Model Projection: Theory and Applications to Fair Machine Learning

Wael Alghamdi*, Shahab Asoodeh*, Hao Wang*, Flavio P. Calmon*,
Dennis Wei†, Karthikeyan Natesan Ramamurthy†
*Harvard University, alghamdi@g.harvard.edu, shahab@seas.harvard.edu, hao_wang@g.harvard.edu, flavio@seas.harvard.edu
†IBM Research, {dwei,knatesa}@us.ibm.com

***Abstract*—We study the problem of finding the element within a convex set of conditional distributions with the smallest f-divergence to a reference distribution. Motivated by applications in machine learning, we refer to this problem as *model projection* since any probabilistic classification model can be viewed as a conditional distribution. We provide conditions under which the existence and uniqueness of the optimal model can be guaranteed and establish strong duality results. Strong duality, in turn, allows the model projection problem to be reduced to a tractable finite-dimensional optimization. Our application of interest is fair machine learning: the model projection formulation can be directly used to design fair models according to different group fairness metrics. Moreover, this information-theoretic formulation generalizes existing approaches within the fair machine learning literature. We give explicit formulas for the optimal fair model and a systematic procedure for computing it.**

I. Introduction

Information projection [1–3] is a fundamental formulation in several applications of information theory. Given a set of probability measures $\mathscr{C}$ and a reference measure P, a distribution $Q \in \mathscr{C}$ is said to be the *projection* of P onto $\mathscr{C}$ if it uniquely achieves the smallest KL-divergence $D_{\mathsf{kl}}(Q\|P)$ among all distributions in $\mathscr{C}$ [2]. Both the minimizing distribution Q and the minimum divergence value are central quantities in large deviation theory [4], universal source compression [5], hypothesis testing [6], and beyond. Existence and uniqueness of the optimal distribution have been studied in [2, 3]. In particular, the optimal distribution has a simple closed-form given by an exponential tilting of the reference distribution P when the set $\mathscr{C}$ is determined by linear inequalities [2]. Even though the information projection is most commonly defined with "distance" measured by the KL-divergence [3, 6–10], it has also been extended to Rényi divergences [11–13] and f-divergences [14, 15].

We study a natural generalization of information projection: finding the "closest" *conditional* distribution (in a prescribed subset $\mathcal{F}$ of all possible conditional distributions) to a reference conditional distribution, where "distance" is measured by averaged (i.e., conditional) f-divergences. Motivated by applications in machine learning, we refer to this setting as *model projection*, since probabilistic classification models (e.g., logistic regression, neural networks with a softmax output layers) which map an input onto a probability distribution over predicted classes can be viewed as a conditional distribution. Analogous to the treatment of information projection, we start by proving the existence and uniqueness of the optimal conditional distribution. We then establish strong duality, which, in turn, leads to an equivalent formulation for obtaining the optimal conditional distribution. This dual formulation is easier to deal with since it converts an optimization with possibly infinitely many primal variables into a tractable, finite-dimensional optimization in Euclidean space. The optimal dual variables, in turn, allow the minimizing conditional distribution to be computed via a generalization of exponential "tilting." For a general f-divergence, one obtains the optimal conditional distribution by tilting the reference distribution by the inverse of the derivative of f. Naturally, this approach reduces to the usual exponential tilting when KL-divergence is the f-divergence of choice.

We provide an application of the model projection theory to fair machine learning. A critical concern when applying probabilistic classifiers to individual-level data is if the classifier may discriminate (e.g., by having a higher error rate) in terms of a (legally) sensitive attribute, such as race, gender, or ethnicity. This concern has recently led to a plethora of research focusing on two questions: (a) how does one "quantify" and "understand" discrimination in typical machine learning algorithms? [16–22] and (b) given a notion of fairness, how does one learn an "optimal" fair model? [23–31]. We refer the reader to a recent survey [32] and the references therein for a more detailed literature review.

We focus on the problem of "projecting" a reference probabilistic classifier to the set of classifiers that satisfy a collection of fairness criteria. When the fairness criteria are given in terms of linear constraints on the classifier—which is the case for several commonly used fairness metrics [see e.g., 27, 28]—this problem can be directly formulated as an optimization via the model projection formulation. We derive both explicit formulas for the optimal fair classifier and a practical pipeline for the design process, thereby generalizing recent methods [see e.g., 29, 30] for fairness assurance.

Strikingly, the model projection formulation implies that the optimal correction for an "unfair" model can be given by a post-processing[1] of the model's output. This follows directly

This work was supported in part by NSF under grant CIF CAREER 1845852.

[1]Broadly speaking, methods that correct a classifier for discrimination can be categorized as pre-processing (changing the input to a model) [25, 33, 34], in-processing (changing the model itself) [17, 24, 35], and post-processing (modifying a model's output) [18, 23] .

from the fact that the projection of a conditional distribution is an f-divergence-dependent tilting. The optimal post-processor only depends on a combination of well-calibrated probabilistic classifiers that predict both an outcome class as well as membership in a protected group. Thus, the model projection theory dictates that the problem of achieving a good fairness-accuracy trade-off can be directly mapped to a task that data scientists could do well: training accurate and well-calibrated prediction models. With these models in hand, an unfair classifier can be corrected by solving the model projection optimization.

All proofs can be found in the extended version at [36].

Notation. We denote $[c] \triangleq \{1, \cdots, c\}$ and use lowercase and uppercase bold letters to represent vectors (e.g., $\boldsymbol{v}$) and matrices (e.g., $\boldsymbol{G}$), respectively. We denote by $\mathbf{0}$ the vector with all entries equal to 0. The i-th coordinate of a vector $\boldsymbol{v}$ is denoted by $\boldsymbol{v}_i$, and the (i,j)-th entry of a matrix $\boldsymbol{G}$ by $\boldsymbol{G}_{i,j}$. The i-th row of $\boldsymbol{G}$ is denoted by $\boldsymbol{G}_{i,:}$. For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^c$, we write $\boldsymbol{a} \leq \boldsymbol{b}$ to indicate that $\boldsymbol{a}_i \leq \boldsymbol{b}_i$ for all $i \in [c]$. Lists of functions are indicated by superscripts. The set of all probability measures definable on a measurable space $(\mathcal{Y}, \Sigma)$ is denoted by $\boldsymbol{\Delta}_{\mathcal{Y}}$. When $\mathcal{Y} = [c]$ is a finite alphabet, $\boldsymbol{\Delta}_{[c]}$ is the probability simplex and we denote it by $\boldsymbol{\Delta}_c$ for short.

II. Model Projection Formulation

In this section, we first recall the definition of information projection and some of its properties. Then we formally introduce model projection, which can be viewed as an extension of information projection. We prove the existence and uniqueness of the optimal model and establish strong duality.

A. Information Projection

For a given reference probability distribution and a set of distributions, information projection seeks to find the "closest" distribution within this set to the reference one. Fix a probability space (Ω, Σ, P). For any subset $\mathscr{C} \subset \boldsymbol{\Delta}_\Omega$, let

$$D_f(\mathscr{C}\|P) \triangleq \inf_{Q \in \mathscr{C}} D_f(Q\|P). \tag{1}$$

Here for a convex $f : (0, \infty) \to \mathbb{R}$ the f-divergence [37, 38] is given by

$$D_f(Q\|P) \triangleq \mathbb{E}_P\left[f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\right] - f(1) \tag{2}$$

whenever Q is absolutely continuous with respect to (w.r.t.) P. We say that a $Q \in \mathscr{C}$ is the D_f-projection of P onto $\mathscr{C}$ if

$$D_f(Q\|P) = D_f(\mathscr{C}\|P) \tag{3}$$

and $D_f(R\|P) > D_f(\mathscr{C}\|P)$ whenever $Q \neq R \in \mathscr{C}$. The existence and uniqueness of the D_f-projection has been established under certain assumptions [14, 15]. Furthermore, an explicit formula for the D_{kl}-projection (also termed I-projection) under linear constraints was proved in [38].

B. Model Projection: Problem Setup

We introduce next the definition of model projection.

Definition 1. Consider a fixed random variable X and a probability space $(\mathcal{X}, \Sigma_1, P_X)$ such that $X \sim P_X$. Moreover, fix both a measurable space $(\mathcal{Y}, \Sigma_2)$ and a conditional distribution $P_{Y|X}$ from $\mathcal{X}$ to $\mathcal{Y}$. For a given convex set $\mathcal{F}$ of conditional distributions from $\mathcal{X}$ to $\mathcal{Y}$, the *model projection* of $P_{Y|X}$ onto $\mathcal{F}$ is given by the unique minimizer (if it exists) of

$$\inf_{W_{Y|X} \in \mathcal{F}} \mathbb{E}_X\left[D_f\left(W_{Y|X}(\cdot|X)\|P_{Y|X}(\cdot|X)\right)\right]. \tag{4}$$

The model projection is the "closest" model to the prescribed model $P_{Y|X}$, where we use the f-divergence to measure the "closeness". The choice of the f-divergence is determined by the application at hand.

In what follows, let $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = [c]$. In this setting, conditional distributions from $\mathcal{X}$ to $\mathcal{Y}$ become simply vector-valued functions. We reserve the letter $\boldsymbol{y} : \mathcal{X} \to \boldsymbol{\Delta}_c$ for $P_{Y|X}$

$$\boldsymbol{y}(x) \triangleq (P_{Y|X}(1|x), \cdots, P_{Y|X}(c|x)), \quad x \in \mathcal{X} \tag{5}$$

and denote an arbitrary conditional distribution from $\mathcal{X}$ to $\mathcal{Y}$ by a vector-valued function $\boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_c$. Then, (4) becomes

$$\inf_{\boldsymbol{h} \in \mathcal{F}} \mathbb{E}_X\left[D_f\left(\boldsymbol{h}(X)\|\boldsymbol{y}(X)\right)\right]. \tag{6}$$

The choice of the constraint set $\mathcal{F}$ is usually application-dependent. Throughout this paper, we consider a special case in which the constraint set is constructed via linear inequalities. In other words, for some given matrix-valued function $\boldsymbol{G} : \mathcal{X} \to \mathbb{R}^{k \times c}$ the constraint set is in the form

$$\mathcal{F} = \{\boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_c \mid \mathbb{E}\left[\boldsymbol{G}(X)\boldsymbol{h}(X)\right] \leq \mathbf{0}\}. \tag{7}$$

C. Connection between Information and Model Projection

We connect model projection (4) with information projection (1) next. Keeping the notation before equation (1), suppose $\Omega = \mathcal{X} \times \mathcal{Y}$ and that $P_{X,Y} \in \boldsymbol{\Delta}_\Omega$ is a probability measure that disintegrates into P_X and $P_{Y|X}$. Let $\mathscr{P} \subset \boldsymbol{\Delta}_\Omega$ be the subset of all probability measures that marginalize to P_X on $\mathcal{X}$, i.e.,

$$\mathscr{P} \triangleq \{Q \in \boldsymbol{\Delta}_\Omega \mid Q(A \times \mathcal{Y}) = P_X(A) \text{ for all } A \times \mathcal{Y} \subset \Sigma\}.$$

Then the model projection (4) is information projection onto a subset of $\mathscr{P}$. In other words, for a set $\mathcal{F}$ of conditional distributions, the model projection of $P_{Y|X}$ onto $\mathcal{F}$ is exactly information projection of $P_{X,Y}$ onto

$$\mathscr{C} \triangleq \{P_X W_{Y|X} \mid W_{Y|X} \in \mathcal{F}\} \subset \mathscr{P}. \tag{8}$$

It is important to note that $\mathscr{P}$ cannot be described by finitely many linear constraints, precisely because a distribution may not be determined by finitely many of its moments. Hence, the results on information projection subject to finitely many linear constraints do not seem applicable to model projection.

On the other direction, observe that model projection subsumes information projection. This fact is rather trivial, since for a singleton $\mathcal{X} = \{x\}$ the set $\Omega = \mathcal{X} \times \mathcal{Y}$ can be identified with $\mathcal{Y}$ via $(x, y) \leftrightarrow y$. Then, P_X is a trivial atom $P_X = \delta_x$ (and $\mathscr{P} = \boldsymbol{\Delta}_\Omega$) so the averaging in (4) collapses into only one term, whose minimization is precisely the problem of information projection.

III. Model Projection Theory

In this section, we first prove the existence and uniqueness of the model projection onto a linear subset under the general f-divergence setting. For the information projection framework with f-divergence measuring "distance", this problem has been studied [14] under the condition $f'(0^+) = -\infty$ to ensure that the projection onto the linear set belongs to the interior of $\mathbf{\Delta}_c$. This condition also appears in our result. Then we compute the model projection by establishing strong duality for a functional optimization over the Banach space $\mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c)$ of continuous conditional distributions.[2]

To start with, we introduce four assumptions, which will be the premises of our main theorems. These assumptions restrict the behavior of the f-divergence, the linear constraints (see (7)), the feasibility set, and the given conditional distribution $\boldsymbol{y}$, respectively. Our optimization is carried over the "interior"

$$\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c) \triangleq \left\{ \boldsymbol{h} \in \mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c) \mid \inf_{j,x} \boldsymbol{h}_j(x) > 0 \right\}. \tag{9}$$

Assumption I:

(a) The function $f : (0, \infty) \to \mathbb{R}$ is twice continuously-differentiable, $f(1) = 0, f'(0^+) = -\infty$, and $f''(t) > 0$ for every $t > 0$.
(b) The functions $\boldsymbol{G}_{i,j} : \mathcal{X} \to \mathbb{R}$ (for $(i, j) \in [k] \times [c]$) are bounded, differentiable, and have bounded gradients.
(c) There exists at least one conditional distribution $\boldsymbol{h} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$ satisfying $\mathbb{E}[\boldsymbol{G}(X)\boldsymbol{h}(X)] < \mathbf{0}$.
(d) The conditional distribution $\boldsymbol{y}$ belongs to $\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$, and each $\boldsymbol{y}_j$ has continuous and bounded partial derivatives.

Theorem 1. *Under Assumption I, there exists a unique* $\boldsymbol{h}^{\mathrm{opt}} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$ *solving the model projection problem*

$$\begin{aligned} \min_{\boldsymbol{h} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)} \quad & \mathbb{E}\left[D_f(\boldsymbol{h}(X) \| \boldsymbol{y}(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[\boldsymbol{G}(X)\boldsymbol{h}(X)\right] \leq \mathbf{0}. \end{aligned} \tag{10}$$

Theorem 1 guarantees the existence and uniqueness of the optimal model $\boldsymbol{h}^{\mathrm{opt}}$. In fact, this optimal model owns an explicit formula utilizing the convex conjugate of the f-divergence. Recall that the convex conjugate D_f^{conj} is defined as

$$D_f^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p}) \triangleq \sup_{\boldsymbol{q} \in \mathbf{\Delta}_c} \boldsymbol{v}^T \boldsymbol{q} - D_f(\boldsymbol{q} \| \boldsymbol{p}). \tag{11}$$

The formula of the optimal model shows that the model projection onto a set constructed by linear constraints can be obtained by tilting the reference model, where the tilting is expressible in terms of $\boldsymbol{v} : \mathcal{X} \times \mathbb{R}^k \to \mathbb{R}^c$ defined by

$$\boldsymbol{v}(x; \boldsymbol{\lambda}) \triangleq -\boldsymbol{G}(x)^T \boldsymbol{\lambda}. \tag{12}$$

Under Assumption I-(a), the derivative f' is strictly increasing, so one can define its inverse $\phi : (-\infty, M) \to (0, \infty)$ by

$$\phi(u) \triangleq (f')^{-1}(u), \tag{13}$$

where $M = \sup_{t>0} f'(t)$.

Theorem 2. *Under Assumption I, we have the formula*

$$\boldsymbol{h}_j^{\mathrm{opt}}(x) = \boldsymbol{y}_j(x)\phi(\gamma(x) + \boldsymbol{v}_j(x; \boldsymbol{\lambda}^*)), \quad (j, x) \in [c] \times \mathcal{X} \tag{14}$$

where the function $\gamma : \mathcal{X} \to \mathbb{R}$ *is uniquely defined by*

$$\mathbb{E}_{j \sim \boldsymbol{y}(x)} \phi(\gamma(x) + \boldsymbol{v}_j(x; \boldsymbol{\lambda}^*)) = 1, \quad x \in \mathcal{X}, \tag{15}$$

and $\boldsymbol{\lambda}^* \geq \mathbf{0}$ *is any solution to the convex optimization problem*

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathbb{E}\left[D_f^{\mathrm{conj}}(\boldsymbol{v}(X; \boldsymbol{\lambda}), \boldsymbol{y}(X)) \right]. \tag{16}$$

Remark 1. If $\mathcal{X}$ is finite, then Theorems 1 and 2 hold without the differentiability assumptions on the $\boldsymbol{G}_{i,j}$ and on the $\boldsymbol{y}_j$.

The duality approach reduces the infinite-dimensional optimization (10) into a tractable finite-dimensional one (16). Note that in our setting, a simple application of duality is inaccessible. The primal optimization (10) is equivalent to

$$\inf_{\boldsymbol{h} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathbb{E}\left[D_f(\boldsymbol{h}(X) \| \boldsymbol{y}(X)) + \boldsymbol{\lambda}^T \boldsymbol{G}(X)\boldsymbol{h}(X) \right], \tag{17}$$

which is not necessarily equal to the dual optimization

$$\sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\boldsymbol{h} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)} \mathbb{E}\left[D_f(\boldsymbol{h}(X) \| \boldsymbol{y}(X)) + \boldsymbol{\lambda}^T \boldsymbol{G}(X)\boldsymbol{h}(X) \right]. \tag{18}$$

The difficulty here is that the space $\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$ is not pre-compact. The minimax property does not hold in general if neither of the two optimization spaces is precompact. Our approach shows that, nevertheless, one may carve a precompact subset of $\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$ that is guaranteed to contain the sought optimizer. Note that strict convexity of f implies that the unique solution of the inner minimization in the dual (18) at any outer maximizer $\boldsymbol{\lambda}^*$ is in fact the unique solution to the primal problem (17) (i.e., it is the sought model projection of $\boldsymbol{y}$ onto $\mathcal{F} \cap \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$, see (7) and (9)).

Remark 2. Notably, for the KL-divergence, the model projection formula closely resembles that of the information projection. Analogous to the information projection formula under linear constraints, the model projection formula (14) for a fixed $x \in \mathcal{X}$ is an exponential tilt since for $f(t) = t \log t$ we have $\phi(u) = e^{u-1}$. The difference between the two projections is how the tilt is computed (i.e., in the value of the parameters $\boldsymbol{\lambda}^*$) where its value under the model projection setting reflects the fact that we are penalizing the average distance. The optimal parameters $\boldsymbol{\lambda}^*$ for the D_{kl}-projection onto a set constructed by linear inequalities $\int g_i \, dQ \leq 0$ are exactly the minimizers of

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \log \mathbb{E}\left[\mathbb{E}\left[\exp \sum_{i \in [k]} -\boldsymbol{\lambda}_i g_i(X, Y) \,\middle|\, X \right] \right]. \tag{19}$$

On the other hand, by plugging

$$D_{\mathsf{kl}}^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p}) = \log \sum_{j \in [c]} \boldsymbol{p}_j e^{\boldsymbol{v}_j}$$

into (16) the optimal parameters for the model projection

[2] We endow $\mathcal{X} = \mathbb{R}^m$ with the standard topology, and $\mathbf{\Delta}_c \subset \mathbb{R}^c$ with the subspace topology, so continuity of $\boldsymbol{h} : \mathcal{X} \to \mathbf{\Delta}_c$ refers to the usual definition of continuous functions between Euclidean spaces. Then, endowing $\mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c)$ with the sup-norm, $\|\boldsymbol{h}\|_\infty = \sup_{x \in \mathcal{X}} \|\boldsymbol{h}(x)\|$, turns it into a Banach space.

problem are solutions to (writing $g_i(x,y) \triangleq \boldsymbol{G}_{i,y}(x)$)

$$\min_{\boldsymbol{\lambda}\geq \mathbf{0}} \ \mathbb{E}\left[\log \mathbb{E}\left[\exp \sum_{i\in[k]} -\boldsymbol{\lambda}_i g_i(X,Y) \,\middle|\, X\right]\right]. \tag{20}$$

We note that formula (14) is valid for f-divergences beyond KL-divergence. To the best of our knowledge, an analogous formula for the information projection (i.e., for general f-divergences) does not appear in the literature.

IV. Application to Fair Machine Learning

In this section, we aim at designing a fairness-aware classifier. We formalize an optimization for this purpose which coincides with the model projection framework explored in the last section. Prior works attempt to design fair classifiers by implicitly solving a model projection problem, where accuracy is measured by, for example, KL-divergence [29] and cross-entropy [30]. Here we provide a general framework in the setting of multiclass classification, and this approach allows the usage of any f-divergence. In what follows, we formally introduce our formulation.

We consider a (multiclass) classification problem where the goal is to use an $\mathcal{X}$-valued input variable X (e.g., criminal history) to predict a target variable Y (e.g., criminal recidivism) taking values in $[c]$, with c denoting the number of classes. We denote a probabilistic classifier, which can be viewed as a conditional distribution, by $\boldsymbol{h}: \mathcal{X} \to \boldsymbol{\Delta}_c$. Hence, for each $x \in \mathcal{X}$, the classifier $\boldsymbol{h}$ assigns a probability vector $\boldsymbol{h}(x)$ that corresponds to a "belief" of the true value of Y given an observation $X = x$. The predicted output of the classifier $\boldsymbol{h}$ given X is denoted by $\widehat{Y}$. In other words, $\widehat{Y}$ is a $[c]$-valued random variable distributed according to

$$\Pr\left(\widehat{Y} = j \mid X = x\right) = \boldsymbol{h}_j(x), \quad (j,x) \in [c] \times \mathcal{X}. \tag{21}$$

As a measure of fairness, we evaluate the performance disparity w.r.t. a sensitive $[d]$-valued attribute S (e.g., race or gender) which correlates with X but not used as an input for the classification task. Nonetheless, we assume S is accessible when designing the classifier. Our goal is to design a classifier $\boldsymbol{h}^{\mathrm{opt}}: \mathcal{X} \to \boldsymbol{\Delta}_c$ that satisfies certain fairness criteria without compromising accuracy.

We assume that we have in hand a well-calibrated classifier that approximates $P_{Y,S|X}$, i.e. that predicts both group membership S and the true label Y from input variables X. This classifier can be directly marginalized into the following $d+2$ models:

- a label classifier $\boldsymbol{y}: \mathcal{X} \to \boldsymbol{\Delta}_c$ that predicts true label from input variables,

$$\boldsymbol{y}(x) \triangleq (P_{Y|X}(1|x), \cdots, P_{Y|X}(c|x)) \quad \text{for} \quad x \in \mathcal{X}, \tag{22}$$

- a group membership classifier $\boldsymbol{s}: \mathcal{X} \to \boldsymbol{\Delta}_d$ that uses input variables to predict group membership,

$$\boldsymbol{s}(x) \triangleq (P_{S|X}(1|x), \cdots, P_{S|X}(d|x)) \quad \text{for} \quad x \in \mathcal{X}, \tag{23}$$

Fairness Criterion	Expression
Statistical parity	$\left\lvert \frac{\Pr(\hat{Y}=\hat{y}\mid S=s)}{\Pr(\hat{Y}=\hat{y})} - 1 \right\rvert \leq \alpha$
Equalized odds	$\left\lvert \frac{\Pr(\hat{Y}=\hat{y}\mid Y=y, S=s)}{\Pr(\hat{Y}=\hat{y}\mid Y=y)} - 1 \right\rvert \leq \alpha$
Overall accuracy equality	$\left\lvert \frac{\Pr(\hat{Y}=Y\mid S=s)}{\Pr(\hat{Y}=Y)} - 1 \right\rvert \leq \alpha$

Table 1: Fairness criteria and their corresponding expressions. Here $\alpha > 0$ is a prescribed constant, and having a metric be satisfied amounts to having the corresponding inequalities hold *for every* $s \in [d]$ and $y, \hat{y} \in [c]$.

- a set of disparate treatment classifiers $\boldsymbol{y}^{(s)}: \mathcal{X} \to \boldsymbol{\Delta}_c$ that predict true label from input variables for each group $s \in [d]$,

$$\boldsymbol{y}^{(s)}(x) \triangleq (P_{Y|X,S=s}(1|x), \cdots, P_{Y|X,S=s}(c|x)) \tag{24}$$

for every $(s,x) \in [d] \times \mathcal{X}$.

In practice, the distribution $P_{Y,S}$ can be reliably estimated as its support size cd is usually small. The classifier that approximates $P_{Y,S|X}$ (and thus $\boldsymbol{y}$, $\boldsymbol{s}$, and $\boldsymbol{y}^{(s)}$) can be produced by training, e.g., a logistic regression. This may lead to a discrepancy between the underlying and the approximated classifiers. How this discrepancy impacts the design of the optimal classifier is still an open question that deserves future work.

A. Fairness Criteria

Many fairness criteria can be written as linear inequalities [see e.g., 27, 28] in terms of the classifier $\boldsymbol{h}$. Consequently, these fairness criteria can be mapped directly to the constraints in our model projection framework. We focus on three commonly-used fairness metrics (see Table 1) and provide their equivalent expressions in linear form in the following lemma.

Lemma 1. *Every fairness criterion listed in Table 1 can be written in the form*

$$\mathbb{E}\left[\langle \delta \boldsymbol{a}^{(i)}(X) - \alpha \boldsymbol{b}^{(i)}(X), \boldsymbol{h}(X)\rangle\right] \leq 0, \quad (i,\delta) \in [\ell] \times \{\pm 1\}$$

for a positive integer ℓ and functions $\boldsymbol{a}^{(i)}: \mathcal{X} \to \mathbb{R}^c$ and $\boldsymbol{b}^{(i)}: \mathcal{X} \to \mathbb{R}^c_{\geq 0}$ that are completely determined by the the classifiers $\{\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(d)}\}$ and the distributions P_S, $P_{S|Y}$, and where the expectation is taken w.r.t. P_X.

We briefly go over the forms of the $\boldsymbol{a}^{(i)}$ and $\boldsymbol{b}^{(i)}$ for the fairness metrics in Table 1. We let $\boldsymbol{e}^{(1)}, \cdots, \boldsymbol{e}^{(c)}$ denote the standard basis vectors of $\mathbb{R}^c$.

a) Statistical parity [21]: measures whether the predicted outcome $\widehat{Y}$ is independent of the sensitive attribute S. For statistical parity, the functions $\boldsymbol{a}^{(i)}$ and $\boldsymbol{b}^{(i)}$ have the forms

$$\boldsymbol{a}^{(s,\hat{y})}(x) = \left(\frac{\boldsymbol{s}_s(x)}{P_S(s)} - 1\right)\boldsymbol{e}^{(\hat{y})} \quad \text{and} \quad \boldsymbol{b}^{(s,\hat{y})}(x) = \boldsymbol{e}^{(\hat{y})}.$$

There are $2d \cdot c$ constraints since $(s,\hat{y}) \in [d] \times [c]$.

b) Equalized odds [18]: requires the predicted outcome $\widehat{Y}$ and the sensitive attribute S to be independent conditioned on the true label Y. When the classification task is binary, the equalized odds becomes the equality of false positive rate and false negative rate [20] over all sensitive groups. For equalized odds,

$$\boldsymbol{a}^{(s,\hat{y},y)}(x) = \left(\frac{\boldsymbol{s}_s(x)\boldsymbol{y}_y^{(s)}(x)}{P_{S|Y}(s|y)} - \boldsymbol{y}_y(x)\right)\boldsymbol{e}^{(\hat{y})},$$
$$\boldsymbol{b}^{(s,\hat{y},y)}(x) = \boldsymbol{y}_y(x)\boldsymbol{e}^{(\hat{y})}.$$

There are $2d \cdot c^2$ constraints.

c) Overall accuracy equality [21]: requires the accuracy of the predictive model to be the same across all sensitive groups. In this case,

$$\boldsymbol{a}^{(s)}(x) = \frac{\boldsymbol{s}_s(x)}{P_S(s)}\boldsymbol{y}^{(s)}(x) - \boldsymbol{y}(x) \quad \text{and} \quad \boldsymbol{b}^{(s)}(x) = \boldsymbol{y}(x).$$

There are $2d$ constraints.

B. Discrimination Correction

Here we consider designing a fair classifier via a discrimination-correction optimization that is a special instance of the model projection problem. Equipped with Lemma 1, we formulate the discrimination-correction optimization problem using f-divergence as a measure of "closeness":

$$\begin{aligned}\min_{\boldsymbol{h}\in\mathcal{C}_+(\mathcal{X},\boldsymbol{\Delta}_c)} \ & \mathbb{E}\left[D_f(\boldsymbol{h}(X)\|\boldsymbol{y}(X))\right], \\ \text{s.t. } & \mathbb{E}\left[\left\langle \delta\boldsymbol{a}^{(i)}(X) - \alpha\boldsymbol{b}^{(i)}(X), \boldsymbol{h}(X)\right\rangle\right] \le 0,\end{aligned} \tag{25}$$

where $\alpha > 0$ and the functions $\boldsymbol{a}^{(i)}$ and $\boldsymbol{b}^{(i)}$ (for $i \in [\ell]$) are all determined by the pre-specified fairness requirements, and $\delta \in \{\pm 1\}$. Recall that $\boldsymbol{G}$ is a matrix with 2ℓ rows encoding the constraints, i.e.,

$$\boldsymbol{G} = \left(\delta\boldsymbol{a}^{(i)} - \alpha\boldsymbol{b}^{(i)}\right)^T_{(\delta,i)\in\{\pm1\}\times[\ell]}, \tag{26}$$

and $\boldsymbol{v}(x;\boldsymbol{\lambda}) = -\boldsymbol{G}(x)^T\boldsymbol{\lambda}$ (see (12)). Consequently, Theorems 1 and 2 together guarantee the existence and uniqueness of the optimal classifier and they also give a way for designing such classifier (see (14)). For the sake of illustration, we give the following formula for the optimal classifier when accuracy is measured in terms of the KL-divergence. It is worth noting that this formula also appears in [29], but no explicit formula for the optimal dual parameter $\boldsymbol{\lambda}^*$ is presented therein.

Corollary 1. *Assume the KL-divergence is used in (25). Then, under Assumption I, the optimal fair classifier is given by*

$$\boldsymbol{h}_j^{\mathrm{opt}}(x) \propto \boldsymbol{y}_j(x)e^{\boldsymbol{v}_j(x;\boldsymbol{\lambda}^*)} \tag{27}$$

where $\boldsymbol{\lambda}^*$ is any solution to the convex optimization problem

$$\min_{\boldsymbol{\lambda}\ge\boldsymbol{0}} \ \mathbb{E}\left[\log \mathbb{E}_{j\sim\boldsymbol{y}(X)}\left[e^{\boldsymbol{v}_j(X;\boldsymbol{\lambda})}\right]\right]. \tag{28}$$

Remark 3. Assumption I is satisfied for the fairness criteria considered in this paper as soon as $\min_{s,y} P_{S|Y}(s|y) > 0$, and $\boldsymbol{y}, \boldsymbol{s}$, and the $\boldsymbol{y}^{(s)}$ satisfy Assumption I-(d). This is true since Assumption I-(a) is satisfied for the KL-divergence, Assumption I-(b) will also be satisfied in view of the formulas for the fairness constraints given in Section IV-A, and Assumption I-(c) is satisfied as the uniform classifier is strictly feasible.

The way we design the fair classifier falls into the post-processing category. This is because the optimal fair classifier is a tilting of the label classifier (see Theorem 2 and Corollary 1). Notably, the formulation (25) does not *a priori* assume a post-processing design procedure. Nevertheless, the optimal classifier turns out to own an optimality guarantee among all classifiers.

We point out that the formulation in [30] presents a special case of the model projection theory using cross-entropy as the f-divergence of choice and assuming Y and S are binary. While computationally lightweight, the experiments in [30, Section 6] demonstrate that the model projection formulation may perform favorably compared to state-of-the-art fairness intervention mechanisms. Here, we provide a general theoretical work that allows usage of a wide class of f-divergences. We refer the reader to [30, Section 6] for numerical results and comparisons, and omit further experiments due to space constraints.

C. Finite-Sample Considerations

The model projection framework gives an explicit way for designing a fairness-aware classifier by first training a classifier $P_{Y,S|X}$, and then solving a convex program to obtain the dual parameter. Therefore, there are only two challenges for a complete design process of a discrimination-correction classifier: 1) obtaining a well-calibrated classifier $P_{Y,S|X}$, and 2) solving the dual convex program (16). This subsection tackles the second challenge, under the assumption that the first challenge is addressed.

The convex program relies on the underlying data distribution. In practice, with finitely-many samples, one can solve the dual convex program using an empirical objective function. Keeping the assumption that the classifier $P_{Y,S|X}$ is known, and letting $\{X_i\}_{i\in[n]}$ be i.i.d. samples drawn from P_X, we show the following generalization bound for the dual problem (16).

Theorem 3. *Let $\boldsymbol{G}$ be given by equation (26), U be a $[c]$-valued random variable such that $U|X = x$ is uniform for every x, and denote*

$$\theta \triangleq \frac{cD_f(P_XP_{U|X}\|P_{X,Y})}{-\max_{i\in[2\ell]}\mathbb{E}\left[\boldsymbol{G}_{i,:}(X)\boldsymbol{1}\right]}, \tag{29}$$

$L \triangleq \sup_{x\in\mathcal{X}}\|\boldsymbol{G}(x)\|_1$, *and* $\zeta \triangleq L/\theta$. *Let $\boldsymbol{\lambda}_n$ be the unique solution to*

$$\min_{\substack{\boldsymbol{\lambda}\ge\boldsymbol{0}\\ \|\boldsymbol{\lambda}\|_1\le\theta}} \frac{1}{n}\sum_{i\in[n]} D_f^{\mathrm{conj}}(\boldsymbol{v}(X_i;\boldsymbol{\lambda}),\boldsymbol{y}(X_i)) + \frac{\zeta}{\sqrt{n}}\|\boldsymbol{\lambda}\|_2^2.$$

Then, with probability at least $1-\delta$,

$$\begin{aligned}&\mathbb{E}\left[D_f^{\mathrm{conj}}(\boldsymbol{v}(X;\boldsymbol{\lambda}_n),\boldsymbol{y}(X))\right] \\ &\le \min_{\boldsymbol{\lambda}\ge\boldsymbol{0}}\mathbb{E}\left[D_f^{\mathrm{conj}}(\boldsymbol{v}(X;\boldsymbol{\lambda}),\boldsymbol{y}(X))\right] + \frac{10L\theta}{\delta\sqrt{n}}.\end{aligned} \tag{30}$$

REFERENCES

[1] N. N. Chentsov, "Nonsymmetrical distance between probability distributions, entropy and the theorem of pythagoras," *Mathematical notes of the Academy of Sciences of the USSR*, vol. 4, no. 3, pp. 686–691, Sep 1968. [Online]. Available: https://doi.org/10.1007/BF01116448

[2] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 02 1975. [Online]. Available: https://doi.org/10.1214/aop/1176996454

[3] I. Csiszar and F. Matúš, "Information projections revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1474–1490, June 2003.

[4] A. Dembo and O. Zeitouni, "Refinements of the gibbs conditioning principle," *Probability theory and related fields*, vol. 104, no. 1, pp. 1–14, 1996.

[5] X. Yang and A. R. Barron, "Minimax compression and large alphabet approximation through poissonization and tilting," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2866–2884, 2017.

[6] I. Csiszar, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, no. 3, pp. 768–793, 08 1984. [Online]. Available: https://doi.org/10.1214/aop/1176993227

[7] F. Topsøe, "Information-theoretical optimization techniques," *Kybernetika*, vol. 15, no. 1, pp. (8)–27, 1979. [Online]. Available: http://eudml.org/doc/27475

[8] A. R. Barron, "Limits of information, markov chains, and projection," in *2000 IEEE International Symposium on Information Theory (Cat. No.00CH37060)*, June 2000, pp. 25–.

[9] N. Slonim, "The information bottleneck : Theory and applications," 2006.

[10] R. M. Bell and T. M. Cover, "Competitive optimality of logarithmic investment," *Mathematics of Operations Research*, vol. 5, no. 2, pp. 161–166, 1980.

[11] M. Ashok Kumar and I. Sason, "Projection theorems for the Rényi divergence on α -convex sets," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 4924–4935, Sep. 2016.

[12] M. A. Kumar and R. Sundaresan, "Minimization problems based on relative α-entropy i: Forward projection," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5063–5080, Sep. 2015.

[13] ——, "Minimization problems based on relative α-entropy ii: Reverse projection," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5081–5095, Sep. 2015.

[14] I. Csiszár, "Generalized projections for non-negative functions," in *Proceedings of 1995 IEEE International Symposium on Information Theory*, Sep. 1995, pp. 6–.

[15] I. Csiszár, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol. 68, no. 1, pp. 161–186, Mar 1995. [Online]. Available: https://doi.org/10.1007/BF01874442

[16] H. Wang, B. Ustun, and F. P. Calmon, "On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning," in *Proceedings of 2018 IEEE International Symposium on Information Theory*, 2018, pp. 126–130.

[17] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 560–568.

[18] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[19] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[20] 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.

[21] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, 2018.

[22] I. Žliobaitė, "Measuring discrimination in algorithmic decision making," *Data Min. Knowl. Discov.*, vol. 31, no. 4, pp. 1060–1089, Jul. 2017. [Online]. Available: https://doi.org/10.1007/s10618-017-0506-1

[23] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Conference on Learning Theory*, 2017, pp. 1920–1953.

[24] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 107–118.

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

[26] H. Wang, B. Ustun, and F. P. Calmon, "Repairing without retraining: Avoiding disparate impact with counterfactual distributions," in *International Conference on Machine Learning*, 2019.

[27] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*, 2018, pp. 60–69.

[28] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees," in *Conference on Fairness, Accountability, and Transparency*, 2019, pp. 319–328.

[29] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," *arXiv preprint arXiv:1901.04966*, 2019.

[30] D. Wei, K. N. Ramamurthy, and F. P. Calmon, "Optimized score transformation for fair classification," in *23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[31] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," in *Proceedings of 2018 IEEE International Symposium on Information Theory*, 2018, pp. 176–180.

[32] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *ArXiv*, vol. abs/1810.08810, 2018.

[33] 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.

[34] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[35] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, 2017, pp. 962–970.

[36] W. Alghamdi, S. Asoodeh, H. Wang, F. P. Calmon, D. Wei, and K. N. Ramamurthy, "Model projection: Theory and applications to fair machine learning," 2020. [Online]. Available: https://github.com/WaelAlghamdi/ModelProjection

[37] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of Royal Statistics*, vol. 28, pp. 131–142, 1966.

[38] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.