclidean distance, leading to a method where the main complex problems can be approximated. This approach has been used in [26], [27] for point-to-point channels and certain broadcast channels. Furthermore,

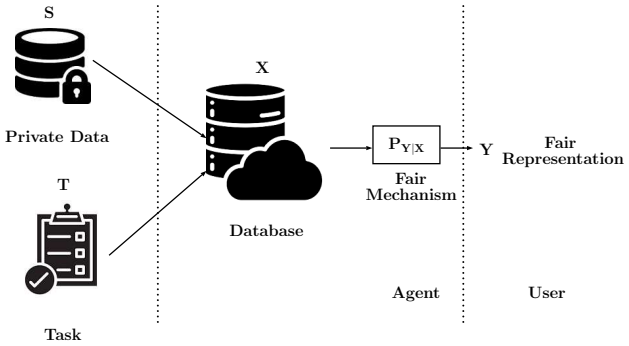

Fig. 1. Fair mechanism design under bounded point-wise statistical parity. The goal is to design Y using database X that is useful for the task T, is compressed, and leaks within a controlled threshold of the latent sensitive attribute S. Here, the agent has no direct access to S and T.

similar techniques have been used in the privacy setting in [9], [11], [12], [28]–[31]. Under the outlined fairness requirements of the representations, we consider a problem that solves a trade-off between utility, compression rate, and bounded point-wise statistical parity: To ensure that the representation Y has as much impact as possible, we maximize the information it contains about the task T. Moreover, we impose a minimum level of compression r to the output representation, that is, $I(Y;X) \leq r$.

Prior works on fairness mechanism design: The concept of *fair representations* was introduced in [6], marking a significant advancement in the field of algorithmic fairness through the capabilities of deep learning. Subsequent research has been predominantly characterized by adversarial [5], [6], [32] and variational [2], [33]–[35] methodologies. The theoretical exploration of the trade-offs between utility and fairness is further studied in [4]. In [2], the authors study an information-theoretic fairness problem. They introduce the CFB (Conditional Fairness Bottleneck), which quantifies the trades-off between the utility, fairness, and *compression* of the representations in terms of mutual information. This additional criterion has since been used in subsequent studies [7], [35], [36]. [7], [35], [36]. More specifically, in [2], the representation design is based on a variational approach. In contrast with [2], in [1], the authors propose a simple and constructive theoretical approach to design fair representations with perfect demographic (statistical) parity. They obtain upper and lower bounds on a trade-off between the utility, fairness, and compression rate of the representation. To achieve the lower bounds, they use randomization techniques and extended versions of *Functional Representation Lemma* and *Strong Functional Representation Lemma* derived in [9]. In [22], the problem in [1] has been extended by replacing zero demographic parity by using a bounded mutual information as the fairness constraint. In [18], a binary classification problem subject to both differential privacy and fairness constraints is studied. In [37], the role of data processing in the fairness–accuracy trade-off, using equalized odds as the fairness criterion, has been studied.

Prior works on privacy mechanism design: In [13], the authors introduce the notion of *secrecy by design* and apply it to privacy mechanism design and lossless compression problems. They find bounds on the trade-off between perfect privacy and utility using the *Functional Representation Lemma* (FRL). In [9], these results are generalized to an arbitrary privacy leakage. To attain privacy mechanism, extensions of the FRL and SFRL have been used. In [38], a compression design problem under the secrecy constraints is studied and bounds on the average length of the encoded message are derived. A similar approach to that of the present paper was used in [31], where an information-geometric perspective on Local Information Privacy (LIP), Max-lift, and Local Differential Privacy (LDP) was considered. In [10], the authors study the privacy-utility trade-off under perfect privacy condition. In [11], [12], this is generalized by relaxing the perfect privacy constraint and allowing some small bounded leakage. In both [11] and [12], point-wise (per-letter) measures have been used in the leakage constraint. Furthermore, the privacy design is based on the information geometry concept which allows them to study the problem geometrically. The notion of the *Privacy Funnel* is introduced in [14], where the privacy-utility trade-off considers the log-loss as the privacy and distortion metrics. In [17], the authors studied the fundamental limits of the privacy and utility trade-off under an estimation-theoretic approach. In [16], [39], the trade-off between privacy and utility is studied considering the equivocation as the privacy measure and the expected distortion as a utility.

Contributions: In this paper, we propose a simple and constructive theoretical approach to design fair representations under bounded point-wise statistical parity. We study the problem of finding a trade-off between utility, fairness, and compression rate, which is described in (3). As outlined earlier, we use a point-wise measure for the fairness constraint. We show that when fairness threshold is small, the main optimization problem can be approximated by a quadratic one. This leads to a simple design based on finding the maximum singular value and the corresponding singular vector of a matrix. We compare the obtained results with the optimal solution in a numerical example.

II. System model and Problem Formulation

In this section, we introduce the problem of designing a compressed representation of data under bounded point-wise fairness constraint. Let the latent sensitive attribute S, database (useful data) X, and task T be discrete random variables (RVs) defined on alphabets $\mathcal{S}$, $\mathcal{X}$, and $\mathcal{T}$, of finite cardinalities $|\mathcal{S}|$, $|\mathcal{X}|$, and $|\mathcal{T}|$, respectively. The marginal probability distributions of S, X, and T are denoted by $P_S \in \mathbb{R}^{|\mathcal{S}|}$, $P_X \in \mathbb{R}^{|\mathcal{X}|}$, and $P_T \in \mathbb{R}^{|\mathcal{T}|}$, respectively. Furthermore, $P_{S,X,T} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{X}|\times|\mathcal{T}|}$ denotes the joint distribution of S, X and T. Let the matrices $P_{S|X} \in \mathbb{R}^{|\mathcal{S}\times|\mathcal{X}|}$ and $P_{T|X} \in \mathbb{R}^{|\mathcal{T}\times|\mathcal{X}|}$ with elements $P_{S=s|X=x}$ and $P_{T=t|X=x}$ denote the conditional distributions of S given X and T given X. Here, for simplicity of design, we assume that $P_{S|X}$ is invertible, implying $|\mathcal{X}| = |\mathcal{S}|$. However, we will later discuss how this assumption can be generalized. Let $\sqrt{P_S} \in \mathbb{R}^{|\mathcal{S}|}$ denote a vector with entries $\sqrt{P_S(s)}$, and let $[\sqrt{P_S}] \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote a diagonal matrix with diagonal elements $\sqrt{P_S(s)}$. We use the same notation for $[\sqrt{P_T}]$ and $[\sqrt{P_X}]$. Let the discrete RV $Y \in \mathcal{Y}$ denote the fair representation. We emphasize that here we assume that the Markov chains $S - X - Y$ and $T - X - Y$ hold, as the agent has no direct access to S and T. The goal is to design a mapping $P_{Y|X} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ that maximizes the information it keeps about the task T, while maintaining a minimum level of compression r and satisfying the bounded point-wise statistical parity, that is $I(Y;X) \leq r$ and for all $y \in \mathcal{Y}$, we have $\chi^2(P_{S|y}; P_S) \leq \epsilon$. More specifically, the representation Y should satisfy the fairness constraint defined as

$$\chi^2(P_{S|y}; P_S) = \|\frac{P_{S|y}(s) - P_S(s)}{P_S(s)}\|^2 = \sum_s \left(\frac{P_{S|y}(s) - P_S(s)}{P_S(s)}\right)^2 \leq \epsilon^2, \ \forall y, \tag{1}$$

where $\|\cdot\|$ corresponds to the Euclidean norm. Intuitively, for small ϵ, (1) means that the two distributions (vectors) $P_{S|Y=y}$ and P_S are close to each other. The closeness of $P_{S|Y=y}$ and P_S allows us to locally approximate $I(T;Y)$ and $I(X;Y)$ which leads to an approximation of (3). For more details on the point-wise measure, see [11], [12].

Remark 1. *In [28], we used bounded mutual information as the statistical parity constraint. Furthermore, bounded statistical parity has been used in [18], [24], [25], where we have*

$$|P(Y = y|S = s_1) - P(Y = y|S = s_2)| \leq \epsilon, \forall s_1, s_2 \in \mathcal{S}. \tag{2}$$

Here, we use a similar but slightly different measure as defined in (1)*. Using a point-wise measure can be motivated by* (2)*. The relation between* (1) *and bounded mutual information has been studied in [11, Remark 3].*

Hence, the fair representation design problem can be stated as follows:

$$g^r_{\epsilon,\chi^2}(P_{S,X,T}) \triangleq \sup_{\substack{P_{Y|X}: S-X-Y,\ T-X-Y,\\ \chi^2(P_{S|y};P_S)\leq\epsilon^2,\ \forall y,\\ I(X;Y)\leq r}} I(Y;T). \tag{3}$$

The constraint $\chi^2(P_{S|y}; P_S) \leq \epsilon^2, \ \forall y$, corresponds to the bounded point-wise statistical parity, and $I(X;Y) \leq r$ represents the compression rate constraint.

Remark 2. *We define the utility–fairness–compression trade-off, subject to bounded mutual information and compression constraints, as follows.*

$$g^r_{\epsilon}(P_{S,X,T}) \triangleq \sup_{\substack{P_{Y|X}: S-X-Y,\ T-X-Y,\\ I(Y;S)\leq\epsilon^2,\\ I(X;Y)\leq r}} I(Y;T). \tag{4}$$

Then, by using [11, Remark 3], we have $g^r_{\epsilon,\chi^2}(P_{S,X,T}) \leq g^r_{\epsilon}(P_{S,X,T})$. In other words, one benefit of solving (3) *is to find a lower bound on* (4) *which is hard to solve. The trade-off in* (4) *is equivalent to the privacy-utility trade-off with a rate constraint studied in [3] and it has been shown that for sufficiently large r ($r \geq H(X)$), the problem leads to a linear program. Finally, if we remove the Markov chains $S - X - Y$ and $T - X - Y$ in* (3)*, we get the problem studied in [28].*

Remark 3. *In this work, we assume that $P_{S|X}$ is invertible; however, this assumption can be generalized using the techniques in [12]. Here, we focus on invertible case for design simplicity and leave the generalization to future work.*

III. Main Results

In this section, we first derive an approximation to (3). We then find lower bounds for this approximation and discuss their tightness. To this end, using (1), we can rewrite the conditional distribution $P_{S|Y=y}$ as a perturbation of P_S. Hence, for any $y \in \mathcal{Y}$, we can write $P_{S|Y=y} = P_S + \epsilon \cdot J_y$, where $J_y \in \mathbb{R}^{\mathcal{S}}$ is a perturbation vector that has the following three properties

$$\sum_s J_y(s) = 0, \quad \forall y, \tag{5}$$

$$\sum_y P_Y(y) J_y(s) = 0, \ \forall s, \tag{6}$$

$$\|[\sqrt{P_S}^{-1}] J_y\|^2 = \sum_x \frac{J_y(s)^2}{P_S(s)} \leq 1. \tag{7}$$

The first two properties ensure that $P_{S|Y=y}$ is a valid probability distribution [11], [12], and the third property follows from (1). Furthermore, (6) implies $\sum_u P_Y(y) J_y = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}|}$. Next, we approximate $I(Y;X)$ and $I(Y;T)$ using the Markov chains and the perturbation vector J_y. In the next results, we use the Bachmann-Landau notation where $o(\epsilon)$ describes the asymptotic behavior of a function $f : \mathbb{R}^+ \to \mathbb{R}$ which satisfies $\frac{f(\epsilon)}{\epsilon} \to 0$ as $\epsilon \to 0$.

Lemma 1. *For all* $\epsilon < \frac{|\sigma_{min}(P_{S|X})| \min_{x \in \mathcal{X}} P_X(x)}{\sqrt{\max_{s \in \mathcal{S}} P_S(s)}}$, *we have*

$$\begin{aligned} I(X;Y) &= \frac{1}{2}\epsilon^2 \sum_y P_y \|[\sqrt{P_X}^{-1}] P_{S|X}^{-1} J_y\|^2 + o(\epsilon^2) \\ &\simeq \frac{1}{2}\epsilon^2 \sum_y P_y \|[\sqrt{P_X}^{-1}] P_{S|X}^{-1} J_y\|^2, \end{aligned} \tag{8}$$

and for all $\epsilon < \frac{\min_{t \in \mathcal{T}} P_T(t)}{|\sigma_{max}(P_{T|X} P_{S|X}^{-1})| \sqrt{\max_{s \in \mathcal{X}} P_S(s)}}$, *we have*

$$\begin{aligned} I(T;Y) &= \frac{1}{2}\epsilon^2 \sum_y P_y \|[\sqrt{P_T}^{-1}] P_{T|X} P_{S|X}^{-1} J_y\|^2 + o(\epsilon^2) \\ &\simeq \frac{1}{2}\epsilon^2 \sum_y P_y \|[\sqrt{P_T}^{-1}] P_{T|X} P_{S|X}^{-1} J_y\|^2. \end{aligned} \tag{9}$$

Proof. To approximate $I(X;Y)$, using the Markov chain $S - X - Y$ and invertible $P_{S|X}$ we can rewrite $P_{X|Y=y}$ as perturbations of P_X as follows

$$\begin{aligned} P_{X|Y=y} &= P_{S|X}^{-1}[P_{S|Y=y} - P_S] + P_X \\ &= \epsilon \cdot P_{S|X}^{-1} J_y + P_X. \end{aligned}$$

Then, by using local approximation of the KL-divergence which is based on the second order Taylor expansion of $\log(1+x)$ we get

$$\begin{aligned} I(Y;X) &= \sum_y P_Y(y) D(P_{X|Y=y} || P_X) \\ &= \sum_y P_Y(y) \sum_x P_{X|Y=y}(x) \log\left(1 + \epsilon \frac{P_{S|X}^{-1} J_y(x)}{P_X(x)}\right) \\ &= \frac{1}{2}\epsilon^2 \sum_y P_Y(y) \sum_x \frac{(P_{S|X}^{-1} J_y(x))^2}{P_X(x)} + o(\epsilon^2). \end{aligned}$$

For more detail see [11, Proposition 3]. Finally, for approximating $I(X;Y)$ we must have $|\epsilon\frac{P_{S|X}^{-1}J_y(x)}{P_x(x)}|<1$ for all x and y. A sufficient condition for ϵ to satisfy this inequality is to have $\epsilon<\frac{|\sigma_{\min}(P_{S|X})|\min_{x\in\mathcal{X}}P_X(x)}{\sqrt{\max_{s\in\mathcal{S}}P_S(s)}}$, since in this case we have

$$\begin{aligned}\epsilon^2|P_{S|X}^{-1}J_y(x)|^2&\leq\epsilon^2\left\|P_{S|X}^{-1}J_y\right\|^2\\&\leq\epsilon^2\sigma_{\max}^2\left(P_{S|X}^{-1}\right)\|J_y\|^2\\&\overset{(a)}{\leq}\epsilon^2\max_{s\in\mathcal{S}}P_S(s)\frac{1}{\sigma_{\min}^2(P_{S|X})}\\&<\min_{x\in\mathcal{X}}P_X^2(x).\end{aligned}$$

which implies

$$|\epsilon\frac{P_{S|X}^{-1}J_y(x)}{P_X(x)}|<1.$$

The step (a) follows from $\sigma_{\max}^2\left(P_{S|X}^{-1}\right)=\frac{1}{\sigma_{\min}^2\left(P_{S|X}\right)}$ and $\|J_y\|^2\leq\max_{s\in\mathcal{S}}P_S(s)$. The latter inequality follows from (7) since we have

$$\begin{aligned}\frac{\|J_y\|^2}{\max_{s\in\mathcal{S}}P_S(s)}&\leq\sum_{s\in\mathcal{S}}\frac{J_y^2(s)}{P_S(s)}\\&\leq 1.\end{aligned}$$

To approximate $I(T;Y)$, using the Markov chain $T-X-Y$ we can rewrite $P_{T|Y=y}$ as perturbations of P_T as follows

$$P_{T|Y=y}=P_{T|X}P_{X|y}=\epsilon\cdot P_{T|X}P_{S|X}^{-1}J_y+P_T.$$

Then, by using the local approximation of the KL-divergence we have

$$\begin{aligned}I(Y;T)&=\sum_y P_Y(y)D(P_{T|Y=y}||P_T)\\&=\sum_y P_Y(y)\sum_t P_{T|Y=y}(t)\log\left(1+\epsilon\frac{P_{T|X}P_{S|X}^{-1}J_y(t)}{P_T(t)}\right)\\&=\frac{1}{2}\epsilon^2\sum_y P_Y(y)\sum_t\frac{(P_{T|X}P_{S|X}^{-1}J_y)^2}{P_T}+o(\epsilon^2).\end{aligned}$$

Due to the Taylor expansion, we must have $|\epsilon\frac{P_{T|X}P_{S|X}^{-1}J_y(t)}{P_T(t)}|<1$ for all t and y. A sufficient condition for ϵ to satisfy this inequality is to have $\epsilon<\frac{\min_{t\in\mathcal{T}}P_T(t)}{|\sigma_{\max}(P_{T|X}P_{S|X}^{-1})|\sqrt{\max_{s\in\mathcal{X}}P_S(s)}}$, since in this case we have

$$\begin{aligned}\epsilon^2|P_{T|X}P_{S|X}^{-1}J_y(t)|^2&\leq\epsilon^2\left\|P_{T|X}P_{S|X}^{-1}J_y\right\|^2\\&\leq\epsilon^2\sigma_{\max}^2\left(P_{T|X}P_{S|X}^{-1}\right)\|J_y\|^2\\&\overset{(a)}{\leq}\epsilon^2\max_{s\in\mathcal{S}}P_S(s)\sigma_{\max}^2(P_{T|X}P_{S|X}^{-1})\\&<\min_{t\in\mathcal{T}}P_T^2(t),\end{aligned}$$

which implies

$$|\epsilon\frac{P_{S|T}^{-1}J_y(t)}{P_T(t)}|<1.$$

The step (a) follows from $\|J_y\|^2\leq\max_{s\in\mathcal{S}}P_S(s)$.

□

In the next result, we present an approximation on (3). Before stating the next result, let us define

$$\begin{aligned}L_y&\triangleq[\sqrt{P_S}^{-1}]J_y\in\mathbb{R}^{|\mathcal{Y}|},\\W^{T;Y}&\triangleq[\sqrt{P_T}^{-1}]P_{T|X}P_{S|X}^{-1}[\sqrt{P_S}]\in\mathbb{R}^{|\mathcal{T}|\times|\mathcal{Y}|},\end{aligned}$$

and

$$W^{X;Y} \triangleq [\sqrt{P_X}^{-1}]P_{S|X}^{-1}[\sqrt{P_S}] \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}.$$

Furthermore, let

$$c_1 \triangleq \frac{\min_{t\in\mathcal{T}} P_T(t)}{|\sigma_{\max}(P_{T|X}P_{S|X}^{-1})|\sqrt{\max_{s\in\mathcal{X}} P_S(s)}},$$

and

$$c_2 \triangleq \frac{|\sigma_{\min}(P_{S|X})|\min_{x\in\mathcal{X}} P_X(x)}{\sqrt{\max_{s\in\mathcal{S}} P_S(s)}}.$$

Theorem 1. *For all $\epsilon < \min\{c_1, c_2\}$,* (3) *can be approximated by the following problem*

$$(3) \simeq P_2 \triangleq \max_{\substack{P_y, L_y : L_y \perp \sqrt{P_S}, \\ \sum_y P_y L_y = 0, \\ \|L_y\|^2 \le 1, \\ \frac{1}{2}\epsilon^2 \sum_y P_y \|W^{X;Y} L_y\|^2 \le r}} \frac{1}{2}\epsilon^2 \sum_y P_y \|W^{T;Y} L_y\|^2. \tag{10}$$

Proof: The proof follows by Lemma 1. The constraints $L_y \perp \sqrt{P_S}$, $\sum_y P_y L_y = 0$, and $\|L_y\|^2 \le 1$ are followed by (5), (6), and (7), for more detail see [11].

Remark 4. *As ϵ decreases, the approximation becomes tighter. This follows since using Taylor expansion, as ϵ decreases the error term becomes smaller.*

Remark 5. *The compression-rate constraint*

$$I(X;Y) \le r$$

can be rewritten as

$$\frac{1}{2}\epsilon^2 \sum_y P_y \|W^{X;Y} L_y\|^2 \le r - o(\epsilon^2).$$

Since the compression threshold r is typically much larger than the fairness constraint parameter ϵ, we neglect the $o(\epsilon^2)$ term relative to r. We can also derive an upper bound on the error term and use it here; we leave this to future work.

Next, we present key properties of the matrices $W^{X;Y}$ and $W^{T;Y}$ that we use to derive lower bounds on (10).

Proposition 1. *Each of the matrices $W^{T;Y}$ and $W^{X;Y}$ has singular value* 1*, with corresponding singular vector $\sqrt{P_S}$.*

Proof. We have

$$\begin{aligned}
&(W^{X;Y})^T W^{X;Y}\sqrt{P_S} \\
&= [\sqrt{P_S}](P_{S|X}^T)^{-1}[\sqrt{P_X}^{-1}][\sqrt{P_X}^{-1}]P_{S|X}^{-1}[\sqrt{P_S}]\sqrt{P_S} \\
&= [\sqrt{P_S}](P_{S|X}^T)^{-1}[\sqrt{P_X}^{-1}][\sqrt{P_X}^{-1}]P_X \\
&= [\sqrt{P_S}](P_{S|X}^T)^{-1}\mathbf{1} = [\sqrt{P_S}]\mathbf{1} \\
&= \sqrt{P_S},
\end{aligned}$$

where $\mathbf{1}$ denotes a vector in which all elements are equal to 1. Furthermore, we have

$$\begin{aligned}
&(W^{T;Y})^T W^{T;Y}\sqrt{P_S} \\
&= [\sqrt{P_S}](P_{T|X}P_{S|X}^{-1})^T[\sqrt{P_T}^{-1}][\sqrt{P_T}^{-1}]P_T \\
&= [\sqrt{P_S}](P_{T|X}P_{S|X}^{-1})^T\mathbf{1} = [\sqrt{P_S}]\mathbf{1} \\
&= \sqrt{P_S}.
\end{aligned}$$

□

Next, we find lower bounds on (10) and discuss their tightness. To this end, let $\sigma_{\max}^{T;Y}$ and $\sigma_{\max_2}^{T;Y}$ denote the largest and second-largest singular values of the matrix $W^{T;Y}$, with corresponding singular vectors $L_\sigma^{W^{T;Y}}$ and $L_{\sigma_2}^{W^{T;Y}}$, respectively, where $\|L_\sigma^{W^{T;Y}}\| = 1$ and $\|L_{\sigma_2}^{W^{T;Y}}\| = 1$.

Theorem 2. *If $\sigma_{\max}^{T;Y} > 1$, we have*

$$P_2 \geq \frac{1}{2}\epsilon^2 \left(\frac{\sigma_{\max}^{T;Y}}{K}\right)^2, \tag{11}$$

where $1 \leq K$ is the smallest constant that satisfies

$$\frac{1}{2}\epsilon^2 \|W^{X;Y} L_\sigma^{W^{T;Y}}\|^2 \leq rK^2.$$

Moreover, when $\frac{1}{2}\epsilon^2 \|W^{X;Y} L_\sigma^{W^{T;Y}}\|^2 \leq r$, then $K = 1$. If $\sigma_{\max}^{T;Y} = 1$, then

$$P_2 \geq \frac{1}{2}\epsilon^2 \left(\frac{\sigma_{\max_2}^{T;Y}}{K}\right)^2, \tag{12}$$

where $1 \leq K$ is the smallest constant that satisfies

$$\frac{1}{2}\epsilon^2 \|W^{X;Y} L_{\sigma_2}^{W^{T;Y}}\|^2 \leq rK^2.$$

Finally, when $|\mathcal{S}| = 2$, if $\sigma_{\max}^{T;Y} > 1$, then we have

$$P_2 = \frac{1}{2}\epsilon^2 \left(\frac{\sigma_{\max}^{T;Y}}{K}\right)^2,$$

otherwise,

$$P_2 = \frac{1}{2}\epsilon^2 \left(\frac{\sigma_{\max_2}^{T;Y}}{K}\right)^2.$$

Proof. To derive the first lower bound, let $\sigma_{\max}^{T;Y} > 1$. Using Proposition 1, we have $L_\sigma^{W^{T;Y}} \perp \sqrt{P_S}$ since $\sqrt{P_S}$ is a singular vector of $W^{T;Y}$. Let Y be a uniform binary RV. Furthermore, let $L_1 = \frac{L_\sigma^{W^{T;Y}}}{K}$ and $L_2 = \frac{-L_\sigma^{W^{T;Y}}}{K}$. Clearly, $L_1 \perp \sqrt{P_S}$ and $L_2 \perp \sqrt{P_S}$ and $P_Y(1)L_1 + P_Y(2)L_2 = 0$. Moreover, we have

$$\frac{1}{2}\epsilon^2 \left(\sum_y P_Y \|W^{T;Y} L_y\|^2\right) = \frac{1}{2}\epsilon^2 \left(\frac{\sigma_{\max}^{T;Y}}{K}\right)^2.$$

The proof of (12) is similar. To prove the last argument, note that when $|\mathcal{S}| = 2$, the matrix $W^{T;Y}$ has two singular vectors, one of which is $\sqrt{P_S}$. Hence, the only feasible direction is the other singular vector which completes the proof. □

Remark 6. *In Theorem 2, K is a scaling factor that ensures $L_\sigma^{W^{T;Y}}$ is feasible under the compression rate constraint. In other words, if $L_\sigma^{W^{T;Y}}$ does not satisfy the rate constraint, we scale it down. We note that if $\frac{1}{2}\epsilon^2 \|L_\sigma^{W^{T;Y}}\|^2 \leq r$, we cannot scale $L_\sigma^{W^{T;Y}}$ up, since $\|L_\sigma^{W^{T;Y}}\| = 1$.*

Corollary 1. *Let ϵ be fixed. Then, by using Remark 6, when $r < \frac{1}{2}\epsilon^2 \|L_\sigma^{W^{T;Y}}\|^2$, we get*

$$K = \sqrt{\frac{\frac{1}{2}\epsilon^2 \|W^{X;Y} L_\sigma^{W^{T;Y}}\|^2}{r}} > 1,$$

which results

$$P_2 \geq r \left(\frac{\sigma_{\max}^{T;Y}}{\|W^{X;Y} L_\sigma^{W^{T;Y}}\|^2}\right)^2. \tag{13}$$

When $|\mathcal{S}| = 2$, (13) is tight. In other words, when r is smaller than a threshold, the lower bound on P_2 increases linearly with respect to r.

After finding L_u and P_U that attain the lower bounds in Theorem 2, we can find the joint distribution P_{STXY} as

$$\begin{aligned}
P_{S|Y=0} &= P_S + \epsilon[\sqrt{P_S}]L_1, \\
P_{S|Y=1} &= P_S + \epsilon[\sqrt{P_S}]L_2, \\
P_{T|Y=0} &= P_T + \epsilon P_{T|X} P_{S|X}^{-1}[\sqrt{P_S}]L_1, \\
P_{T|Y=1} &= P_T + \epsilon P_{T|X} P_{S|X}^{-1}[\sqrt{P_S}]L_2, \\
P_{X|Y=0} &= P_X + \epsilon P_{S|X}^{-1}[\sqrt{P_S}]L_1, \\
P_{X|Y=1} &= P_X + \epsilon P_{S|X}^{-1}[\sqrt{P_S}]L_2,
\end{aligned}$$

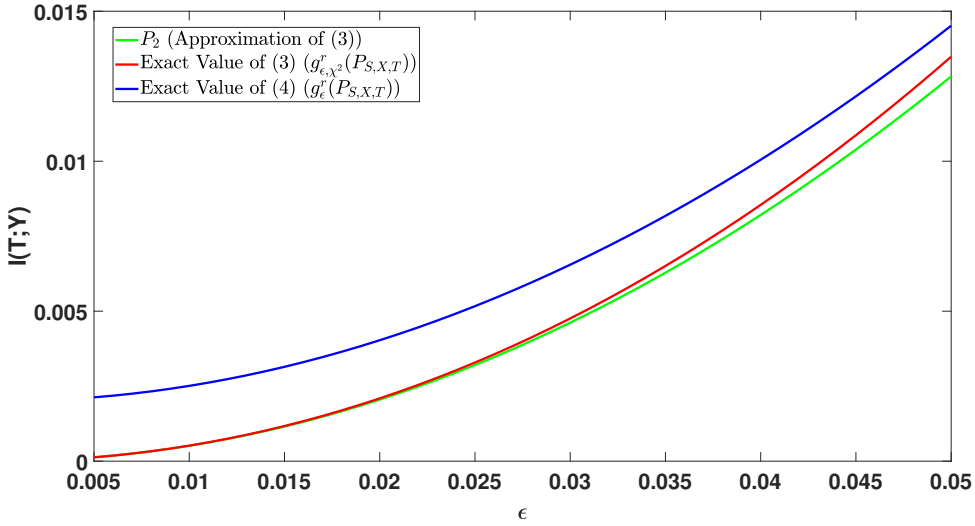

Fig. 2. Comparing the proposed approach in this paper with the optimal solutions of (3) and (4). We can see that in the high privacy regimes, P_2 is close to the exact solution of (3) which is found by exhaustive search and $g^r_{\epsilon}(P_{S,X,T})$ dominates $g^r_{\epsilon,\chi^2}(P_{S,X,T})$.

where L_1, L_2 and marginal distribution of Y are obtained in Theorem 2. Finally,

$$P_{STXY}(x, y, , s, t) = P_{XST}(x, s, t)P_{Y|X}(y|x),$$

where $P_{Y|X}$ can be calculated by using $P_{X|Y}$ and P_Y obtained in Theorem 2.

IV. Numerical Example:

In this section, we present a numerical example to evaluate the proposed approach and compare the results with the optimal solutions of (3) and (4).

Example: Let $P_{T|X} = \begin{bmatrix} \frac{1}{4} & \frac{2}{5} \\ \frac{3}{4} & \frac{3}{5} \end{bmatrix}$, and P_X be given as $[\frac{1}{4}, \frac{3}{4}]^T$. Furthermore, let $P_{S|X} = \begin{bmatrix} 0.275 & 0.32 \\ 0.725 & 0.68 \end{bmatrix}$. Thus, we find $P_T = P_{T|X}P_X = [0.3625, 0.6375]^T$ and $P_S = P_{S|X}P_X = [0.3088, 0.6913]^T$. Furthermore, we find

$$W^{T;Y} = \begin{bmatrix} 2.4610 & -0.9206 \\ -1.1599 & 1.7355 \end{bmatrix},$$

and

$$W^{X;Y} = \begin{bmatrix} -16.7931 & 11.8246 \\ -10.3371 & -5.8669 \end{bmatrix}.$$

The singular values of $W^{T;Y}$ are 3.2034 and 1 with corresponding right singular vectors

$$[-0.8314, 0.5557]^T,$$

and

$$[0.5557, 0.8314]^T,$$

respectively. Moreover, the singular values of $W^{X;Y}$ are 23.7087 and 1 with corresponding right singular vectors

$$[0.8314, -0.5557]^T,$$

and

$$[0.5557, 0.8314]^T,$$

respectively. Here, we let $\epsilon \in [0.005, 0.05]$ and $r = 0.75$. We have

$$\|W^{X;Y} L_\sigma^{W^{T;Y}}\|^2 = 562.1029,$$

and we can check that for $\epsilon \in [0.005, 0.05]$ the compression rate constraint is satisfied, since for $\epsilon = 0.05$, we have

$$\frac{1}{2}\epsilon^2 \|W^{X;Y} L_\sigma^{W^{T;Y}}\|^2 = 0.7026.$$

Here, L_1 equals to $[-0.8314, 0.5557]^T$ and the compression rate constraint is satisfied for L_1 which yields $K = 1$. Using Theorem 2, we have

$$P_2 = \frac{1}{2}\epsilon^2(3.2034)^2,$$

and the lower bound is tight since $|\mathcal{S}| = 2$. In Fig. 2, we compare P_2 and the optimal solutions of (3) and (4) using exhaustive search. We recall that P_2 is the approximate of (3) and here achieves the lower bound in Theorem 2. We can see that $g^r_\epsilon(P_{S,X,T})$ dominates $g^r_{\epsilon,\chi^2}(P_{S,X,T})$, and the gap between exact values of (3) and (4) is decreasing as ϵ increases. Finally, the gap between P_2 and $g^r_{\epsilon,\chi^2}(P_{S,X,T})$ is small in the high privacy regimes. Intuitively, the blue curve dominates (3), since in (3) we impose the point-wise χ^2-criterion instead of bounded mutual information. Furthermore, when $r \geq H(X) = 0.8113$, we can use the linear program from [3] to obtain the optimal solution to (4). Finally, let us fix $\epsilon = 0.05$. Then, for $r \leq 0.7026$, we can use (13).

V. CONCLUSION

We have shown that concepts from information geometry can be used to simplify an information-theoretic fair mechanism design problem under the bounded point-wise statistical parity. When a small ϵ is allowed, simple approximate solutions are derived. Specifically, we look for vectors satisfying the fairness constraints of having the largest Euclidean norm, leading to finding the largest principle singular value and vector of a matrix. The proposed approach establishes a useful and general design framework, which has been used in other privacy design problems such as considering point-wise χ^2 and ℓ_1-criterion in the literature.

REFERENCES

[1] A. Zamani, B. Rodríguez-Gálvez, and M. Skoglund, "On information theoretic fairness: Compressed representations with perfect demographic parity," in *2024 IEEE Information Theory Workshop (ITW)*, 2024, pp. 25–30.

[2] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, "A variational approach to privacy and fairness," in *2021 IEEE Information Theory Workshop (ITW)*. IEEE, 2021, pp. 1–6.

[3] S. Sreekumar and D. Gündüz, "Optimal privacy-utility trade-off under a rate constraint," in *2019 IEEE International Symposium on Information Theory*, July 2019, pp. 2159–2163.

[4] H. Zhao and G. J. Gordon, "Inherent tradeoffs in learning fair representations," *Journal of Machine Learning Research*, vol. 23, no. 57, pp. 1–26, 2022.

[5] H. Zhao, A. Coston, T. Adel, and G. J. Gordon, "Conditional learning of fair representations," in *International Conference on Learning Representations*, 2019.

[6] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[7] A. Gronowski, W. Paul, F. Alajaji, B. Gharesifard, and P. Burlina, "Classification utility, fairness, and compactness via tunable information bottleneck and rényi measures," *IEEE Transactions on Information Forensics and Security*, 2023.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[9] A. Zamani, T. J. Oechtering, and M. Skoglund, "On the privacy-utility trade-off with and without direct access to the private data," *IEEE Transactions on Information Theory*, vol. 70, no. 3, pp. 2177–2200, 2024.

[10] B. Rassouli and D. Gündüz, "On perfect privacy," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 177–191, 2021.

[11] A. Zamani, T. J. Oechtering, and M. Skoglund, "A design framework for strongly χ^2-private data disclosure," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2312–2325, 2021.

[12] A. Zamani, T. J. Oechtering, and M. Skoglund, "Data disclosure with non-zero leakage and non-invertible leakage matrix," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 165–179, 2022.

[13] Y. Y. Shkel, R. S. Blum, and H. V. Poor, "Secrecy by design with applications to privacy and compression," *IEEE Transactions on Information Theory*, vol. 67, no. 2, pp. 824–843, 2021.

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

[15] H. Yamamoto, "A rate-distortion problem for a communication system with a secondary decoder to be hindered," *IEEE Transactions on Information Theory*, vol. 34, no. 4, pp. 835–842, 1988.

[16] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[17] H. Wang, L. Vo, F. P. Calmon, M. Médard, K. R. Duffy, and M. Varia, "Privacy with estimation guarantees," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8025–8042, Dec 2019.

[18] H. Ghoukasian and S. Asoodeh, "Differentially private fair binary classifications," in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 611–616.

[19] S. Asoodeh, W.-N. Chen, F. P. Calmon, and A. Özgür, "Differentially private federated learning: An information-theoretic perspective," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 344–349.

[20] B. Isik, W.-N. Chen, A. Özgür, T. Weissman, and A. No, "Exact optimality of communication-privacy-utility tradeoffs in distributed mean estimation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37 761–37 785, 2023.

[21] Y. Liu, W.-N. Chen, A. Özgür, and C. T. Li, "Universal exact compression of differentially private mechanisms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 91 492–91 531, 2024.

[22] A. Zamani, A. Changizi, R. Thobaben, and M. Skoglund, "Information-theoretic fairness with a bounded statistical parity constraint," in *2025 23rd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2025, pp. 1–8.

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

[24] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[25] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 13–18.

[26] S. Borade and L. Zheng, "Euclidean information theory," in *2008 IEEE International Zurich Seminar on Communications*, 2008, pp. 14–17.

[27] S. L. Huang and L. Zheng, "Linear information coupling problems," in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 1029–1033.

[28] A. Zamani, T. J. Oechtering, and M. Skoglund, "Bounds for privacy-utility trade-off with per-letter privacy constraints and non-zero leakage," in *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 13–18.

[29] E. M. Athanasakos, N. Kalouptsidis, and H. Manjunath, "Local approximation of secrecy capacity," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024, pp. 745–749.

[30] J. Liao, L. Sankar, V. Y. F. Tan, and F. du Pin Calmon, "Hypothesis testing under mutual information privacy constraints in the high privacy regime," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pp. 1058–1071, 2018.

[31] A. Zamani, P. Sadeghi, and M. Skoglund, "An information geometric approach to local information privacy with applications to max-lift and local differential privacy," *arXiv preprint arXiv:2501.11757*, 2025.

[32] H. Edwards and A. Storkey, "Censoring representations with an adversary," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1–14.

[33] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *International conference on machine learning*. PMLR, 2019, pp. 1436–1445.

[34] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *International Conference on Learning Representations (ICLR)*, 2016.

[35] U. Gupta, A. M. Ferber, B. Dilkina, and G. Ver Steeg, "Controllable guarantees for fair outcomes via contrastive information estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7610–7619.

[36] J. M. de Freitas and B. C. Geiger, "Funck: Information funnels and bottlenecks for invariant representation learning," *arXiv preprint arXiv:2211.01446*, 2022.

[37] S. Khodadadian, A. Ghassami, and N. Kiyavash, "Impact of data processing on fairness in supervised learning," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2643–2648.

[38] A. Zamani and M. Skoglund, "Variable-length coding with zero and non-zero privacy leakage," *Entropy*, vol. 27, no. 2, p. 124, 2025.

[39] 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*, vol. 29, no. 6, pp. 918–923, 1983.