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 from the fact that the projection of a conditional distribution is

[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] .

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 should 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 of this paper at [36].

Notation. We denote $[c] \triangleq \{1, \cdots, c\}$ and use lowercase and uppercase bold letters to represent vectors (e.g., $\mathbf{v}$) and matrices (e.g., $\mathbf{G}$), respectively. We denote by $\mathbf{0}$ the vector with all entries equal to 0. The i-th coordinate of a vector $\mathbf{v}$ is denoted by $\mathbf{v}_i$, and the (i,j)-th entry of a matrix $\mathbf{G}$ by $\mathbf{G}_{i,j}$. The j-th column of $\mathbf{G}$ is denoted by $\mathbf{G}_{:,j}$. For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^c$, we write $\mathbf{a} \leq \mathbf{b}$ to indicate that $\mathbf{a}_i \leq \mathbf{b}_i$ 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 $\mathbf{\Delta}_{\mathcal{Y}}$. When $\mathcal{Y} = [c]$ is a finite alphabet, $\mathbf{\Delta}_{[c]}$ is the probability simplex and we denote it by $\mathbf{\Delta}_c$ 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 $\mathscr{C} \subset \mathbf{\Delta}_\Omega$, 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 is proved [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 $\mathbf{y} : \mathcal{X} \to \mathbf{\Delta}_c$ for $P_{Y|X}$

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

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

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

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 $\mathbf{G} : \mathcal{X} \to \mathbb{R}^{c \times k}$ the constraint set is in the form

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

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 $P_{X,Y} \in \mathbf{\Delta}_\Omega$ is a probability measure that disintegrates into P_X and $P_{Y|X}$. Let $\mathscr{P} \subset \mathbf{\Delta}_\Omega$ be the subset of all probability measures that marginalize to P_X on $\mathcal{X}$, i.e.,

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

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 $\mathscr{P} = \mathbf{\Delta}_\Omega$) 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 $\mathbf{y}$, respectively. Our optimization is carried over the "interior"

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

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 $\mathbf{G}_{i,j} : \mathcal{X} \to \mathbb{R}$ (for $(i, j) \in [k] \times [c]$) are bounded, differentiable, and have bounded gradients.
(c) There exists at least one conditional distribution $\mathbf{h} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$ satisfying $\mathbb{E}\left[\mathbf{h}(X)^T \mathbf{G}(X)\right] < \mathbf{0}$.
(d) The conditional distribution $\mathbf{y}$ belongs to $\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$, and each $\mathbf{y}_j$ has bounded partial derivatives.

Under Assumption I-(a), the derivative f' is strictly increasing, so one can define its inverse $\phi : (-\infty, M) \to (0, \infty)$, $\phi(u) \triangleq (f')^{-1}(u)$, where $M = \sup_{t>0} f'(t)$.

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

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

Theorem 1 guarantees the existence and uniqueness of the optimal model $\mathbf{h}^{\mathrm{opt}}$. 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

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

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 $\mathbf{v} : \mathcal{X} \times \mathbb{R}^k \to \mathbb{R}^c$ defined by

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

[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 $\mathbf{h} : \mathcal{X} \to \mathbf{\Delta}_c$ refers to the usual definition of continuous functions between Euclidean spaces. Then, endowing $\mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c)$ with the sup-norm, $\|\mathbf{h}\|_\infty = \sup_{x \in \mathcal{X}} \|\mathbf{h}(x)\|$, turns it into a Banach space.

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

$$\mathbf{h}_j^{\mathrm{opt}}(x) = \mathbf{y}_j(x) \phi(\gamma(x) + \mathbf{v}_j(x; \boldsymbol{\lambda}^*)), \quad (j, x) \in [c] \times \mathcal{X} \tag{13}$$

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

$$\mathbb{E}_{j \sim \mathbf{y}(x)} \phi(\gamma(x) + \mathbf{v}_j(x; \boldsymbol{\lambda}^*)) = 1, \quad x \in \mathcal{X}, \tag{14}$$

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

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathbb{E}\left[D_f^{\mathrm{conj}}(\mathbf{v}(X; \boldsymbol{\lambda}), \mathbf{y}(X))\right]. \tag{15}$$

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

The duality approach reduces the infinite-dimensional optimization (10) into a tractable finite-dimensional one (15). Note that in our setting, a simple application of duality is inaccessible. The primal optimization (10) is equivalent to

$$\inf_{\mathbf{h} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathbb{E}\left[D_f(\mathbf{h}(X) \| \mathbf{y}(X)) + \mathbf{h}(X)^T \mathbf{G}(X) \boldsymbol{\lambda}\right], \tag{16}$$

which is not necessarily equal to the dual optimization

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

The difficulty here is that the space $\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$ is not precompact. 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 (17) at any outer maximizer $\boldsymbol{\lambda}^*$ is in fact the unique solution to the primal problem (16) (i.e., it is the sought model projection of $\mathbf{y}$ 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 (13) 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 over $\mathscr{C}$ are exactly the minimizers of (writing $g_i(x, y) \triangleq \mathbf{G}_{i,y}(x)$)

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \log \mathbb{E}\left[\mathbb{E}\left[\exp \sum_{i \in [k]} \boldsymbol{\lambda}_i g_i(X, Y) \,\middle|\, X\right]\right]. \tag{18}$$

On the other hand, by plugging

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

into (15) the optimal parameters for the model projection

problem are solutions to

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

We note that formula (13) 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 $\mathbf{h}:\mathcal{X}\to\mathbf{\Delta}_c$. Hence, for each $x\in\mathcal{X}$, the classifier $\mathbf{h}$ assigns a probability vector $\mathbf{h}(x)$ that corresponds to a "belief" of the true value of Y given an observation $X=x$. The predicted output of the classifier $\mathbf{h}$ given X is denoted by $\widehat{Y}$. In other words, $\widehat{Y}$ is a $[c]$-valued random variable distributed according to

$$\Pr\left(\widehat{Y}=j \mid X=x\right)=\mathbf{h}_j(x), \quad (j,x)\in[c]\times\mathcal{X}. \quad (20)$$

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 $\mathbf{h}^{\mathrm{opt}}:\mathcal{X}\to\mathbf{\Delta}_c$ 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 $\mathbf{y}:\mathcal{X}\to\mathbf{\Delta}_c$ that predicts true label from input variables,

$$\mathbf{y}(x)\triangleq(P_{Y|X}(1|x),\cdots,P_{Y|X}(c|x)) \quad \text{for} \quad x\in\mathcal{X}, \quad (21)$$

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

$$\mathbf{s}(x)\triangleq(P_{S|X}(1|x),\cdots,P_{S|X}(d|x)) \quad \text{for} \quad x\in\mathcal{X}, \quad (22)$$

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 $\mathbf{y}^{(s)}:\mathcal{X}\to\mathbf{\Delta}_c$ that predict true label from input variables for each group $s\in[d]$,

$$\mathbf{y}^{(s)}(x)\triangleq(P_{Y|X,S=s}(1|x),\cdots,P_{Y|X,S=s}(c|x)) \quad (23)$$

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 $\mathbf{y},\mathbf{s}$, and $\mathbf{y}^{(s)}$) 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 $\mathbf{h}$. 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*

$$\mathbb{E}\left[\langle\delta\mathbf{a}^{(i)}(X)-\alpha\mathbf{b}^{(i)}(X),\mathbf{h}(X)\rangle\right]\leq 0, \quad (i,\delta)\in[\ell]\times\{\pm1\}$$

for a positive integer ℓ and functions $\mathbf{a}^{(i)}:\mathcal{X}\to\mathbb{R}^c$ and $\mathbf{b}^{(i)}:\mathcal{X}\to\mathbb{R}^c_{\geq 0}$ that are completely determined by the the classifiers $\{\mathbf{y},\mathbf{s},\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(d)}\}$ 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 $\mathbf{a}^{(i)}$ and $\mathbf{b}^{(i)}$ for the fairness metrics in Table 1. We let $\mathbf{e}^{(1)},\cdots,\mathbf{e}^{(c)}$ 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 $\mathbf{a}^{(i)}$ and $\mathbf{b}^{(i)}$ have the forms

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

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,

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

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,

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

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":

$$\begin{aligned}\min_{\mathbf{h}\in\mathcal{C}_+(\mathcal{X},\mathbf{\Delta}_c)} \ & \mathbb{E}\left[D_f(\mathbf{h}(X)\|\mathbf{y}(X))\right], \\ \text{s.t. } & \mathbb{E}\left[\langle \delta\mathbf{a}^{(i)}(X) - \alpha\mathbf{b}^{(i)}(X), \mathbf{h}(X)\rangle\right] \le 0,\end{aligned} \tag{24}$$

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

$$\mathbf{G} = \left(\delta\mathbf{a}^{(i)} - \alpha\mathbf{b}^{(i)}\right)_{(\delta,i)\in\{\pm1\}\times[\ell]}, \tag{25}$$

and $\mathbf{v}(x;\boldsymbol{\lambda}) = -\mathbf{G}(x)\boldsymbol{\lambda}$ (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 (13)). 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 (24). Then, under Assumption I, the optimal fair classifier is given by*

$$\mathbf{h}_j^{\text{opt}}(x) \propto \mathbf{y}_j(x)e^{\mathbf{v}_j(x;\boldsymbol{\lambda}^*)} \tag{26}$$

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

$$\min_{\boldsymbol{\lambda}\ge\mathbf{0}} \mathbb{E}\left[\log \mathbb{E}_{j\sim\mathbf{y}(X)}\left[e^{\mathbf{v}_j(X;\boldsymbol{\lambda})}\right]\right]. \tag{27}$$

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 $\mathbf{y}, \mathbf{s}$, and the $\mathbf{y}^{(s)}$ 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 (24) 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 (15). 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 (15).

Theorem 3. *Let $\mathbf{G}$ be given by equation (25), U be a $[c]$-valued random variable such that $U|X = x$ is uniform for every x, and denote*

$$\theta \triangleq \frac{cD_f(P_XP_{U|X}\|P_{X,Y})}{-\max_{j\in[2\ell]}\mathbb{E}\left[\mathbf{1}^T\mathbf{G}_{:,j}(X)\right]}, \tag{28}$$

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

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

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

$$\begin{aligned}&\mathbb{E}\left[D_f^{\text{conj}}(\mathbf{v}(X;\boldsymbol{\lambda}_n),\mathbf{y}(X))\right] \\ &\le \min_{\boldsymbol{\lambda}\ge\mathbf{0}}\mathbb{E}\left[D_f^{\text{conj}}(\mathbf{v}(X;\boldsymbol{\lambda}),\mathbf{y}(X))\right] + \frac{10L\theta}{\delta\sqrt{n}}.\end{aligned} \tag{29}$$

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

[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.

[39] A. J. Kurdila and M. Zabarankin, *Convex functional analysis*. Springer Science & Business Media, 2006.

[40] B. Hajek and M. Raginsky, "Statistical learning theory," *Lecture Notes*, vol. 387, 2019.

APPENDIX

We consider a more general statement here which will naturally imply Theorems 1 and 2. Then, we prove Lemma 1, and we end with a proof of Theorem 3.

A. Notation

For the starting points below, we assume only that $\mathcal{X}$ is a topological space (i.e., we do not assume $\mathcal{X} = \mathbb{R}^m$ yet). Let $\mathcal{C}(\mathcal{X})$ denote the Banach space of continuous and bounded functions

$$\mathcal{C}(\mathcal{X}) \triangleq \left\{ h : \mathcal{X} \to \mathbb{R}^c \mid h \text{ continuous and } \sup_{x \in \mathcal{X}} \|h(x)\|_1 < \infty \right\}, \tag{30}$$

which is Banach when equipped with the sup norm, i.e., for $h \in \mathcal{C}(\mathcal{X})$

$$\|h\|_\infty \triangleq \sup_{x \in \mathcal{X}} \|h(x)\|_1. \tag{31}$$

Consider the following optimization problem

$$\begin{aligned} \min_{h \in \mathcal{C}(\mathcal{X})} \quad & \mathbb{E}\left[F(X, h(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[G_i(X, h(X))\right] \leq 0, \quad i \in [k]. \end{aligned} \tag{32}$$

where F and $G_1, \cdots, G_k$ are functions defined on $\mathcal{X} \times \mathbb{R}^c$ and taking values in $\mathbb{R} \cup \{\infty\}$. We denote by $\mathcal{C}(\mathcal{X}, \mathcal{Z}) \subset \mathcal{C}(\mathcal{X})$, for $\mathcal{Z} \subset \mathbb{R}^c$, the subset of functions taking values in $\mathcal{Z}$, i.e.,

$$\mathcal{C}(\mathcal{X}, \mathcal{Z}) \triangleq \{h \in \mathcal{C}(\mathcal{X}) \mid h(x) \in \mathcal{Z} \text{ for every } x \in \mathcal{X}\}. \tag{33}$$

Note that $\mathcal{C}(\mathcal{X}, \mathcal{Z})$ is closed or convex if $\mathcal{Z}$ is closed (in $\mathbb{R}^c$) or convex, respectively. Therefore, $\mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c)$ is a convex Banach space (for any $\mathcal{X}$). However, it is not compact in general. Therefore, it might not be straightforward to tackle restricted optimization problem (32) even when restricted to only $\mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c)$. Therefore, we tackle (32) indirectly via solving a much more restricted problem of the form

$$\begin{aligned} \min_{h \in \mathcal{K}} \quad & \mathbb{E}\left[F(X, h(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[G_i(X, h(X))\right] \leq 0, \quad i \in [k] \end{aligned} \tag{34}$$

for a compact subset $\mathcal{K} \subset \mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c)$ then showing that the problem (34) produces a global optimizer. We also consider ε-truncations of the simplex

$$\mathbf{\Delta}_c^\varepsilon \triangleq \mathbf{\Delta}_c \cap [\varepsilon, 1]^c \tag{35}$$

and the corresponding space

$$\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c) \triangleq \bigcup_{\varepsilon > 0} \mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c^\varepsilon). \tag{36}$$

We set

$$\mathbf{\Delta}_c^+ \triangleq \mathbf{\Delta}_c \cap (0, 1]^c. \tag{37}$$

We let $\mathcal{F}$ denote the feasibility region in (32), i.e.,

$$\mathcal{F} \triangleq \left\{ h \in \mathcal{C}(\mathcal{X}) \mid \max_{i \in [k]} \mathbb{E}\left[G_i(X, h(X))\right] \leq 0 \right\}. \tag{38}$$

We denote by $\mathcal{S}$ the strict-feasibilitiy region, i.e.,

$$\mathcal{S} \triangleq \left\{ h \in \mathcal{C}(\mathcal{X}) \mid \max_{i \in [k]} \mathbb{E}\left[G_i(X, h(X))\right] < 0 \right\}. \tag{39}$$

We let $\mathcal{D}$ be the set of functions in $\mathcal{C}(\mathcal{X})$ at which the objective function and the constraints are integrable[3]

$$\mathcal{D} \triangleq \left\{ h \in \mathcal{C}(\mathcal{X}) \mid \max\left(\mathbb{E}\left[|F(X, h(X))|\right], \max_{i \in [k]} \mathbb{E}\left[|G_i(X, h(X))|\right] \right) < \infty \right\}. \tag{40}$$

For a function $\psi : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$, the domain of ψ is the set of points at which ψ is defined and finite

$$\textbf{dom } \psi \triangleq \{v \in \mathcal{V} \mid \psi(v) < \infty\}. \tag{41}$$

[3]We say that a function $V : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is integrable if $\mathbb{E}\left[|V(X)|\right] < \infty$.

We define the intersection of domains

$$D \triangleq \bigcap_{x\in\mathcal{X}} \{p \in \mathbb{R}^c \mid \max(F(x,p), G_1(x,p), \cdots, G_k(x,p)) < \infty\}. \tag{42}$$

We also consider the domain

$$D' \triangleq \bigcap_{x\in\mathcal{X}} \{p \in [0,1]^c \mid \max(G_1(x,p), \cdots, G_k(x,p)) < \infty\} \tag{43}$$

and the intersection of domains

$$\mathcal{D}' \triangleq \{h \in \mathcal{C}(\mathcal{X}, [0,1]^c) \mid \max(\mathbb{E}\left[|G_1(X,h(X))|\right], \cdots, \mathbb{E}\left[|G_k(X,h(X))|\right]) < \infty\}. \tag{44}$$

We denote the convex hull and closure of a set $\mathcal{A}$ by $\mathrm{co}(\mathcal{A})$ and $\overline{\mathcal{A}}$, respectively. Abusing notation, we will also denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. We denote the indicator function of a set $\mathcal{U} \subset \mathcal{C}(\mathcal{X})$ by $\mathbb{I}_{\mathcal{U}}$

$$\mathbb{I}_{\mathcal{U}}(h) \triangleq \begin{cases} 0 & \text{if } h \in \mathcal{U}, \\ \infty & \text{otherwise.} \end{cases} \tag{45}$$

We define extended functionals $\mathcal{A}, \mathcal{B}_1, \cdots, \mathcal{B}_k : \mathcal{C}(\mathcal{X}) \to \overline{\mathbb{R}}$ by

$$\mathcal{A}(h) \triangleq \mathbb{E}\left[F(X,h(X))\right] + \mathbb{I}_{\mathcal{D}}(h), \tag{46}$$

$$\mathcal{B}_i(h) \triangleq \mathbb{E}\left[G_i(X,h(X))\right] + \mathbb{I}_{\mathcal{D}}(h), \quad i \in [k]. \tag{47}$$

In these definitions, it is understood that the value ∞ is assigned outside the set $\mathcal{D}$ regardless of whether the original function is defined and regardless of its value if it is defined there, e.g., if $h \in \mathcal{C}(\mathcal{X})$ is such that $F(\cdot, h(\cdot))$ is not integrable or if its integral is $-\infty$ then $\mathcal{A}(h)$ is defined to be ∞ because $h \notin \mathcal{D}$.

For a function $\psi : \mathcal{V} \to \mathcal{W}^n$, we let $\psi_i : \mathcal{V} \to \mathcal{W}$ denote its i-th part, i.e., $\psi(v) = (\psi_1(v), \cdots, \psi_n(v))$. Furthermore, if $\mathcal{W} = \mathcal{Z}^m$, we let $\psi_{i,j}(v) \in \mathcal{Z}$ denote the j-th coordinate of $\psi_i(v)$, i.e., $\psi_i(v) = (\psi_{i,1}(v), \cdots, \psi_{i,m}(v))$. For $\beta : \mathbb{R}^n \to \mathbb{R}$ and $\ell \in [n]$, the notation $\partial_\ell \beta$ will refer to the partial derivative of β with respect to its ℓ-th input.

Recall the definition of the convex conjugate (see equation (11)).

Definition 2 (Convex Conjugate)**.** The convex conjugate of a proper[4] function $W : \mathbf{\Delta}_c \to \overline{\mathbb{R}}$ is the function $W^{\mathrm{conj}} : \mathbb{R}^c \to \overline{\mathbb{R}}$ defined by

$$W^{\mathrm{conj}}(\mathbf{v}) \triangleq \sup_{\mathbf{p}\in\mathbf{\Delta}_c} \langle \mathbf{v}, \mathbf{p} \rangle - W(\mathbf{p}). \tag{48}$$

For a fixed $x \in \mathcal{X}$, we denote the convex conjugate of $F(x,\cdot)|_{\mathbf{\Delta}_c}$ at $\mathbf{v}$ by $F^{\mathrm{conj}}(x, \mathbf{v})$.

We will prove results under some subset of assumptions that we introduce here and in the beginning of the following section. The first set of assumptions has to do with the well-definededness of our optimization problem, and it will be sufficient to develop the general theory.

Assumption I'.

(a) The set $\mathcal{D}$ is nonempty.
(b) For each $J \in \{F, G_1, \cdots, G_k\}$, the function $\inf_{h\in\mathcal{D}} J(\cdot, h(\cdot))$ is lower bounded by an integrable function.
(c) For each $J \in \{F, G_1, \cdots, G_k\}$ and $x \in \mathcal{X}$, the function $J(x,\cdot)$ is lower-semicontinuous.
(d) For each $x \in \mathcal{X}$, the function $F(x,\cdot)$ is strictly convex. For each $(i,x) \in [k] \times \mathcal{X}$, the function $G_i(x,\cdot)$ is convex.

Note that the Assumption I'.b is satisfied if, e.g., the functions $F, G_1, \cdots, G_k$, are lower bounded by a constant.

Under Assumption 1, the optimization problem (34), over a nonempty convex and compact $\mathcal{K}$, has a unique solution. We state this result here, and relegate the proof to Appendix C.

Lemma 2. *Suppose Assumption I'.a-c holds. For a nonempty compact set $\mathcal{K} \subset \mathcal{D}$, the following optimization problem has a minimizer*

$$\begin{aligned} \min_{h\in\mathcal{K}} \quad & \mathcal{A}(h), \\ \text{s.t.} \quad & \mathcal{B}_i(h) \le 0, \quad i \in [k]. \end{aligned} \tag{49}$$

If, in addition, $\mathcal{K}$ is convex and Assumption 1.d holds, then the minimizer is unique.

[4]We say W is proper if **dom** W is nonempty.

Next, we show how a unique minimizer of (34) can be obtained from the dual problem. This procedure is possible thanks to Sion's minimax theorem. It will be useful to introduce the following quantities. First, the following term will bound the norm of optimal dual variables corresponding to the dual of the optimization problem (32).

Definition 3. For $q \in \mathcal{S} \cap \mathcal{D}$, we define

$$\theta_q \triangleq \frac{\mathcal{A}(q) - \inf_{h \in \mathcal{D}} \mathcal{A}(h)}{-\max_{i \in [k]} \mathcal{B}_i(q)}. \tag{50}$$

Next, we define the Lagrangian of the optimization problem (32).

Definition 4. Define the Lagrangian function $\mathcal{L} : \mathcal{D} \times \mathbb{R}^k_{\geq 0} \to \mathbb{R}$ by

$$\mathcal{L}(h, \boldsymbol{\lambda}) \triangleq \mathbb{E}\left[F(X, h(X)) + \sum_{i \in [k]} \lambda_i G_i(X, h(X))\right] = \mathcal{A}(h) + \sum_{i \in [k]} \lambda_i \mathcal{B}_i(h). \tag{51}$$

We use the following notation for what will be shown to be a class of models that contains the optimal model.

Definition 5. For fixed $\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}$ and $\mathcal{Z} \subset D$, define $q^{\mathcal{Z}}_{\boldsymbol{\lambda}} : \mathcal{X} \to \mathcal{Z}$ by

$$q^{\mathcal{Z}}_{\boldsymbol{\lambda}}(x) \triangleq \operatorname*{argmin}_{p \in \mathcal{Z}} F(x, p) + \sum_{i \in [k]} \lambda_i G_i(x, p), \quad x \in \mathcal{X}, \tag{52}$$

if the minimization in (52) has a unique solution[5] for every $x \in \mathcal{X}$.

B. Proof of Theorems 1 and 2

The main theorem underlying our results is as follows.

Theorem 4. *Suppose Assumption I' holds. Let $\mathcal{Z} \subset D$ be convex and compact such that $\mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{S}$ is nonempty, say $v \in \mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{S}$, and set $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0} \mid \|\boldsymbol{\lambda}\|_1 \leq \theta_v\}$. If*

$$\mathcal{Q} = \{q^{\mathcal{Z}}_{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda} \in \Lambda\} \tag{53}$$

is precompact and $\mathcal{Q} \subset \mathcal{C}(\mathcal{X}, \mathcal{Z}) \subset \mathcal{D}$, then the problem

$$\begin{aligned} \min_{h \in \mathcal{C}(\mathcal{X}, \mathcal{Z})} \quad & \mathbb{E}\left[F(X, h(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[G_i(X, h(X))\right] \leq 0, \quad i \in [k] \end{aligned} \tag{54}$$

has a unique solution, and this solution is $q^{\mathcal{Z}}_{\boldsymbol{\lambda}^*}$ where $\boldsymbol{\lambda}^*$ is any solution of

$$\sup_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}(q^{\mathcal{Z}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}). \tag{55}$$

We apply Theorem 4 to the problem of model projection. An intermediary step is that in which separability of the objective function F and linearity of the constraining functions G_i are assumed. More precisely, we introduce the following assumptions.

Assumption II'.

(a) For each $x \in \mathcal{X}$, the function $F(x, \cdot)$ is separable and can be written as

$$F(x, p) = \sum_{j \in [c]} f_j(x, p_j) \tag{56}$$

for continuously differentiable strictly convex functions $f_j(x, \cdot)$ satisfying $\lim_{t_0 \to 0^+} \frac{\partial f_j}{\partial t}(x, t_0) = -\infty$.

(b) For each fixed $(i, x) \in [k] \times \mathcal{X}$ the function $G_i(x, \cdot)$ is linear, i.e.,

$$G_i(x, q) = q^T \mathbf{g}^{(i)}(x). \tag{57}$$

Further, for each $i \in [k]$ the function $\mathbf{g}^{(i)} : \mathcal{X} \to \mathbb{R}^c$ is continuous. We denote

$$\mathbf{G} = \left(\mathbf{g}^{(1)}, \cdots, \mathbf{g}^{(k)}\right). \tag{58}$$

Note that Assumption 2.a implies that $t_0 \mapsto (\partial f_j / \partial t)(x, t_0)$ is strictly increasing for fixed $(j, x) \in [c] \times \mathcal{X}$. We let φ_j denote its inverse, and we formally introduce the constants γ next. Suppose that, for each fixed $(j, x) \in [c] \times \mathcal{X}$, the function

[5]One way to guarantee the well-definedness of $q^{\mathcal{Z}}_{\boldsymbol{\lambda}}$, for any fixed $\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}$, is to ensure $\mathcal{Z}$ is a nonempty convex and compact set, each $F(x, \cdot)$ be lower-semicontinuous and strictly convex, and each $G_i(x, \cdot)$ be lower-semicontinuous and convex. Indeed, under such assumptions, each mapping $p \mapsto F(x, p) + \sum_{i \in [k]} \lambda_i G_i(x, p)$ is lower-semicontinuous and strictly convex, which is then uniquely minimized over the convex and compact set $\mathcal{Z}$.

$f_j(x,\cdot):(0,1)\to\mathbb{R}$ is strictly convex and continuously differentiable, and that $\partial_{m+1}f_j(x,0^+)=-\infty$. Then the range of $\partial_{m+1}f_j(x,\cdot)$ over $(0,1)$ takes the form $(-\infty,r_{j,x})$ for some $r_{j,x}\in(-\infty,\infty]$. Therefore, $\partial_{m+1}f_j(x,\cdot)$ is invertible and its inverse $\varphi_j(x,\cdot):(-\infty,r_{j,x})\to(0,1)$ is continuous, strictly increasing, and satisfies $\varphi_j(x,-\infty)=0$ and $\varphi_j(x,r_{j,x}^-)=1$. Therefore, for any $\mathbf{a}\in\mathbb{R}^c$ the mapping

$$\gamma\mapsto\sum_{j\in[c]}\varphi_j\left(x,\gamma+a_j\right) \tag{59}$$

is a strictly increasing continuous bijection from an interval $\mathcal{I}_1=(-\infty,\tau_1)$ to another $\mathcal{I}_2=(0,\tau_2)$ where $\tau_2>1$. We define $\gamma:\mathcal{X}\times\mathbb{R}^k\to\mathbb{R}$ implicitly by

$$\sum_{j\in[c]}\varphi_j\left(x,\gamma(x,\boldsymbol{\lambda})+\mathbf{v}_j(x;\boldsymbol{\lambda})\right)=1. \tag{60}$$

Note that we allow $\boldsymbol{\lambda}$ with negative coordinates in the definition of $\gamma(x,\boldsymbol{\lambda})$. Recall that we set $\mathbf{v}(x;\boldsymbol{\lambda})=-\mathbf{G}(x)\boldsymbol{\lambda}$.

In the remainder of the appendices, we will take the f_j to be the following functions. For any $(j,x,t)\in[c]\times\mathcal{X}\times[0,1]$,

$$f_j(x,t)\triangleq\mathbf{y}_j(x)f\left(\frac{t}{\mathbf{y}_j(x)}\right). \tag{61}$$

Then, $F(x,\mathbf{p})=\sum_{j\in[c]}f_j(x,\mathbf{p}_j)$ satisfies

$$F(x,\mathbf{p})=D_f(\mathbf{p}\|\mathbf{y}(x)). \tag{62}$$

We will repeatedly make use of the following bound on the values of φ_j.

Lemma 3. *Fix $\mathbf{y}\in\boldsymbol{\Delta}_c^+$, and let $f:[0,\infty)\to\overline{\mathbb{R}}$ be strictly convex and continuously differentiable over $(0,\infty)$ such that $f'(0^+)=-\infty$ and denote the inverse of its derivative by ϕ. For each $j\in[c]$, define $f_j:[0,1]\to\overline{\mathbb{R}}$ by $f_j(t)=y_jf(t/y_j)$, and let $\varphi_j:(-\infty,f'(1/y_j)]\to(0,1]$ be the inverse of f_j'. Let $\mathbf{w}\in\mathbb{R}^c$ and $U\in\mathbb{R}_{\geq 0}$ be such that $\|\mathbf{w}\|_\infty\leq U$, and let $\eta\in\mathbb{R}$ be the unique real such that $\sum_{j\in[c]}\varphi_j(\eta+w_j)=1$. Then,*

$$\min_{j\in[c]}\varphi_j(\eta+w_j)\geq\phi\left(f'\left(\frac{1}{c}\right)-2U\right)\min_{j\in[c]}y_j. \tag{63}$$

Proof. For each $j\in[c]$, let $\beta_j=\eta+w_j$. Then, $|\beta_i-\beta_j|\leq 2U$ for every $(i,j)\in[c]^2$. We must have that, for some $a\in[c]$,

$$\varphi_a(\beta_a)\geq 1/c. \tag{64}$$

Therefore, $\beta_a\geq f_a'(1/c)$. But, $f_a'(1/c)=f'(1/(cy_j))\geq f'(1/c)$. Then,

$$\min_{j\in[c]}\beta_j\geq f'\left(\frac{1}{c}\right)-2U. \tag{65}$$

Finally, as $\varphi_j(v)=y_j\phi(v)$,

$$\min_{j\in[c]}\varphi_j(\beta_j)\geq\min_{j\in[c]}\varphi_j\left(\min_{i\in[c]}\beta_i\right)\geq\min_{j\in[c]}\varphi_j\left(f'\left(\frac{1}{c}\right)-2U\right)=\min_{j\in[c]}y_j\phi\left(f'\left(\frac{1}{c}\right)-2U\right), \tag{66}$$

as desired. □

In view of this result, we will employ the following notation. Write

$$y_{\min}\triangleq\inf_{x,j}\mathbf{y}_j(x), \tag{67}$$

and, for $\theta>0$, let

$$t_{\min}(\theta)\triangleq\phi\left(f'\left(\frac{1}{c}\right)-2\theta-1\right)y_{\min} \tag{68}$$

and

$$u_{\min}(\theta)\triangleq f'\left(\frac{1}{c}\right)-2\theta-1. \tag{69}$$

We use $2\theta+1$ instead of 2θ to obtain a strict inequality

$$\varphi_j(x,\gamma(x,\boldsymbol{\lambda})+\mathbf{v}_j(x;\boldsymbol{\lambda}))>t_{\min}(\|\boldsymbol{\lambda}\|) \tag{70}$$

The following regularity conditions guarantee that an optimizer over a compact set $\mathcal{K}\subset\mathcal{C}(\mathbb{R}^m,\boldsymbol{\Delta}_c)$ is also a global optimizer. Note that we introduce the following definition only for the case $\mathcal{X}=\mathbb{R}^m$.

Definition 6. Assume $\mathcal{X} = \mathbb{R}^m$. We call the functions f_j and $\mathbf{G}$ *regular* if

(a) every function $f_j(x, \cdot)$ is twice continuously differentiable and, for every $\varepsilon > 0$,

$$\inf_{(j,x,t)\in[c]\times\mathbb{R}^m\times(\varepsilon,1)} \partial_{m+1}^2 f_j(x,t) > 0, \tag{71}$$

(b) the partial derivatives $\partial_\ell \partial_{m+1} f_j(x,t)$ and $\partial_\ell \mathbf{G}_{i,j}(x)$ exist and are continuous, and for every $\varepsilon > 0$,

$$\sup_{(\ell,i,j,x,t)\in[m]\times[k]\times[c]\times\mathbb{R}^m\times(\varepsilon,1)} \max\left(|\partial_\ell \partial_{m+1} f_j(x,t)|, |\mathbf{G}_{i,j}(x)|, |\partial_\ell \mathbf{G}_{i,j}(x)|\right) < \infty, \tag{72}$$

(c) the functions $\partial_{m+1} f_j(\cdot, t)$ are continuous for every $t \in (0, 1]$.

We show that the regularity conditions on the f_j and $\mathbf{G}$ yield Lipschitzness of φ_j and local Lipschitzness of the γ_j. This in turn will yield precompactness the set $\mathcal{Q}$ given in equation (53) in Theorem 4. The key tool we employ is utilizing a simplified version of the implicit function theorem, where the simplicity is due to the triviality of gluing. The proof of the following precompactness result is given in Appendix E.

Theorem 5. *Under Assumption I and II', for any $\theta \in \mathbb{R}_{\geq 0}$ the set*

$$\mathcal{Q} = \left\{ q_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c} \mid \boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}, \|\boldsymbol{\lambda}\|_1 \leq \theta \right\} \tag{73}$$

is a precompact subset of $\mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_c)$.

The explicit formula for $\mathbf{h}^{\mathrm{opt}}$ is a direct consequence of the formula for $q_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c}$, which we give next and prove in Appendix F

Lemma 4. *Let $f : [0, \infty) \to \overline{\mathbb{R}}$ be a strictly convex continuously differentiable function over[6] $(0, \infty)$ such that $f(1) = 0$ and $f'(0^+) = -\infty$, and let ϕ be the inverse of f'. Fix $\mathbf{q} \in \boldsymbol{\Delta}_c^+$, and define $F : [0, 1]^c \to \overline{\mathbb{R}}$ by*

$$F(\mathbf{p}) = \mathbb{E}_{i\sim\mathbf{q}}\left[f\left(\frac{p_i}{q_i}\right)\right]. \tag{74}$$

Then, the convex conjugate of F is defined over all of $\mathbb{R}^c$ and satisfies

$$F^{\mathrm{conj}}(\mathbf{v}) = \mathbb{E}_{i\sim\mathbf{q}}\left[v_i \phi(\gamma(\mathbf{v}) + v_i) - f(\phi(\gamma(\mathbf{v}) + v_i))\right] \tag{75}$$

where $\gamma : \mathbb{R}^c \to \mathbb{R}$ is the unique function satisfying

$$\mathbb{E}_{i\sim\mathbf{q}} \phi(\gamma(\mathbf{v}) + v_i) = 1, \quad \mathbf{v} \in \mathbb{R}^c. \tag{76}$$

Further, for any $\mathbf{v} \in \mathbb{R}^c$ and $j \in [c]$

$$\mathbf{q}_j^{\mathrm{conj}}(\mathbf{v}) = \phi(\gamma(\mathbf{v}) + v_i). \tag{77}$$

Corollary 2. *Under Assumption I and II', the j-th coordinate of $q_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c}(x)$ (see (52)) is $\varphi_j\left(x, \gamma(x, \boldsymbol{\lambda}) + \mathbf{v}_j(x; \boldsymbol{\lambda})\right)$.*

Note that

$$\varphi_j(x, u) = \mathbf{y}_j(x)\phi(u). \tag{78}$$

The final ingredient in the proof is a direct consequence of Lemma 3.

Corollary 3. *Under Assumption I and II', for any $\boldsymbol{\lambda} \geq \mathbf{0}$ and $\varepsilon \in [0, t_{\min}(\|\boldsymbol{\lambda}\|))$*

$$q_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c^\varepsilon} = q_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c}. \tag{79}$$

Now, we are ready to finish the proof of both Theorems 1 and 2. We operate under Assumption I, and note that the model projection problem we consider, then, satisfies Assumption II'. We apply the general results with $\mathcal{Z} = \boldsymbol{\Delta}_c^\varepsilon$ for all small enough ε.

By continuity of f,

$$D \supset \boldsymbol{\Delta}_c^+. \tag{80}$$

Further, for any $\varepsilon \in (0, 1)$,

$$\mathcal{D} \supset \mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_c^\varepsilon), \tag{81}$$

[6] It is assumed that $f(0) = f(0^+)$.

so $\mathcal{D} \supset \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$. Fix $\widetilde{\mathbf{h}} \in \mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$ such that $\mathbb{E}\left[\widetilde{\mathbf{h}}(X)^T \mathbf{G}(X)\right] < \mathbf{0}$, i.e., $\widetilde{\mathbf{h}} \in \mathcal{S}$. Let ε be small enough that $\widetilde{\mathbf{h}} \in \mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c^{\varepsilon})$. Denote $\widetilde{\theta} = \theta_{\widetilde{\mathbf{h}}}$. Fix $\theta \geq \widetilde{\theta}$. Decrease, if necessary, the value of ε so that $\varepsilon < t_{\min}(\theta)$. Then, by Corollary 3,

$$q_{\boldsymbol{\lambda}}^{\mathbf{\Delta}_c^{\varepsilon}} = q_{\boldsymbol{\lambda}}^{\mathbf{\Delta}_c} \tag{82}$$

for all $\boldsymbol{\lambda}$ with $\|\boldsymbol{\lambda}\| \leq \theta$.

By Theorem 5, we have precompactness of the set

$$\mathcal{Q} \triangleq \{q_{\boldsymbol{\lambda}}^{\mathbf{\Delta}_c} \mid \boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 \leq \theta\} \tag{83}$$

and that $\mathcal{Q} \subset \mathcal{C}(\mathbb{R}^m, \mathbf{\Delta}_c)$. But, by (82),

$$\mathcal{Q} = \{q_{\boldsymbol{\lambda}}^{\mathbf{\Delta}_c^{\varepsilon}} \mid \boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 \leq \theta\}. \tag{84}$$

Then, $\mathcal{Q} \subset \mathcal{C}(\mathbb{R}^m, \mathbf{\Delta}_c^{\varepsilon})$. Precompactness of $\mathcal{Q}$, then, implies by Theorem 4 (using $\mathcal{Z} = \mathbf{\Delta}_c^{\varepsilon}$) that the problem

$$\begin{aligned} \min_{\mathbf{h} \in \mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c^{\varepsilon})} \quad & \mathbb{E}\left[D_f(\mathbf{h}(X) \| \mathbf{y}(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[\mathbf{h}(X)^T \mathbf{G}(X)\right] \leq \mathbf{0} \end{aligned} \tag{85}$$

has the unique solution $q_{\boldsymbol{\lambda}^*}^{\mathbf{\Delta}_c}$ for any $\boldsymbol{\lambda}^*$ solving

$$\inf_{\boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\| \leq \widetilde{\theta}} \mathbb{E}\left[D_f^{\text{conj}}(\mathbf{v}(X; \boldsymbol{\lambda}), \mathbf{y}(X))\right] \tag{86}$$

where we used the fact that

$$\mathcal{L}(q_{\boldsymbol{\lambda}}^{\mathbf{\Delta}_c}, \boldsymbol{\lambda}) = -\mathbb{E}\left[D_f^{\text{conj}}\left(\mathbf{v}(X; \boldsymbol{\lambda}), \mathbf{y}(X)\right)\right]. \tag{87}$$

By Corollary 4, we may remove the condition $\|\boldsymbol{\lambda}| \leq \widetilde{\theta}$. As the solution $q_{\boldsymbol{\lambda}^*}^{\mathbf{\Delta}_c}$ does not depend on ε, and as ε is arbitrary, we may extend the optimization to be over all of $\mathcal{C}_+(\mathcal{X}, \mathbf{\Delta}_c)$. Finally, the proof is complete in view of the equation of $q_{\boldsymbol{\lambda}^*}^{\mathbf{\Delta}_c}$ as given by Corollary 2.

C. *Proof of Lemma 2*

We prove the existence of a minimizer first. Then we treat uniqueness.

Existence of a minimizer. Suppose that Assumption 1.a-c holds, and fix a compact set $\mathcal{K} \subset \mathcal{D}$. We show that the objective function is lower-semicontinuous on $\mathcal{K}$ and that the feasibility set $\mathcal{K} \cap \mathcal{F}$ is compact, which together yield via the extreme value theorem the existence of a minimizer. Thus, let us show that the mappings $\mathcal{A}, \mathcal{B}_1, \cdots, \mathcal{B}_k$ are lower-semicontinuous on $\mathcal{K}$. Lower-semicontinuity of the $\mathcal{B}_i$ will yield that the feasibility set $\mathcal{K} \cap \mathcal{F}$ of (49) is compact.

Fix $J \in \{F, G_1, \cdots, G_k\}$, and we will show that the mapping $h \mapsto \mathbb{E}\left[J(X, h(X))\right] + \mathbb{I}_{\mathcal{D}}(h)$ is lower-semicontinuous when restricted to $\mathcal{K}$. As $\mathcal{K} \subset \mathcal{D}$ by assumption, this mapping is just $h \mapsto \mathbb{E}\left[J(X, h(X))\right]$. As $\mathcal{K}$ is a metric space, lower-semincontinuity on $\mathcal{K}$ is equivalent to sequential-lower-semicontinuity [39, Theorem 7.1.2]. Fix a convergent sequence $h_n \to h$ in $\mathcal{K}$ (i.e., $\sup_{x \in \mathcal{X}} \|h_n(x) - h(x)\|_1 \to 0$ as $n \to \infty$). By Assumption 1.b, we may apply Fatou's lemma to obtain

$$\liminf_{n \to \infty} \mathbb{E}\left[J(X, h_n(X))\right] \geq \mathbb{E}\left[\liminf_{n \to \infty} J(X, h_n(X))\right]. \tag{88}$$

Uniform convergence $h_n \to h$ implies, in particular, pointwise convergence: $h_n(x) \to h(x)$ for every $x \in \mathcal{X}$. Therefore, by lower-semicontinuity of each $J(x, \cdot)$ (Assumption 1.c)

$$\mathbb{E}\left[\liminf_{n \to \infty} J(X, h_n(X))\right] \geq \mathbb{E}\left[J(X, h(X))\right]. \tag{89}$$

Therefore,

$$\liminf_{n \to \infty} \mathbb{E}\left[J(X, h_n(X))\right] \geq \mathbb{E}\left[J(X, h(X))\right], \tag{90}$$

and lower-semicontinuity of $\mathcal{A}, \mathcal{B}_1, \cdots, \mathcal{B}_k$ on $\mathcal{K}$ follows. In particular, the lower-level sets

$$\mathcal{V}_i \triangleq \{h \in \mathcal{K} \mid \mathbb{E}\left[G_i(X, h(X))\right] \leq 0\} \tag{91}$$

are closed[7] [39, Theorem 7.1.1]. Therefore, the feasibility set $\mathcal{F} \cap \mathcal{K} = \bigcap_{i \in [k]} \mathcal{V}_i$ is closed. By compactness of $\mathcal{K}$, the feasibility set $\mathcal{F} \cap \mathcal{K}$ is compact too. Finally, lower-semicontinuity of $\mathcal{A}$ on $\mathcal{K}$ and compactness of the feasibility set $\mathcal{F} \cap \mathcal{K}$ yield the existence of a minimizer [39, Theorem 7.3.1].

Uniqueness of the minimizer. Now, suppose that $\mathcal{K}$ is also convex, and that Assumption 1.d holds too. Since expectation is a linear operator, $h \mapsto \mathbb{E}\left[F(X, h(X))\right]$ is strictly convex, and each $h \mapsto \mathbb{E}\left[G_i(X, h(X))\right]$ is convex. Hence, the lower-level

[7] The $\mathcal{V}_i$ are closed both in $\mathcal{K}$ and in $\mathcal{C}(\mathcal{X})$, as the compact set $\mathcal{K}$ is closed in the Hausdorff space $\mathcal{C}(\mathcal{X})$.

sets (91) are convex which implies that the feasibility set $\mathcal{K} \cap \mathcal{F}$ is convex. Thus, the optimization problem (49) has a unique minimizer.

D. Proving Theorem 4

Definition 7. For a given $\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}$ and a subset $\mathcal{K} \subset \mathcal{D}$, define the function in $\mathcal{K}$ that achieves the minimal value of the Lagrangian by

$$h^{\mathcal{K}}_{\boldsymbol{\lambda}} \triangleq \operatorname*{argmin}_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{92}$$

if there is such a unique function.

Theorem 6. *Suppose Assumption 1.a-d holds, and fix a nonempty compact and convex $\mathcal{K} \subset \mathcal{D}$. For every $\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}$, the function $\mathcal{L}(\cdot, \boldsymbol{\lambda})$ has a unique minimizer over $\mathcal{K}$, i.e., $h^{\mathcal{K}}_{\boldsymbol{\lambda}}$ in (92) is well-defined. In addition, if $\boldsymbol{\lambda}^*$ satisfies*

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^*) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{93}$$

then $h^{\mathcal{K}}_{\boldsymbol{\lambda}^*}$ is the unique solution for problem (34).

Proof. Since $\mathcal{K}$ is compact and $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$ is strictly convex and lower-semicontinuous for any fixed $\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}$, there is a unique minimizer of $\mathcal{L}(h, \boldsymbol{\lambda})$ over $\mathcal{K}$. Hence, $h^{\mathcal{K}}_{\boldsymbol{\lambda}}$ is well-defined and satisfies

$$\mathcal{L}(h^{\mathcal{K}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{94}$$

Next, we prove strong duality for (34). Again, the mapping $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$ is strictly convex and lower-semicontinuous for each fixed $\boldsymbol{\lambda}$. Also, $\boldsymbol{\lambda} \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$ is concave for each fixed h (as it is affine). Therefore, by Sion's minimax theorem and the compactness of $\mathcal{K}$,

$$\inf_{h \in \mathcal{K}} \sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{95}$$

Let h^* denote the unique solution of (34), whose existence and uniqueness are guaranteed by Lemma 2. We have that

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \mathcal{L}(h^*, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{96}$$

Combining (96), (95), and (93) together, we have

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \mathcal{L}(h^*, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^*). \tag{97}$$

Furthermore, since

$$\mathcal{L}(h^*, \boldsymbol{\lambda}^*) \leq \sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \mathcal{L}(h^*, \boldsymbol{\lambda}) \quad \text{and} \quad \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^*) \leq \mathcal{L}(h^*, \boldsymbol{\lambda}^*), \tag{98}$$

then we have

$$\mathcal{L}(h^*, \boldsymbol{\lambda}^*) \leq \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^*) \leq \mathcal{L}(h^*, \boldsymbol{\lambda}^*) \tag{99}$$

which implies $\mathcal{L}(h^*, \boldsymbol{\lambda}^*) = \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^*)$. Therefore, by strict convexity of $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda}^*)$, $h^* = h^{\mathcal{K}}_{\boldsymbol{\lambda}^*}$. □

Next, we prove the existence of a $\boldsymbol{\lambda}^*$ satisfying (93) in Theorem 6 whenever $\mathcal{K} \cap \mathcal{S} \neq \emptyset$. It will be convenient to introduce the following quantity, which will be used to bound the searching space of dual variable.

Definition 8. For a subset $\mathcal{K} \subset \mathcal{D}$, we define

$$\theta(\mathcal{K}) \triangleq \inf_{q \in \mathcal{K} \cap \mathcal{S}} \frac{\mathcal{A}(q) - \inf_{h \in \mathcal{K}} \mathcal{A}(h)}{-\max_{i \in [k]} \mathcal{B}_i(q)}. \tag{100}$$

We note that under Assumption 1.a-b, if $\mathcal{K} \subset \mathcal{D}$ is such that $\mathcal{K} \cap \mathcal{S}$ is nonempty, then $\theta(\mathcal{K}) \in \mathbb{R}_{\geq 0}$. Indeed, fix an integrable $L : \mathcal{X} \to \mathbb{R}$ such that

$$L(x) \leq \inf_{h \in \mathcal{D}} F(x, h(x)) \tag{101}$$

for every $x \in \mathcal{X}$. Then, for any $q \in \mathcal{K} \cap \mathcal{S}$

$$-\infty < \mathbb{E}\left[L(X)\right] \leq \inf_{h \in \mathcal{D}} \mathcal{A}(h) \leq \inf_{h \in \mathcal{K}} \mathcal{A}(h) \leq \mathcal{A}(q) < \infty. \tag{102}$$

Thus, $\inf_{h\in\mathcal{K}} \mathcal{A}(h) \in \mathbb{R}$. Hence, by definition of $\mathcal{D}$ and because $\mathcal{K}\cap\mathcal{S} \subset \mathcal{D}$, we obtain $\theta(\mathcal{K}) \in \mathbb{R}_{\geq 0}$.

Theorem 7. *Suppose Assumption 1.a-b holds, and fix $\mathcal{K} \subset \mathcal{D}$. If $\mathcal{K}\cap\mathcal{S}$ is nonempty, then*

$$\sup_{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}\\ \|\boldsymbol{\lambda}\|_1\leq\theta(\mathcal{K})}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}), \tag{103}$$

there exists a $\boldsymbol{\lambda}^*$ that achieves the supremum in the left-hand-side in (103), and any such maximizer satisfies $\|\boldsymbol{\lambda}^*\|_1 \leq \theta(\mathcal{K})$.

Proof. If the equality

$$\inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0}) = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) \tag{104}$$

holds, then the desired equality (103) also holds and $\boldsymbol{\lambda}=\mathbf{0}$ achieves the supremum. Thus, for the remainder of the proof, we assume that (104) does not hold, i.e.,

$$\inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0}) < \sup_{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}). \tag{105}$$

For any $\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}$ and $q\in\mathcal{S}$, by the definition of $\mathcal{S}$ (see (39)), $\mathbb{E}\left[G_i(X,q(X))\right] < 0$ for all $i\in[k]$. Then, for any $q\in\mathcal{K}\cap\mathcal{S}$

$$\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) \leq \inf_{h\in\mathcal{K}\cap\mathcal{S}} \mathcal{L}(h,\boldsymbol{\lambda}) \leq \mathcal{L}(q,\boldsymbol{\lambda}) = \mathcal{A}(q) + \sum_{i\in[k]} \lambda_i \mathcal{B}_i(q) \leq \mathcal{A}(q) + \|\boldsymbol{\lambda}\|_1 \max_{i\in[k]} \mathcal{B}_i(q) \tag{106}$$

where we used the fact that $q\in\mathcal{K}\cap\mathcal{S}\subset\mathcal{K}\subset\mathcal{D}$. Thus, we have

$$\|\boldsymbol{\lambda}\|_1 \leq \frac{\mathcal{A}(q) - \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda})}{-\max_{i\in[k]} \mathcal{B}_i(q)}. \tag{107}$$

Now, if $\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}$ satisfies both $\|\boldsymbol{\lambda}\|_1 > \theta(\mathcal{K})$ and $\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) \geq \inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0})$, then, we must have (because $\mathcal{L}(h,\mathbf{0}) = \mathbb{E}\left[F(X,h(X)\right] = \mathcal{A}(h)$ for $h\in\mathcal{D}$)

$$\theta(\mathcal{K}) < \|\boldsymbol{\lambda}\|_1 \leq \frac{\mathcal{A}(q) - \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda})}{-\max_{i\in[k]} \mathcal{B}_i(q)} \leq \frac{\mathcal{A}(q) - \inf_{h\in\mathcal{K}} \mathcal{A}(h)}{-\max_{i\in[k]} \mathcal{B}_i(q)} \tag{108}$$

for every $q\in\mathcal{K}\cap\mathcal{S}$. Taking the infimum over all $q\in\mathcal{K}\cap\mathcal{S}$, we obtain

$$\theta(\mathcal{K}) < \|\boldsymbol{\lambda}\|_1 \leq \theta(\mathcal{K}), \tag{109}$$

which is absurd. Thus, every $\boldsymbol{\lambda}$ that satisfies $\|\boldsymbol{\lambda}\|_1 > \theta(\mathcal{K})$ must have $\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) < \inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0})$. Taking the supremum over all such $\boldsymbol{\lambda}$ implies

$$\sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}\\ \|\boldsymbol{\lambda}\|_1>\theta(\mathcal{K})}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) \leq \inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0}) < \sup_{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}). \tag{110}$$

In particular, the desired equality (103) holds.

Finally, being the pointwise infimum of linear (in particular, upper-semicontinuous) functions in $\boldsymbol{\lambda}$, $\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda})$ is upper-semicontinuous. Hence, having $\theta(\mathcal{K}) < \infty$ would imply that at least one $\boldsymbol{\lambda}^*$ maximizing the dual optimization problem (103) exists. By inequality (110), $\|\boldsymbol{\lambda}^*\|_1 \leq \theta(\mathcal{K})$ for any such maximizer $\boldsymbol{\lambda}^*$. □

Though this theorem gives a way to bound the value of the dual parameter $\boldsymbol{\lambda}$, the upper bound $\theta(\mathcal{K})$ might not be computable. In particular, computing $\theta(\mathcal{K})$ requires global information about $\mathcal{K}$. Nevertheless, note that removing the outer infimum in the definition of $\theta(\mathcal{K})$ still yields a finite upper bound. Further, relaxing the inner infimum to be over the domain $\mathcal{D}$ also gives a finite upper bound (under Assumption 1.b).

Under Assumption 1.b, θ_q is always finite. Also, $\theta(\mathcal{K}) \leq \theta_q$ whenever $\mathcal{K}\subset\mathcal{D}$ and $q\in\mathcal{K}\cap\mathcal{S}$. Thus, Theorem 7 immediately implies the following result.

Corollary 4. *Suppose Assumption 1.a-b holds, and fix $\mathcal{K}\subset\mathcal{D}$. If $\mathcal{K}\cap\mathcal{S}$ is nonempty and $q\in\mathcal{K}\cap\mathcal{S}$, then*

$$\sup_{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^k_{\geq 0}\\ \|\boldsymbol{\lambda}\|_1\leq\theta_q}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}), \tag{111}$$

and the supremum is achievable. Furthermore, all maximizers have 1-norm at most θ_q.

Next, we give a more tractable way of expressing $h^{\mathcal{K}}_{\boldsymbol{\lambda}}$. It will be useful to introduce the following class of functions.

Theorem 8. *Suppose Assumption 1.a-d holds. Fix a nonempty convex and compact subset $\mathcal{Z} \subset D$, and a nonempty convex and compact subset $\mathcal{K} \subset \mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{D}$. For any $\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}$, if $q^{\mathcal{Z}}_{\boldsymbol{\lambda}} \in \mathcal{K}$, then $h^{\mathcal{K}}_{\boldsymbol{\lambda}} = q^{\mathcal{Z}}_{\boldsymbol{\lambda}}$.*

Proof. For each $x \in \mathcal{X}$, let $\mathcal{R}_x \subset \mathbb{R}^c$ denote the image of $\mathcal{K}$ under the mapping $h \mapsto h(x)$, i.e.,

$$\mathcal{R}_x \triangleq \{h(x) \mid h \in \mathcal{K}\}. \tag{112}$$

We have, by assumption, $\bigcup_{x \in \mathcal{X}} \mathcal{R}_x \subset \mathcal{Z}$. Fix $\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}$, and write

$$L(x, q) = F(X, q) + \sum_{i \in [k]} \lambda_i G_i(X, q) \tag{113}$$

for short. Then, for any $(x, h) \in \mathcal{X} \times \mathcal{K}$

$$L(x, h(x)) \geq \inf_{p \in \mathcal{K}} L(x, p(x)) \geq \inf_{r \in \mathcal{R}_x} L(x, r) \geq \inf_{q \in \mathcal{Z}} L(x, q) = L(x, q^{\mathcal{Z}}_{\boldsymbol{\lambda}}(x)). \tag{114}$$

Assume that $q^{\mathcal{Z}}_{\boldsymbol{\lambda}} \in \mathcal{K}$. Then, taking the expectation of the two far ends of (114) then the infimum for $h \in \mathcal{K}$ we get

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) \geq \mathcal{L}(q^{\mathcal{Z}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}). \tag{115}$$

However, it is also true that

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) \leq \mathcal{L}(q^{\mathcal{Z}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}). \tag{116}$$

Therefore, we get the equality

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(q^{\mathcal{Z}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}). \tag{117}$$

By strict convexity of $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$, and by definition of $h^{\mathcal{K}}_{\boldsymbol{\lambda}}$, we have $h^{\mathcal{K}}_{\boldsymbol{\lambda}} = q^{\mathcal{Z}}_{\boldsymbol{\lambda}}$. □

Proof of Theorem 4. Write $\theta = \theta_v$, and note that $\theta \in \mathbb{R}_{\geq 0}$. Let $u \in \mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{F}$ be arbitrary. Consider the two sets

$$\mathcal{K} = \overline{\operatorname{co}(\mathcal{H} \cup \{v\})}, \tag{118}$$

$$\mathcal{K}' = \overline{\operatorname{co}(\mathcal{H} \cup \{u, v\})}. \tag{119}$$

The sets $\mathcal{K}$ and $\mathcal{K}'$ are convex and compact, and they satisfy $\mathcal{K}, \mathcal{K}' \subset \mathcal{C}(\mathcal{X}, \mathcal{Z})$ because $\mathcal{C}(\mathcal{X}, \mathcal{Z})$ is convex and closed and $\mathcal{H} \subset \mathcal{C}(\mathcal{X}, \mathcal{Z})$ by assumption. If $\boldsymbol{\lambda} \in \Lambda$, then by definition $q^{\mathcal{Z}}_{\boldsymbol{\lambda}}$ is an element in both $\mathcal{K}$ and $\mathcal{K}'$, hence by Theorem 8

$$h^{\mathcal{K}}_{\boldsymbol{\lambda}} = q^{\mathcal{Z}}_{\boldsymbol{\lambda}} = h^{\mathcal{K}'}_{\boldsymbol{\lambda}}. \tag{120}$$

By Corollary 4,

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0} \\ \|\boldsymbol{\lambda}\|_1 \leq \theta}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{121}$$

and the same is true for $\mathcal{K}'$

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0}} \inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0} \\ \|\boldsymbol{\lambda}\|_1 \leq \theta}} \inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{122}$$

By definition, $\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(h^{\mathcal{K}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda})$ and $\inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(h^{\mathcal{K}'}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda})$.

Therefore, for any $\boldsymbol{\lambda} \in \Lambda$

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(q^{\mathcal{Z}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{123}$$

Thus, the problems (121) and (122) are equivalent to each other, and they are equivalent to

$$\sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^k_{\geq 0} \\ \|\boldsymbol{\lambda}\|_1 \leq \theta}} \mathcal{L}(q^{\mathcal{Z}}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}). \tag{124}$$

Furthermore, there is a $\boldsymbol{\lambda}^*$ achieving this supremum. In addition, by Theorem 6, for any such $\boldsymbol{\lambda}^*$ we have that $q^{\mathcal{Z}}_{\boldsymbol{\lambda}^*}$ is the unique solution to both $\inf_{h \in \mathcal{K} \cap \mathcal{F}} \mathbb{E}\left[F(X, h(X))\right]$ and $\inf_{h \in \mathcal{K}' \cap \mathcal{F}} \mathbb{E}\left[F(X, h(X))\right]$. Now,

$$\mathbb{E}\left[F(X, q^{\mathcal{Z}}_{\boldsymbol{\lambda}^*}(X))\right] = \inf_{h \in \mathcal{K}' \cap \mathcal{F}} \mathbb{E}\left[F(X, h(X))\right] \leq \mathbb{E}\left[F(X, u(X))\right]. \tag{125}$$

Therefore, by arbitrariness of u,

$$\mathbb{E}\left[F(X, q^{\mathcal{Z}}_{\boldsymbol{\lambda}^*}(X))\right] = \inf_{u \in \mathcal{C}(\mathcal{X},\mathcal{Z}) \cap \mathcal{F}} \mathbb{E}\left[F(X, u(X))\right]. \tag{126}$$

Finally, uniqueness follows by convexity of the set $\mathcal{F}$ and strict convexity of the function $\mathcal{A}|_{\mathcal{C}(\mathcal{X},\mathcal{Z})}$. □

E. Proof of Theorem 5

We note that Assumption I implies regularity of $f_j(x,t) = \mathbf{y}_j(x) f(t/\mathbf{y}_j(x))$ and $\mathbf{G}$. To see this, note that $\partial^2_{m+1} f_j(x,t) = f''(t/\mathbf{y}_j(x))/\mathbf{y}_j(x)$. By continuity of f'', condition (a) is satisfied. Also,

$$\partial_\ell \partial_{m+1} f_j(x,t) = \frac{-t\partial_\ell \mathbf{y}_j(x)}{\mathbf{y}_j(x)^2} f''\left(\frac{t}{\mathbf{y}_j(x)}\right) \tag{127}$$

and again continuity of f'' implies that condition (b) is also satisfied.

We employ the following version of the implicit function theorem.

Theorem 9 (Implicit Function Theorem)**.** *Let $\Omega \subset \mathbb{R}^e \times \mathbb{R}$ be an open set, denote by $U \subset \mathbb{R}^e$ and $V \subset \mathbb{R}$ its projections, and let $C : \Omega \to \mathbb{R}$ be a differentiable function. If there exists a unique function $c : U \to V$ satisfying both $(a, c(a)) \in \Omega$ and $C(a, c(a)) = 0$ for every $a \in U$, and if $\partial_{e+1} C(a, c(a)) \neq 0$ for every $a \in U$, then c is differentiable and $\partial_i c(a) = (-\partial_i C/\partial_{e+1} C)|_{(a,c(a))}$ for every $(i, a) \in [e] \times U$.*

We begin by deriving upper bounds on the partial derivatives of the φ_j and γ. Then, we conclude from Lipshcitzness of the φ_j and γ total boundedness of $\mathcal{Q}$ via compactness of $\boldsymbol{\Delta}_c$. As a by-product, it will follow that $\mathcal{Q}$ consists of continuous functions, i.e., that $\mathcal{Q} \subset \mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_c)$. For convenience of notation, we will show precompactness when $\|\boldsymbol{\lambda}\|_1$ is restricted to be at most $\theta - 1$ for some $\theta > 1$.

Fix $j \in [c]$, and we will show an upper bound on the partial derivatives of φ_j. Set

$$\Omega_j \triangleq \left\{(x,u) \in \mathbb{R}^m \times \mathbb{R} \;\middle|\; u_{\min}(\theta) < u < \partial_{m+1} f_j(x, 1^-)\right\}. \tag{128}$$

By the assumption of continuity of $\partial_{m+1} f_j(\cdot, 1^-)$, the set Ω_j is open; indeed, Ω_j is the intersection of the preimage of the open set $(0,\infty)$ under the continuous map $(x,u) \mapsto \partial_{m+1} f_j(x, 1^-) - u$ with the open set $\mathbb{R}^m \times (u_{\min}(\theta), \infty)$. Define $\rho_j : \Omega_j \times (0,1) \to \mathbb{R}$ by

$$\rho_j(x,u,t) = \partial_{m+1} f_j(x,t) - u. \tag{129}$$

For any $(x,u) \in \Omega$, there exists a unique $t \in (0,1)$ such that $\rho_j(x,u,t) = 0$, namely, $t = \varphi_j(x,u)$. In other words, $\varphi_j(x,u)$ is defined via

$$\rho_j(x, u, \varphi_j(x,u)) = 0. \tag{130}$$

By assumption on f_j, all partial derivative of ρ_j exist and are continuous. Therefore, ρ_j is differentiable. Further, by regularity of f_j, $\partial_{m+2}\rho_j(x,u,t) \neq 0$. Hence, by the implicit function theorem, φ_j is differentiable and its partial derivatives are given by

$$\partial_{m+1}\varphi_j(x,u) = -\frac{\partial_{m+1}\rho_j(x,u,\varphi_j(x,u))}{\partial_{m+2}\rho_j(x,u,\varphi_j(x,u))} = \frac{1}{\partial^2_{m+1} f_j(x, \varphi_j(x,u))}, \tag{131}$$

$$\partial_\ell \varphi_j(x,u) = -\frac{\partial_\ell \rho_j(x,u,\varphi_j(x,u))}{\partial_{m+2}\rho_j(x,u,\varphi_j(x,u))} = \frac{-\partial_{\ell,m+1} f_j(x, \varphi_j(x,u))}{\partial^2_{m+1} f_j(x, \varphi_j(x,u))}, \tag{132}$$

for every $(x,u) \in \Omega_j$, where $\ell \leq m$. Because φ_j is differentiable it is also continuous. Further, by assumption of regularity, we have the bound

$$\max_{r \in [m+1]} \max_{j \in [c]} \sup_{(x,u) \in \Omega_j} |\partial_r \varphi_j(x,u)| \leq A \tag{133}$$

for some positive constants A.

Next, we show an upper bound on partial derivative of γ. Let $\varepsilon < t_{\min}(\theta)$ be small enough so that

$$\inf_{x, \|\boldsymbol{\lambda}\|_1 \leq \theta} \left(\min_{j \in [c]} \left(\partial_{m+1} f_j(x, 1^-) + \sum_{i \in [k]} \lambda_i g_{i,j}(x) \right) - \max_{j \in [c]} \left(\partial_{m+1} f_j(x, \varepsilon) + \sum_{i \in [k]} \lambda_i g_{i,j}(x) \right) \right) > 0, \tag{134}$$

and set

$$\Omega = \left\{ (x, \boldsymbol{\lambda}, u) \in \mathbb{R}^m \times \mathbb{R}^k \times \mathbb{R} \;\middle|\; \max_{j \in [c]} \left(\partial_{m+1} f_j(x, \varepsilon) + \sum_{i \in [k]} \lambda_i g_{i,j}(x) \right) < u < \min_{j \in [c]} \left(\partial_{m+1} f_j(x, 1^-) + \sum_{i \in [k]} \lambda_i g_{i,j}(x) \right) \right\}. \tag{135}$$

Similarly to the Ω_j, the set Ω is open. Note that for any $(x, \boldsymbol{\lambda}) \in \mathbb{R}^m \times \mathbb{R}^k$ with $\|\boldsymbol{\lambda}\|_1 \leq \theta$, we have $(x, \boldsymbol{\lambda}, \gamma(x, \boldsymbol{\lambda})) \in \Omega$. For each $j \in [c]$, define $\psi_j : \Omega \to (0, 1)$ by

$$\psi_j(x, \boldsymbol{\lambda}, u) = \varphi_j \left(x, u - \sum_{i \in [k]} \lambda_i g_{i,j}(x) \right). \tag{136}$$

Define $\eta : \Omega \to (-1, c)$ by

$$\eta(x, \boldsymbol{\lambda}, u) = -1 + \sum_{j \in [c]} \psi_j(x, \boldsymbol{\lambda}, u). \tag{137}$$

Then, $\gamma(x, \boldsymbol{\lambda})$ is defined by

$$\eta(x, \boldsymbol{\lambda}, \gamma(x, \boldsymbol{\lambda})) = 0. \tag{138}$$

As we have shown that each φ_j is differentiable, and as each partial derivative $\partial_\ell g_{i,j}$ is assumed to exist and be continuous, the function η is differentiable. Further, we may compute the partial derivatives of η by the chain rule

$$\partial_{m+k+1}\eta(x, \boldsymbol{\lambda}, u) = \sum_j \partial_{m+k+1}\psi_j(x, \boldsymbol{\lambda}, u) = \sum_j \partial_{m+1}\varphi_j \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) \tag{139}$$

$$= \sum_j \frac{1}{\partial^2_{m+1} f_j \left(x, \varphi \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) \right)}, \tag{140}$$

$$\partial_\ell \eta(x, \boldsymbol{\lambda}, u) \overset{\ell \leq m}{=} \sum_j \left(\partial_\ell \varphi_j \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) - \left(\sum_i \lambda_i \partial_\ell g_{i,j}(x) \right) \partial_{m+1}\varphi_j \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) \right) \tag{141}$$

$$= -\sum_j \frac{\partial_{\ell, m+1} f_j \left(x, \varphi_j \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) \right) + \sum_i \lambda_i \partial_\ell g_{i,j}(x)}{\partial^2_{m+1} f_j \left(x, \varphi \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) \right)}, \tag{142}$$

$$\partial_{m+\ell}\eta(x, \boldsymbol{\lambda}, u) \overset{1 \leq \ell \leq k}{=} \sum_j -g_{\ell,j}(x) \partial_{m+1}\varphi_j \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) = \sum_j \frac{-g_{\ell,j}(x)}{\partial^2_{m+1} f_j \left(x, \varphi \left(x, u - \sum_i \lambda_i g_{i,j}(x) \right) \right)}. \tag{143}$$

Therefore, by the implicit function theorem, we have that γ is differentiable and

$$\partial_\ell \gamma(x, \boldsymbol{\lambda}) \overset{\ell \leq m}{=} \frac{-\partial_\ell \eta(x, \boldsymbol{\lambda}, \gamma(x, \boldsymbol{\lambda}))}{\partial_{m+k+1}\eta(x, \boldsymbol{\lambda}, \gamma(x, \boldsymbol{\lambda}))} = \frac{\sum_j \frac{\partial_{\ell, m+1} f_j \left(x, \varphi_j \left(x, \gamma(x, \boldsymbol{\lambda}) - \sum_i \lambda_i g_{i,j}(x) \right) \right) + \sum_i \lambda_i \partial_\ell g_{i,j}(x)}{\partial^2_{m+1} f_j \left(x, \varphi \left(x, \gamma(x, \boldsymbol{\lambda}) - \sum_i \lambda_i g_{i,j}(x) \right) \right)}}{\sum_j \frac{1}{\partial^2_{m+1} f_j \left(x, \varphi \left(x, \gamma(x, \boldsymbol{\lambda}) - \sum_i \lambda_i g_{i,j}(x) \right) \right)}}, \tag{144}$$

$$\partial_{m+\ell}\gamma(x, \boldsymbol{\lambda}) \overset{1 \leq \ell \leq k}{=} \frac{\sum_j \frac{g_{\ell,j}(x)}{\partial^2_{m+1} f_j \left(x, \varphi \left(x, \gamma(x, \boldsymbol{\lambda}) - \sum_i \lambda_i g_{i,j}(x) \right) \right)}}{\sum_j \frac{1}{\partial^2_{m+1} f_j \left(x, \varphi \left(x, \gamma(x, \boldsymbol{\lambda}) - \sum_i \lambda_i g_{i,j}(x) \right) \right)}}. \tag{145}$$

Thus, by assumption of regularity

$$\sup_{r,x} |\partial_r \gamma(x, \boldsymbol{\lambda})| \leq B(2 + \|\boldsymbol{\lambda}\|_1) \tag{146}$$

for some positive constant B.

Define functions $\mathbf{p}_{\boldsymbol{\lambda}} \in \mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_c)$, one for each $\boldsymbol{\lambda} \in \mathbb{R}^k$, as follows. For each $(x, \boldsymbol{\lambda}) \in \mathbb{R}^m \times \mathbb{R}^k$, let $\mathbf{p}_{\boldsymbol{\lambda}}(x) \in \boldsymbol{\Delta}_c$ be the probability vector whose j-th coordinate is

$$\varphi_j \left(x, \gamma(x, \boldsymbol{\lambda}) - \sum_{i \in [k]} \lambda_i g_{i,j}(x) \right). \tag{147}$$

When $\boldsymbol{\lambda} \geq \mathbf{0}$, we get $q^{\boldsymbol{\Delta}_c}_{\boldsymbol{\lambda}} = \mathbf{p}_{\boldsymbol{\lambda}}$. Let $\mathcal{Q}' = \{\mathbf{p}_{\boldsymbol{\lambda}} \mid \|\boldsymbol{\lambda}\|_1 \leq \theta\}$.

We have the Lipshitz conditions

$$|\varphi_j(x, u) - \varphi_j(x, u')| \leq A\sqrt{m+1}|u - u'| \tag{148}$$

$$|\gamma(x, \boldsymbol{\lambda}) - \gamma(x, \boldsymbol{\lambda}')| \leq B(2 + \theta)\sqrt{m+k}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_1^2 \tag{149}$$

for every $x \in \mathbb{R}^m$, u, u' such that $(x, u), (x, u') \in \Omega_j$, and $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathcal{B}_1(\mathbf{0}, \theta)$. Let

$$L = \max \left(A\sqrt{m+1}, B(2+\theta)\sqrt{m+k} \right). \tag{150}$$

Fix $\nu > 0$, and set $\delta = \min(1, \nu/(Lc(L+A)))$. Let $N \in \mathbb{N}$ and $\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_N \in \mathcal{B}_1(\mathbf{0}, \theta)$ be such that the balls $\mathcal{B}_1(\boldsymbol{\lambda}_r, \delta)$ cover $\mathcal{B}_1(\mathbf{0}, \theta)$. Fix $\mathbf{p}_{\boldsymbol{\lambda}} \in \mathcal{Q}'$. Let $r \in [N]$ be such that $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_r\|_1 \leq \delta$. Then, for every $x \in \mathbb{R}^m$,

$$\|\mathbf{p}_{\boldsymbol{\lambda}}(x) - \mathbf{p}_{\boldsymbol{\lambda}_r}(x)\|_1 = \sum_{j\in[c]} \left| \varphi_j \left(x, \gamma(x, \boldsymbol{\lambda}) - \sum_{i\in[k]} \lambda_i g_{i,j}(x) \right) - \varphi_j \left(x, \gamma(x, \boldsymbol{\lambda}_r) - \sum_{i\in[k]} \lambda_{r,i} g_{i,j}(x) \right) \right| \tag{151}$$

$$\leq L \sum_{j\in[c]} \left| \gamma(x, \boldsymbol{\lambda}) - \gamma(x, \boldsymbol{\lambda}_r) + \sum_{i\in[k]} (\lambda_{r,i} - \lambda_i) g_{i,j}(x) \right| \tag{152}$$

$$\leq Lc \left(|\gamma(x, \boldsymbol{\lambda}) - \gamma(x, \boldsymbol{\lambda}_r)| + A\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_r\|_1 \right) \tag{153}$$

$$\leq Lc(L\delta^2 + A\delta) \leq \varepsilon. \tag{154}$$

Therefore, $\mathcal{Q}'$ is totally bounded. Hence, $\mathcal{Q}$ is totally bounded too. As $\mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_c)$ is a complete metric space, $\mathcal{Q}$ is precompact.

F. The convex conjugate: Proof of Lemma 4

By definition of the convex conjugate (Definition 2), for any $\mathbf{v} \in \mathbb{R}^c$

$$F^{\text{conj}}(\mathbf{v}) = \sup_{\mathbf{p}\in\boldsymbol{\Delta}_c} \mathbf{v}^T\mathbf{p} - F(\mathbf{p}) = -\inf\{F(\mathbf{p}) - \mathbf{v}^T\mathbf{p} \mid \mathbf{p} \in [0,1]^c, \mathbf{1}^T\mathbf{p} = 1\}. \tag{155}$$

Fix $\mathbf{v}$. Let $\eta_{\mathbf{v}} \triangleq \min_{j\in[c]} f'(1/q_j) - v_j$. For any $\gamma \in (-\infty, \eta_{\mathbf{v}})$, define $\mathbf{p}(\gamma) \in \boldsymbol{\Delta}_c^+$ by

$$p_j(\gamma) \triangleq q_j \phi(\gamma + v_j). \tag{156}$$

Note that both f' and ϕ are strictly increasing, continuous functions so for any $\gamma \in (-\infty, \eta_{\mathbf{v}})$

$$0 = \lim_{t\to-\infty} p_j(t) < p_j(\gamma) < q_j\phi(\eta_{\mathbf{v}} + v_j) \leq q_j\phi(f'(1/q_j)) = 1 \tag{157}$$

for every $j \in [c]$. Let $a \in [c]$ be such that $\eta_{\mathbf{v}} = f'(1/q_a) - v_a$. We have that

$$\lim_{\gamma\to\eta_{\mathbf{v}}} p_a(\gamma) = q_a \lim_{u\to f'(1/q_a)} \phi(u) = 1, \tag{158}$$

so

$$\lim_{\gamma\to\eta_{\mathbf{v}}} \sum_{j\in[c]} p_j(\gamma) > 1. \tag{159}$$

On the other hand,

$$\lim_{\gamma\to-\infty} \sum_{j\in[c]} p_j(\gamma) = 0. \tag{160}$$

The intermediate value theorem implies that $\gamma(\mathbf{v})$ as given in (76) is well-defined.

Introducing a Lagrange multiplier η

$$F^{\text{conj}}(\mathbf{v}) = -\inf_{\mathbf{p}\in[0,1]^c} \sup_{\eta\in\mathbb{R}} F(\mathbf{p}) - \mathbf{v}^T\mathbf{p} - \eta(\mathbf{1}^T\mathbf{p} - 1). \tag{161}$$

Define $g_{\mathbf{v}} : \mathbb{R} \to \mathbb{R} \cup \{\pm\infty\}$ by

$$g_{\mathbf{v}}(\eta) \triangleq \inf_{\mathbf{p}\in[0,1]^c} F(\mathbf{p}) - \mathbf{v}^T\mathbf{p} - \eta(\mathbf{1}^T\mathbf{p} - 1). \tag{162}$$

Note that

$$g_{\mathbf{v}}(\gamma(\mathbf{v})) = F(\mathbf{p}(\mathbf{v})) - \mathbf{v}^T\mathbf{p}(\mathbf{v}). \tag{163}$$

Indeed, we have $(0,1]^c \subset \mathbf{dom}\ F$ and

$$\nabla \left(F(\mathbf{p}) - \mathbf{v}^T\mathbf{p} - \gamma(\mathbf{v})(\mathbf{1}^T\mathbf{p} - 1) \right)\big|_{\mathbf{p}=\mathbf{p}(\gamma(\mathbf{v}))} = \left(f' \left(\frac{p_j(\gamma(\mathbf{v}))}{q_j} \right) - v_j - \gamma(\mathbf{v}) \right)_{j\in[c]} = \mathbf{0}, \tag{164}$$

so (163) follows by convexity of F. Then,

$$F^{\text{conj}}(\mathbf{v}) = -\inf_{\mathbf{p}\in[0,1]^c} \sup_{\eta\in\mathbb{R}} F(\mathbf{p}) - \mathbf{v}^T\mathbf{p} - \eta(\mathbf{1}^T\mathbf{p} - 1) \tag{165}$$

$$\leq -\sup_{\eta\in\mathbb{R}} \inf_{\mathbf{p}\in[0,1]^c} F(\mathbf{p}) - \mathbf{v}^T\mathbf{p} - \eta(\mathbf{1}^T\mathbf{p} - 1) \tag{166}$$

$$= -\sup_{\eta\in\mathbb{R}} g_{\mathbf{v}}(\eta) \tag{167}$$

$$\leq -g_{\mathbf{v}}(\gamma(\mathbf{v})) \tag{168}$$

$$= \mathbf{v}^T\mathbf{p}(\mathbf{v}) - F(\mathbf{p}(\mathbf{v})). \tag{169}$$

Therefore, formula (75) holds.

Further, by strict convexity of F, $\mathbf{p}(\mathbf{v})$ is the unique minimizer of $F(\mathbf{h}) - \mathbf{v}^T\mathbf{h}$ for $\mathbf{h} \in \mathbf{\Delta}_c^+$. We show that $\mathbf{q}^{\text{conj}}(\mathbf{v}) = \mathbf{p}(\mathbf{v})$. If $f(0^+) = \infty$, then F takes the value ∞ on the relative boundary $\mathbf{\Delta}_c \setminus \mathbf{\Delta}_c^+$ of $\mathbf{\Delta}_c$, so $\mathbf{p}(\mathbf{v})$ is the unique minimizer of $F(\mathbf{h}) - \mathbf{v}^T\mathbf{h}$ over $\mathbf{h} \in \mathbf{\Delta}_c$, i.e., $\mathbf{q}^{\text{conj}}(\mathbf{v}) = \mathbf{p}(\mathbf{v})$. Assume $f(0^+) < \infty$. Then, F is convex over $\mathbf{\Delta}_c$. Let $G(\mathbf{h}) = F(\mathbf{h}) - \mathbf{v}^T\mathbf{h}$. For $\mathbf{h} \in \mathbf{\Delta}_c$ such that $G(\mathbf{h}) \leq G(\mathbf{p}(\mathbf{v}))$, the point $\frac{1}{2}(\mathbf{p}(\mathbf{v}) + \mathbf{h})$ lies in $\mathbf{\Delta}_c^+$ and satisfies

$$G\left(\frac{1}{2}(\mathbf{p}(\mathbf{v}) + \mathbf{h})\right) \leq \frac{1}{2}(G(\mathbf{p}(\mathbf{v})) + G(\mathbf{h})) \leq G(\mathbf{p}(\mathbf{v})), \tag{170}$$

so by uniqueness of $\mathbf{p}(\mathbf{v})$, we must have $\mathbf{h} = \mathbf{p}(\mathbf{v})$. Therefore, $\mathbf{p}(\mathbf{v})$ is the unique minimizer of G over $\mathbf{\Delta}_c$ when $f(0^+) < \infty$ too, and $\mathbf{q}^{\text{conj}}(\mathbf{v}) = \mathbf{p}(\mathbf{v})$, completing the proof of equation (77) and the lemma.

G. *Proof of Lemma 1*

Recall that $(S, X, Y, \hat{Y})$ form a Markov chain in the order $(S, Y) - X - \hat{Y}$. Hence,

$$P_{S,X,Y,\hat{Y}}(s, x, y, \hat{y}) = P_{S,Y}(s, y)P_{X|S,Y}(x|s, y)P_{\hat{Y}|X}(\hat{y}|x). \tag{171}$$

1) The Statistical Parity requirement can be written as

$$\left|\frac{1}{P_S(s)} \frac{\mathbb{E}\left[\mathbf{s}_s(X)\mathbf{h}_{\hat{y}}(X)\right]}{\mathbb{E}\left[\mathbf{h}_{\hat{y}}(X)\right]} - 1\right| \leq \alpha, \tag{172}$$

for all $(s, \hat{y}) \in [d] \times [c]$. This is equivalent to

$$\mathbb{E}\left[\langle \delta\mathbf{a}^{(s,\hat{y})}(X) - \alpha\mathbf{b}^{(\hat{y})}(X), \mathbf{h}(X)\rangle\right] \leq 0, \quad \delta \in \{\pm 1\},\ (s, \hat{y}) \in [d] \times [c] \tag{173}$$

where

$$\mathbf{a}^{(s,\hat{y})}(x) \triangleq \left(\frac{\mathbf{s}_s(x)}{P_S(s)} - 1\right)\mathbf{e}^{(\hat{y})}, \tag{174}$$

$$\mathbf{b}^{(\hat{y})}(x) \triangleq \mathbf{e}^{(\hat{y})}. \tag{175}$$

Here $\mathbf{e}^{(\hat{y})}$ is a vector of size c with one on the $\hat{y}$-th coordinate and zero elsewhere.

2) The Equalized Odds requirement can be written as

$$\left|\frac{1}{P_{S|Y}(s|y)} \frac{\mathbb{E}\left[\mathbf{s}_s(X)\mathbf{y}_y^{(s)}(X)\mathbf{h}_{\hat{y}}(X)\right]}{\mathbb{E}\left[\mathbf{y}_y(X)\mathbf{h}_{\hat{y}}(X)\right]} - 1\right| \leq \alpha, \tag{176}$$

for all $(s, \hat{y}, y) \in [d] \times [c] \times [c]$. This is equivalent to

$$\mathbb{E}\left[\langle \delta\mathbf{a}^{(s,\hat{y},y)}(X) - \alpha\mathbf{b}^{(\hat{y},y)}(X), \mathbf{h}(X)\rangle\right] \leq 0, \quad \delta \in \{\pm 1\},\ (s, \hat{y}, y) \in [d] \times [c] \times [c] \tag{177}$$

where

$$\mathbf{a}^{(s,\hat{y},y)}(x) = \left(\frac{\mathbf{s}_s(x)\mathbf{y}_y^{(s)}(x)}{P_{S|Y}(s|y)} - \mathbf{y}_y(x)\right)\mathbf{e}^{(\hat{y})}, \tag{178}$$

$$\mathbf{b}^{(\hat{y},y)}(x) = \mathbf{y}_y(x)\mathbf{e}^{(\hat{y})}. \tag{179}$$

3) Overall Accuracy Equality condition can be written as

$$\left| \frac{1}{P_S(s)} \frac{\mathbb{E}\left[\langle \mathbf{s}_s(X)\mathbf{y}^{(s)}(X), \mathbf{h}(X)\rangle\right]}{\mathbb{E}\left[\langle \mathbf{y}(X), \mathbf{h}(X)\rangle\right]} - 1 \right| \leq \alpha, \tag{180}$$

for all $s \in [d]$. This is equivalent to

$$\mathbb{E}\left[\langle \delta\mathbf{a}^{(s)}(X) - \alpha\mathbf{b}(X), \mathbf{h}(X)\rangle\right] \leq 0, \quad \varepsilon \in \{\pm 1\},\ s \in [d] \tag{181}$$

where

$$\mathbf{a}^{(s)}(x) = \frac{\mathbf{s}_s(x)}{P_S(s)}\mathbf{y}^{(s)}(x) - \mathbf{y}(x), \tag{182}$$

$$\mathbf{b}(x) = \mathbf{y}(x). \tag{183}$$

H. Proof of Theorem 3

We apply Theorem 13.2 in [40] to prove our theorem. To verify the assumptions therein, we show that the search space of the minimization we aim to solve can be restricted into a convex and norm-bounded subset. Furthermore, we prove that, for a fixed $x \in \mathcal{X}$, the mapping $\boldsymbol{\lambda} \mapsto D_f^{\text{conj}}(\mathbf{v}(x;\boldsymbol{\lambda}), \mathbf{y}(x))$ (i.e., the loss function $f \mapsto \ell(f, z)$ in [40]) is convex and Lipschitz. Before we state the proof, recall that $\mathbf{v}(x;\boldsymbol{\lambda}) = -\mathbf{G}(x)\boldsymbol{\lambda}$.

a) Convexity of the mapping: By the definition of D_f^{conj}, for each fixed $x \in \mathcal{X}$, the function $D_f^{\text{conj}}(-\mathbf{G}(x)\boldsymbol{\lambda}, \mathbf{y}(x))$ is a pointwise maximum of linear functions in $\boldsymbol{\lambda}$. Therefore, $\boldsymbol{\lambda} \mapsto D_f^{\text{conj}}(-\mathbf{G}(x)\boldsymbol{\lambda}, \mathbf{y}(x))$ is convex.

b) Search space: Due to the linear formulation of the fairness constraints, we have that the uniform classifier is strictly feasible, i.e., $\mathbb{E}\left[\mathbf{1}^T\mathbf{G}(X)\right] < \mathbf{0}$. Furthermore, by Assumption I-(d), we have $\inf_{x,j} \mathbf{y}_j(x) > 0$. Therefore, $P_X P_{U|X} \ll P_{X,Y}$ and $D_f(P_X P_{U|X} \| P_{X,Y}) < \infty$. Now Corollary 4 guarantees that

$$\inf_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathbb{E}\left[D_f^{\text{conj}}(\mathbf{v}(X;\boldsymbol{\lambda}), \mathbf{y}(X))\right] = \inf_{\substack{\boldsymbol{\lambda} \geq \mathbf{0} \\ \|\boldsymbol{\lambda}\|_1 \leq \theta}} \mathbb{E}\left[D_f^{\text{conj}}(\mathbf{v}(X;\boldsymbol{\lambda}), \mathbf{y}(X))\right] \tag{184}$$

where

$$\theta \triangleq \frac{cD_f(P_X P_{U|X} \| P_{X,Y})}{-\max_{j \in [2\ell]} \mathbb{E}\left[\mathbf{1}^T \mathbf{G}_{:,j}(X)\right]}, \tag{185}$$

since the constant θ_q therein can be further bounded by θ.

c) Lipschitzianity of the mapping: First, we have that

$$\nabla_{\mathbf{v}} D_f^{\text{conj}}(\mathbf{v}, \mathbf{y}(x)) = \mathbf{q}^{\text{conj}}(\mathbf{v}) \in \boldsymbol{\Delta}_c. \tag{186}$$

Then, $\nabla_{\boldsymbol{\lambda}} D_f^{\text{conj}}(-\mathbf{G}(x)\boldsymbol{\lambda}, \mathbf{y}(x)) = -\mathbf{G}(x)^T \mathbf{q}^{\text{conj}}(\mathbf{v})$ where $\mathbf{v} = -\mathbf{G}(x)\boldsymbol{\lambda}$. Therefore, $\boldsymbol{\lambda} \mapsto D_f^{\text{conj}}(\mathbf{v}(x;\boldsymbol{\lambda}), \mathbf{y}(x))$ is Lipshitz with Lipschitz constant $\|\mathbf{G}(x)\|_1$: by the mean-value theorem, for any $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$, there is a $\mathbf{v}'' = -\mathbf{G}(x)\boldsymbol{\lambda}''$ such that

$$|D_f^{\text{conj}}(\mathbf{v}(x;\boldsymbol{\lambda}), \mathbf{y}(x)) - D_f^{\text{conj}}(\mathbf{v}(x;\boldsymbol{\lambda}'), \mathbf{y}(x))| = |\nabla_{\boldsymbol{\lambda}} D_f^{\text{conj}}(\mathbf{v}(x;\boldsymbol{\lambda}''), \mathbf{y}(x))^T(\boldsymbol{\lambda} - \boldsymbol{\lambda}')| \tag{187}$$

$$\leq \|\mathbf{G}(x)^T \mathbf{q}^{\text{conj}}(\mathbf{v}'')\|_2 \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2 \tag{188}$$

$$\leq \|\mathbf{G}(x)\|_1 \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2. \tag{189}$$

Hence, the mapping $\boldsymbol{\lambda} \mapsto D_f^{\text{conj}}(\mathbf{v}(x;\boldsymbol{\lambda}), \mathbf{y}(x))$ is Lipshitz with Lipshitz constant $\sup_{x \in \mathcal{X}} \|\mathbf{G}(x)\|_1$.